

THE UNIVERSITY OF CHICAGO

STATISTICAL LEARNING AND OPTIMIZATION UNDER DISTRIBUTION SHIFT

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE UNIVERSITY OF CHICAGO
BOOTH SCHOOL OF BUSINESS
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY
BOXIN ZHAO

CHICAGO, ILLINOIS

AUGUST 2025

To my mentors, collaborators, friends and family — for your guidance, inspiration,
companionship and support throughout this journey.

“All models are wrong, some are useful.” — George E. P. Box

CONTENTS

LIST OF FIGURES	viii
LIST OF TABLES	x
ACKNOWLEDGMENTS	xi
ABSTRACT	xii
1 INTRODUCTION	1
1.1 Summary of Contributions	3
1.2 Authorship and Other Contributions	6
2 TRANS-GLASSO: A TRANSFER LEARNING APPROACH TO PRECISION MA- TRIX ESTIMATION	8
2.1 Introduction	8
2.1.1 Related Work	9
2.1.2 Organization and Notation	10
2.2 Problem Setup	12
2.2.1 Comparison of Similarity Assumptions	13
2.3 Trans-Glasso Algorithm	15
2.3.1 Initialization via Multi-Task Learning	15
2.3.2 Refinement via Differential Network Estimation	16
2.4 Implementation in Practice	17
2.4.1 Optimization Algorithms	17
2.4.2 Hyperparameter Selection	27
2.4.3 Identifying the Informative Set	27
2.5 Theoretical Analysis	28
2.5.1 Analysis of Trans-MT-Glasso	29
2.5.2 Analysis of Trans-Glasso	30
2.5.3 Upper Bound for Expected Error	32
2.5.4 Minimax Lower Bounds and Optimality	35
2.5.5 Minimax Optimal Rate for Differential Network Estimation	38
2.5.6 Graph Recovery Guarantee for Gaussian Graphical Models	41
2.6 Simulations	44
2.6.1 Data Generation Models	44
2.6.2 Experimental Design	45
2.6.3 Comparison Methods	47
2.6.4 Results	48
2.6.5 Ablation Study of the Refinement Step	49
2.6.6 Experiments with Unknown Informative Set	50
2.7 Real-World Data Analysis	52
2.7.1 Gene Networks Data for Brain Tissues	53

2.7.2	Protein Networks Data for AML	54
2.8	Conclusion	55
2.9	Technical Details	56
2.9.1	Preliminary Lemmas	56
2.9.2	Proof of Theorem 1	64
2.9.3	Proof of Theorem 5	74
2.9.4	Proof of Lemma 9	77
2.9.5	Proof of Lemma 10	77
2.9.6	Proof of Theorem 11	80
2.9.7	Corollary 28 and Its Proof	82
3	SMART: A SPECTRAL TRANSFER APPROACH TO MULTI-TASK LEARNING	85
3.1	Introduction	85
3.1.1	Related Work	87
3.1.2	Organization and Notation	89
3.2	Problem Setup	90
3.3	SMART Algorithm	95
3.3.1	Optimization Algorithms	97
3.3.2	Hyperparameter Selection	103
3.4	Theoretical Analysis	103
3.4.1	Technical Conditions	105
3.4.2	Theoretical Guarantees	107
3.5	Simulation	112
3.5.1	Data Generation Process	113
3.5.2	Experimental Design	114
3.5.3	Baseline Methods	115
3.5.4	Results	117
3.6	Real-World Data Analysis	117
3.7	Conclusion	123
3.8	Technical Details	124
3.8.1	Preliminary Lemmas	124
3.8.2	Proof of Theorem 30	144
3.8.3	Additional Useful Lemmas	153
4	ADAPTIVE CLIENT SAMPLING IN FEDERATED LEARNING VIA ONLINE LEARNING WITH BANDIT FEEDBACK	155
4.1	Introduction	155
4.1.1	Related Work	157
4.1.2	Organization and Notation	160
4.2	Adaptive Client Sampling	161
4.2.1	Client Sampling as an Online Learning Problem	161
4.2.2	OSMD Sampler	165
4.3	Application of OSMD Sampler on Federated Optimization Algorithms	167
4.3.1	Convergence Analysis of Mini-batch SGD with OSMD Sampler	172

4.3.2	Convergence Analysis of FedAvg with OSMD Sampler	174
4.4	Regret Analysis of OSMD Sampler	177
4.5	Adaptive-OSMD Sampler	179
4.6	Extensions of Adaptive-OSMD Sampler	184
4.6.1	Adaptive-OSMD Sampler with Doubling Trick	184
4.6.2	Adaptive Sampling Without Replacement	186
4.7	Simulation Experiments	188
4.7.1	Adaptive-OSMD Sampler vs Uniform Sampling	190
4.7.2	Adaptive-OSMD Sampler vs MABS vs VRB vs Avare	191
4.7.3	Robustness of Adaptive-OSMD Sampler to the Choice of α	191
4.7.4	Dynamic Sampling Distribution v.s. Fixed Sampling Distribution	192
4.7.5	Comparison Between Sampling with Replacement and Sampling without Replacement	193
4.8	Real Data Experiment	194
4.9	Conclusion	197
4.9.1	Client Sampling with Personalized FL Objective	198
4.9.2	Sampling with Physical Constraint in FL System	201
4.10	Technical Details	201
4.10.1	Useful Lemmas	202
4.10.2	Proof of Proposition 37	207
4.10.3	Proof of Theorem 38	208
4.10.4	Proof of Theorem 39	214
4.10.5	Proof of Theorem 40	227
4.10.6	Proof of Theorem 42	232
4.10.7	Proof of Theorem 43	236
4.10.8	Proof of Theorem 44	239
5	CONCLUSION	241
	REFERENCES	244

LIST OF FIGURES

2.1	Illustration of Assumption 1. The target precision matrix, $\Omega^{(0)}$, is shown alongside two source precision matrices, $\Omega^{(1)}$ and $\Omega^{(2)}$. Black crosses represent the shared entries across the matrices, while colored shapes indicate individual, unique entries.	13
2.2	Simulation results of Trans-Glasso for Model I.	46
2.3	Simulation results of Trans-Glasso for Model II.	47
2.4	Simulation results of Trans-Glasso for Model III.	48
2.5	Ablation study results of Trans-Glasso for Model I.	49
2.6	Ablation study results of Trans-Glasso for Model II.	50
2.7	Ablation study results of Trans-Glasso for Model III.	51
2.8	Simulation results of Trans-Glasso-CV when the informative set $\mathcal{A}$ is unknown. .	52
2.9	Cross-validation prediction error of different methods on GTEx brain tissue datasets, relative to Glasso-Target.	53
2.10	Protein networks for four AML subtypes. Red edges are shared across subtypes; blue edges are unique to specific subtypes.	54
3.1	Simulation results of SMART for Model I.	116
3.2	Simulation results of SMART for Model II.	118
3.3	Simulation results of SMART for Model III.	119
3.4	Cell counts for each annotated cell type in the single-cell multi-modal dataset. .	120
4.1	The training loss (top row) and cumulative regret (bottom row) are compared for the Adaptive-OSMD Sampler, Uniform Sampler, and Optimal Sampler, under $\sigma = 1.0$, $\sigma = 3.0$, and $\sigma = 10.0$. Solid lines represent the mean values, while the shaded regions indicate mean $\pm$ standard deviation across independent runs. . .	190
4.2	The training loss (top row) and cumulative regret (bottom row) are compared across the Adaptive-OSMD Sampler, MABS, VRB, and Avare methods for $\sigma = 1.0$, $\sigma = 3.0$, and $\sigma = 10.0$. Solid lines represent the mean values, while shaded regions indicate mean $\pm$ standard deviation across independent runs.	191
4.3	The training loss (top row) and cumulative regret (bottom row) are shown for the Adaptive-OSMD Sampler with different values of α under $\sigma = 1.0$, $\sigma = 3.0$, and $\sigma = 10.0$. Solid lines represent the mean values, while shaded regions indicate mean $\pm$ standard deviation across independent runs.	192
4.4	The training loss (top row) and cumulative regret (bottom row) are compared between the Adaptive-OSMD Sampler and p^{IS} for $\nu = 1.0$, $\nu = 3.0$, and $\nu = 10.0$. Solid lines represent the mean values, while shaded regions illustrate mean $\pm$ standard deviation across independent runs.	193
4.5	The training loss (top row) and cumulative regret (bottom row) are compared for the Adaptive-OSMD Sampler with replacement and without replacement, across $\sigma = 1.0$, $\sigma = 3.0$, and $\sigma = 10.0$. Solid lines represent the mean values, while shaded regions indicate mean $\pm$ standard deviation across independent runs. . .	194
4.6	The sample size distribution in the training set across clients.	195

4.7	Comparison of Adaptive-OSMD Sampler, Uniform Sampler, and other online learning samplers on real datasets. Each column corresponds to a different dataset. The top row shows training loss, and the bottom row shows validation accuracy. Results are averaged over five independent runs.	196
-----	--	-----

LIST OF TABLES

1.1	Adaptation framework across the three dissertation chapters.	3
3.1	Summary statistics of prediction errors across 100 repeated experiments on the single-cell multi-modal dataset.	122

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my advisors, Professors Mladen Kolar and Cong Ma. They taught me how to identify and tackle challenging problems, communicate research effectively, and—most importantly—grow into an independent researcher. I am especially thankful to Professor Kolar, who introduced me to academic research at a time when I knew very little. Throughout my doctoral studies, he has provided unwavering guidance across every dimension of my PhD. I am also deeply grateful to Professor Ma—one of the most insightful theorists I have had the privilege to work with—for helping me appreciate the elegance of theory and uncover practical insights from abstract results. His thoughtful feedback and guidance have left a lasting impact on my development.

In addition, I am grateful to Professors Tengyuan Liang and Sanmi Koyejo for generously serving on my dissertation committee and contributing their invaluable time and expertise. I also thank Professor Bryon Aragam for organizing the machine learning reading group, which proved to be a valuable and enriching part of my PhD experience.

During my PhD studies, I had the privilege of collaborating with many brilliant researchers. I am especially grateful to Jinchi Lv, Raul Castro Fernandez, Y. Samuel Wang, Lingxiao Wang, Filip Hanzely, Ziqi Liu, and Katherine Tsai, whose insights and inspiration have been invaluable to the development of this work. I also wish to thank my fellow PhD friends—Dake Zhang, Boxiang Lyu, Jizhou Liu, Zhouyu Shen, Wenxuan Guo, Ming Gao, Percy S. Zhai, Chang Deng, Jiguang Li, Shunzhuang Huang, and many others—for their unwavering camaraderie and support. The friendships we built have been a true blessing, and I remain deeply grateful.

Finally, I would like to express my deepest personal gratitude to my parents—my father, Junxiang Zhao, and my mother, Baohua Wang. Your unconditional love has been my greatest source of strength. I am also profoundly grateful to my girlfriend, Jiayue Cheng. Your presence by my side has been my deepest comfort.

ABSTRACT

Modern machine learning models are increasingly deployed in settings where the classical assumption of independent and identically distributed (i.i.d.) data is violated due to distribution shifts across domains, populations, or time. This dissertation develops new methodologies in statistical learning and optimization that address such shifts through the unifying principle of *adaptation*—distinguishing between components that remain stable (preserved) and those that vary (adapted) across environments.

Chapter 2 introduces TRANS-GLASSO, a two-stage procedure for precision matrix estimation under distribution shift. The method exploits a shared sparsity pattern across domains as the preserved structure and adjusts for domain-specific deviations via differential network estimation. We establish non-asymptotic error bounds, prove minimax optimality, and demonstrate the method’s practical effectiveness on both simulated and real-world data.

Chapter 3 presents SMART, a spectral regularization framework for multi-task learning. Assuming that the singular subspaces of the regression matrix are preserved while the projection weights vary across tasks, SMART estimates the target model via a nonconvex optimization problem regularized by source-informed subspaces. The method is supported by theoretical guarantees and achieves strong empirical performance.

Chapter 4 introduces an online client sampling method for federated optimization under data heterogeneity. It leverages the slowly varying informativeness of clients as the preserved structure and dynamically adapts the sampling distribution. Formulated as an online learning problem with bandit feedback, the algorithm builds on Online Stochastic Mirror Descent and achieves consistent improvements over uniform sampling in both theory and practice.

Collectively, these contributions advance a unified framework for learning and optimization under distribution shift by systematically decomposing each problem into preserved and adapted components. This framework enables the principled design of algorithms that are both theoretically grounded and practically effective in the presence of heterogeneous data.

CHAPTER 1

INTRODUCTION

The classical assumption that data are independently and identically distributed (i.i.d.) serves as a foundational pillar of traditional statistical learning and stochastic optimization theory. However, in real-world applications, this assumption is frequently violated. Practical problems—ranging from molecular network estimation under varying conditions [Theodoris et al., 2023], to financial forecasting amid changing market regimes [Chen et al., 2024], to federated learning across heterogeneous client populations [McMahan et al., 2017]—often involve substantial *distribution shifts* across environments, time, or populations. These shifts present fundamental challenges for both modeling and optimization, as they can significantly undermine the reliability and generalizability of learning algorithms.

This dissertation is motivated by a central question:

How can statistical and optimization methods adapt to distribution shift by identifying what remains invariant and what must be updated?

We adopt a unifying lens of *adaptation*, in which a model or algorithm is guided by structural priors that distinguish between components of the learning problem that remain invariant—the *preserved components*—and those that must vary across domains—the *adapted components*. To illustrate this principle, consider the high-dimensional transfer learning framework proposed by Li et al. [2022] for linear regression under distribution shift. Let $(X^{(s)}, Y^{(s)})$ and $(X^{(t)}, Y^{(t)})$ denote the source and target data, respectively, generated according to the linear models:

$$Y^{(s)} = \left(X^{(s)}\right)^{\top} \beta^{(s)} + \varepsilon^{(s)}, \quad Y^{(t)} = \left(X^{(t)}\right)^{\top} \beta^{(t)} + \varepsilon^{(t)},$$

where $\beta^{(s)}, \beta^{(t)} \in \mathbb{R}^p$ are high-dimensional regression vectors, and $\varepsilon^{(s)}, \varepsilon^{(t)}$ are independent noise terms. The key assumption that enables transfer is that the source and target

coefficients are close in norm:

$$\|\beta^{(s)} - \beta^{(t)}\|_q \leq \delta,$$

for some $q \in \{1, 2\}$ and small $\delta > 0$. This assumption encodes a notion of parametric proximity, enabling information from the source domain to be effectively transferred to the target via regularization. In this framework, the *preserved component* is the shared structure between $\beta^{(s)}$ and $\beta^{(t)}$, as reflected by their small difference, while the *adapted component* is the residual discrepancy that must be corrected using target-specific data.

This structural approach to adaptation—leveraging the preserved component while selectively adjusting the adapted component—has informed methodological advances across a range of domains. For example, in functional linear models, Lin and Reimherr [2024] assume that the coefficient function differs by a smooth offset, enabling adaptation through correction within a shared reproducing kernel Hilbert space (RKHS). In nonparametric classification, Cai and Wei [2021] exploit common low-noise structures across domains while adapting to domain-specific variations in the regression function. In contextual bandits, Cai et al. [2024a] regularize the target policy toward a source policy class, thereby preserving structural priors while allowing for task-specific exploration. Across these diverse settings, structural similarity provides a principled foundation for effective and efficient transfer.

While the idea of structural adaptation is not new, this dissertation advances the concept by extending it to novel problem settings in creative and nontrivial ways. Rather than applying existing techniques wholesale, each chapter introduces a tailored methodological framework that reinterprets the notions of preservation and adaptation to suit the specific structure of the problem at hand. In doing so, we demonstrate that the power of adaptation lies not only in its generality, but also in the flexibility with which it can be operationalized to address domain-specific challenges.

We formalize adaptation under distribution shift as a two-part process:

1. **Preservation:** Identify components of the model or optimization problem that remain

stable across environments.

2. **Adaptation:** Adjust the variable components using data, guided by prior knowledge or regularization derived from the preserved structure.

This principle underpins each of the three core chapters of the dissertation, as summarized in the table below:

Chapter	Setting	Preserved Component	Adapted Component	Adaptation Method
2	Precision matrix estimation	Shared component of the precision matrix	Unique component of the precision matrix	Multi-task learning + refinement
3	Multi-task learning	Spectral subspaces	Projection weights	Spectral regularization
4	Federated optimization	Relative informativeness of clients (slowly varying)	Client sampling distribution	Online learning (OSMD-based)

Table 1.1: Adaptation framework across the three dissertation chapters.

This structured perspective unifies a diverse set of technical challenges—spanning precision matrix estimation, multi-task learning, and adaptive client sampling—under a common theoretical and algorithmic goal: learning what changes and what remains invariant in the presence of distribution shift. The next section summarizes the main contributions of the dissertation, highlighting how this lens of adaptation leads to novel methodologies and rigorous theoretical guarantees across domains.

1.1 Summary of Contributions

This section provides a high-level overview of the three core chapters of the dissertation. Each chapter introduces a methodological contribution that addresses distribution shift through the lens of adaptation, focusing on a distinct modeling or algorithmic setting. The summaries highlight, for each case, the preserved and adapted components, the adaptation strategy employed, and the theoretical and empirical results that support the proposed approach.

Chapter 2: Structural Adaptation in Precision Matrix Estimation

This chapter addresses the problem of *precision matrix estimation* under distribution shift, motivated by the task of learning Gaussian graphical models in heterogeneous biological contexts—such as cancer subtypes—where the target domain often suffers from limited sample sizes due to the rarity of the condition. Such data scarcity poses significant challenges for accurate estimation of high-dimensional network structures, unless one can leverage auxiliary information from better-sampled source domains.

We approach this challenge through the lens of *structural adaptation*. The *preserved component* is the set of *shared entries* in the precision matrices across domains, which capture common conditional dependencies likely to hold in related biological settings. The *adapted component* is the *task-specific deviation*, representing structural differences unique to the target domain that must be accounted for to ensure accurate estimation.

To this end, we propose **Trans-Glasso**, a two-step transfer learning procedure that exploits this structural decomposition:

- In the first step, we obtain initial estimators for both the source and target domains using a *multi-task graphical lasso*, which jointly estimates the shared structure and domain-specific deviations.
- In the second step, we refine the target estimate via *differential network estimation*, which corrects residual discrepancies by estimating the difference between the target and source networks.

Trans-Glasso integrates the preserved and adapted components into a coherent estimation framework. We derive non-asymptotic *error bounds*, establish *minimax optimality*, and validate the method through simulations and applications to real-world biological data. Empirically, Trans-Glasso consistently improves estimation accuracy in small-sample regimes while remaining robust to potential negative transfer.

Chapter 3: Spectral Adaptation in Multi-Task Learning

We next examine the problem of *multi-task learning*, where the regression coefficient matrix is assumed to be low-rank, capturing latent structures such as gene–protein regulatory modules. Under distribution shift, the primary challenge is to effectively transfer information from a well-sampled source task to a sparsely observed target task.

In this setting, we adopt the assumption that the *singular subspaces* of the regression matrix are preserved across tasks, while the *projection weights*—which determine the contribution of each singular direction—are task-specific and must be adapted. Accordingly, the *preserved component* is the subspace structure inferred from the source task, and the *adapted component* is the projection weights required to reconstruct the target regression matrix.

To address this challenge, we propose **SMART** (Spectral Multi-task Adaptive Regression Transfer), which estimates the target regression matrix through a nonconvex optimization problem regularized by the source-informed spectral subspaces. We develop an efficient *ADMM-based algorithm* to solve this problem, and establish *finite-sample error bounds* via a convexity-assisted theoretical analysis.

SMART consistently outperforms existing baselines in both synthetic experiments and real-world applications involving single-cell multi-modal datasets, enabling accurate recovery of regression structure in rare or underrepresented cell types.

Chapter 4: Sampling Adaptation in Federated Optimization

This chapter shifts the focus from statistical estimation to *federated optimization*, where data are decentralized across clients with heterogeneous distributions. In this setting, a global model is updated through iterative communication with a subset of clients, each holding potentially non-i.i.d. local data. A central challenge is to adaptively select which clients participate in each round when their contributions to model improvement are initially unknown and may vary over time.

We frame this challenge through the lens of adaptation. The *preserved component* is the *relative informativeness of clients*, which is assumed to evolve gradually over time. While the influence of individual clients may fluctuate, the overall informativeness structure remains approximately stable across training rounds. The *adapted component* is the *client sampling distribution*, which dynamically adjusts to prioritize the most beneficial clients for updates.

To address this problem, we formulate client selection as an *online learning problem with bandit feedback* and develop an adaptive sampling algorithm based on *Online Stochastic Mirror Descent (OSMD)*. This approach balances exploration and exploitation to learn sampling distributions that minimize regret with respect to dynamic performance benchmarks.

We provide theoretical guarantees on both regret and convergence rates of the proposed method, and validate its practical utility through experiments on synthetic data and real-world federated learning tasks.

Together, these three chapters demonstrate how the principle of adaptation—by disentangling preserved and adapted components—can guide the development of statistically and computationally efficient methods across a range of learning settings. The work reflects a combination of original research and collaborative efforts, as detailed in the authorship statement below.

1.2 Authorship and Other Contributions

Portions of this dissertation are based on collaborative work conducted during the author’s PhD studies. Chapter 2 is based on Zhao et al. [2024a], co-authored with Cong Ma and Mladen Kolar. Chapter 4 is adapted from Zhao et al. [2025], co-authored with Lingxiao Wang, Ziqi Liu, Zhiqiang Zhang, Jun Zhou, Chaochao Chen, and Mladen Kolar. The work presented in Chapter 3 is original to this dissertation and was developed in collaboration with Jinchi Lv and Mladen Kolar.

Beyond the scope of this dissertation, the author has contributed to several additional

research projects during the PhD program. These include work on functional graphical models [Zhao et al., 2019, 2022, 2024c, Tsai et al., 2024], federated learning [Hanzely et al., 2023], optimization [Zhao et al., 2023b], causal discovery [Zhao et al., 2024b], matrix completion [Wang et al., 2023], and data markets [Zhao et al., 2023a]. Although these projects are not included in the main body of the dissertation, they reflect the broader research interests cultivated throughout the author’s doctoral training.

CHAPTER 2

TRANS-GLASSO: A TRANSFER LEARNING APPROACH TO PRECISION MATRIX ESTIMATION

2.1 Introduction

Estimating the precision matrix, i.e., the inverse covariance matrix, is a fundamental task in statistical analysis and has broad applications, including in portfolio optimization, speech recognition, and genomics [Lauritzen, 1996]. The precision matrix is closely tied to Gaussian graphical models: estimating the support of the precision matrix corresponds to uncovering the network structure of conditional dependencies between multivariate normal variables [Lauritzen, 1996]. However, estimating a precision matrix accurately is often challenging when the sample size is small compared to the dimension—a typical scenario in high-dimensional settings.

In many applications, sample sizes are constrained for the target study of interest, yet data from related studies may be available. Transfer learning [Pan and Yang, 2009] provides a promising approach in these scenarios by leveraging information from related source studies to improve estimation accuracy in the target study. For example, in gene expression studies across different tissue types, sample sizes may be small for specific tissues, but data from related tissues can help improve estimates [Li et al., 2023a]. Similarly, protein network studies for different cancer subtypes can benefit from transfer learning, as leveraging data from related subtypes can enhance estimation for a particular subtype with limited data [Peterson et al., 2015].

A critical aspect of transfer learning is establishing similarity between the target and source tasks. Here, we assume that most entries of the target precision matrix are shared with those of the source matrices, with only a few differences. Based on this assumption, we propose Trans-Glasso, a novel two-step transfer learning method for precision matrix

estimation. First, we obtain initial estimators through a multi-task learning objective that captures shared and unique dependencies across datasets. Second, we refine these estimators using differential network estimation to adjust for differences between the target and source matrices [Zhao et al., 2014, Yuan et al., 2017].

We provide a theoretical analysis of Trans-Glasso, deriving non-asymptotic error bounds and establishing that the method achieves minimax optimality in a wide range of parameter regimes. Through extensive simulations, we demonstrate that Trans-Glasso outperforms several baseline methods, particularly in scenarios where the target sample size is small. We also apply Trans-Glasso to gene networks across brain tissues and protein networks for various cancer subtypes, showing its practical effectiveness in biological applications. Additionally, as a byproduct of our analysis, we derive the minimax optimal rate for differential network estimation, to our knowledge, the first of its kind.

2.1.1 *Related Work*

Precision matrix estimation. Estimation of sparse precision matrices in a single study is well-studied. Common methods include penalized M-estimator [Yuan and Lin, 2007, Friedman et al., 2008, Rothman et al., 2008, Lam and Fan, 2009, Ravikumar et al., 2011] and constrained L_1 minimization [Cai et al., 2011, 2016, Ren et al., 2015]. There is also extensive literature on multi-task precision matrix estimation, which estimates multiple related but non-identical precision matrices from multiple studies [Guo et al., 2011, Danaher et al., 2014, Zhu et al., 2014]. See Tsai et al. [2022] for a survey. While related to transfer learning, multi-task learning aims to estimate parameters of all studies, whereas transfer learning only focuses on the target study.

Transfer learning. Transfer learning has a long history [Pan and Yang, 2009]. Recently, interest in transfer learning for statistical problems has grown. Li et al. [2022], He et al. [2024b] studied high-dimensional linear regression. The fused regularizer in He et al. [2024b]

is similar to our multi-task objective; however, He et al. [2024b] focuses on linear regression, while we focus on precision matrix estimation, making the techniques different. Li et al. [2023b], Tian and Feng [2023] studied high-dimensional generalized linear regression. Pathak et al. [2022], Ma et al. [2023], Wang [2023], Ge et al. [2024] addressed covariate shift. Cai and Wei [2021] studied nonparametric classification and Cai et al. [2024a] investigated multi-armed bandit problems. Lin and Reimherr [2022], Cai et al. [2024b] studied transfer learning for functional data analysis.

The most closely related work is Li et al. [2023a], which also investigates transfer learning for precision matrix estimation. The key difference is the similarity assumption. Li et al. [2023a] assumes that the divergence matrices between the target and source precision matrices are sparse. We assume that most entries of the target precision matrix are shared across source precision matrices, with few different entries. Although the assumption in Li et al. [2023a] is motivated by the KL divergence between Gaussian distributions, ours is a structural assumption, making it applicable beyond Gaussian data and easier to interpret. Consequently, our method differs significantly from that of Li et al. [2023a].

Differential Network Estimation. Our approach leverages differential network estimation techniques, which aim to directly estimate the difference between two precision matrices without the need to estimate the individual ones [Zhao et al., 2014, Yuan et al., 2017, Liu et al., 2014, Ma et al., 2021]. Fazayeli and Banerjee [2016] explored this concept in the context of Ising models. Additionally, Zhao et al. [2019, 2022] extended differential network estimation methods to functional data, while Tugnait [2023] broadened its application to multi-attribute data.

2.1.2 Organization and Notation

The rest of the chapter is organized as follows. In Section 2.2, we introduce the problem setup. In Section 2.3, we introduce the proposed methodology. We then describe how to

implement our method in practice in Section 2.4. The theoretical results are developed in Section 2.5. Besides, we implement extensive simulation experiments in Section 2.6. Furthermore, in Section 2.7, we apply our method on two real-world datasets. Finally, we conclude this chapter with Section 2.8. Technical proofs are presented in Section 2.9.

Notation. In this chapter, we adopt the following notations and conventions. For a vector $v \in \mathbb{R}^d$, we use $\|v\|_p$ to denote its L_p -norm. More specifically, we have $\|v\|_p = (\sum_{i=1}^d |v_i|^p)^{\frac{1}{p}}$ for $1 \leq p \leq \infty$, where $\|v\|_\infty = \max_i |v_i|$. For a matrix $A \in \mathbb{R}^{d \times d}$, we use $|\cdot|$ to denote its elementwise norm and $\|\cdot\|$ to denote its operator norm. For example, $|A|_1 = \sum_{i=1}^d \sum_{j=1}^d |A_{ij}|$, $|A|_0 = \sum_{i=1}^d \sum_{j=1}^d \mathbb{1}\{A_{ij} \neq 0\}$, $|A|_\infty = \max_{1 \leq i, j \leq d} |A_{ij}|$; $\|A\|_1 = \max_{1 \leq j \leq d} \sum_{i=1}^d |A_{ij}|$, $\|A\|_\infty = \max_{1 \leq i \leq d} \sum_{j=1}^d |A_{ij}|$, and $\|A\|_2$ denote the largest singular value of A . We use $\|A\|_F = (\sum_{i=1}^d \sum_{j=1}^d |A_{ij}|^2)^{1/2}$ to denote the Frobenius norm of A . In addition, we use $\langle A, B \rangle = \text{tr}(A^\top B) = \sum_{i,j} A_{ij} B_{ij}$ for $A, B \in \mathbb{R}^{d \times d}$ to define the inner product between two matrices. We use $\text{vec}(A)$ to denote the d^2 -vector obtained by stacking the columns of A . When A is symmetric, we let $\gamma_{\min}(A)$ and $\gamma_{\max}(A)$ denote its smallest and largest eigenvalue. For $A \in \mathbb{R}^{n_1 \times n_2}$ and $B \in \mathbb{R}^{m_1 \times m_2}$, we let $A \otimes B = [A_{ij} B_{lm}]_{i,j,l,m} \in \mathbb{R}^{n_1 m_1 \times n_2 m_2}$ denote the Kronecker product of two matrices. We define $\mathbb{S}^{d \times d}$ as the set of symmetric matrices with dimension d . The universal constants may vary from one line to another without further clarification. Finally, following Ravikumar et al. [2011], we say a random vector $X \in \mathbb{R}^d$ with $\mathbb{E}[X] = 0$ and $\Sigma = \text{Cov}(X)$ is sub-Gaussian if there exists a constant $\sigma > 0$ such that

$$\mathbb{E} \left[\exp \left(\lambda X_j / \sqrt{\Sigma_{jj}} \right) \right] \leq \exp \left(\sigma^2 \lambda^2 / 2 \right) \quad \text{for all } \lambda \in \mathbb{R} \text{ and } 1 \leq j \leq d.$$

In addition, we use the following standard notation in this chapter. For two positive sequences $\{f(n)\}_{n \geq 1}$ and $\{g(n)\}_{n \geq 1}$, $f(n) = O(g(n))$ or $f(n) \lesssim g(n)$ means that there exists a universal constant $c > 0$ such that $f(n) \leq cg(n)$ holds for sufficiently large n ; $f(n) = \Omega(g(n))$ or $f(n) \gtrsim g(n)$ means that there exists a universal constant $c > 0$ such that

$f(n) \geq cg(n)$ holds for sufficiently large n ; $f(n) = \Theta(g(n))$ or $f(n) \asymp g(n)$ means that there exist universal constants $c_1, c_2 > 0$ such that $c_1g(n) \leq f(n) \leq c_2g(n)$ holds for sufficiently large n ; $f(n) = o(g(n))$ indicates that $f(n)/g(n) \rightarrow 0$ as $n \rightarrow \infty$.

2.2 Problem Setup

Imagine that we observe n_0 i.i.d. samples $\{\mathbf{x}_i^{(0)}\}_{i=1}^{n_0} := \mathcal{D}_0$ from a sub-Gaussian target distribution $\mathcal{P}_0$. Each sample $\mathbf{x}_i^{(0)} \in \mathbb{R}^d$ is assumed to have zero mean and covariance matrix $\Sigma^{(0)}$. Our goal is to estimate the target precision matrix $\Omega^{(0)} = \left\{\Sigma^{(0)}\right\}^{-1}$. Additionally, we have access to K sub-Gaussian source distributions $\{\mathcal{P}_k\}_{k=1}^K$, each with n_k i.i.d. samples $\{\mathbf{x}_i^{(k)}\}_{i=1}^{n_k} := \mathcal{D}_k$. For $1 \leq k \leq K$, each $\mathbf{x}_i^{(k)} \in \mathbb{R}^d$ also has zero mean, covariance matrix $\Sigma^{(k)}$, and corresponding precision matrix $\Omega^{(k)} = \left\{\Sigma^{(k)}\right\}^{-1}$. The goal is to leverage samples from both the target and source distributions to accurately estimate $\Omega^{(0)}$.

To facilitate transfer learning, we assume structural similarity between the target and source precision matrices, whereby most entries in $\Omega^{(0)}$ are shared with those in $\Omega^{(k)}$, with relatively few differences. This assumption enables us to efficiently utilize source samples to enhance the estimation of the target precision matrix. Formally, we characterize the relationship between the target and source precision matrices by the following assumption.

Assumption 1. *For each $0 \leq k \leq K$, there exists a shared component $\Omega^\star$ and a unique component $\Gamma^{(k)\star}$ with disjoint supports such that*

$$\Omega^{(k)} = \Omega^\star + \Gamma^{(k)\star}, \quad (2.1)$$

where $|\Omega^\star|_0 \leq s$, $|\Gamma^{(k)\star}|_0 \leq h$, and $|\Gamma^{(k)\star}|_1 \leq M_\Gamma$. The parameters s and h satisfy $s \gg h$, indicating that the majority of the structure is shared, while unique components are minimal.

See Figure 2.1 for a visual illustration of Assumption 1. Assumption 1 is inspired by the assumptions widely used in the differential network estimation literature [Zhao et al., 2014]. To see this connection, for each $1 \leq k \leq K$, define $\Psi^{(k)} = \Omega^{(k)} - \Omega^{(0)}$ to be the

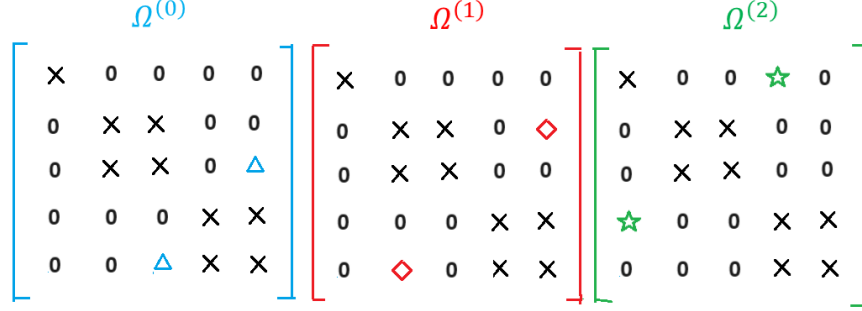


Figure 2.1: Illustration of Assumption 1. The target precision matrix, $\Omega^{(0)}$, is shown alongside two source precision matrices, $\Omega^{(1)}$ and $\Omega^{(2)}$. Black crosses represent the shared entries across the matrices, while colored shapes indicate individual, unique entries.

differential network between the target $\Omega^{(0)}$ and the source $\Omega^{(k)}$. Two immediate implications of Assumption 1 are

$$\left| \Psi^{(k)} \right|_0 \leq 2h, \quad \text{and} \quad \left| \Psi^{(k)} \right|_1 \leq 2M_\Gamma, \quad \text{for all } 1 \leq k \leq K.$$

This matches exactly Condition 1 in the paper Zhao et al. [2014].

In addition, Assumption 1 is naturally interpretable within Gaussian graphical models. Suppose we have an undirected graphical model $G = (V, E)$ where nodes represent variables and edges represent conditional dependencies. In this model, an edge exists between nodes i and j if and only if $\Omega_{ij}^{(k)} \neq 0$ [Lauritzen, 1996]. Under Assumption 1, the target precision matrix $\Omega^{(0)}$ and the source matrices $\{\Omega^{(k)}\}$ share a large subset of edges, with only a small number of unique edges in each source, corresponding to sparse deviations $\Psi^{(k)}$.

2.2.1 Comparison of Similarity Assumptions

In this section, we provide a detailed comparison between our structural assumption (Assumption 1) and the similarity assumption based on divergence measures proposed in Li et al. [2023a]. Specifically, Li et al. [2023a] introduce the divergence matrix $\Upsilon^{(k)} = \Omega^{(0)}\Sigma^{(k)} - I_d$ and assume that $\Upsilon^{(k)}$ is column-wise sparse under the L_p -norm. Equivalently, this condition can be rewritten as $\Psi^{(k)} = -\Upsilon^{(k)}\Omega^{(k)}$. While both our assumption and that of Li et al.

[2023a] aim to capture the similarity between the target and source precision matrices, they are not directly comparable and do not imply one another. Nonetheless, our structural assumption offers a more intuitive interpretation of shared structure and applies more broadly beyond Gaussian graphical models, as it does not rely on distributional assumptions specific to Gaussian data. In the following, we present several comparisons to further clarify the differences between the two assumptions.

First, while $\Upsilon^{(k)}$ is motivated by the Kullback–Leibler divergence between Gaussian distributions, Assumption 1 is structural and applies to general distributions. Second, our assumption admits a more natural interpretation within Gaussian graphical models, whereas the similarity condition in Li et al. [2023a] is less directly connected to the graphical model structure. Third, a technical advantage of our assumption lies in its flexibility in characterizing sparsity relationships. In particular, sparsity in $\Psi^{(k)}$ does not necessarily imply sparsity in $\Upsilon^{(k)}$, nor vice versa; however, additional sparsity assumptions on either $\Sigma^{(k)}$ or $\Omega^{(k)}$ can help bridge this gap. Specifically, we observe that

$$\left| \Psi^{(k)} \right|_0 \leq \sum_{i,j} \mathbb{1} \left\{ \sum_{l=1}^d \mathbb{1} \left\{ \Upsilon_{il}^{(k)} \neq 0 \right\} \mathbb{1} \left\{ \Omega_{lj}^{(k)} \neq 0 \right\} \geq 1 \right\}.$$

This relationship suggests that if both $\Upsilon^{(k)}$ and $\Omega^{(k)}$ are sparse, then $\Psi^{(k)}$ is also sparse. Similarly, if both $\Psi^{(k)}$ and $\Sigma^{(k)}$ are sparse, then $\Upsilon^{(k)}$ must be sparse. The key distinction is that sparsity in $\Upsilon^{(k)}$ implies sparsity in $\Psi^{(k)}$ only when $\Omega^{(k)}$ is sparse, whereas sparsity in $\Psi^{(k)}$ implies sparsity in $\Upsilon^{(k)}$ only when $\Sigma^{(k)}$ is sparse. In graphical models, sparsity assumptions on the precision matrix $\Omega^{(k)}$ are more commonly imposed than those on the covariance matrix $\Sigma^{(k)}$, making sparsity in $\Psi^{(k)}$ a weaker and often more natural assumption than sparsity in $\Upsilon^{(k)}$.

Nevertheless, we do not claim that Trans-Glasso consistently outperforms Trans-CLIME, the method proposed in Li et al. [2023a], and leave a full theoretical comparison of the two

approaches for future research.

2.3 Trans-Glasso Algorithm

In this section, we introduce Trans-Glasso, a transfer learning method for precision matrix estimation. It comprises two steps: (1) initializing precision matrix estimators via a multi-task learning problem and (2) refining them with differential network estimation to account for structural differences between target and source matrices.

2.3.1 Initialization via Multi-Task Learning

To leverage shared structure across the target and source matrices, we begin by jointly estimating precision matrices for both the target and source distributions. Based on Assumption 1, we employ a multi-task variant of the graphical lasso estimator [Friedman et al., 2008], which we refer to as Trans-MT-Glasso (**T**ransfer **M**ulti-**T**ask **G**raphical **lasso**).

Let $\widehat{\Sigma}^{(k)} = \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbf{x}_i^{(k)} \mathbf{x}_i^{(k)\top}$ denote the sample covariance matrix for $0 \leq k \leq K$. We define $\Theta := \left(\Omega, \left\{ \Gamma^{(k)} \right\}_{k=0}^K \right)$ for the shared component and the sparse unique components. The Trans-MT-Glasso objective is then given by

$$\widehat{\Theta} = \left(\widehat{\Omega}, \left\{ \widehat{\Gamma}^{(k)} \right\}_{k=0}^K \right) \in \arg \min_{\Theta \in \mathcal{C}(M_{\text{op}})} \{ \mathcal{L}(\Theta) + \lambda_M \Phi(\Theta) \}, \quad (2.2)$$

where

$$\mathcal{L}(\Theta) := \sum_{k=0}^K \alpha_k \left\{ \left\langle \Omega + \Gamma^{(k)}, \widehat{\Sigma}^{(k)} \right\rangle - \log \det \left(\Omega + \Gamma^{(k)} \right) \right\}, \quad (2.3)$$

$$\Phi(\Theta) := |\Omega|_1 + \sum_{k=0}^K \sqrt{\alpha_k} \left\| \Gamma^{(k)} \right\|_1, \quad \text{and} \quad (2.4)$$

$$\mathcal{C}(M_{\text{op}}) := \left\{ \Theta = \left(\Omega, \left\{ \Gamma^{(k)} \right\}_{k=0}^K \right) : \Omega + \Gamma^{(k)} \succ 0 \text{ and } \max_{0 \leq k \leq K} \left\| \Omega + \Gamma^{(k)} \right\|_2 \leq M_{\text{op}} \right\}. \quad (2.5)$$

Here, $\alpha_k = n_k/N$ with $N = \sum_{k=0}^K n_k$ controls the contribution of each source, $\lambda_M > 0$ is a regularization parameter, and $M_{\text{op}} > 0$ is a predefined constant. The constraint $\|\Omega + \Gamma^{(k)}\|_2 \leq M_{\text{op}}$ is primarily included to facilitate theoretical analysis. While we theoretically need to set M_{op} to be a sufficiently large constant, as detailed in a later section, in practice, we simply set $M_{\text{op}} = \infty$ to effectively remove this constraint.

The first term of (2.2) measures parameter fitness with observed data, while the second term $\Phi(\Theta)$ promotes sparsity in both the shared and individual components. The sparsity penalization level for $\Gamma^{(k)}$ is proportional to $\sqrt{n_k/N}$, a factor that is crucial to balancing contributions from the target and sources. See Section 2.5.1 for details.

In the end, we construct the initial estimators of $\Omega^{(k)}$ as $\check{\Omega}^{(k)} = \hat{\Omega} + \hat{\Gamma}^{(k)}$ for $0 \leq k \leq K$.

2.3.2 Refinement via Differential Network Estimation

To further enhance accuracy, we refine these initial estimators by estimating the differential networks $\Psi^{(k)} = \Omega^{(k)} - \Omega^{(0)}$, which capture the structural differences between each source and the target. This refinement step corrects for potential biases in the initial estimates.

Estimating $\Psi^{(k)}$ has been extensively studied in the differential network estimation literature [Zhao et al., 2014, Yuan et al., 2017], with a variety of good estimators. For instance, Yuan et al. [2017] proposed estimation by solving

$$\hat{\Psi}^{(k)} \in \arg \min_{\Psi} L_D \left(\Psi; \hat{\Sigma}^{(0)}, \hat{\Sigma}^{(k)} \right) + \lambda_{\Psi}^{(k)} \|\Psi\|_1, \quad (2.6)$$

where

$$L_D \left(\Psi; \hat{\Sigma}^{(0)}, \hat{\Sigma}^{(k)} \right) = \frac{1}{4} \left(\left\langle \hat{\Sigma}^{(0)} \Psi, \Psi \hat{\Sigma}^{(k)} \right\rangle + \left\langle \hat{\Sigma}^{(k)} \Psi, \Psi \hat{\Sigma}^{(0)} \right\rangle \right) - \left\langle \Psi, \hat{\Sigma}^{(0)} - \hat{\Sigma}^{(k)} \right\rangle,$$

and $\lambda_{\Psi}^{(k)}$ is a tuning parameter.

For our purpose, any reasonable differential network estimator can correct for the bias.

Thus, we treat differential network estimation as a black-box algorithm, and we obtain the estimates $\widehat{\Psi}^{(k)}$.

With the initial estimator $\check{\Omega}^{(k)}$ and refined differential network estimators $\widehat{\Psi}^{(k)}$, we construct the final transfer learning estimator for $\Omega^{(0)}$ as

$$\widehat{\Omega}^{(0)} = \sum_{k=0}^K \alpha_k \left(\check{\Omega}^{(k)} - \widehat{\Psi}^{(k)} \right), \quad (2.7)$$

where $\widehat{\Psi}^{(0)} = 0$ by definition. The final estimator Trans-Glasso (**T**ransfer learning **G**raphical **lasso**) integrates both shared information and source-specific refinements, yielding a transfer learning approach that leverages structural similarities across datasets for improved precision matrix estimation. Compared to Li et al. [2023a], our method is more sample efficient as it does not require sample splitting between the steps.

2.4 Implementation in Practice

In this section, we provide practical guidelines for implementing Trans-Glasso, including optimization techniques, hyperparameter selection, and a method for identifying the informative set when not all sources are useful.

2.4.1 Optimization Algorithms

To implement Trans-Glasso, we first solve the Trans-MT-Glasso objective from Equation (2.2) to obtain initial estimates for Ω^* and $\Gamma^{(k)*}$. This is a constrained optimization problem that can be efficiently solved using the Alternating Direction Method of Multipliers (ADMM) [Boyd et al., 2011]. In this section, we slightly abuse notation by using superscripts to denote the iteration round and subscripts to represent the population.

Define

$$X = \begin{bmatrix} \Omega_0 \\ \Omega_1 \\ \vdots \\ \Omega_K \end{bmatrix}, \quad Y = \begin{bmatrix} \Omega \\ \Gamma_0 \\ \Gamma_1 \\ \vdots \\ \Gamma_K \end{bmatrix}, \quad \text{and} \quad B = \begin{bmatrix} I_d & I_d & 0 & \cdots & 0 \\ I_d & 0 & I_d & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ I_d & 0 & 0 & \cdots & I_d \end{bmatrix},$$

where $X \in \mathbb{R}^{(K+1)d \times d}$, $Y \in \mathbb{R}^{(K+2)d \times d}$, $B \in \mathbb{R}^{(K+1)d \times (K+2)d}$. With this notation, the objective in Equation (2.2) can be reformulated as

$$\text{minimize } f(X) + g(Y), \quad \text{subject to } X = BY,$$

where

$$f(X) = \sum_{k=0}^K f_k(\Omega_k), \quad f_k(\Omega_k) = \alpha_k \left\{ -\log \det(\Omega_k) + \left\langle \widehat{\Sigma}^{(k)}, \Omega_k \right\rangle \right\} + \mathbb{I}(\Omega_k \succ 0), \quad 0 \leq k \leq K,$$

$$g(Y) = \lambda_M |\Omega|_1 + \lambda_M \sum_{k=0}^K \sqrt{\alpha_k} |\Gamma_k|_1,$$

where $\mathbb{I}(\Omega \succ 0) = 0$ if $\Omega \succ 0$ and $\mathbb{I}(\Omega \succ 0) = \infty$ otherwise.

The augmented Lagrangian for this problem is:

$$L_\rho(X, Y, Z) = f(X) + g(Y) + \rho \langle Z, X - BY \rangle + \frac{\rho}{2} \|X - BY\|_F^2,$$

where $Z^\top = [Z_0^\top, Z_1^\top, \dots, Z_K^\top] \in \mathbb{R}^{d \times (K+1)d}$ is a dual variable and ρ is a penalty parameter.

After initializing $Y^{(0)}$ and $Z^{(0)}$ such that $\Omega^{(0)}$, $\Gamma_k^{(0)}$, and $Z_k^{(0)}$ are symmetric, ADMM

iteratively updates:

$$X^{(t)} = \arg \min_X L_\rho \left(X, Y^{(t-1)}, Z^{(t-1)} \right), \quad (2.8)$$

$$Y^{(t)} \in \arg \min_Y L_\rho \left(X^{(t)}, Y, Z^{(t-1)} \right), \quad (2.9)$$

$$Z^{(t)} = Z^{(t-1)} + \rho \left(X^{(t)} - BY^{(t)} \right). \quad (2.10)$$

Note that (2.10) is equivalent to $Z_k^{(t)} = Z_k^{(t-1)} + \rho \left(\Omega_k^{(t)} - \Omega^{(t)} - \Gamma_k^{(t)} \right)$ for $0 \leq k \leq K$. We then discuss how to compute the updating steps (2.8) and (2.9).

Solving (2.8) is equivalent to solving

$$\begin{aligned} \Omega_k^{(t)} &= \arg \min_{\Omega_k \succ 0} \left\{ f_k(\Omega_k) + \rho \left\langle Z_k^{(t-1)}, \Omega_k - \Omega^{(t-1)} - \Gamma_k^{(t-1)} \right\rangle \right. \\ &\quad \left. + \frac{\rho}{2} \left\| \Omega_k - \Omega^{(t-1)} - \Gamma_k^{(t-1)} \right\|_F^2 \right\} \\ &= \arg \min_{\Omega_k \succ 0} \left\{ -\alpha_k \log \det(\Omega_k) + \frac{\rho}{2} \left\| \Omega_k - \tilde{C}_k^{(t-1)} \right\|_F^2 \right\}, \end{aligned} \quad (2.11)$$

where $\tilde{C}_k^{(t-1)} = -Z_k^{(t-1)} + \left(\Omega^{(t-1)} + \Gamma_k^{(t-1)} \right) - (\alpha_k/\rho) \hat{\Sigma}^{(k)}$ for $0 \leq k \leq K$. Taking the gradient with respect to Ω_k in (2.11) and setting it to zero gives

$$-\alpha_k \Omega_k^{-1} + \rho \Omega_k - \rho \tilde{C}_k^{(t-1)} = 0. \quad (2.12)$$

The matrix $\Omega_k^{(t)}$ is obtained by finding $\Omega_k \succ 0$ that satisfies (2.12). Let $\rho \tilde{C}_k^{(t-1)} = U \Lambda U^\top$, $\Lambda = \text{diag} \left(\{\lambda_i\}_{i=1}^d \right)$, be the eigenvalue decomposition of $\rho \tilde{C}_k^{(t-1)}$. Following [Boyd et al., 2011, Section 6.5], we have

$$\Omega_k^{(t)} = U \text{diag} \left(\left\{ \frac{\lambda_i + \sqrt{\lambda_i^2 + 4\rho\alpha_k}}{2\rho} \right\}_{i=1}^d \right) U^\top.$$

On the other hand, computing (2.9) is equivalent to solving

$$\begin{aligned} \Omega^{(t)}, \{\Gamma_k^{(t)}\} &\in \arg \min_{\Omega, \{\Gamma^{(k)}\}} \left\{ \rho \sum_{k=0}^K \left\langle Z_k^{(t-1)}, \Omega_k^{(t)} - \Omega - \Gamma^{(k)} \right\rangle + \frac{\rho}{2} \sum_{k=0}^K \left\| \Omega_k^{(t)} - \Omega - \Gamma^{(k)} \right\|_F^2 \right. \\ &= \arg \min_{\Omega, \{\Gamma^{(k)}\}} \left\{ \frac{\rho}{2} \sum_{k=0}^K \left\| \Omega + \Gamma^{(k)} - \check{C}_k^{(t)} \right\|_F^2 + \lambda_M |\Omega|_1 + \lambda_M \sum_{k=0}^K \sqrt{\alpha_k} \left| \Gamma^{(k)} \right|_1 \right\} \end{aligned}$$

where $\check{C}_k^{(t)} = \Omega_k^{(t)} + Z_k^{(t-1)}$. Given $c = (c_1, \dots, c_K)$, let $\mathcal{S}(c)$ be defined as $\mathcal{S}(c) = (x^*, y^*)$

where

$$(x^*, y^*) \in \arg \min_{(x, y)} \left\{ \frac{\rho}{2} \sum_{k=0}^K (x + y_k - c_k)^2 + \lambda_M |x| + \lambda_M \sum_{k=0}^K \sqrt{\alpha_k} |y_k| \right\} \quad (2.13)$$

with $y = (y_1, \dots, y_K)$. With $\Omega^{(t)} = (\Omega_{jl}^{(t)})_{1 \leq j, l \leq d}$, $\Gamma_k^{(t)} = (\Gamma_{k,jl}^{(t)})_{1 \leq j, l \leq d}$, and $\check{C}_k^{(t)} = (\check{C}_{k,jl}^{(t)})_{1 \leq j, l \leq d}$, we have

$$\Omega_{jl}^{(t)}, \{\Gamma_{k,jl}^{(t)}\}_{k=1}^K = \mathcal{S} \left(\left\{ \check{C}_{k,jl}^{(t)} \right\}_{k=1}^K \right), \quad \text{for } 1 \leq j, l \leq d.$$

To solve (2.13), we iteratively update x or y while fixing the other until convergence. For $c \in \mathbb{R}$ and $\lambda \geq 0$, let

$$\text{ST}_\lambda(c) = \begin{cases} c - \lambda & \text{if } c > \lambda, \\ 0 & \text{if } |c| \leq \lambda, \\ c + \lambda & \text{if } c < -\lambda \end{cases} \quad (2.14)$$

be the soft-thresholding function. After initializing $x^{(0)}, y^{(0)}$, we repeat the following process

until convergence:

$$\begin{aligned} x^{(r)} &= \arg \min_x \left\{ \frac{\rho}{2} \sum_{k=0}^K \left(x + y_k^{(r-1)} - c_k \right)^2 + \lambda_M |x| + \lambda_M \sum_{k=0}^K \sqrt{\alpha_k} |y_k^{(r-1)}| \right\} \\ &= \text{ST}_{\lambda_M/(\rho(K+1))} \left(\frac{1}{K+1} \sum_{k=0}^K \left(c_k - y_k^{(r-1)} \right) \right), \end{aligned}$$

and

$$\begin{aligned} y^{(r)} &= \arg \min_y \left\{ \frac{\rho}{2} \sum_{k=0}^K \left(x^{(r)} + y_k - c_k \right)^2 + \lambda_M |x^{(r)}| + \lambda_M \sum_{k=0}^K \sqrt{\alpha_k} |y_k| \right\} \\ \Leftrightarrow y_k^{(r)} &= \text{ST}_{\lambda_M \sqrt{\alpha_k}/\rho} \left(c_k - x^{(r)} \right) \quad \text{for } 0 \leq k \leq K. \end{aligned}$$

We next discuss the stopping criterion for the ADMM algorithm to solve Trans-MT-Glasso objective (2.2) and the stopping criterion for subproblem (2.13).

Stopping criterion for ADMM. Following [Boyd et al., 2011, Section 3.3.1], let $\epsilon^{\text{abs}} > 0$ be an absolute tolerance and $\epsilon^{\text{rel}} > 0$ be a relative tolerance, we then define the feasibility tolerance for primal feasibility condition $\epsilon^{\text{pri}} > 0$ and the feasibility tolerance for dual feasibility condition $\epsilon^{\text{dual}} > 0$ at iteration t as

$$\begin{aligned} \epsilon^{\text{pri}} &= \epsilon^{\text{abs}} d\sqrt{K+1} + \epsilon^{\text{rel}} \max \left\{ \left(\sum_{k=0}^K \left\| \Omega_k^{(t)} \right\|_{\text{F}}^2 \right)^{\frac{1}{2}}, \left(\sum_{k=0}^K \left\| \Omega^{(t)} + \Gamma_k^{(t)} \right\|_{\text{F}}^2 \right)^{\frac{1}{2}} \right\}, \\ \epsilon^{\text{dual}} &= \epsilon^{\text{abs}} d\sqrt{K+1} + \epsilon^{\text{rel}} \left(\sum_{k=0}^K \left\| Z_k^{(t)} \right\|_{\text{F}}^2 \right)^{\frac{1}{2}}. \end{aligned}$$

Besides, let

$$r^{\text{pri}} = \left(\sum_{k=0}^K \left\| \Omega_k^{(t)} - \left(\Omega^{(t)} + \Gamma_k^{(t)} \right) \right\|_{\text{F}}^2 \right)^{\frac{1}{2}} \quad \text{and}$$

$$r^{\text{dual}} = \rho \left(\sum_{k=0}^K \left\| \left(\Omega^{(t)} + \Gamma_k^{(t)} \right) - \left(\Omega^{(t-1)} + \Gamma_k^{(t-1)} \right) \right\|_{\text{F}}^2 \right)^{\frac{1}{2}}$$

be the primal and dual residuals at iteration t . We then stop the iteration if

$$r^{\text{pri}} \leq \epsilon^{\text{pri}} \quad \text{and} \quad r^{\text{dual}} \leq \epsilon^{\text{dual}}.$$

Stopping criterion for sub problem (2.13). The optimality conditions of problem (2.13) are

$$0 \in \rho \sum_{k=0}^K (x^{\star} + y_k^{\star} - c_k) + \lambda_{\text{M}} \partial |x^{\star}|, \quad (2.15)$$

$$0 \in \rho (x^{\star} + y^{\star} - c_k) + \lambda_{\text{M}} \sqrt{\alpha_k} \partial |y_k^{\star}| \quad \text{for all } k = 0, 1, \dots, K. \quad (2.16)$$

By the definition of $y^{(r)}$, we always have $x^{(r)}, y^{(r)}$ satisfying (2.16). Besides, by the definition of $x^{(r)}$, we have

$$0 \in \rho \sum_{k=0}^K \left(x^{(r)} + y_k^{(r-1)} - c_k \right) + \lambda_{\text{M}} \partial |x^{(r)}|. \quad (2.17)$$

Thus, when

$$\rho \sum_{k=0}^K \left(y_k^{(r)} - y_k^{(r-1)} \right) = 0,$$

Algorithm 1 ADMM for Trans-MT-GLasso

1: **Input:** $\{\widehat{\Sigma}^{(k)}\}_{k=0}^K$, λ_M , ρ , ϵ^{abs} , ϵ^{rel} and $\{\alpha_k\}_{k=0}^K$.

2: **Initialize:** Let $\Omega^{(0)} = I_d$, $\Gamma_k^{(0)} = 0$ and $Z_k^{(0)} = I_d$ for all $0 \leq k \leq K$. Let $r^{\text{pri}} = r^{\text{dual}} = \infty$ and $\epsilon^{\text{pri}} = \epsilon^{\text{dual}} = 0$. Let $t = 0$.

3: **while** $r^{\text{pri}} > \epsilon^{\text{pri}}$ or $r^{\text{dual}} > \epsilon^{\text{dual}}$ **do**

4: $t \leftarrow t + 1$.

5: **for** $k = 0, 1, \dots, K$ **do**

6: Let

$$\tilde{C}_k^{(t-1)} = -Z_k^{(t-1)} + \left(\Omega^{(t-1)} + \Gamma_k^{(t-1)} \right) - (\alpha_k / \rho) \widehat{\Sigma}^{(k)},$$

and compute the eigenvalue decomposition of $\rho \tilde{C}_k^{(t-1)}$ as

$$\rho \tilde{C}_k^{(t-1)} = U \Lambda U^\top, \quad \Lambda = \text{diag} \left(\{\lambda_i\}_{i=1}^d \right).$$

7: Let

$$\Omega_k^{(t)} = U \text{diag} \left(\left\{ \frac{\lambda_i + \sqrt{\lambda_i^2 + 4\rho\alpha_k}}{2\rho} \right\}_{i=1}^d \right) U^\top.$$

8: Let

$$\check{C}_k^{(t)} = \Omega_k^{(t)} + Z_k^{(t-1)}.$$

9: **end for**

10: Solve

$$\Omega_{jl}^{(t)}, \left\{ \Gamma_{k,jl}^{(t)} \right\}_{k=1}^K = \mathcal{S} \left(\left\{ \check{C}_{k,jl}^{(t)} \right\}_{k=1}^K \right) \quad \text{for all } 1 \leq j, l \leq d.$$

by Algorithm 1.

11: Let

$$Z_k^{(t)} = Z_k^{(t-1)} + \rho \left(\Omega_k^{(t)} - \Omega^{(t)} - \Gamma_k^{(t)} \right) \text{ for all } 0 \leq k \leq K.$$

12: Let

$$r^{\text{pri}} = \left(\sum_{k=0}^K \left\| \Omega_k^{(t)} - \left(\Omega^{(t)} + \Gamma_k^{(t)} \right) \right\|_F^2 \right)^{\frac{1}{2}},$$

$$r^{\text{dual}} = \rho \left(\sum_{k=0}^K \left\| \left(\Omega^{(t)} + \Gamma_k^{(t)} \right) - \left(\Omega^{(t-1)} + \Gamma_k^{(t-1)} \right) \right\|_F^2 \right)^{\frac{1}{2}},$$

and

$$\epsilon^{\text{pri}} = \epsilon^{\text{abs}} d \sqrt{K+1} + \epsilon^{\text{rel}} \max \left\{ \left(\sum_{k=0}^K \left\| \Omega_k^{(t)} \right\|_F^2 \right)^{\frac{1}{2}}, \left(\sum_{k=0}^K \left\| \Omega^{(t)} + \Gamma_k^{(t)} \right\|_F^2 \right)^{\frac{1}{2}} \right\},$$

$$\epsilon^{\text{dual}} = \epsilon^{\text{abs}} d \sqrt{K+1} + \epsilon^{\text{rel}} \left(\sum_{k=0}^K \left\| Z_k^{(t)} \right\|_F^2 \right)^{\frac{1}{2}}.$$

13: **end while**

14: **Output:** $\check{\Omega}^{(k)} = \Omega^{(t)} + \Gamma_k^{(t)}$ for all $k = 0, 1, \dots, K$.

Algorithm 2 Solver of sub problem (2.13)

1: **Input:** $c = (c_0, c_1, \dots, c_K)$; Initial $x^{(0)}$ and $\{y_k^{(0)}\}_{k=0}^K$; λ_M , ρ and $\{\alpha_k\}_{k=0}^K$.
2: **Initialize:** $r^{\text{sub}} = \infty$ and $\epsilon^{\text{sub}} = 0$. Define $\text{ST}_\lambda(\cdot)$ as in (2.14). Let $r = 0$.
3: **while** $r^{\text{sub}} > \epsilon^{\text{sub}}$ **do**
4: $r \leftarrow r + 1$.
5: Let

$$x^{(r)} = \text{ST}_{\lambda_M/(\rho(K+1))} \left(\frac{1}{K+1} \sum_{k=0}^K (c_k - y_k^{(r-1)}) \right).$$

6: Let

$$y_k^{(r)} = \text{ST}_{\lambda_M \sqrt{\alpha_k}/\rho} (c_k - x^{(r)}) \quad \text{for all } 0 \leq k \leq K.$$

7: Let

$$r^{\text{sub}} = \rho \left| \sum_{k=0}^K (y_k^{(r)} - y_k^{(r-1)}) \right|,$$

$$\epsilon^{\text{sub}} = (K+1)\epsilon^{\text{abs}} + \epsilon^{\text{rel}} \rho \max \left\{ \sum_{k=0}^K |y_k^{(r)}|, \sum_{k=0}^K |y_k^{(r-1)}| \right\}.$$

8: **end while**

9: **Output:** $x^{(r)}$ and $\{y_k^{(r)}\}_{k=0}^K$.

we will have $x^{(r)}, y^{(r)}$ satisfy (2.17). Let $\epsilon^{\text{abs}} > 0$ be an absolute tolerance and $\epsilon^{\text{rel}} > 0$ be a relative tolerance; we then stop at iteration r if

$$r^{\text{sub}} = \rho \left| \sum_{k=0}^K (y_k^{(r)} - y_k^{(r-1)}) \right| \leq (K+1)\epsilon^{\text{abs}} + \epsilon^{\text{rel}} \max \left\{ \sum_{k=0}^K |y_k^{(r)}|, \sum_{k=0}^K |y_k^{(r-1)}| \right\} := \epsilon^{\text{sub}}.$$

The final ADMM algorithm and the algorithm to solve the subproblem are summarized in Algorithm 1 and Algorithm 2.

To refine the initial estimators, we solve the D-Trace loss objective (2.6) for differential network estimation. Here, we use a different proximal gradient descent algorithm [Parikh and Boyd, 2014] following Zhao et al. [2022, 2019].

For simplicity, let $L_D(\Psi) := L_D(\Psi; \widehat{\Sigma}^{(0)}, \widehat{\Sigma}^{(k)})$ for the chosen k . In iteration t , we

update $\Psi^{(t-1)}$ by solving

$$\Psi^{(t)} = \arg \min_{\Psi} \left\{ \frac{1}{2} \left\| \Psi - \left(\Psi^{(t-1)} - \eta \nabla L_D \left(\Psi^{(t-1)} \right) \right) \right\|_F^2 + \eta \cdot \lambda_{\Psi}^{(k)} \|\Psi\|_1 \right\}, \quad (2.18)$$

where η is a user-specified step size. Note that $\nabla L_D(\cdot)$ is Lipschitz continuous with constant $\|\widehat{\Sigma}^{(0)} \otimes \widehat{\Sigma}^{(k)}\|_2 = \|\widehat{\Sigma}^{(0)}\|_2 \|\widehat{\Sigma}^{(k)}\|_2$. Thus, for $0 < \eta \leq \|\widehat{\Sigma}^{(0)}\|_2^{-1} \|\widehat{\Sigma}^{(k)}\|_2^{-1}$, the proximal gradient method converges [Beck and Teboulle, 2009]. The update in (2.18) has a closed-form solution:

$$\Psi_{jl}^{(t)} = \left[\left| A_{jl}^{(t-1)} \right| - \lambda_{\Psi}^{(k)} \eta \right]_+ \cdot A_{jl}^{(t-1)} / \left| A_{jl}^{(t-1)} \right|, \quad 1 \leq j, l \leq d, \quad (2.19)$$

where $A^{(t-1)} = \Psi^{(t-1)} - \eta \nabla L_D(\Psi^{(t-1)})$ and $x_+ = \max\{0, x\}$, $x \in \mathbb{R}$.

We then discuss the stopping criterion for the proximal gradient method algorithm.

Stopping criterion for proximal gradient method. If $\Gamma^{(t)}$ is the solution for (2.6), the optimization criterion requires that

$$0 \in \nabla L_D(\Gamma^{(t)}) + \lambda_{\Gamma}^{(k)} \cdot \partial \left| \Gamma^{(t)} \right|_1.$$

Note that by the definition of $\Gamma^{(t)}$ in (2.18), we have

$$0 \in \Gamma^{(t)} - \left(\Gamma^{(t-1)} - \eta \nabla L_D \left(\Gamma^{(t-1)} \right) \right) + \eta \lambda_{\Gamma}^{(k)} \cdot \partial \left| \Gamma^{(t)} \right|_1, \quad (2.20)$$

which implies that

$$0 \in \nabla L_D(\Gamma^{(t)}) + \lambda_{\Gamma}^{(k)} \cdot \partial \left| \Gamma^{(t)} \right|_1 + \frac{1}{\eta} \left(\Gamma^{(t)} - \Gamma^{(t-1)} \right) - \left[\nabla L_D \left(\Gamma^{(t)} \right) - \nabla L_D \left(\Gamma^{(t-1)} \right) \right].$$

Algorithm 3 Proximal Gradient Method for D-Trace Loss

1: **Input:** $\widehat{\Sigma}^{(0)}, \widehat{\Sigma}^{(k)}, \lambda_{\Gamma}^{(k)}, \eta, \epsilon^{\text{abs}}$ and ϵ^{rel} .
2: **Initialize:** $\Gamma^{(0)} = 0, r^D = \infty, \epsilon^D = 0$ and $t = 0$.
3: **while** $r^D > \epsilon^D$ **do**
4: $t \leftarrow t + 1$
5: Let

$$\begin{aligned} A^{(t-1)} &= \Gamma^{(t-1)} - \eta \nabla L_D \left(\Gamma^{(t-1)} \right) \\ &= \Gamma^{(t-1)} - \eta \left\{ \frac{1}{2} \left(\widehat{\Sigma}^{(k)} \Gamma^{(t-1)} \widehat{\Sigma}^{(0)} + \widehat{\Sigma}^{(0)} \Gamma^{(t-1)} \widehat{\Sigma}^{(k)} \right) - \left(\widehat{\Sigma}^{(0)} - \widehat{\Sigma}^{(k)} \right) \right\}. \end{aligned}$$

6: Let

$$\Gamma_{jl}^{(t)} = \left[|A_{jl}^{(t-1)}| - \lambda_{\Gamma}^{(k)} \eta \right]_+ \cdot A_{jl}^{(t-1)} / |A_{jl}^{(t-1)}|, \quad 1 \leq j, l \leq d.$$

7: Let

$$r^D = \left\| \frac{1}{\eta} \left(\Gamma^{(t)} - \Gamma^{(t-1)} \right) - \frac{1}{2} \widehat{\Sigma}^{(0)} \left(\Gamma^{(t)} - \Gamma^{(t-1)} \right) \widehat{\Sigma}^{(k)} - \frac{1}{2} \widehat{\Sigma}^{(k)} \left(\Gamma^{(t)} - \Gamma^{(t-1)} \right) \widehat{\Sigma}^{(0)} \right\|_{\text{F}},$$

and

$$\epsilon^D = \epsilon^{\text{abs}} d + \epsilon^{\text{rel}} \times \max \left\{ \frac{1}{\eta} \left\| \Gamma^{(t)} - \Gamma^{(t-1)} \right\|_{\text{F}}, \left\| \frac{1}{2} \widehat{\Sigma}^{(0)} \left(\Gamma^{(t)} - \Gamma^{(t-1)} \right) \widehat{\Sigma}^{(k)} \right\|_{\text{F}} \right\}.$$

8: **end while**

9: **Output:** $\widehat{\Psi}^{(k)} = \Gamma^{(t)}$.

Thus, when

$$\begin{aligned} r^D &= \left\| \frac{1}{\eta} \left(\Gamma^{(t)} - \Gamma^{(t-1)} \right) - \left[\nabla L_D \left(\Gamma^{(t)} \right) - \nabla L_D \left(\Gamma^{(t-1)} \right) \right] \right\|_{\text{F}} \\ &= \left\| \frac{1}{\eta} \left(\Gamma^{(t)} - \Gamma^{(t-1)} \right) - \frac{1}{2} \widehat{\Sigma}^{(0)} \left(\Gamma^{(t)} - \Gamma^{(t-1)} \right) \widehat{\Sigma}^{(k)} - \frac{1}{2} \widehat{\Sigma}^{(k)} \left(\Gamma^{(t)} - \Gamma^{(t-1)} \right) \widehat{\Sigma}^{(0)} \right\|_{\text{F}} \end{aligned}$$

is close to zero, we then have (2.20) hold approximately. Therefore, given an absolute tolerance $\epsilon^{\text{abs}} > 0$ and a relative tolerance $\epsilon^{\text{rel}} > 0$, we then stop at iteration t if

$$r^D \leq \epsilon^{\text{abs}} d + \epsilon^{\text{rel}} \times \max \left\{ \frac{1}{\eta} \left\| \Gamma^{(t)} - \Gamma^{(t-1)} \right\|_{\text{F}}, \left\| \frac{1}{2} \widehat{\Sigma}^{(0)} \left(\Gamma^{(t)} - \Gamma^{(t-1)} \right) \widehat{\Sigma}^{(k)} \right\|_{\text{F}} \right\}.$$

The detailed algorithm is described in Algorithm 3.

2.4.2 Hyperparameter Selection

This section covers the selection of hyperparameters, specifically λ_M in Trans-MT-Glasso (2.2) and $\lambda_\Psi^{(k)}$ in D-Trace loss (2.6).

We choose $\lambda_\Psi^{(k)}$ for $k \in [K]$ to minimize the Bayesian information criterion (BIC) of D-Trace loss:

$$\begin{aligned} \text{BIC}_\Psi^{(k)} = (n_0 + n_k) & \left\| \frac{1}{2} \left(\widehat{\Sigma}^{(0)} \widehat{\Psi}^{(k)} \widehat{\Sigma}^{(k)} + \widehat{\Sigma}^{(k)} \widehat{\Psi}^{(k)} \widehat{\Sigma}^{(0)} \right) - \widehat{\Sigma}^{(0)} + \widehat{\Sigma}^{(k)} \right\|_F \\ & + \log(n_0 + n_k) \cdot \left| \widehat{\Psi}^{(k)} \right|_0, \end{aligned} \quad (2.21)$$

following Yuan et al. [2017]. After selecting $\lambda_\Psi^{(k)}$ and obtaining $\widehat{\Psi}^{(k)}$ for all $k \in [K]$, $\widehat{\Omega}^{(0)}$ depends on λ_M . Recall that $N = \sum_{k=0}^K n_k$. We choose λ_M to minimize the BIC of Trans-Glasso, defined as

$$\text{BIC}_{\text{Trans}} = N \cdot \left[\left\langle \widehat{\Sigma}^{(0)}, \widehat{\Omega}^{(0)} \right\rangle - \log \det \left(\widehat{\Omega}^{(0)} \right) \right] + \log N \cdot \left| \widehat{\Omega}^{(0)} \right|_0. \quad (2.22)$$

2.4.3 Identifying the Informative Set

In practice, it is not necessarily true that all source distributions are structurally similar to the target. We propose a data-driven method to estimate the informative set $\mathcal{A} \subseteq [K]$.

We obtain differential network estimations $\widehat{\Psi}^{(k)}$ for all $k \in [K]$, with the hyperparameter $\lambda_\Psi^{(k)}$ chosen by minimizing the BIC criterion in (2.21). We then rank sources according to the sparsity level of $\widehat{\Psi}^{(k)}$. Let R_k be the rank of the source k . For any $1 \leq k_1, k_2 \leq K$, $|\widehat{\Psi}^{(k_1)}|_0 \leq |\widehat{\Psi}^{(k_2)}|_0$ implies $R_{k_1} \leq R_{k_2}$. After ranking sources, we input samples into Trans-Glasso and determine the number of sources based on the cross-validation (CV) error. For $K_{\text{chosen}} = 0, 1, \dots, K$, we select sources k with $R_k \leq K_{\text{chosen}}$. When $K_{\text{chosen}} = 0$, we obtain $\widehat{\Omega}^{(0)}$ from graphical lasso [Friedman et al., 2008] using the target data alone. We then compute the CV error of K_{chosen} by the following procedure:

- (i) We randomly split the *target samples* into M -fold.
- (ii) For $m = 1, \dots, M$, we select the m -th fold as the validation set, $\mathcal{D}_{\text{val}}$, and the rest as the training set. We input the training set and chosen source samples into Trans-Glasso to obtain $\hat{\Omega}^{(0)}$. We compute the CV error for the m -th fold as

$$\text{CV}_m = \frac{1}{2d} \left\{ \frac{1}{|\mathcal{D}_{\text{val}}|} \sum_{i \in \mathcal{D}_{\text{val}}} \text{tr} \left(\mathbf{x}_i^{(0)} \mathbf{x}_i^{(0)\top} \hat{\Omega}^{(0)} \right) - \log \det \left(\hat{\Omega}^{(0)} \right) \right\} + \frac{1}{2} \log \pi, \quad (2.23)$$

and define $\text{CV}(K_{\text{chosen}}) = \frac{1}{M} \sum_{m=1}^M \text{CV}_m$.

We set the estimated informative set as $\hat{\mathcal{A}} = \{k \in [K] : R_k \leq K_{\text{chosen}}^*\}$, where $K_{\text{chosen}}^* = \arg \min_{k=0,1,\dots,K} \text{CV}(k)$. Source samples from $\hat{\mathcal{A}}$ are used to estimate $\Omega^{(0)}$. If $\hat{\mathcal{A}} = \emptyset$, we obtain $\hat{\Omega}^{(0)}$ via graphical lasso on target samples. We note that sample splitting is not necessary between estimating $\mathcal{A}$ and $\Omega^{(0)}$. This procedure is called Trans-Glasso-CV. We evaluate its empirical performance in Section 2.6 and defer the study of its theoretical properties to future research. In the next section, we establish the theoretical results under the assumption that $\mathcal{A} = [K]$.

2.5 Theoretical Analysis

In this section, we establish theoretical guarantees for the Trans-Glasso algorithm. We begin by analyzing the initial estimation step using Trans-MT-Glasso, followed by an error bound for the complete Trans-Glasso estimator. Finally, we derive a minimax lower bound, demonstrating that Trans-Glasso is minimax optimal in a wide range of parameter regimes.

To simplify the theoretical statements, we assume the following condition throughout this section.

Assumption 2. *Assume that*

$$M_{\Sigma} := \max_{0 \leq k \leq K} \left| \Sigma^{(k)} \right|_{\infty} = O(1) \quad \text{and} \quad M_{\Omega} := \max_{0 \leq k \leq K} \left\| \Omega^{(k)} \right\|_2 = O(1). \quad (2.24)$$

2.5.1 Analysis of Trans-MT-Glasso

We first provide error bounds for the initial multi-task estimation step, Trans-MT-Glasso. This method estimates both the shared precision matrix component $\Omega^\star$ and the deviation matrices $\Gamma^{(k)\star}$ based on the structural similarity outlined in Assumption 1.

The following theorem provides a high probability upper bound on the Frobenius norm error for the Trans-MT-Glasso estimator. Recall that $N = \sum_{k=0}^K n_k$ and let $\bar{n} = N/(K+1)$.

Theorem 1. *Suppose Assumptions 1 and 2 hold. Fix a failure probability $\delta \in (0, 1]$. Suppose that the local sample size is large enough so that*

$$\min_{0 \leq k \leq K} n_k \geq 2 \log(2(K+2)d^2/\delta). \quad (2.25)$$

Set $M_{\text{op}} \geq M_\Omega$ and the penalty parameter λ_M such that

$$\lambda_M \geq 160M_\Sigma \sqrt{\frac{\log(2(K+2)d^2/\delta)}{2N}}. \quad (2.26)$$

Then with probability at least $1 - \delta$, the estimator satisfies

$$\sum_{k=0}^K \alpha_k \|\check{\Omega}^{(k)} - \Omega^{(k)}\|_{\text{F}}^2 \leq \frac{18(s + (K+1)h)\lambda_M^2}{\kappa^2}, \quad (2.27)$$

where $\kappa = (2M_\Omega + M_{\text{op}})^{-2}$.

Note that the loss function defined in (2.3) is not strongly convex with respect to the Euclidean norm. This prevents us from obtaining an error guarantee for the individual precision matrix. Nevertheless, we make a key observation that the loss function exhibits strong convexity with respect to the weighted norm used in (2.27). A detailed proof is in Section 2.9.2.

When we choose $\lambda_M \asymp \sqrt{(\log d)/N}$, the rate shown in Theorem 1 consists of two parts. The first part, which is of the order $(s \log d)/N$, refers to the estimation error of the shared

component. In words, Trans-Glasso uses all the samples to estimate the shared component. The second part, of the order $(Kh \log d)/N$, relates to the estimation error of the individual components, i.e., on average, there are N/K samples to estimate each individual component.

2.5.2 Analysis of Trans-Glasso

After the initial estimates are obtained via Trans-MT-Glasso, the differential network estimation step refines these estimates by isolating the deviations $\Psi^{(k)}$. This yields the final Trans-Glasso estimator $\widehat{\Omega}^{(0)}$.

As discussed in Section 2.3, any differential network estimator can be used in Step 2 for refinement. The differential network estimates $\widehat{\Psi}^{(k)}$ are treated as the result of a black-box algorithm, obeying

$$\left\| \widehat{\Psi}^{(k)} - \Psi^{(k)} \right\|_{\text{F}} \lesssim g_{\text{F}}^{(k)}(n_0, n_k, d, h, M_{\Gamma}, \delta) := g_{\text{F}}^{(k)} \quad (2.28)$$

holds simultaneously for all $k = 1, \dots, K$ with probability at least $1 - \delta$. We now establish a non-asymptotic error bound for this estimator, which combines the initial estimation error with the error from differential network estimation.

Corollary 2. *Let $\widehat{\Omega}^{(0)}$ be the Trans-Glasso estimator obtained in Equation (2.7). Under the same conditions as in Theorem 1, and assuming Equation (2.28) holds for the differential network estimators. Then with probability at least $1 - 2\delta$, one has*

$$\left\| \widehat{\Omega}^{(0)} - \Omega^{(0)} \right\|_{\text{F}}^2 \lesssim \left(\frac{s}{N} + \frac{h}{n} \right) \log(2(K+2)d^2/\delta) + \sum_{k=0}^K \alpha_k g_{\text{F}}^{(k)}.$$

The error rates depend on differential network estimators' performance. Next, we provide specific error rates using the D-Trace loss estimator.

A differential network estimator: D-Trace loss minimization

We characterize $g_F^{(k)}$ in the case when the D-Trace loss estimator in (2.6) is used. We use the tighter D-Trace loss estimator analysis from Zhao et al. [2019, 2022], Tugnait [2023].

We then have the following lemma providing a high-probability error upper bound for the D-Trace loss estimator. The lemma is derived directly from Corollary 28 in Section 2.9.7.

Lemma 3. *Suppose that Assumptions 1-2 hold. Assume that*

$$\min_{0 \leq k \leq K} n_k \gg h^2 \log \left(2(K+1)d^2/\delta \right).$$

Set

$$\lambda_\Psi^{(k)} = C \sqrt{\frac{\log \left(2(K+1)d^2/\delta \right)}{\min\{n_k, n_0\}}} \quad \text{for all } k \in [K]$$

for some large constant $C > 0$. Then

$$\left\| \widehat{\Psi}^{(k)} - \Psi^{(k)} \right\|_F \lesssim \sqrt{h} M_\Gamma \sqrt{\frac{\log \left(2(K+1)d^2/\delta \right)}{\min\{n_k, n_0\}}}$$

holds simultaneously for all $k \in [K]$ with probability at least $1 - \delta$.

By Lemma 3, we have

$$g_F^{(k)} = \sqrt{h} M_\Gamma \sqrt{\frac{\log \left(2(K+1)d^2/\delta \right)}{\min\{n_k, n_0\}}}$$

for the D-Trace loss estimator. Plugging the above results into Corollary 2, we have the following corollary.

Corollary 4. *Let $\widehat{\Omega}^{(0)}$ be obtained by Trans-Glasso (2.7) with the D-Trace loss estimator used in Step 2. Assuming the assumptions in Theorem 1 and Lemma 3 hold. For a given*

$\delta \in (0, 1]$, letting

$$\lambda_M \asymp \sqrt{\frac{\log(2(K+2)d^2/\delta)}{N}}, \quad \lambda_{\Psi}^{(k)} \asymp M_{\Gamma} \sqrt{\frac{\log(2(K+1)d^2/\delta)}{\min\{n_k, n_0\}}} \quad \text{for all } k \in [K],$$

we have that

$$\left\| \widehat{\Omega}^{(0)} - \Omega^{(0)} \right\|_{\text{F}} \lesssim \left(\sqrt{\frac{s}{N}} + (1 + M_{\Gamma}) \sqrt{\frac{h}{\bar{n}}} + M_{\Gamma} \sqrt{\frac{h}{n_0}} \right) \sqrt{\log(2(K+2)d^2/\delta)}$$

holds with probability at least $1 - 2\delta$.

The estimation error consists of three parts: shared component estimation, individual component estimation, and differential network estimation. If $\bar{n} \geq n_0$ and M_{Γ} is bounded by a universal constant, the error scales as $\sqrt{\frac{s \log d}{N}} + \sqrt{\frac{h \log d}{n_0}}$. Note that the rate obtained when using only target samples is on the order of $\sqrt{\frac{s \log d}{n_0}} + \sqrt{\frac{h \log d}{n_0}}$. Thus, when $s \gg h$ and $s/N \ll h/n_0$, transfer learning can significantly reduce the error rate.

2.5.3 Upper Bound for Expected Error

In this section, we develop theoretical guarantees for the expected error. We start with the analysis of the Trans-MT-Glasso. The subsequent theorem provides the expected error measured in the Frobenius norm.

Theorem 5. *Suppose that Assumption 1 and Assumption 2 hold. Assume that $2(K+2) \leq d^{\tau_1}$ and $N \leq d^{\tau_2}$, for some universal constants $\tau_1, \tau_2 > 0$. In addition, assume that*

$$\frac{\log d}{\min_{0 \leq k \leq K} n_k} \lesssim 1.$$

Let $\bar{n} = N/(K+1)$, $M_{\text{op}} \geq M_{\Omega}$ and $M_{\text{op}} = O(1)$. If $\lambda_M \asymp \sqrt{\log d/N}$, we have

$$\mathbb{E} \left[\sum_{k=0}^K \alpha_k \left\| \widehat{\Omega}_k - \Omega^{(k)} \right\|_{\text{F}}^2 \right] \lesssim \left(\frac{s}{N} + \frac{h}{\bar{n}} \right) \log d.$$

Refer to the proof in Section 2.9.3. The rate described in Theorem 5 consists of two parts. The first part, which is of the order $(s \log d)/N$, refers to the estimation of the shared component. The second part, of the order $(h \log d)/\bar{n}$, relates to the estimation of the individual components.

As discussed in Section 2.5.2, the differential network estimate $\widehat{\Psi}^{(k)}$ is considered the result of a black-box algorithm, with the presumption that its estimation errors are appropriately controlled. Let $\mathcal{B}_F(R) := \left\{ A \in \mathbb{R}^{d \times d} : \|A\|_F \leq R \right\}$ be the ball with radius R . For $\tau > 0$ and $0 \leq k \leq K$, we define

$$\widehat{\Psi}_{\text{proj}}^{(k)}(\tau) \in \arg \min_{\Gamma \in \mathcal{B}_F(d\tau)} \left\| \Gamma - \widehat{\Psi}^{(k)} \right\|_F. \quad (2.29)$$

$\widehat{\Psi}_{\text{proj}}^{(k)}(\tau)$ is utilized solely for theoretical reasons. Suppose that by choosing τ appropriately, we have

$$\mathbb{E} \left[\left\| \widehat{\Psi}_{\text{proj}}^{(k)}(\tau) - \Psi^{(k)} \right\|_F^2 \right] \lesssim \bar{g}_F^{(k)}(n_0, n_k, d, h, M_\Gamma) := \bar{g}_F^{(k)} \quad \text{for all } 0 \leq k \leq K. \quad (2.30)$$

We define

$$\widehat{\Omega}_{\text{proj}}^{(0)}(\tau) = \sum_{k=0}^K \alpha_k \left(\check{\Omega}^{(k)} - \widehat{\Psi}_{\text{proj}}^{(k)}(\tau) \right), \quad (2.31)$$

$\widehat{\Psi}_{\text{proj}}^{(k)}(\tau)$ is defined in (2.29). Combining (2.30) with Theorem 5 yields the following corollary with the expected error upper bound for Trans-Glasso.

Corollary 6. *If the conditions of Theorem 5 are satisfied and τ is chosen so that (2.30) holds, we have*

$$\mathbb{E} \left[\left\| \widehat{\Omega}_{\text{proj}}^{(0)}(\tau) - \Omega^{(0)} \right\|_F^2 \right] \lesssim \left(\frac{s}{N} + \frac{h}{\bar{n}} \right) \log d + \sum_{k=0}^K \alpha_k \bar{g}_F^{(k)}.$$

We then characterize $\bar{g}_F^{(k)}$ in the case where the D-Trace loss estimator in (2.6) is used.

Recall that $\widehat{\Psi}_{\text{proj}}^{(k)}(\tau)$ is defined by (2.30).

Theorem 7. *Suppose that Assumption 1 and Assumption 2 hold. Assume that $2(K+1) \leq d^{\tau_1}$, $N \leq d^{\tau_2}$ and $M_\Gamma \leq d^{\tau_3}$ for some universal constants $\tau_1, \tau_2, \tau_3 > 0$. Let*

$$\lambda_\Gamma^{(k)} \asymp M_\Gamma \sqrt{\frac{\log d}{\min\{n_k, n_0\}}}.$$

Then, for any $\tau \geq \tau_3$, we have

$$\mathbb{E} \left[\left\| \widehat{\Psi}_{\text{proj}}^{(k)}(\tau) - \Psi^{(k)} \right\|_F^2 \right] \lesssim \frac{M_\Gamma^2 h \log d}{\min\{n_k, n_0\}}.$$

Proof. Note that $\|A\|_F \leq |A|_1$ for any matrix A . Thus we have $\|\Psi^{(k)}\|_F \leq \|\Psi^{(k)}\|_1 \leq 2M_\Gamma$.

The rest of the proof is similar to the proof of Theorem 5. $\square$

By Theorem 7, we have

$$\bar{g}_F^{(k)} = \frac{M_\Gamma^2 h \log d}{\min\{n_k, n_0\}}$$

for D-Trace loss estimator. Plugging the above results into Corollary 6, we then have the following corollary.

Corollary 8. *Let $\widehat{\Omega}^{(0)}$ be obtained by Trans-Glasso (2.7) with the D-Trace loss estimator used in Step 1. Suppose that Assumption 1 and Assumption 2 hold, and that the conditions in Theorem 5 and Theorem 7 are satisfied. If $2(K+2) \leq d^{\tau_1}$, $N \leq d^{\tau_2}$, $M_\Gamma \leq d^{\tau_3}$ for some universal constants $\tau_1, \tau_2, \tau_3 > 0$, and*

$$\lambda_M \asymp \sqrt{\frac{\log d}{N}}, \quad \lambda_\Gamma^{(k)} \asymp M_\Gamma \sqrt{\frac{\log d}{\min\{n_k, n_0\}}} \quad \text{for all } k \in [K],$$

then for any $\tau \geq \tau_3$, we have

$$\mathbb{E} \left[\left\| \widehat{\Omega}_{\text{proj}}^{(0)}(\tau) - \Omega^{(0)} \right\|_F^2 \right] \lesssim \left(\frac{s}{N} + (1 + M_\Gamma^2) \cdot \frac{h}{n} + M_\Gamma^2 \cdot \frac{h}{n_0} \right) \log d. \quad (2.32)$$

The estimation error in (2.32) is comprised of three parts: shared component estimation, individual component estimation, and differential network estimation. If $\bar{n} \geq n_0$ and M_Γ is bounded by a universal constant, the error scales as $\frac{s \log d}{N} + \frac{h \log d}{n_0}$. When using only target samples, the lowest error rate achievable is $\frac{(s+h) \log d}{n_0}$ as stated later in Lemma 9. Therefore, if $N \gg n_0$, the error rate can be significantly reduced compared to the optimal rate obtained with only the target samples. Moreover, as demonstrated in the next section, the rate $\frac{s \log d}{N} + \frac{h \log d}{n_0}$ is minimax optimal under certain conditions.

2.5.4 Minimax Lower Bounds and Optimality

To evaluate the theoretical performance of Trans-Glasso, we derive the minimax lower bound for estimating the target precision matrix $\Omega^{(0)}$ over the parameter space defined by Assumptions 1–2. More precisely, we define the relevant parameter space:

$$\begin{aligned} \mathcal{G}(s, h) := & \left\{ \left\{ \Omega^{(k)} \right\}_{k=0}^K : \Omega^{(k)} \succ 0, \Omega^{(k)} = \Omega^\star + \Gamma^{(k)\star}, \text{supp}(\Omega^\star) \cap \text{supp}(\Gamma^{(k)\star}) = \emptyset \forall k, \right. \\ & |\Omega^\star|_0 \leq s, \max_{0 \leq k \leq K} |\Gamma^{(k)\star}|_0 \leq h, \\ & \max_{0 \leq k \leq K} |\Gamma^{(k)\star}|_1 \leq M_\Gamma, \max_{0 \leq k \leq K} \|\Omega^{(k)}\|_2 \leq M_\Omega, \\ & \left. \max_{0 \leq k \leq K} |\Sigma^{(k)}|_\infty \leq M_\Sigma \right\}, \end{aligned} \tag{2.33}$$

where $M_\Gamma > 0$, $M_\Omega, M_\Sigma > 1$ are universal constants.

Intuitively, the performance limit of any transfer learning estimator is dictated by the information-theoretic lower bounds of estimating two parts, namely the shared component and the individual component. Hence, to derive the minimax lower bound for the transfer learning estimator, we need to provide lower bounds for estimating these two parts.

Lower bound for estimating shared component. The following theorem provides the minimax lower bound for estimating the shared component when all the distributions are the same.

Lemma 9. Assume that we have n i.i.d. samples $X_1, \dots, X_n$ from $N(0, \Omega^{-1})$, where

$$\Omega \in \mathcal{G}_1 = \left\{ \Omega \in \mathbb{S}^{d \times d} : \Omega \succ 0, |\Omega|_0 \leq s, 0 < c_1 \leq \gamma_{\min}(\Omega) \leq \gamma_{\max}(\Omega) \leq c_2 < \infty \right\},$$

and c_1, c_2 are universal constants. In addition, assume that $s \geq d \geq c' n^\beta$ for some universal constants $\beta > 1$ and $c' > 0$, and

$$\lceil s/d \rceil = o\left(\frac{n}{(\log d)^{\frac{3}{2}}}\right).$$

We then have

$$\inf_{\hat{\Omega}} \sup_{\Omega \in \mathcal{G}_1} \mathbb{E} \left[\left\| \hat{\Omega} - \Omega \right\|_F^2 \right] \gtrsim \frac{s \log d}{n}.$$

The proof is based on [Cai et al., 2016, Theorem 6.1]. See Section 2.9.4 for more details. By Lemma 9, it is easy to see that the squared Frobenius error of estimating the shared component is lower bounded by $s \log d / N$ where N is the total number of samples.

Lower bound for estimating the individual components. We also provide a lower bound for estimating the individual component. When $\Omega^{(k)} = I_d$ for all $1 \leq k \leq K$ and $\Omega^{(0)} = I_d + \Delta$ with $\text{diag}(\Delta) = 0$, the source samples are not helpful to estimate $\Omega^{(0)}$ at all. Thus, the minimax lower bound for estimating Δ provides a valid minimax lower bound for the transfer learning problem.

Lemma 10. Assume that we have n i.i.d. samples $X_1, \dots, X_n$ from $N(0, \Omega^{-1})$, where

$$\begin{aligned} \Omega \in \mathcal{G}_2 = \left\{ \Omega \in \mathbb{S}^{d \times d} : \Omega \succ 0, \Omega = I_d + \Delta, \Delta_{jj} = 0 \text{ for all } 1 \leq j \leq d, \right. \\ \left. |\Delta|_0 \leq h, |\Delta|_1 \leq C_\Gamma, 0 < c_1 \leq \gamma_{\min}(\Omega) \leq \gamma_{\max}(\Omega) \leq c_2 < \infty \right\}, \end{aligned} \quad (2.34)$$

where $C_\Gamma > 0$, $c_1 < 1$ and $c_2 > 1$ are constants. In addition, assume that

$$d \geq 4h, h \log d \geq 8 \log 3, \frac{h \log d}{n} \leq \min \left\{ 2, 8(1 - c_1)^2, 8(1 - c_2)^2 \right\}, h \sqrt{\frac{\log d}{n}} \leq 4C_\Gamma. \quad (2.35)$$

We then have

$$\inf_{\hat{\Omega}} \sup_{\Omega \in \mathcal{G}_2} \mathbb{E} \left[\left\| \hat{\Omega} - \Omega \right\|_F^2 \right] \gtrsim \frac{h \log d}{n}.$$

See Section 2.9.5 for the proof, which relies on a novel construction of the packing set of the parameter space and on the celebrated Fano's method [Wainwright, 2019, Section 15.3].

Combining pieces together. Combining Lemma 9 and Lemma 10, we have the following lower bound for any transfer learning estimator.

Theorem 11. *Suppose that we have n_k i.i.d. samples from a sub-Gaussian distribution $\mathcal{P}_k$ with zero mean and precision matrix $\Omega^{(k)}$ for all $0 \leq k \leq K$. Besides, assume that $s \geq d \geq c' N^\beta$ for some universal constants $\beta > 1$, $c' > 0$ and*

$$\lceil s/d \rceil = o \left(\frac{N}{(\log d)^{\frac{3}{2}}} \right).$$

In addition, assume that

$$d \geq 4h, h \log d \geq 8 \log 3, h \sqrt{\frac{\log d}{n_0}} \leq 4M_\Gamma, \frac{h \log d}{n_0} \leq \min \left\{ 2, 8(1 - M_\Omega)^2, 8 \left(1 - \frac{1}{M_\Sigma} \right)^2 \right\},$$

where $M_\Gamma > 0$, $M_\Omega, M_\Sigma > 1$ are universal constants defined in (2.33). We then have

$$\inf_{\hat{\Omega}} \sup_{\{\Omega^{(k)}\}_{k=0}^K \in \mathcal{G}(s, h)} \mathbb{E} \left[\left\| \hat{\Omega} - \Omega^{(0)} \right\|_F^2 \right] \gtrsim \frac{s \log d}{N} + \frac{h \log d}{n_0}.$$

Theorem 11 demonstrates that Trans-Glasso achieves minimax optimality for the parameter space specified in (2.33) when $\bar{n} \geq n_0$. The obtained minimax optimal rate is reasonable

given N samples for estimating the shared component with s non-zero entries and n_0 samples for estimating the individual component with h non-zero entries. Furthermore, from a practical viewpoint, the rate suggests that the target sample size only needs to be sufficiently large in relation to the sparsity level h of the individual component. In contrast, if we only have target samples, the target sample size needs to be large enough to match the sparsity level $s + h$ of the entire precision matrix, which can be significantly larger.

2.5.5 Minimax Optimal Rate for Differential Network Estimation

Differential network estimation aims to estimate the difference between two precision matrices without first estimating the individual precision matrices. Although existing studies focus on providing upper bounds for this problem [Zhao et al., 2014, Yuan et al., 2017, Zhao et al., 2022], there is no known matching lower bound, making the minimax optimal rate an open question. As a byproduct of our analysis, in this section, we provide a minimax optimal rate for the differential network estimation problem under certain conditions. To the best of our knowledge, this is the first minimax optimal guarantee towards this direction.

We start by formulating the problem setup. Note that we reintroduce some notation used in this section to make it self-contained, and one should not confuse it with the notation used in the other parts of the chapter. Suppose that we have n_X i.i.d. samples $X_1, \dots, X_{n_X} \sim N(0, \Omega_X^{-1})$ and n_Y i.i.d. samples $Y_1, \dots, Y_{n_Y} \sim N(0, \Omega_Y^{-1})$. Let $\Delta := \Omega_Y - \Omega_X$ be the differential network between Ω_X and Ω_Y . Our goal is to use samples from two populations to estimate Δ . In addition, we assume that Ω_X and Ω_Y belong to the following parameter space.

$$(\Omega_X, \Omega_Y) \in \mathcal{M} := \left\{ \Omega_X, \Omega_Y \in \mathbb{S}^{d \times d} : c_1 \leq \gamma_{\min}(\Omega_X) \leq \gamma_{\max}(\Omega_X) \leq c_2, \right. \\ \left. c_1 \leq \gamma_{\min}(\Omega_Y) \leq \gamma_{\max}(\Omega_Y) \leq c_2, |\Delta|_0 \leq h, |\Delta|_1 \leq C_\Gamma \right\}, \quad (2.36)$$

where $c_1, c_2, C_\Gamma > 0$ are positive universal constants. The parameter space defined in (2.36)

requires that the smallest eigenvalues and the largest eigenvalues of Ω_X and Ω_Y are lower bounded and upper bounded, respectively. Besides, we also require that the sparsity level of the differential network is bounded by h , and the L_1 norm of the differential network is bounded by a universal constant.

We first present the minimax lower bound for the above estimation problem. We state the result in the following theorem.

Theorem 12. *Assume that*

$$d \geq 4h, h \log d \geq 8 \log 3, \frac{h \log d}{n} \leq \min \left\{ 2, 8(1 - c_1)^2, 8(1 - c_2)^2 \right\}, h \sqrt{\frac{\log d}{n}} \leq 4C_\Gamma. \quad (2.37)$$

We then have

$$\min_{\hat{\Delta}} \sup_{(\Omega_X, \Omega_Y) \in \mathcal{M}} \mathbb{E} \left[\left\| \hat{\Delta} - \Delta \right\|_F^2 \right] \gtrsim \frac{h \log d}{\min \{n_X, n_Y\}}$$

Proof. Let $\Omega_X = I_d$, note that

$$(I_d, \Omega_Y) \in \mathcal{M} \Leftarrow \Omega_Y \in \mathcal{G}_2,$$

where $\mathcal{G}_2$ is defined in (2.34). Now that samples from population X are useless for estimating Ω_Y , and we can only rely on samples from population Y to estimate Ω_Y . By Lemma 10, we then have

$$\min_{\hat{\Delta}} \sup_{(\Omega_X, \Omega_Y) \in \mathcal{M}} \mathbb{E} \left[\left\| \hat{\Delta} - \Delta \right\|_F^2 \right] \geq \min_{\hat{\Omega}_Y} \sup_{\Omega_Y \in \mathcal{G}_2} \mathbb{E} \left[\left\| \hat{\Omega}_Y - \Omega_Y \right\|_F^2 \right] \gtrsim \frac{h \log d}{n_Y}.$$

Similarly, we can show that

$$\min_{\hat{\Delta}} \sup_{(\Omega_X, \Omega_Y) \in \mathcal{M}} \mathbb{E} \left[\left\| \hat{\Delta} - \Delta \right\|_F^2 \right] \gtrsim \frac{h \log d}{n_X}.$$

Combine the above two inequalities, we have the final result. $\square$

Next, we derive the matching upper bound. Let $\widehat{\Delta}$ be the D-Trace loss estimator defined in (2.6), and $\check{\Delta}(\tau)$ be the truncated version of $\widehat{\Delta}$ as defined in (2.29). Following directly from Theorem 7, we then have the following corollary.

Corollary 13. *Assume that $C_\Gamma \leq d^{\tau_3}$, where C_Γ is the same universal constant used in (2.36). Let*

$$\lambda_\Gamma \asymp \sqrt{\frac{\log d}{\min\{n_X, n_Y\}}},$$

where λ_Γ is the penalization parameter of D-Trace loss. Then for any $\tau \geq \tau_3$, we have

$$\sup_{(\Omega_X, \Omega_Y) \in \mathcal{M}} \mathbb{E} \left[\left\| \check{\Delta}(\tau) - \Delta \right\|_F^2 \right] \lesssim \frac{h \log d}{\min\{n_X, n_Y\}}.$$

Combine Theorem 12 and Corollary 13, we then have the final main result of this section.

Theorem 14. *Assume that the conditions of Theorem 12 and Corollary 13 hold. We have*

$$\min_{\widehat{\Delta}} \sup_{(\Omega_X, \Omega_Y) \in \mathcal{M}} \mathbb{E} \left[\left\| \widehat{\Delta} - \Delta \right\|_F^2 \right] \asymp \frac{h \log d}{\min\{n_X, n_Y\}}.$$

Theorem 14 highlights a distinctive feature of the differential network estimation problem: the minimax rate depends solely on the sparsity level h of the differential network Δ , and is independent of the sparsity levels of the individual precision matrices Ω_X and Ω_Y . This suggests that directly estimating the difference Δ can be statistically more efficient than estimating Ω_X and Ω_Y separately and then taking their difference, especially when Ω_X and Ω_Y are themselves dense but Δ remains sparse.

This phenomenon reflects the fact that the target parameter Δ may have a simpler structure than its components, and that appropriate estimators—such as the D-Trace loss estimator analyzed here—can directly exploit this structure. In contrast, indirect approaches that estimate Ω_X and Ω_Y separately may suffer from error accumulation, particularly when the individual precision matrices are high-dimensional and complex.

Overall, the established minimax rate provides a theoretical justification for the use of direct estimation methods for differential networks, and underscores the importance of tailoring estimation procedures to the specific structure of the parameter of interest.

2.5.6 Graph Recovery Guarantee for Gaussian Graphical Models

While this chapter primarily focuses on precision matrix estimation, it is also of great interest to explore the implications of Trans-Glasso for graph recovery in the Gaussian Graphical Models (GGM), given the close relationship between the precision matrix and GGM. In this section, we introduce a graph estimation procedure based on thresholding the output of Trans-Glasso. We then establish a theoretical guarantee for the accuracy of the proposed graph estimator. Finally, we demonstrate its advantages over the optimal graph estimator that relies solely on target data.

Let $E^{(0)} = [E_{jl}^{(0)}]_{1 \leq j, l \leq d} \in \{0, 1\}^{d \times d}$ denote the adjacency matrix of the undirected Gaussian graphical model for the target data. Specifically, for any $j \neq l$, we set $E_{jl}^{(0)} = 1$ if nodes j and l are connected and $E_{jl}^{(0)} = 0$ otherwise. By definition, $E_{jj}^{(0)} = 0$ for all $1 \leq j \leq d$. Thus, estimating the Gaussian graphical model reduces to estimating $E^{(0)}$. Given the well-established relationship between the Gaussian graphical model and the precision matrix [Lauritzen, 1996], we have, for any $j \neq l$,

$$E_{jl}^{(0)} = \begin{cases} 1 & \text{if } \Omega_{jl}^{(0)} \neq 0, \\ 0 & \text{if } \Omega_{jl}^{(0)} = 0. \end{cases}$$

We propose a graph estimator based on thresholding the output of Trans-Glasso, $\hat{\Omega}^{(0)}$. Specifically, let $\hat{E}^{(0)} = [\hat{E}_{jl}^{(0)}]_{1 \leq j, l \leq d} \in \{0, 1\}^{d \times d}$ denote the estimated adjacency matrix.

We set $\widehat{E}_{jj}^{(0)} = 0$ for all $1 \leq j \leq d$ and, for any $j \neq l$, define

$$\widehat{E}_{jl}^{(0)} = \begin{cases} 1 & \text{if } |\widehat{\Omega}_{jl}^{(0)}| > \epsilon_n, \\ 0 & \text{if } |\widehat{\Omega}_{jl}^{(0)}| \leq \epsilon_n, \end{cases}$$

where ϵ_n is a thresholding parameter.

We measure the graph estimation error using the Hamming distance between $E^{(0)}$ and $\widehat{E}^{(0)}$. Specifically, we define

$$\text{Err}(\widehat{E}^{(0)}) := \frac{1}{2} \sum_{1 \leq j, l \leq d} \mathbb{1} \left\{ \widehat{E}_{jl}^{(0)} \neq E_{jl}^{(0)} \right\}.$$

Additionally, let $\omega_{\min} = \min_{(j,l): E_{jl}^{(0)}=1} |\Omega_{jl}^{(0)}| > 0$ denote the minimum signal strength in the graphical model. The following deterministic bound then holds for $\text{Err}(\widehat{E}^{(0)})$.

Proposition 15. *If $0 < \epsilon_n \leq \frac{1}{2}\omega_{\min}$, then*

$$\text{Err}(\widehat{E}^{(0)}) \leq \frac{1}{2\epsilon_n^2} \left\| \widehat{\Omega}^{(0)} - \Omega^{(0)} \right\|_F^2.$$

Moreover, if the right-hand side of the inequality is smaller than 1, perfect graph recovery is achieved, i.e., $\widehat{E}^{(0)} = E^{(0)}$.

Proof. Let

$$S_{\text{I}} = \left\{ (j, l) : E_{jl}^{(0)} = 0, \widehat{E}_{jl}^{(0)} = 1 \right\} \quad \text{and} \quad S_{\text{II}} = \left\{ (j, l) : E_{jl}^{(0)} = 1, \widehat{E}_{jl}^{(0)} = 0 \right\}$$

denote the sets of type-I error and type-II error. We then have

$$\text{Err}(\widehat{E}^{(0)}) = \frac{1}{2} \sum_{1 \leq j, l \leq d} \mathbb{1} \left\{ \widehat{E}_{jl}^{(0)} \neq E_{jl}^{(0)} \right\} = \frac{1}{2} (|S_{\text{I}}| + |S_{\text{II}}|).$$

Note that for any $(j, l) \in S_{\text{I}}$, we have

$$\left| \widehat{\Omega}_{jl}^{(0)} - \Omega_{jl}^{(0)} \right| = \left| \widehat{\Omega}_{jl}^{(0)} \right| > \epsilon_n. \quad (2.38)$$

Besides, for any $(j, l) \in S_{\text{II}}$, we have

$$\left| \widehat{\Omega}_{jl}^{(0)} - \Omega_{jl}^{(0)} \right| \geq \left| \Omega_{jl}^{(0)} \right| - \left| \widehat{\Omega}_{jl}^{(0)} \right| \geq \left| \Omega_{jl}^{(0)} \right| - \epsilon_n \geq \omega_{\min} - \epsilon_n \geq \epsilon_n, \quad (2.39)$$

where the last inequality follows from the assumption that $\epsilon_n \leq \frac{1}{2}\omega_{\min}$.

Thus, combining (2.38) and (2.39), we have

$$\begin{aligned} \left\| \widehat{\Omega}^{(0)} - \Omega^{(0)} \right\|_{\text{F}}^2 &\geq \sum_{(j,l) \in S_{\text{I}}} \left(\widehat{\Omega}_{jl}^{(0)} - \Omega_{jl}^{(0)} \right)^2 + \sum_{(j,l) \in S_{\text{II}}} \left(\widehat{\Omega}_{jl}^{(0)} - \Omega_{jl}^{(0)} \right)^2 \\ &\geq \epsilon_n^2 (|S_{\text{I}}| + |S_{\text{II}}|) \\ &= 2\epsilon_n^2 \cdot \text{Err} \left(\widehat{E}^{(0)} \right), \end{aligned}$$

which then implies the final result. $\square$

To illustrate the advantage of Trans-Glasso in graph recovery over methods that rely solely on target data, we consider a special case of the parameter space defined in (2.33), where $\Omega_{jj}^{(0)} = 1$ for all $1 \leq j \leq d$. In this scenario, the sample-optimal algorithm for graph recovery proposed by Misra et al. [2020], which exclusively uses target data, requires

$$n_0 \gg (s + h) \log d / \omega_{\min}^2 \quad (2.40)$$

to achieve perfect graph recovery.

On the other hand, applying Proposition 15 and Corollary 4, we choose ϵ_n such that $\epsilon_n \asymp \omega_{\min}$ and $\epsilon_n \leq \frac{1}{2}\omega_{\min}$. Given that the conditions of Corollary 4 hold, and assuming $\bar{n} \geq n_0$ and $K \leq d^\tau$ for some universal constant $\tau > 0$, achieving perfect graph recovery

only requires n_0 and N to be sufficiently large such that

$$n_0 \gg h \log d / \omega_{\min}^2 \quad \text{and} \quad N \gg s \log d / \omega_{\min}^2. \quad (2.41)$$

Comparing (2.40) with (2.41), we observe that the target sample complexity requirement for n_0 in Trans-Glasso is significantly lower than that of algorithms relying solely on target samples when $s \gg h$. This highlights the advantage of transfer learning in Gaussian graphical model estimation.

2.6 Simulations

In this section, we demonstrate the empirical performance of Trans-Glasso through a series of simulations. We evaluate its accuracy in comparison with several baseline methods under different settings, including varying sample sizes and sparsity levels.

We set the dimensionality $d = 100$ and the number of source distributions $K = 5$ for all experiments. In each experiment, we vary parameters such as the target sample size n_0 , the source sample size $n_k = n_{\text{source}}$, and the sparsity level h to assess the robustness of Trans-Glasso across diverse conditions. The data are simulated under three different models, each reflecting a specific structure for the shared and individual components of the precision matrices.

2.6.1 Data Generation Models

We generate data from three distinct models to assess the flexibility of Trans-Glasso. Each model starts with a shared component, followed by individual components. The final precision matrices are made positive definite by adding a diagonal matrix. Specifically, each model is set as follows.

- **Model I:** The shared component is a banded matrix with bandwidth 1, where each entry $\tilde{\Omega}_{ij} = 5 \times 0.6^{|i-j|} \mathbb{1}(|i-j| \leq 1)$ for $1 \leq i, j \leq d$. For a given h and each

$k = 0, 1, \dots, K$, we uniformly choose $\lceil h/2 \rceil$ entries (i, j) such that $1 \leq i \leq \lfloor \frac{d}{2} \rfloor$ and $\lfloor \frac{d}{2} \rfloor + 1 \leq j \leq d$, denoted as $\mathcal{S}_{\Gamma^{(k)}, \text{up}}$. We let $\tilde{\Gamma}_{ij}^{(k)} = u_{ij} \mathbb{1}\{(i, j) \in \mathcal{S}_{\Gamma^{(k)}, \text{up}}\}$, $1 \leq i, j \leq d$, where u_{ij} 's are from $\text{Unif}[-3, 3]$. Then $\Gamma^{(k)\star} = \tilde{\Gamma}^{(k)} + (\tilde{\Gamma}^{(k)})^\top$. Finally, we let $\Omega^{(k)} = \tilde{\Omega} + \Gamma^{(k)\star} + \sigma I_d$, where σ ensures $\gamma_{\min}(\Omega^{(k)}) \geq 0.1$ for $0 \leq k \leq K$.

- **Model II:** Model II is similar to Model I but with a wider bandwidth of 5, introducing a more connected structure in the shared component.
- **Model III:** We generate the shared component from an Erdos–Renyi graph. Specifically, let $\tilde{\Omega}_{ii} = 5$, $1 \leq i \leq d$, and $U_{ij} \sim \text{Bernoulli}(0.02)$, $1 \leq i < j \leq d$. If $U_{ij} = 1$, let $\tilde{\Omega}_{ij} = \tilde{\Omega}_{ji} \sim \text{Unif}[-3, 3]$; otherwise, set $\tilde{\Omega}_{ij} = \tilde{\Omega}_{ji} = 0$. Let $\mathcal{S}_{\tilde{\Omega}} = \{(i, j) \in [d] \times [d] : \tilde{\Omega}_{ij} \neq 0\}$ be the support of $\tilde{\Omega}$. For given h and $0 \leq k \leq K$, uniformly choose h entries ($h+1$ if h is odd) from $[d] \times [d] \setminus \tilde{\Omega}$, denoted as $\mathcal{S}_{\Gamma^{(k)}}$, such that $(i, j) \in \mathcal{S}_{\Gamma^{(k)}}$ if and only if $(j, i) \in \mathcal{S}_{\Gamma^{(k)}}$. Let $\Gamma_{ij}^{(k)\star} = u_{ij} \mathbb{1}\{(i, j) \in \mathcal{S}_{\Gamma^{(k)}}\}$, $1 \leq i, j \leq d$, where $u_{ij} \sim \text{Unif}[-3, 3]$. Finally, let $\Omega^{(k)} = \tilde{\Omega} + \Gamma^{(k)\star} + \sigma I_d$, where σ ensures $\gamma_{\min}(\Omega^{(k)}) \geq 0.1$ for $0 \leq k \leq K$.

2.6.2 Experimental Design

For each model, we conduct four sets of experiments to systematically evaluate the performance of Trans-Glasso under various scenarios. Each experiment is independently repeated 30 times to compute stable averages and standard errors.

1. **Experiment 1:** Vary the target sample size n_0 , while keeping the source sample size n_{source} and sparsity level h fixed.
2. **Experiment 2:** Vary the source sample size n_{source} , while fixing n_0 and h .
3. **Experiment 3:** Simultaneously scale n_0 , n_{source} , and h proportionally to examine the scalability of the method.
4. **Experiment 4:** Fix n_0 and n_{source} , and increase h to assess robustness as the sparsity level of the differential structure increases.

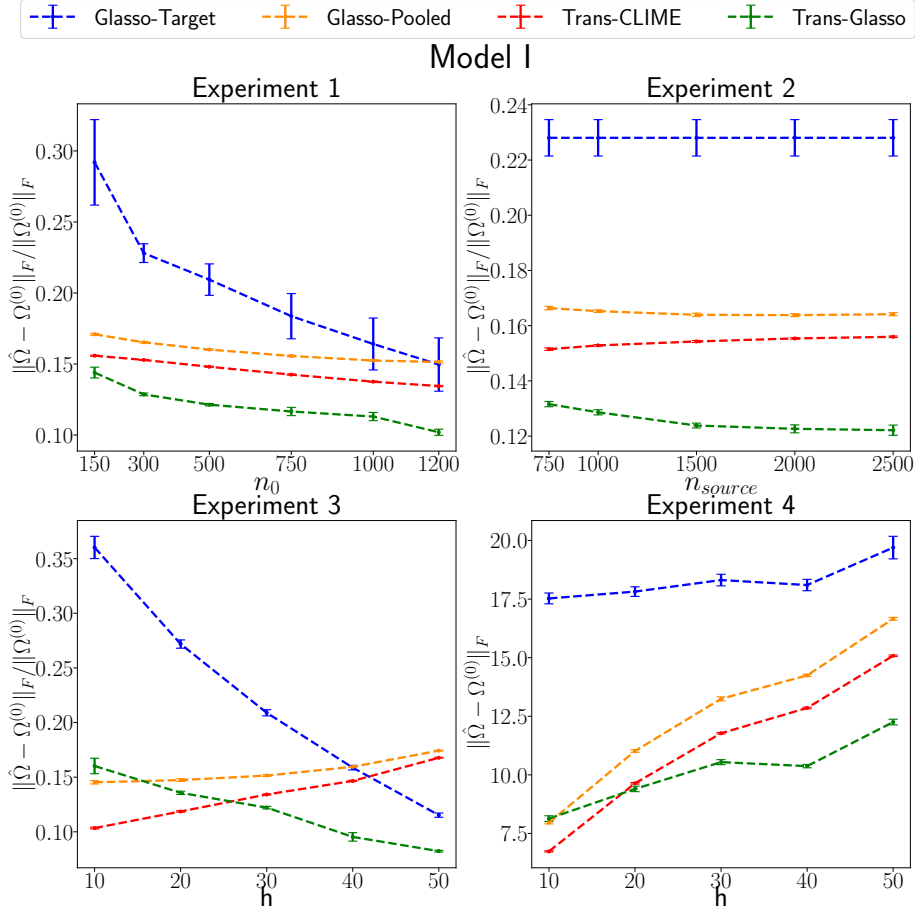


Figure 2.2: Simulation results of Trans-Glasso for Model I.

The specific simulation settings for each model are summarized below:

Model I. The default configuration is $n_0 = 300$, $n_{\text{source}} = 1000$, and $h = 40$. In Experiment 3, we set $n_{\text{source}} = 3n_0$, with n_0 adjusted according to the following schedule: $n_0 = 70$ for $h = 10$, $n_0 = 150$ for $h = 20$, $n_0 = 300$ for $h = 30$, $n_0 = 600$ for $h = 40$, and $n_0 = 1200$ for $h = 50$.

Model II. The default configuration is $n_0 = 750$, $n_{\text{source}} = 2000$, and $h = 40$. In Experiment 3, we again set $n_{\text{source}} = 3n_0$, with n_0 chosen as follows: $n_0 = 100$ for $h = 20$, $n_0 = 200$ for $h = 30$, $n_0 = 300$ for $h = 40$, $n_0 = 500$ for $h = 50$, $n_0 = 800$ for $h = 60$, $n_0 = 1000$ for $h = 70$, and $n_0 = 1200$ for $h = 80$.

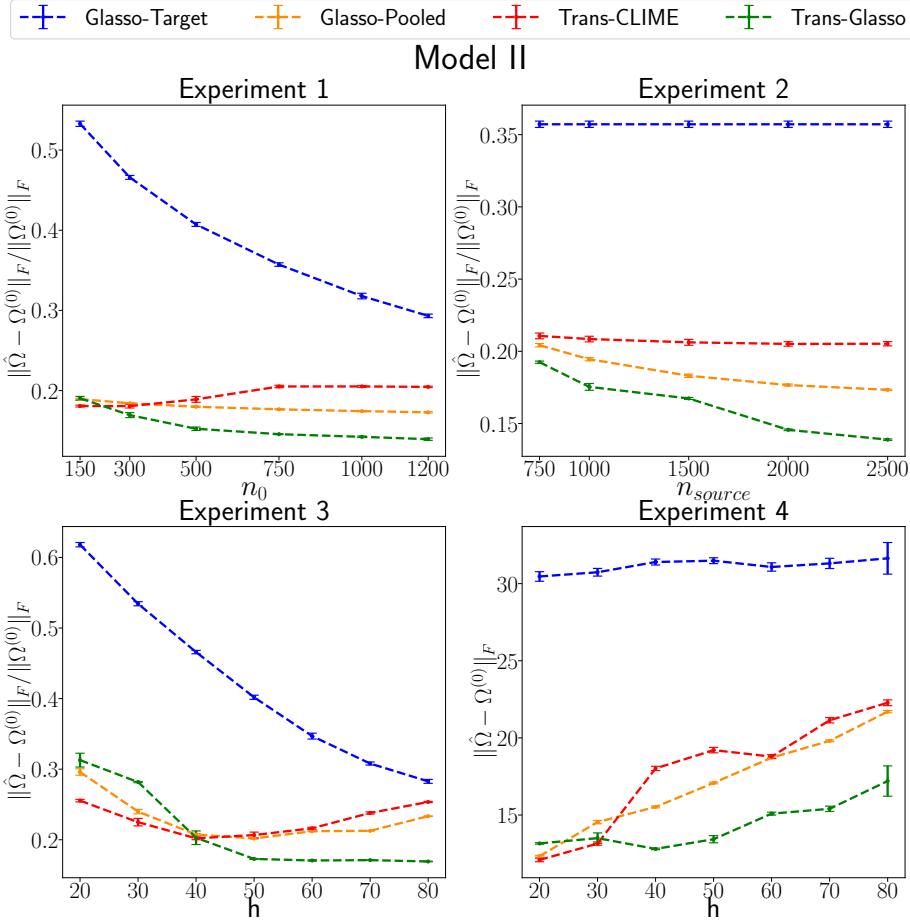


Figure 2.3: Simulation results of Trans-Glasso for Model II.

Model III. The default configuration is $n_0 = 150$, $n_{\text{source}} = 1000$, and $h = 40$. In Experiment 3, we set $n_{\text{source}} = 4n_0$, with n_0 specified as follows: $n_0 = 15$ for $h = 20$, $n_0 = 30$ for $h = 30$, $n_0 = 80$ for $h = 40$, $n_0 = 300$ for $h = 50$, and $n_0 = 1000$ for $h = 60$.

2.6.3 Comparison Methods

We compare Trans-Glasso with the following baseline methods:

- **Glasso-Target:** Applies graphical lasso [Friedman et al., 2008] to target data only.
- **Glasso-Pooled:** Pools target and source data, then applies graphical lasso.
- **Trans-CLIME:** A transfer learning method by Li et al. [2023a] that assumes a sparse divergence matrix across sources.

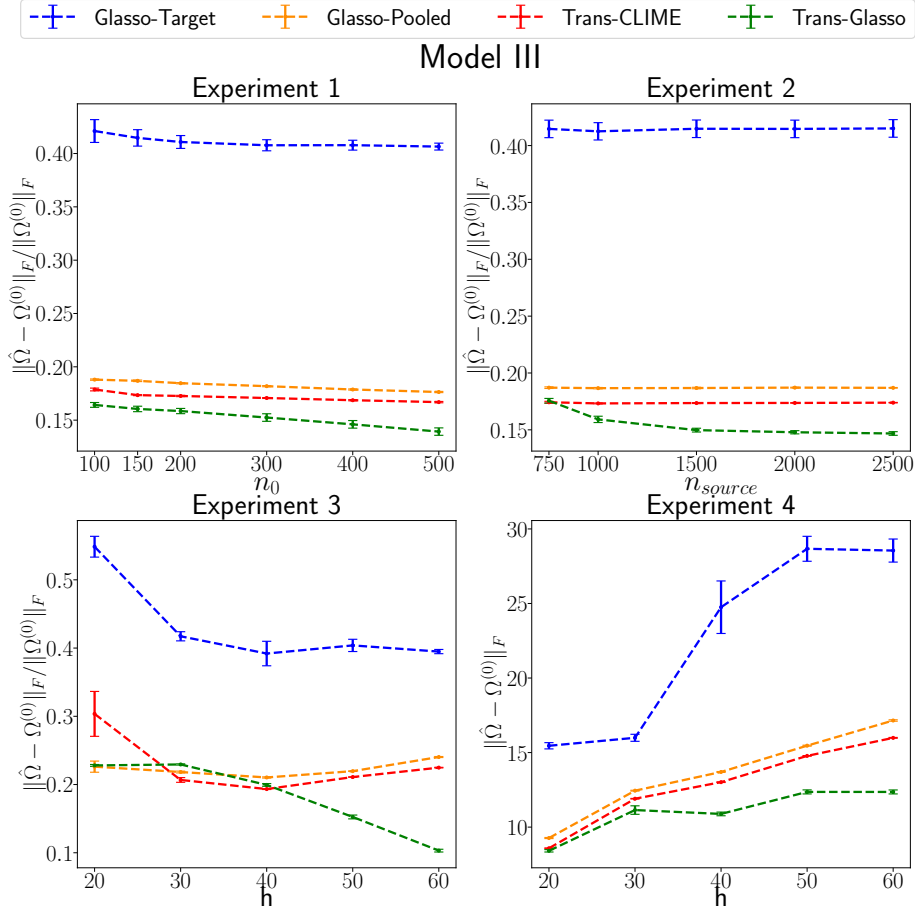


Figure 2.4: Simulation results of Trans-Glasso for Model III.

2.6.4 Results

The results for Models I – III are shown in Figure 2.2–2.4. Each dot represents the empirical mean across 30 repetitions and the vertical bar represents $\text{Mean} \pm \frac{2}{\sqrt{30}} \times \text{Standard Error}$. Trans-Glasso generally outperforms baseline methods. In Experiment 3, Trans-Glasso shows consistency across all models, whereas Glasso-Pooled and Trans-CLIME do not. Trans-CLIME performs better with small h , but its performance deteriorates as h increases. This is because a small increase in h can significantly increase the sparsity of the divergence matrix when the covariance matrix is not sparse. Therefore, when the precision matrix is sparse but the covariance matrix is not, Trans-Glasso is more reliable and robust.

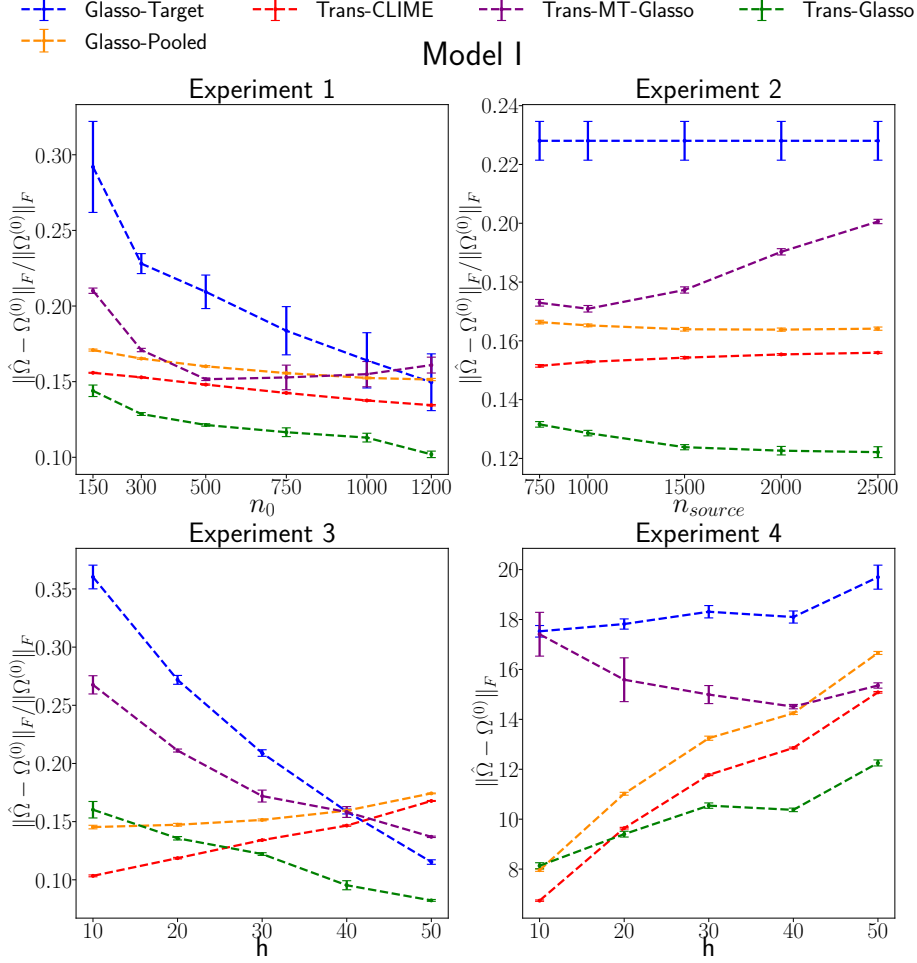


Figure 2.5: Ablation study results of Trans-Glasso for Model I.

2.6.5 Ablation Study of the Refinement Step

In this section, we empirically evaluate the effect of the refinement step in the Trans-Glasso procedure. Specifically, we compare the final estimator $\hat{\Omega}^{(0)}$ with the intermediate estimator $\check{\Omega}^{(0)}$, which is directly obtained from the initial Trans-MT-Glasso step without refinement. This comparison allows us to isolate and assess the contribution of the refinement step, which leverages differential network estimation to improve accuracy.

The experiments are conducted under the same simulation settings described in Section 2.6, and include all three data generation models.

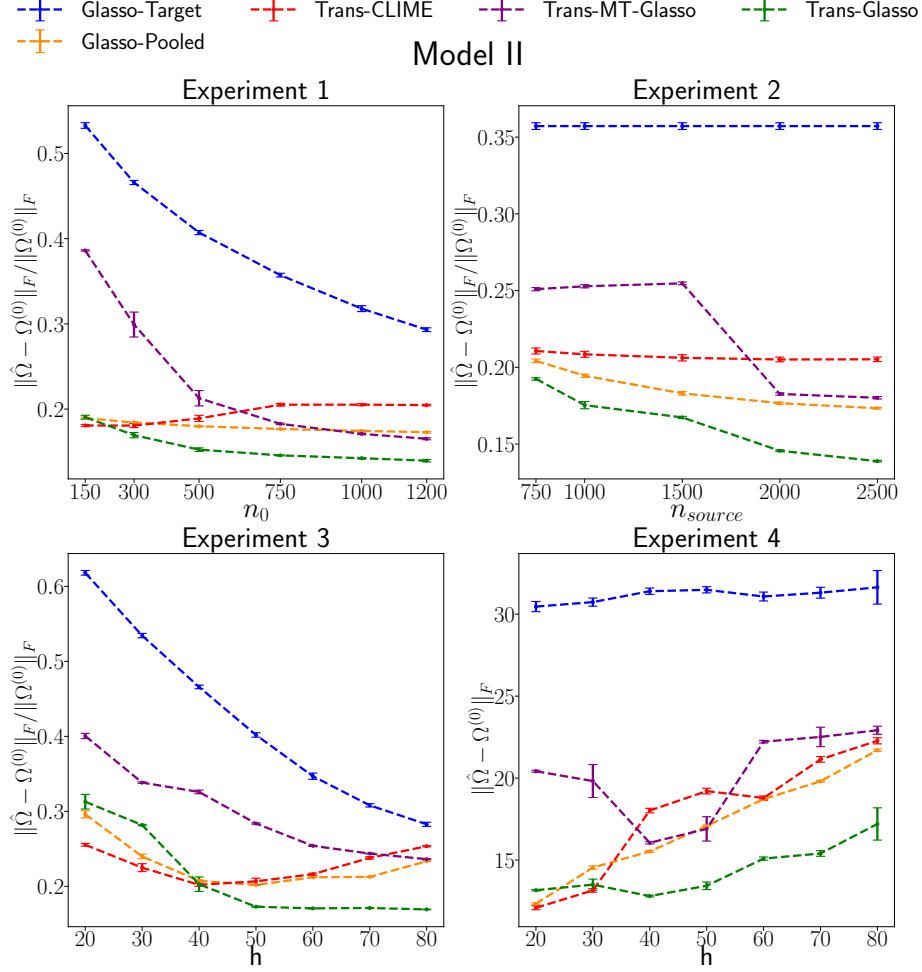


Figure 2.6: Ablation study results of Trans-Glasso for Model II.

The results are summarized in Figures 2.5–2.7. As shown, Trans-Glasso consistently outperforms Trans-MT-Glasso across all scenarios, demonstrating that the refinement step substantially enhances estimation accuracy. This highlights the practical importance of incorporating the differential network correction into the overall procedure.

2.6.6 Experiments with Unknown Informative Set

We further conduct simulation experiments where the informative set $\mathcal{A}$ is unknown, using the same three models as in the previous sections. We partition $[K]$ into $[K] = \mathcal{A} \cup \mathcal{A}^c$, where $k \in \mathcal{A}$ represents informative tasks with smaller sparsity, while $k \in \mathcal{A}^c$ corresponds

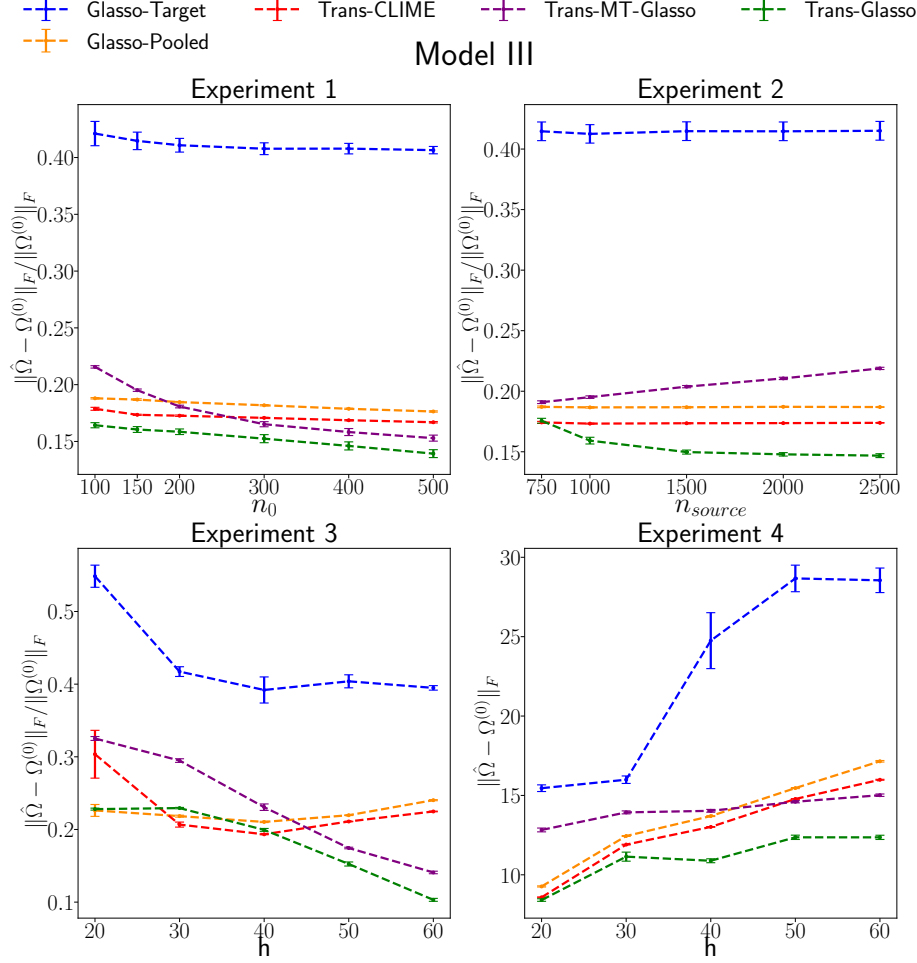


Figure 2.7: Ablation study results of Trans-Glasso for Model III.

to non-informative tasks with larger sparsity levels. Specifically, for each model, we set:

- **Model I:** $h = 20$ for $k \in \mathcal{A}$ and $h = 600$ for $k \in \mathcal{A}^c$;
- **Model II:** $h = 30$ for $k \in \mathcal{A}$ and $h = 600$ for $k \in \mathcal{A}^c$;
- **Model III:** $h = 10$ for $k \in \mathcal{A}$ and $h = 300$ for $k \in \mathcal{A}^c$.

The sample sizes are set as follows: $(n_0, n_{\text{source}}) = (300, 1000)$ for Model I, $(750, 2000)$ for Model II, and $(300, 1000)$ for Model III. We apply the Trans-Glasso-CV algorithm introduced in Section 2.4.3, and compare its performance against several baseline methods. To evaluate performance under varying levels of informativeness, we vary the size of $\mathcal{A}$ from 0 to K .

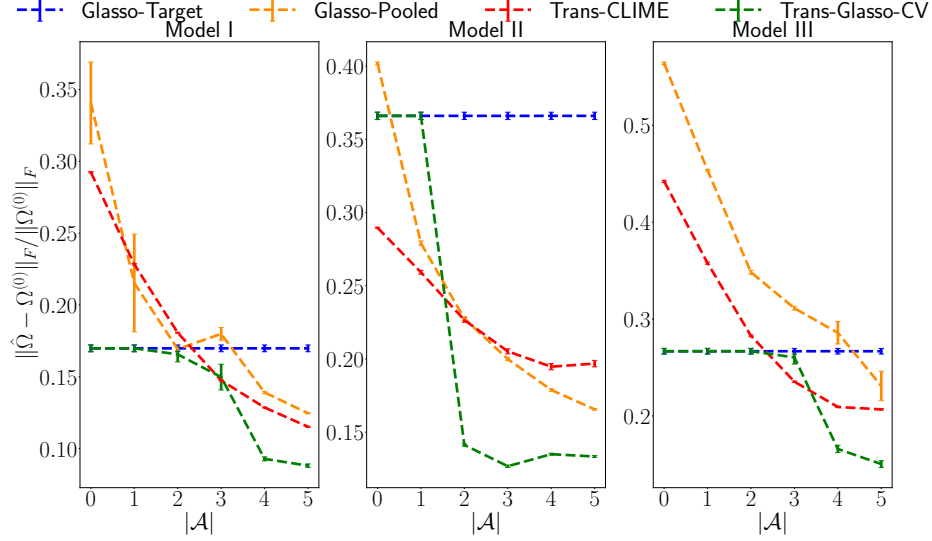


Figure 2.8: Simulation results of Trans-Glasso-CV when the informative set $\mathcal{A}$ is unknown.

Each experiment is independently repeated 30 times using different random seeds to ensure stable averages and standard errors.

The results, presented in Figure 2.8, demonstrate that Trans-Glasso-CV consistently outperforms the competing methods. Notably, it never performs worse than Glasso-Target, suggesting that it successfully avoids “negative transfer.” In contrast, both Glasso-Pooled and Trans-CLIME can suffer from performance degradation relative to Glasso-Target. Furthermore, as the size of $\mathcal{A}$ increases, Trans-Glasso-CV increasingly leverages the informative tasks to achieve improved estimation accuracy.

2.7 Real-World Data Analysis

We apply the Trans-Glasso algorithm to two real-world datasets. In Section 2.7.1, we use it on gene networks with different brain tissues. In Section 2.7.2, we use it on protein networks for various cancer subtypes.

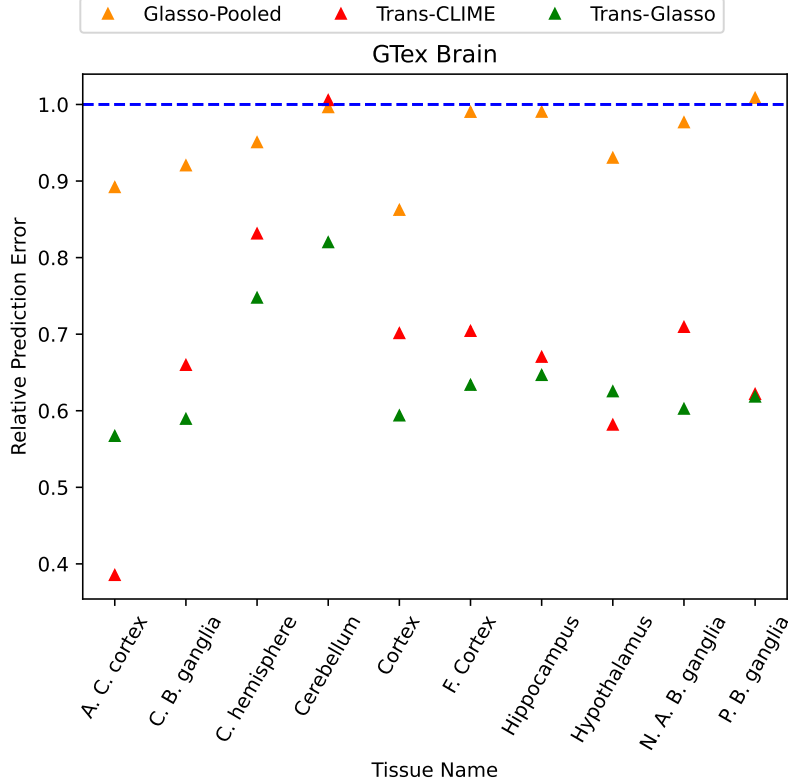


Figure 2.9: Cross-validation prediction error of different methods on GTEx brain tissue datasets, relative to Glasso-Target.

2.7.1 Gene Networks Data for Brain Tissues

We apply Trans-Glasso to detect gene networks across tissues using the Genotype-Tissue Expression (GTEx) data ¹. Following Li et al. [2023a], we focus on genes involved in central nervous system neuron differentiation (GO:0021953) across 13 brain tissues, treating one as the target and the remaining 12 as sources. To avoid small sample sizes, we exclude three tissues from the target set, using only 10. A complete list of tissues is in Table 3 of Li et al. [2023a]. We remove genes with missing values and compare Trans-Glasso to baselines via cross-validation prediction error (2.23).

Figure 2.9 presents the results, benchmarking Trans-Glasso against Glasso-Target, Glasso-Pooled, and Trans-CLIME, as in Section 2.6. Prediction errors are reported relative to

1. <https://gtexportal.org/home/>

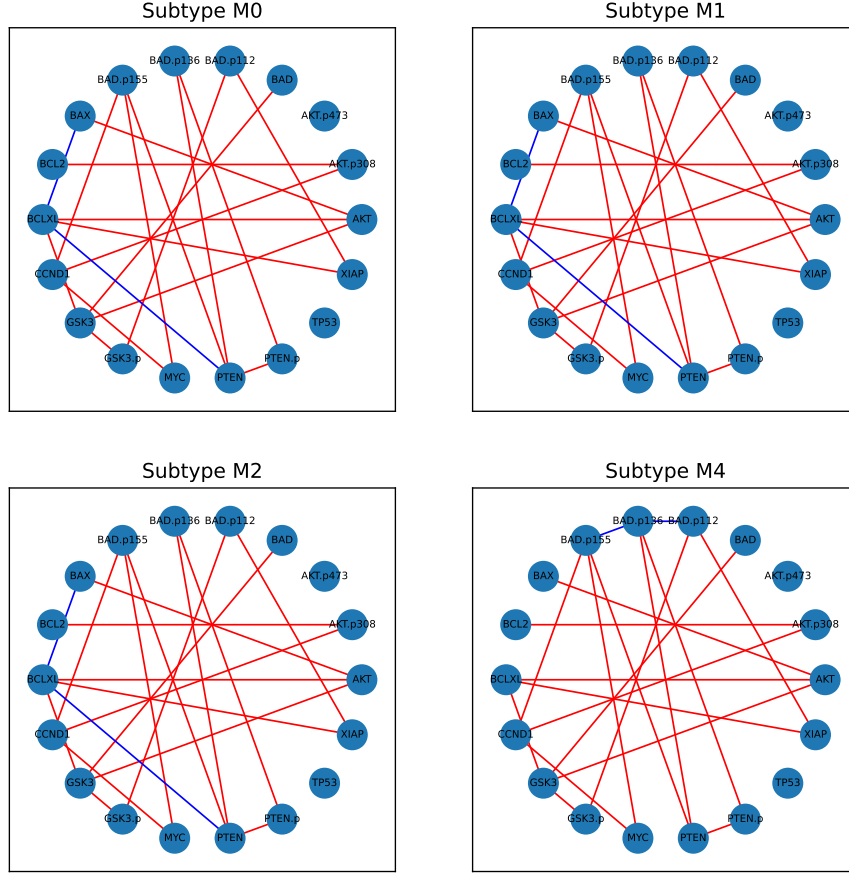


Figure 2.10: Protein networks for four AML subtypes. Red edges are shared across subtypes; blue edges are unique to specific subtypes.

Glasso-Target for cross-tissue comparison. As shown in Figure 2.9, Trans-Glasso consistently outperforms baselines, with a relative prediction error well below 1, demonstrating superior performance and robustness to negative transfer. In contrast, Glasso-Pooled and Trans-CLIME sometimes perform comparably or worse than Glasso-Target.

2.7.2 Protein Networks Data for AML

We apply our method to a protein network dataset for Acute Myeloid Leukemia (AML) subtypes. Understanding protein relationships is crucial in cancer studies, and graphical models aid in constructing these networks. Following Peterson et al. [2015], we analyze protein lev-

els from 213 newly diagnosed AML patients,² classified by the FAB system. While protein interactions may vary across subtypes, shared AML-related processes suggest common connections, making transfer learning beneficial. We focus on 18 proteins involved in apoptosis and cell cycle regulation across four subtypes: M0 (17 subjects), M1 (34), M2 (68), and M4 (59) [Peterson et al., 2015, Kanehisa et al., 2012].

For each subtype, we apply Trans-Glasso and extract the final estimated graph by selecting 20 edges with the largest absolute values in the estimated precision matrix, ensuring comparability with Peterson et al. [2015]. The results in Figure 2.10 reveal substantial overlap with Peterson et al. [2015], though our graphs exhibit greater structural similarity across subtypes. M0, M1, and M2 share identical structures, while M4 differs by two edges and shows stronger BAD family protein connections. This aligns with Tzifi et al. [2012], which reported higher BAD protein expression in AML subtypes M4, M5, and M6 [Tzifi et al., 2012, Table 1], suggesting more active interactions.

2.8 Conclusion

We introduce Trans-Glasso, a novel transfer learning approach for precision matrix estimation that mitigates small target sample limitations by leveraging related source data. Trans-Glasso follows a two-step process: initial estimation via multi-task learning, followed by refinement through differential network estimation. This method achieves minimax optimality across various parameter regimes. Extensive simulations show that Trans-Glasso consistently outperforms baselines, demonstrating robustness and adaptability, especially in high-dimensional settings with limited target samples.

Future directions include extending Trans-Glasso to a broader class of graphical models, such as Gaussian copula models [Liu et al., 2009, 2012a], transelliptical models [Liu et al., 2012b], functional graphical models [Qiao et al., 2019, Tsai et al., 2024, Zhao et al., 2024c], and Ising models [Kuang et al., 2017]. Another promising direction is the development

2. Dataset provided as a supplement to Kornblau et al. [2009].

of inferential procedures within the transfer learning framework. In addition, it would be valuable to explore transfer learning for directed graphical models [Glymour et al., 2019], particularly in light of recent advances in continuous optimization-based approaches [Zheng et al., 2018, Deng et al., 2023a,b, 2024].

2.9 Technical Details

This section presents the technical proofs for the main theorems in this chapter, as well as further details on the optimization algorithms.

2.9.1 Preliminary Lemmas

We first collect several inequalities related to matrix norms.

Lemma 16. *For any two matrices $A, B \in \mathbb{R}^{d \times d}$, we have*

$$|AB|_\infty \leq \min \{ \|A\|_\infty |B|_\infty, \|B\|_1 |A|_\infty \}. \quad (2.42)$$

Proof. For $1 \leq j \leq d$, denote by $B_{:j}$ the j -th column of B . We then have

$$\begin{aligned} |AB|_\infty &= \max_{1 \leq j \leq d} \|AB_{:j}\|_\infty \\ &\leq \max_{1 \leq j \leq d} \|A\|_\infty \|B_{:j}\|_\infty \\ &= \|A\|_\infty \max_{1 \leq j \leq d} \|B_{:j}\|_\infty \\ &= \|A\|_\infty |B|_\infty. \end{aligned}$$

The second bound follows from the identities $|AB|_\infty = |B^\top A^\top|_\infty$ and $|B^\top|_\infty = |B|_1$. Applying the same argument as above to $B^\top A^\top$ yields the desired result. Combining both bounds completes the proof.

□

Lemma 17. For any matrix $A \in \mathbb{R}^{m \times n}$ and any vector $v \in \mathbb{R}^n$, we have $\|Av\|_\infty \leq |A|_\infty \|v\|_1$.

Proof. Let $A_i.$ denote the i -th row of A , we then have

$$\|Av\|_\infty = \max_{1 \leq i \leq m} |\langle A_i., v \rangle| \leq \max_{1 \leq i \leq m} \|A_i.\|_\infty \|v\|_1 = |A|_\infty \|v\|_1,$$

which completes the proof. $\square$

The next lemmas require the following definition of the sub-Gaussian random variable.

Definition 1 (Sub-Gaussian Random Variable, Definition 2 of Ravikumar et al. [2011]). We say that a random variable $X \in \mathbb{R}$ is sub-Gaussian with parameter σ if

$$\mathbb{E}[\exp(\lambda X)] \leq \exp\left(\sigma^2 \lambda^2 / 2\right) \quad \text{for all } \lambda \in \mathbb{R}.$$

Lemma 18. Assume that we obtain samples from K distributions, each with a mean of zero and a covariance matrix of $\Sigma^{(k)}$. Let $\{X_i^{(k)} = (X_{i1}^{(k)}, \dots, X_{id}^{(k)})^\top\}_{i=1}^{n_k}$ represent n_k independently distributed samples from the k -th distribution. In addition, we assume that $X_{ij}^{(k)} / \sqrt{\Sigma_{jj}^{(k)}}$ is sub-Gaussian with parameter σ as defined in Definition 1. Let

$$\widehat{\Sigma}^{(k)} = \frac{1}{n_k} \sum_{i=1}^{n_k} X_i^{(k)} X_i^{(k)\top}, \quad \widehat{\Sigma} = \sum_{k=1}^K \alpha_k \widehat{\Sigma}^{(k)}, \quad \text{and} \quad \Sigma^* = \sum_{k=1}^K \alpha_k \Sigma^{(k)},$$

where $\alpha_k = n_k / N$ and $N = \sum_{k=1}^K n_k$. For a fixed $\delta \in (0, 1]$, we have

$$\mathbb{P} \left\{ \left| \widehat{\Sigma} - \Sigma^* \right|_\infty > 16(1 + 4\sigma^2) \left\{ \max_{1 \leq k \leq K} \left| \Sigma^{(k)} \right|_\infty \right\} \max \left\{ \sqrt{\frac{\log(2d^2/\delta)}{2N}}, \frac{\log(2d^2/\delta)}{N} \right\} \right\} \leq \delta.$$

Proof. Let $\overline{X}_{ij}^{(k)} = X_{ij}^{(k)} / \sqrt{\Sigma_{jj}^{(k)}}$, $i = 1, \dots, n_k$, $k = 1, \dots, K$, $j = 1, \dots, d$. Define $U_{i,jl}^{(k)} = \overline{X}_{ij}^{(k)} + \overline{X}_{il}^{(k)}$, $V_{i,jl}^{(k)} = \overline{X}_{ij}^{(k)} - \overline{X}_{il}^{(k)}$, and $\rho_{jl}^{(k)} = \Sigma_{jl}^{(k)} / \sqrt{\Sigma_{jj}^{(k)} \cdot \Sigma_{ll}^{(k)}}$. Then it follows from

Proposition 2.9 and (2.18) in Wainwright [2019], that

$$\begin{aligned}\mathbb{E} \left[\exp \left\{ \lambda \left[\left(U_{i,jl}^{(k)} \right)^2 - 2 \left(1 + \rho_{jl}^{(k)} \right) \right] \right\} \right] &\leq \exp \left(\frac{\lambda^2 \nu^2}{2} \right), \quad \text{and} \\ \mathbb{E} \left[\exp \left\{ \lambda \left[\left(V_{i,jl}^{(k)} \right)^2 - 2 \left(1 - \rho_{jl}^{(k)} \right) \right] \right\} \right] &\leq \exp \left(\frac{\lambda^2 \nu^2}{2} \right),\end{aligned}$$

for all $|\lambda| < 1/\phi$ where $\nu = \phi = 16(1 + 4\sigma^2)$. Since

$$\begin{aligned}&\sum_{i=1}^{n_k} \left(\overline{X}_{ij}^{(k)} \overline{X}_{il}^{(k)} - \rho_{jl}^{(k)} \right) \\ &= \frac{1}{4} \sum_{i=1}^{n_k} \left(\left(U_{i,jl}^{(k)} \right)^2 - 2 \left(1 + \rho_{jl}^{(k)} \right) \right) - \frac{1}{4} \sum_{i=1}^{n_k} \left(\left(V_{i,jl}^{(k)} \right)^2 - 2 \left(1 - \rho_{jl}^{(k)} \right) \right),\end{aligned}$$

we have

$$\begin{aligned}&\mathbb{E} \left[\exp \left\{ \lambda \left[\sum_{i=1}^{n_k} \left(\overline{X}_{ij}^{(k)} \overline{X}_{il}^{(k)} - \rho_{jl}^{(k)} \right) \right] \right\} \right] \\ &= \mathbb{E} \left[\exp \left\{ \frac{\lambda}{4} \left[\sum_{i=1}^{n_k} \left(\left(U_{i,jl}^{(k)} \right)^2 - 2 \left(1 + \rho_{jl}^{(k)} \right) \right) \right] \right\} \times \right. \\ &\quad \left. \exp \left\{ -\frac{\lambda}{4} \left[\sum_{i=1}^{n_k} \left(\left(V_{i,jl}^{(k)} \right)^2 - 2 \left(1 - \rho_{jl}^{(k)} \right) \right) \right] \right\} \right] \\ &\leq \left\{ \mathbb{E} \left[\exp \left\{ \frac{\lambda}{2} \left[\sum_{i=1}^{n_k} \left(\left(U_{i,jl}^{(k)} \right)^2 - 2 \left(1 + \rho_{jl}^{(k)} \right) \right) \right] \right\} \right] \right\}^{\frac{1}{2}} \times \\ &\quad \left\{ \mathbb{E} \left[\exp \left\{ -\frac{\lambda}{2} \left[\sum_{i=1}^{n_k} \left(\left(V_{i,jl}^{(k)} \right)^2 - 2 \left(1 - \rho_{jl}^{(k)} \right) \right) \right] \right\} \right] \right\}^{\frac{1}{2}} \\ &\leq \exp \left(\frac{n_k \nu^2 \lambda^2}{8} \right) \quad \text{for all } |\lambda| < \frac{2}{\phi}.\end{aligned}$$

Let $\tau_\infty = \max_{1 \leq k \leq K} |\Sigma^{(k)}|_\infty$. It then follows that

$$\begin{aligned}
& \mathbb{E} \left[\exp \left\{ \lambda \left[\sum_{i=1}^{n_k} \left(X_{ij}^{(k)} X_{il}^{(k)} - \Sigma_{jl}^{(k)} \right) \right] \right\} \right] \\
&= \mathbb{E} \left[\exp \left\{ \lambda \sqrt{\Sigma_{jj}^{(k)}} \sqrt{\Sigma_{ll}^{(k)}} \left[\sum_{i=1}^{n_k} \left(\bar{X}_{ij}^{(k)} \bar{X}_{il}^{(k)} - \rho_{jl}^{(k)} \right) \right] \right\} \right] \\
&\leq \mathbb{E} \left[\exp \left\{ \lambda |\Sigma^{(k)}|_\infty \left[\sum_{i=1}^{n_k} \left(\bar{X}_{ij}^{(k)} \bar{X}_{il}^{(k)} - \rho_{jl}^{(k)} \right) \right] \right\} \right] \\
&\leq \mathbb{E} \left[\exp \left\{ \lambda \tau_\infty \left[\sum_{i=1}^{n_k} \left(\bar{X}_{ij}^{(k)} \bar{X}_{il}^{(k)} - \rho_{jl}^{(k)} \right) \right] \right\} \right] \\
&\leq \exp \left(\frac{n_k \nu^2 \tau_\infty^2 \lambda^2}{8} \right) \quad \text{for all } |\lambda| < \frac{2}{\phi \tau_\infty},
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E} \left[\exp \left\{ \lambda N \left(\widehat{\Sigma}_{jl} - \Sigma_{jl}^* \right) \right\} \right] &= \mathbb{E} \left[\exp \left\{ \lambda \left(\sum_{k=1}^K \sum_{i=1}^{n_k} \left(X_{ij}^{(k)} X_{il}^{(k)} - \Sigma_{jl}^{(k)} \right) \right) \right\} \right] \\
&= \prod_{k=1}^K \mathbb{E} \left[\exp \left\{ \lambda \left(\sum_{i=1}^{n_k} \left(X_{ij}^{(k)} X_{il}^{(k)} - \Sigma_{jl}^{(k)} \right) \right) \right\} \right] \\
&\leq \prod_{k=1}^K \exp \left(\frac{n_k \nu^2 \tau_\infty^2 \lambda^2}{8} \right) \\
&= \exp \left(\frac{N \nu^2 \tau_\infty^2 \lambda^2}{8} \right)
\end{aligned}$$

when $|\lambda| < 2/(\tau_\infty \phi)$. Using Proposition 2.9 in Wainwright [2019], for $t \geq 0$, we have that

$$\mathbb{P} \left\{ \left| \widehat{\Sigma}_{jl} - \Sigma_{jl}^* \right| > t \right\} = \mathbb{P} \left\{ \left| N \left(\widehat{\Sigma}_{jl} - \Sigma_{jl}^* \right) \right| > Nt \right\} \leq 2 \exp \left\{ -\frac{1}{2} \min \left(\frac{4Nt^2}{\nu^2 \tau_\infty^2}, \frac{2Nt}{\phi \tau_\infty} \right) \right\}.$$

A union bound then gives us

$$\mathbb{P} \left\{ \left| \widehat{\Sigma} - \Sigma^* \right|_\infty > t \right\} \leq 2d^2 \exp \left\{ -\frac{1}{2} \min \left(\frac{4Nt^2}{\nu^2 \tau_\infty^2}, \frac{2Nt}{\phi \tau_\infty} \right) \right\}.$$

Rewriting the above equation, we complete the proof. $\square$

Lemma 19. Consider a zero mean random vector $X = (X_1, \dots, X_d)^\top$ with covariance $\Sigma^\star$ such that each $X_j/\sqrt{\Sigma_{jj}^\star}$ is sub-Gaussian with parameter σ as defined in Definition 1. Let $\{X(i)\}_{i=1}^n$ be n i.i.d. copies of X and let $\widehat{\Sigma}_n = \frac{1}{n} \sum_{i=1}^n X(i)X(i)^\top$. For $\delta \in (0, 1]$, we have

$$\mathbb{P} \left\{ \left| \widehat{\Sigma}_n - \Sigma^\star \right|_\infty > 16(1 + 4\sigma^2) |\Sigma^\star|_\infty \max \left\{ \sqrt{\frac{\log(2d^2/\delta)}{2n}}, \frac{\log(2d^2/\delta)}{n} \right\} \right\} \leq \delta.$$

Proof. The proof follows directly from Lemma 18. $\square$

In the next lemma, we use $\mathbb{S}_+^d$ to denote the set of d -dimensional positive definite matrices.

Lemma 20. Let $A \in \mathbb{S}_+^d$. Define $f(\Omega) := \langle \Omega, A \rangle - \log \det \Omega$ for $\Omega \in \mathbb{S}_+^d$. Given $\Omega_0 \in \mathbb{S}_+^d$ and $\Delta_\Omega \in \mathbb{R}^{d \times d}$ such that $\Omega_0 + \Delta_\Omega \in \mathbb{S}_+^d$, there exists $t \in (0, 1)$ such that

$$\begin{aligned} & f(\Omega_0 + \Delta_\Omega) - f(\Omega_0) - \langle \nabla f(\Omega_0), \Delta_\Omega \rangle \\ &= \frac{1}{2} \text{vec}(\Delta_\Omega)^\top \left\{ (\Omega_0 + t\Delta_\Omega)^{-1} \otimes (\Omega_0 + t\Delta_\Omega)^{-1} \right\} \text{vec}(\Delta_\Omega). \end{aligned}$$

Proof. One can view f as a function of $\text{vec}(\Omega)$. Note that $\nabla^2 f(\text{vec}(\Omega)) = \Omega^{-1} \otimes \Omega^{-1}$. The rest of the proof follows Taylor's Theorem [Nocedal and Wright, 2006, Theorem 2.1]. $\square$

The following lemma provides a packing set construction on sparse vectors, which is instrumental for establishing the minimax lower bound via information-theoretic arguments.

Lemma 21. Let $\mathbb{S}^p(s) := \{\theta \in \mathbb{R}^p : \|\theta\|_0 \leq s, \|\theta\|_2 \leq 1\}$. There exists $\{\theta^0, \theta^1, \dots, \theta^M\} \subseteq \mathbb{S}^p(s)$ such that

$$(i) \quad \theta^0 = 0;$$

$$(ii) \quad \left\| \theta^j - \theta^k \right\|_2 \geq \frac{1}{4} \text{ for all } 0 \leq j \neq k \leq M;$$

$$(iii) \quad \log(M+1) \geq \frac{s}{2} \log \left(\frac{d-s}{s} \right).$$

Proof. It follows from Example 15.16 in Wainwright [2019] that there exists a $1/2$ -packing of $\mathbb{S}^p(s)$, which we denote as $\{\tilde{\theta}^0, \tilde{\theta}^1, \dots, \tilde{\theta}^M\}$, such that

$$\left\| \tilde{\theta}^j - \tilde{\theta}^k \right\|_2 \geq \frac{1}{2} \quad \text{for all } 0 \leq j \neq k \leq M \quad \text{and} \quad \log(M+1) \geq \frac{s}{2} \log \left(\frac{d-s}{s} \right). \quad (2.43)$$

Without loss of generality, assume that $\|\tilde{\theta}^0\|_2 \leq \|\tilde{\theta}^j\|_2$ for $1 \leq j \leq M$. Let $\theta^0 = 0$ and let $\theta^j = \tilde{\theta}^j$ for $1 \leq j \leq M$. Then the set $\{\theta^0, \theta^1, \dots, \theta^M\}$ satisfies (i)–(iii). To prove the claim, by our construction of the set and (2.43), we only need to verify that $\|\theta^j - \theta^0\|_2 = \|\theta^j\|_2 \geq \frac{1}{4}$ for all $1 \leq j \leq M$. We prove the result in two cases.

Case 1: $\|\tilde{\theta}^0\|_2 \geq \frac{1}{4}$. Since $\|\tilde{\theta}^0\|_2 \leq \|\tilde{\theta}^j\|_2$, we have that $\|\theta^j\|_2 = \|\tilde{\theta}^j\|_2 \geq \|\tilde{\theta}^0\|_2 \geq \frac{1}{4}$ for $1 \leq j \leq M$.

Case 2: $\|\tilde{\theta}^0\|_2 < \frac{1}{4}$. We have

$$\left\| \theta^j \right\|_2 = \left\| \tilde{\theta}^j \right\|_2 \geq \left\| \tilde{\theta}^j - \tilde{\theta}^0 \right\|_2 - \left\| \tilde{\theta}^0 \right\|_2 > \frac{1}{2} - \frac{1}{4} = \frac{1}{4} \quad \text{for } 1 \leq j \leq M,$$

where the last inequality follows from (2.43).

Combining Case 1 and Case 2, we have proved that $\|\theta^j\|_2 \geq \frac{1}{4}$ for $1 \leq j \leq M$, which completes the proof. □

The next lemma bounds the Kullback–Leibler divergence between Gaussian distributions, aiding the information distance control in the lower bound analysis.

Lemma 22. *Let $m \geq 2$ be an even integer and set $d = 2m$. Let $B \in \mathbb{R}^{m \times m}$ be such that $\|B\|_{\text{F}} \leq 1/4$ and define*

$$\Omega = \begin{bmatrix} I_m & B \\ B^\top & I_m \end{bmatrix}.$$

Then $\Omega \succ 0$ and $D_{KL}(N(0, \Omega^{-1}) \| N(0, I_d)) \leq \frac{16}{15} \|B\|_{\text{F}}^2$.

Proof. We begin with proving $\Omega \succ 0$. By Weyl's inequality, we know that

$$\begin{aligned}\gamma_{\min}(\Omega) &\geq 1 - \left\| \begin{bmatrix} 0 & B \\ B^\top & 0 \end{bmatrix} \right\| \geq 1 - \left\| \begin{bmatrix} 0 & B \\ B^\top & 0 \end{bmatrix} \right\|_{\text{F}} \\ &= 1 - \sqrt{2} \|B\|_{\text{F}} \geq \frac{1}{2}.\end{aligned}$$

As a result, we have $\Omega \succ 0$.

Now we move on to the second claim regarding the KL divergence. Recall that

$$D_{\text{KL}} \left(N \left(0, \Omega^{-1} \right) \parallel N \left(0, I_d \right) \right) = \frac{1}{2} \left[\log \det(\Omega) - d + \text{tr} \left(\Omega^{-1} \right) \right], \quad (2.44)$$

which motivates us to compute $\log \det(\Omega)$ and $\text{tr}(\Omega^{-1})$. Let $B = UDV^\top$ be the singular value decomposition of B with $U, D, V \in \mathbb{R}^{m \times m}$, $U^\top U = UU^\top = I_m$, $V^\top V = VV^\top = I_m$, and $D = \text{diag}(\lambda_1, \dots, \lambda_m)$. We have the following identities.

$$\log \det(\Omega) = \sum_{i=1}^m \log \left(1 - \lambda_i^2 \right). \quad (2.45a)$$

$$\text{tr}(\Omega^{-1}) = \sum_{i=1}^m \frac{2}{1 - \lambda_i^2}. \quad (2.45b)$$

Combining (2.44), (2.45a), and (2.45b), we obtain

$$\begin{aligned}D_{\text{KL}} \left(N \left(0, \Omega^{-1} \right) \parallel N \left(0, I_d \right) \right) &= \frac{1}{2} \left[\sum_{i=1}^m \log \left(1 - \lambda_i^2 \right) + \sum_{i=1}^m \frac{2}{1 - \lambda_i^2} - d \right] \\ &= \sum_{i=1}^m \left[\frac{\log \left(1 - \lambda_i^2 \right)}{2} + \frac{1}{1 - \lambda_i^2} - 1 \right].\end{aligned}$$

Since $\lambda_j^2 \leq \sum_{i=1}^d \lambda_i^2 = \|B\|_{\text{F}}^2 \leq 1/16$, we have $-\lambda_j^2 \geq -1/16$. Also note that $\log(1+x) \leq x$

for all $x \geq -1$. We thus have

$$\begin{aligned} D_{\text{KL}} \left(N \left(0, \Omega^{-1} \right) \parallel N \left(0, I_d \right) \right) &\leq \sum_{i=1}^m \left[\frac{-\lambda_i^2}{2} + \frac{1}{1-\lambda_i^2} - 1 \right] \\ &= \sum_{i=1}^m \frac{\lambda_i^2 + \lambda_i^4}{2(1-\lambda_i^2)} \leq \sum_{i=1}^m \frac{\lambda_i^2}{1-\lambda_i^2} \leq \frac{16}{15} \sum_{i=1}^m \lambda_i^2 = \frac{16}{15} \|B\|_{\text{F}}^2. \end{aligned}$$

This completes the proof.

Proof of Equation (2.45a). Using Section 9.1.2 of Petersen et al. [2008], we have $\det(\Omega) = \det(I_m - B^\top B)$. Since

$$I_m - B^\top B = I_m - V D^2 V^\top = V (I_m - D^2) V^\top, \quad (2.46)$$

we have $\det(\Omega) = \det(I_m - D^2) = \prod_{i=1}^m (1 - \lambda_i^2)$ and hence $\log \det(\Omega) = \sum_{i=1}^m \log(1 - \lambda_i^2)$.

Proof of Equation (2.45b). Using Section 9.1.3 of Petersen et al. [2008], we have

$$\Omega^{-1} = \begin{bmatrix} I_m & B \\ B^\top & I_m \end{bmatrix}^{-1} = \begin{bmatrix} (I_m - BB^\top)^{-1} & -B(I_m - B^\top B)^{-1} \\ - (I_m - B^\top B)^{-1} B^\top & (I_m - B^\top B)^{-1} \end{bmatrix},$$

which implies

$$\text{tr}(\Omega^{-1}) = \text{tr} \left\{ (I_m - BB^\top)^{-1} \right\} + \text{tr} \left\{ (I_m - B^\top B)^{-1} \right\}. \quad (2.47)$$

It follows from (2.46) that $(I_m - B^\top B)^{-1} = V (I_m - D^2)^{-1} V^\top$, and therefore,

$$\text{tr} \left\{ (I_m - BB^\top)^{-1} \right\} = \text{tr} \left\{ V (I_m - D^2)^{-1} V^\top \right\} = \text{tr} \left\{ (I_m - D^2)^{-1} \right\} = \sum_{i=1}^m \frac{1}{1-\lambda_i^2}. \quad (2.48)$$

Similarly, we have $\text{tr}\{(I_m - B^\top B)^{-1}\} = \sum_{i=1}^m \frac{1}{1-\lambda_i^2}$. Take the previous results collectively to yield the claim (2.45b). $\square$

2.9.2 Proof of Theorem 1

We adopt the proof strategy laid out in Wainwright [2019, Chapter 9].

Additional Notation

We first introduce some additional notation that will be helpful in the proof. We define the Hilbert space for the parameters

$$\mathbb{H} := \left\{ \Theta = \left(\Omega, \{\Gamma^{(k)}\}_{k=0}^K \right) : \Omega, \Gamma^{(k)} \in \mathbb{R}^{d \times d} \text{ for all } 0 \leq k \leq K \right\},$$

with the associated inner product

$$\langle \Theta, \Theta' \rangle_{\mathbb{H}} = \langle \Theta, \Theta' \rangle := \langle \Omega, \Omega' \rangle + \sum_{k=0}^K \langle \Gamma^{(k)}, \Gamma^{(k)'} \rangle.$$

The space $\mathbb{H}$ endowed with the inner product $\langle \cdot, \cdot \rangle$ is indeed a Hilbert space. Correspondingly, we have

$$\|\Theta\|_{\mathbb{H}}^2 = \|\Omega\|_{\text{F}}^2 + \sum_{k=0}^K \|\Gamma^{(k)}\|_{\text{F}}^2.$$

We also need the dual norm of $\Phi(\cdot)$, which is defined as

$$\Phi^*(\Theta) := \sup_{\Theta': \Phi(\Theta') \leq 1} \langle \Theta, \Theta' \rangle. \quad (2.49)$$

Recall that $\Theta^* = (\Omega^*, \{\Gamma^{(k)*}\}_{k=0}^K)$ is the true parameter. Let $\mathcal{S}_{\Omega}$ and $\mathcal{S}_{\Gamma^{(k)}}$ be the supports of Ω^* and $\Gamma^{(k)*}$ for all $0 \leq k \leq K$, respectively. Under Assumption 1, the true

parameter Θ^* lies in the following subspace of $\mathbb{H}$:

$$\mathbb{M} := \left\{ \Theta = \left(\Omega, \left\{ \Gamma^{(k)} \right\}_{k=0}^K \right) : \text{supp}(\Omega) \subseteq \mathcal{S}_\Omega, \text{supp} \left[\Gamma^{(k)} \right] \subseteq \mathcal{S}_{\Gamma^{(k)}} \text{ for all } 0 \leq k \leq K \right\}. \quad (2.50)$$

The orthogonal complement of $\mathbb{M}$ is given by

$$\mathbb{M}^\perp := \left\{ \Theta = \left(\Omega, \left\{ \Gamma^{(k)} \right\}_{k=0}^K \right) : \text{supp}(\Omega) \subseteq \mathcal{S}_\Omega^c, \text{supp} \left[\Gamma^{(k)} \right] \subseteq \mathcal{S}_{\Gamma^{(k)}}^c \text{ for all } 0 \leq k \leq K \right\}. \quad (2.51)$$

Clearly, for any $\Theta \in \mathbb{M}$ and $\Theta' \in \mathbb{M}^\perp$, we have that $\langle \Theta, \Theta' \rangle = 0$.

We also need to define the projection of a parameter onto a subspace. For any matrix $B \in \mathbb{R}^{d \times d}$ and any subspace $\mathcal{F} \subseteq \mathbb{R}^{d \times d}$, we define

$$[B]_{\mathcal{F}} = \arg \min_{\tilde{B} \in \mathcal{F}} \left\| \tilde{B} - B \right\|_{\text{F}}.$$

Similarly, for any $\Theta \in \mathbb{H}$ and any subspace $\mathbb{F} \subseteq \mathbb{H}$, we define

$$[\Theta]_{\mathbb{F}} = \arg \min_{\tilde{\Theta} \in \mathbb{F}} \left\| \tilde{\Theta} - \Theta \right\|_{\mathbb{H}}.$$

In addition, for $S \subseteq [d] \times [d]$, we define

$$\mathcal{M}(S) := \left\{ B \in \mathbb{R}^{d \times d} : \text{supp}(B) \subseteq S \right\}.$$

For any $B \in \mathbb{R}^{d \times d}$, we define $[B]_S := [B]_{\mathcal{M}(S)}$. This way, for any $\Theta = (\Omega, \{\Gamma^{(k)}\}_{k=0}^K) \in \mathbb{H}$, we have

$$[\Theta]_{\mathbb{M}} = \left([\Omega]_{\mathcal{S}_\Omega}, \left\{ [\Gamma^{(k)}]_{\mathcal{S}_{\Gamma^{(k)}}} \right\}_{k=0}^K \right) \quad \text{and} \quad [\Theta]_{\mathbb{M}^\perp} = \left([\Omega]_{\mathcal{S}_\Omega^c}, \left\{ [\Gamma^{(k)}]_{\mathcal{S}_{\Gamma^{(k)}}^c} \right\}_{k=0}^K \right).$$

Finally, we define our metric on the estimation error.

For $\Theta = (\Omega, \{\Gamma^{(k)}\}_{k=0}^K) \in \mathbb{H}$, let

$$H(\Theta) := \sum_{k=0}^K \alpha_k \left\| \Omega + \Gamma^{(k)} \right\|_{\mathbb{F}}^2. \quad (2.52)$$

We define

$$\widehat{\Delta} := \widehat{\Theta} - \Theta^* = \left(\widehat{\Omega} - \Omega^*, \left\{ \widehat{\Gamma}^{(k)} - \Gamma^{(k)*} \right\}_{k=0}^K \right),$$

where $\widehat{\Theta}$ is the Trans-Glasso estimator (2.2). Then our goal is to control

$$H(\widehat{\Delta}) = \sum_{k=0}^K \alpha_k \left\| \widehat{\Omega} - \Omega^* + \widehat{\Gamma}^{(k)} - \Gamma^{(k)*} \right\|_{\mathbb{F}}^2 = \sum_{k=0}^K \alpha_k \left\| \check{\Omega}^{(k)} - \Omega^{(k)} \right\|_{\mathbb{F}}^2.$$

Useful lemmas

We begin by showing that $\Phi(\cdot)$, defined in (2.4), is decomposable with respect to $(\mathbb{M}, \mathbb{M}^\perp)$.

Lemma 23. *For any $\Theta \in \mathbb{M}$ and $\Theta' \in \mathbb{M}^\perp$, we have*

$$\Phi(\Theta + \Theta') = \Phi(\Theta) + \Phi(\Theta').$$

Proof. By definition of $\Phi(\cdot)$ and the orthogonality of $\Theta \in \mathbb{M}$ and $\Theta' \in \mathbb{M}^\perp$, we have

$$\begin{aligned} \Phi(\Theta + \Theta') &= \Phi([\Theta]_{\mathbb{M}} + [\Theta']_{\mathbb{M}^\perp}) \\ &= \left| [\Omega]_{\mathcal{S}_\Omega} + [\Omega']_{\mathcal{S}_\Omega^c} \right|_1 + \sum_{k=0}^K \sqrt{\alpha_k} \left| [\Gamma^{(k)}]_{\mathcal{S}_{\Gamma^{(k)}}} + [\Gamma^{(k)'}]_{\mathcal{S}_{\Gamma^{(k)}}^c} \right|_1 \\ &= \left| [\Omega]_{\mathcal{S}_\Omega} \right|_1 + \left| [\Omega']_{\mathcal{S}_\Omega^c} \right|_1 + \sum_{k=0}^K \sqrt{\alpha_k} \left(\left| [\Gamma^{(k)}]_{\mathcal{S}_{\Gamma^{(k)}}} \right|_1 + \left| [\Gamma^{(k)'}]_{\mathcal{S}_{\Gamma^{(k)}}^c} \right|_1 \right) \\ &= \Phi([\Theta]_{\mathbb{M}}) + \Phi([\Theta']_{\mathbb{M}^\perp}) \\ &= \Phi(\Theta) + \Phi(\Theta'). \end{aligned}$$

□

The following lemma relates $\Phi(\Theta)$ with $H(\Theta)$ defined in (2.52) for any $\Theta \in \mathbb{M}$.

Lemma 24. *For any $\Theta \in \mathbb{M}$, we have*

$$\Phi(\Theta) \leq \sqrt{2} \sqrt{s + (K+1)h} \sqrt{H(\Theta)}.$$

Proof. For any $\Theta \in \mathbb{M}$, we have $\Omega = [\Omega]_{\mathcal{S}_\Omega}$ and $\Gamma^{(k)} = [\Gamma^{(k)}]_{\mathcal{S}_{\Gamma^{(k)}}}$. Thus

$$\begin{aligned} \Phi(\Theta) &= \left| [\Omega]_{\mathcal{S}_\Omega} \right|_1 + \sum_{k=0}^K \sqrt{\alpha_k} \left| [\Gamma^{(k)}]_{\mathcal{S}_{\Gamma^{(k)}}} \right|_1 \\ &\leq |\mathcal{S}_\Omega|^{\frac{1}{2}} \left\| [\Omega]_{\mathcal{S}_\Omega} \right\|_{\mathbb{F}} + \sum_{k=0}^K \sqrt{\alpha_k} |\mathcal{S}_{\Gamma^{(k)}}|^{\frac{1}{2}} \left\| [\Gamma^{(k)}]_{\mathcal{S}_{\Gamma^{(k)}}} \right\|_{\mathbb{F}} \quad (\text{Jensen's inequality}) \\ &\stackrel{(ii)}{\leq} \sqrt{s} \left\| [\Omega]_{\mathcal{S}_\Omega} \right\|_{\mathbb{F}} + \sqrt{h} \sum_{k=0}^K \sqrt{\alpha_k} \left\| [\Gamma^{(k)}]_{\mathcal{S}_{\Gamma^{(k)}}} \right\|_{\mathbb{F}}. \quad (\text{by Assumption 1}) \end{aligned}$$

Furthermore, by Jensen's inequality, we have

$$\begin{aligned} \frac{1}{2} \Phi^2(\Theta) &\leq s \left\| [\Omega]_{\mathcal{S}_\Omega} \right\|_{\mathbb{F}}^2 + h \left(\sum_{k=0}^K \sqrt{\alpha_k} \left\| [\Gamma^{(k)}]_{\mathcal{S}_{\Gamma^{(k)}}} \right\|_{\mathbb{F}} \right)^2 \\ &\leq s \left\| [\Omega]_{\mathcal{S}_\Omega} \right\|_{\mathbb{F}}^2 + h(K+1) \sum_{k=0}^K \alpha_k \left\| [\Gamma^{(k)}]_{\mathcal{S}_{\Gamma^{(k)}}} \right\|_{\mathbb{F}}^2 \\ &\leq (s + (K+1)h) \left\{ \left\| [\Omega]_{\mathcal{S}_\Omega} \right\|_{\mathbb{F}}^2 + \sum_{k=0}^K \alpha_k \left\| [\Gamma^{(k)}]_{\mathcal{S}_{\Gamma^{(k)}}} \right\|_{\mathbb{F}}^2 \right\} \\ &= (s + (K+1)h) \sum_{k=0}^K \alpha_k \left(\left\| [\Omega]_{\mathcal{S}_\Omega} \right\|_{\mathbb{F}}^2 + \left\| [\Gamma^{(k)}]_{\mathcal{S}_{\Gamma^{(k)}}} \right\|_{\mathbb{F}}^2 \right). \end{aligned}$$

By the assumption that the supports of $\Omega^\star$ and $\Gamma^{(k)\star}$ are disjoint and the fact that $\Omega = [\Omega]_{\mathcal{S}_\Omega}$

and $\Gamma^{(k)} = \left[\Gamma^{(k)} \right]_{\mathcal{S}_{\Gamma^{(k)}}}$, we then have

$$\begin{aligned} \Phi^2(\Theta) &\leq 2(s + (K+1)h) \sum_{k=0}^K \alpha_k \left\| [\Omega]_{\mathcal{S}_{\Omega}} + \left[\Gamma^{(k)} \right]_{\mathcal{S}_{\Gamma^{(k)}}} \right\|_{\mathbb{F}}^2 \\ &\leq 2(s + (K+1)h) \sum_{k=0}^K \alpha_k \left\| \Omega + \Gamma^{(k)} \right\|_{\mathbb{F}}^2 \\ &= 2(s + (K+1)h) H(\Theta). \end{aligned}$$

□

For $\Delta_{\Theta} = (\Delta_{\Omega}, \{\Delta_{\Gamma^{(k)}}\}_{k=0}^K) \in \mathbb{H}$ such that $\Delta_{\Omega} + \Delta_{\Gamma^{(k)}} + \Omega^{(k)} \succ 0$ for all $0 \leq k \leq K$, we define

$$\mathcal{R}(\Delta_{\Theta}) := \mathcal{L}(\Theta^* + \Delta_{\Theta}) - \mathcal{L}(\Theta^*) - \langle \nabla \mathcal{L}(\Theta^*), \Delta_{\Theta} \rangle$$

to be the residual of $\mathcal{L}(\cdot)$ around Θ^* , where $\mathcal{L}(\cdot)$ is defined in (2.3). The following lemma claims that $\mathcal{R}(\cdot)$ is locally strongly convex with respect to the geometry defined by $H(\cdot)$.

Lemma 25. *Let $\Delta_{\Theta} = (\Delta_{\Omega}, \{\Delta_{\Gamma^{(k)}}\}_{k=0}^K) \in \mathbb{H}$ be such that $\Delta_{\Omega} + \Delta_{\Gamma^{(k)}} + \Omega^{(k)} \succ 0$ and assume that $\|\Delta_{\Omega} + \Delta_{\Gamma^{(k)}}\|_2 \leq M_{\text{op}} + M_{\Omega}$ for all $0 \leq k \leq K$. Then, we have*

$$\mathcal{R}(\Delta_{\Theta}) \geq \frac{\kappa}{2} H(\Delta_{\Theta}),$$

where $\kappa = (2M_{\Omega} + M_{\text{op}})^{-2}$.

Proof. By Lemma 20, we have

$$\begin{aligned} \mathcal{R}(\Delta_{\Theta}) &= \sum_{k=0}^K \frac{\alpha_k}{2} \text{vec}(\Delta_{\Omega} + \Delta_{\Gamma^{(k)}}) \\ &\quad \times \left\{ \left(\Omega^{(k)} + t_k(\Delta_{\Omega} + \Delta_{\Gamma^{(k)}}) \right)^{-1} \otimes \left(\Omega^{(k)} + t_k(\Delta_{\Omega} + \Delta_{\Gamma^{(k)}}) \right)^{-1} \right\} \\ &\quad \times \text{vec}(\Delta_{\Omega} + \Delta_{\Gamma^{(k)}}), \quad (2.53) \end{aligned}$$

where $t_k \in (0, 1)$, $0 \leq k \leq K$. Since $\gamma_{\min}(A^{-1} \otimes A^{-1}) = \|A\|_2^{-2}$ for any $A \succ 0$, we have

$$\begin{aligned}
& \gamma_{\min} \left(\left(\Omega^{(k)} + t_k (\Delta_\Omega + \Delta_{\Gamma^{(k)}}) \right)^{-1} \otimes \left(\Omega^{(k)} + t_k (\Delta_\Omega + \Delta_{\Gamma^{(k)}}) \right)^{-1} \right) \\
&= \left\{ \left\| \Omega^{(k)} + t_k (\Delta_\Omega + \Delta_{\Gamma^{(k)}}) \right\|_2 \right\}^{-2} \\
&\geq \left\{ \left\| \Omega^{(k)} \right\|_2 + t_k \left\| \Delta_\Omega + \Delta_{\Gamma^{(k)}} \right\|_2 \right\}^{-2} \\
&\geq \left\{ \left\| \Omega^{(k)} \right\|_2 + \left\| \Delta_\Omega + \Delta_{\Gamma^{(k)}} \right\|_2 \right\}^{-2} \\
&\geq (2M_\Omega + M_{\text{op}})^{-2},
\end{aligned}$$

where the last line follows the definition of M_Ω in (2.24) and the assumption that

$$\left\| \Delta_\Omega + \Delta_{\Gamma^{(k)}} \right\|_2 \leq M_\Omega + M_{\text{op}} \quad \text{for all } 0 \leq k \leq K.$$

Let $\kappa = (2M_\Omega + M_{\text{op}})^{-2}$. Then

$$\mathcal{R}(\Delta_\Theta) \geq \frac{\kappa}{2} \sum_{k=0}^K \alpha_k \left\| \text{vec}(\Delta_\Omega + \Delta_{\Gamma^{(k)}}) \right\|_2^2.$$

The final result follows by noting that $\left\| \text{vec}(\Delta_\Omega + \Delta_{\Gamma^{(k)}}) \right\|_2 = \left\| \Delta_\Omega + \Delta_{\Gamma^{(k)}} \right\|_{\text{F}}$.

□

The following lemma characterizes the dual norm $\Phi^*(\cdot)$, which plays a key role in deriving the upper bound by controlling the stochastic error term.

Lemma 26. *For the dual norm defined in (2.49), we have*

$$\Phi^*(\Theta) = \max \left\{ |\Omega|_\infty, \max_{0 \leq k \leq K} \frac{|\Gamma^{(k)}|_\infty}{\sqrt{\alpha_k}} \right\}.$$

Proof. For any $\Theta, \Theta' \in \mathbb{H}$, we have

$$\begin{aligned}
\langle \Theta, \Theta' \rangle &= \langle \Omega, \Omega' \rangle + \sum_{k=0}^K \langle \Gamma^{(k)}, \Gamma^{(k)'} \rangle \\
&\leq |\Omega|_\infty |\Omega'|_1 + \sum_{k=0}^K \left| \Gamma^{(k)} \right|_\infty |\Gamma'|_1 \\
&= |\Omega|_\infty |\Omega'|_1 + \sum_{k=0}^K \sqrt{\alpha_k} |\Gamma'|_1 \frac{|\Gamma^{(k)}|_\infty}{\sqrt{\alpha_k}} \\
&\leq |\Omega|_\infty |\Omega'|_1 + \left\{ \max_{0 \leq k \leq K} \frac{|\Gamma^{(k)}|_\infty}{\sqrt{\alpha_k}} \right\} \sum_{k=0}^K \sqrt{\alpha_k} |\Gamma'|_1 \\
&\leq \max \left\{ |\Omega|_\infty, \max_{0 \leq k \leq K} \frac{|\Gamma^{(k)}|_\infty}{\sqrt{\alpha_k}} \right\} \left(|\Omega'|_1 + \sum_{k=0}^K \sqrt{\alpha_k} |\Gamma'|_1 \right) \\
&= \max \left\{ |\Omega|_\infty, \max_{0 \leq k \leq K} \frac{|\Gamma^{(k)}|_\infty}{\sqrt{\alpha_k}} \right\} \Phi(\Theta').
\end{aligned}$$

Thus, we have

$$\Phi^*(\Theta) := \sup_{\Theta': \Phi(\Theta') \leq 1} \langle \Theta, \Theta' \rangle \leq \max \left\{ |\Omega|_\infty, \max_{0 \leq k \leq K} \frac{|\Gamma^{(k)}|_\infty}{\sqrt{\alpha_k}} \right\}.$$

Finally, it is easy to see that the equality is achievable. □

Last but not least, define the event

$$\mathbb{G}(\lambda_M) := \left\{ \frac{\lambda_M}{2} \geq \Phi^*(\nabla \mathcal{L}(\Theta^*)) \right\}. \tag{2.54}$$

The following lemma asserts that this event happens with high probability.

Lemma 27. *When (2.25) holds and λ_M satisfies (2.26), we have that $\mathbb{P}\{\mathbb{G}(\lambda_M)\} \geq 1 - \delta$.*

Proof. Note that

$$\begin{aligned}\nabla_{\Omega} \mathcal{L}(\Theta) &= \sum_{k=0}^K \alpha_k \left(\widehat{\Sigma}^{(k)} - \left(\Omega + \Gamma^{(k)} \right)^{-1} \right), \\ \nabla_{\Gamma^{(k)}} \mathcal{L}(\Theta) &= \alpha_k \left(\widehat{\Sigma}^{(k)} - \left(\Omega + \Gamma^{(k)} \right)^{-1} \right) \quad \text{for all } 0 \leq k \leq K.\end{aligned}$$

Then by Lemma 26, we have

$$\mathbb{G}(\lambda_M) = \left\{ \frac{\lambda_M}{2} \geq \max \left\{ \left| \sum_{k=0}^K \alpha_k \left(\widehat{\Sigma}^{(k)} - \Sigma^{(k)} \right) \right|_{\infty}, \max_{0 \leq k \leq K} \sqrt{\alpha_k} \left| \widehat{\Sigma}^{(k)} - \Sigma^{(k)} \right|_{\infty} \right\} \right\}. \quad (2.55)$$

By Lemma 18 and the union bound, we have

$$\left| \sum_{k=0}^K \alpha_k \left(\widehat{\Sigma}^{(k)} - \Sigma^{(k)} \right) \right|_{\infty} \leq 80 M_{\Sigma} \max \left\{ \sqrt{\frac{\log(2(K+2)d^2/\delta)}{2N}}, \frac{\log(2(K+2)d^2/\delta)}{N} \right\}$$

and

$$\left| \widehat{\Sigma}^{(k)} - \Sigma^{(k)} \right|_{\infty} \leq 80 M_{\Sigma} \max \left\{ \sqrt{\frac{\log(2(K+2)d^2/\delta)}{2n_k}}, \frac{\log(2(K+2)d^2/\delta)}{n_k} \right\}$$

for all $0 \leq k \leq K$ hold simultaneously with probability at least $1 - \delta$. When $\min_{0 \leq k \leq K} n_k$ is large enough such that

$$\frac{\log(2(K+2)d^2/\delta)}{\min_{0 \leq k \leq K} n_k} \leq \frac{1}{2},$$

we then have that

$$\begin{aligned} \left| \sum_{k=0}^K \alpha_k \left(\widehat{\Sigma}^{(k)} - \Sigma^{(k)} \right) \right|_{\infty} &\leq 80 M_{\Sigma} \sqrt{\frac{\log(2(K+2)d^2/\delta)}{2N}}, \\ \left| \widehat{\Sigma}^{(k)} - \Sigma^{(k)} \right|_{\infty} &\leq 80 M_{\Sigma} \sqrt{\frac{\log(2(K+2)d^2/\delta)}{2n_k}} \quad \text{for all } 0 \leq k \leq K. \end{aligned} \quad (2.56)$$

hold simultaneously with probability at least $1 - \delta$. Note that (2.56) implies that

$$\sqrt{\alpha_k} \left| \widehat{\Sigma}^{(k)} - \Sigma^{(k)} \right|_{\infty} \leq 80M_{\Sigma} \sqrt{\frac{\log(2(K+2)d^2/\delta)}{2N}} \quad \text{for all } 0 \leq k \leq K.$$

Thus, when

$$\lambda_M \geq 160M_{\Sigma} \sqrt{\frac{\log(2(K+2)d^2/\delta)}{2N}}, \quad (2.57)$$

by (2.55), we have $\mathbb{G}(\lambda_M)$ hold with probability $1 - \delta$.

□

Remaining proof

Now we are ready to prove Theorem 1. As Lemma 27 asserts that $\mathbb{G}(\lambda_M)$ holds with high probability, thus we only need to prove the conclusion under the assumption that $\mathbb{G}(\lambda_M)$ holds. Throughout the proof, we assume the event $\mathbb{G}(\lambda_M)$ holds.

For any $\Delta_{\Theta} = (\Delta_{\Omega}, \{\Delta_{\Gamma^{(k)}}\}_{k=0}^K) \in \mathbb{H}$, we define the objective difference

$$F(\Delta_{\Theta}) := \mathcal{L}(\Theta^* + \Delta_{\Theta}) - \mathcal{L}(\Theta^*) + \lambda_M \{\Phi(\Theta^* + \Delta_{\Theta}) - \Phi(\Theta^*)\}.$$

Note that by definition, we have $F(\widehat{\Delta}_{\Theta}) \leq 0$. It suffices to show that for all Δ_{Θ} such that $\Delta_{\Theta} + \Theta^* \in \mathcal{C}(M_{\text{op}})$ and $H(\Delta_{\Theta}) > \frac{18(s+(K+1)h)\lambda_M^2}{\kappa^2}$, we have $F(\Delta_{\Theta}) > 0$.

It is easy to see that any such Δ_{Θ} obeys

$$\Delta\Omega + \Delta\Gamma_k + \Omega^{(k)} = \Omega - \Omega^* + \Gamma_k - \Gamma^{(k)*} + \Omega^{(k)} = \Omega + \Gamma_k \succ 0,$$

and for all $0 \leq k \leq K$

$$\begin{aligned} \|\Delta\Omega + \Delta\Gamma_k\|_2 &= \left\| \Omega - \Omega^* + \Gamma_k - \Gamma^{(k)*} \right\|_2 = \left\| \Omega + \Gamma_k - \Omega^{(k)} \right\|_2 \\ &\leq \|\Omega + \Gamma_k\|_2 + \left\| \Omega^{(k)} \right\|_2 \leq M_{\text{op}} + M_{\Omega}, \end{aligned}$$

where the last inequality follows the definition of M_{op} and M_{Ω} . These two taken together allow us to invoke Lemma 25 to obtain

$$F(\Delta_{\Theta}) \geq \langle \nabla \mathcal{L}(\Theta^*), \Delta_{\Theta} \rangle + \frac{\kappa}{2} H(\Delta_{\Theta}) + \lambda_{\mathbb{M}} \{ \Phi(\Theta^* + \Delta_{\Theta}) - \Phi(\Theta^*) \},$$

where $\kappa = (2M_{\Omega} + M_{\text{op}})^{-2}$.

In addition, combining Lemma 23 and the fact that $[\Theta^*]_{\mathbb{M}^{\perp}} = 0$ with Wainwright [2019, Lemma 9.14], we have

$$\Phi(\Theta^* + \Delta_{\Theta}) - \Phi(\Theta^*) \geq \Phi([\Delta_{\Theta}]_{\mathbb{M}^{\perp}}) - \Phi([\Delta_{\Theta}]_{\mathbb{M}}) \quad (2.58)$$

for any $\Delta_{\Theta} \in \mathbb{H}$. This way, we have obtained a lower bound for $\Phi(\Theta^* + \Delta_{\Theta}) - \Phi(\Theta^*)$.

Take the previous two displays together to reach

$$F(\Delta_{\Theta}) \geq \langle \nabla \mathcal{L}(\Theta^*), \Delta_{\Theta} \rangle + \frac{\kappa}{2} H(\Delta_{\Theta}) + \lambda_{\mathbb{M}} \{ \Phi([\Delta_{\Theta}]_{\mathbb{M}^{\perp}}) - \Phi([\Delta_{\Theta}]_{\mathbb{M}}) \}.$$

Recall the definition of Φ^* in (2.49), under the assumption that $\mathbb{G}(\lambda_{\mathbb{M}})$ is true, we have

$$\begin{aligned} F(\Delta_{\Theta}) &\geq \frac{\kappa}{2} H(\Delta_{\Theta}) - |\langle \nabla \mathcal{L}(\Theta^*), \Delta_{\Theta} \rangle| + \lambda_{\mathbb{M}} \{ \Phi([\Delta_{\Theta}]_{\mathbb{M}^{\perp}}) - \Phi([\Delta_{\Theta}]_{\mathbb{M}}) \} \\ &\geq \frac{\kappa}{2} H(\Delta_{\Theta}) - \Phi(\Delta_{\Theta}) \Phi^*(\nabla \mathcal{L}(\Theta^*)) + \lambda_{\mathbb{M}} \{ \Phi([\Delta_{\Theta}]_{\mathbb{M}^{\perp}}) - \Phi([\Delta_{\Theta}]_{\mathbb{M}}) \} \\ &\geq \frac{\kappa}{2} H(\Delta_{\Theta}) - \frac{\lambda_{\mathbb{M}}}{2} \Phi(\Delta_{\Theta}) + \lambda_{\mathbb{M}} \{ \Phi([\Delta_{\Theta}]_{\mathbb{M}^{\perp}}) - \Phi([\Delta_{\Theta}]_{\mathbb{M}}) \} \\ &= \frac{\kappa}{2} H(\Delta_{\Theta}) - \frac{\lambda_{\mathbb{M}}}{2} \{ \Phi([\Delta_{\Theta}]_{\mathbb{M}}) + \Phi([\Delta_{\Theta}]_{\mathbb{M}^{\perp}}) \} + \lambda_{\mathbb{M}} \{ \Phi([\Delta_{\Theta}]_{\mathbb{M}^{\perp}}) - \Phi([\Delta_{\Theta}]_{\mathbb{M}}) \} \\ &= \frac{\kappa}{2} H(\Delta_{\Theta}) - \frac{\lambda_{\mathbb{M}}}{2} \{ 3\Phi([\Delta_{\Theta}]_{\mathbb{M}}) - \Phi([\Delta_{\Theta}]_{\mathbb{M}^{\perp}}) \} \\ &\geq \frac{\kappa}{2} H(\Delta_{\Theta}) - \frac{3\lambda_{\mathbb{M}}}{2} \Phi([\Delta_{\Theta}]_{\mathbb{M}}). \end{aligned} \quad (2.59)$$

By Lemma 24, we have

$$\Phi([\Delta_\Theta]_{\mathbb{M}}) \leq \sqrt{s + (K+1)h} \sqrt{H([\Delta_\Theta]_{\mathbb{M}})} \leq \sqrt{2} \sqrt{s + (K+1)h} \sqrt{H(\Delta_\Theta)},$$

where the last inequality follows from

$$H(\Delta_\Theta) = H([\Delta_\Theta]_{\mathbb{M}}) + H([\Delta_\Theta]_{\mathbb{M}^\perp}) \geq H([\Delta_\Theta]_{\mathbb{M}}).$$

Plugging the above inequality into (2.59), we arrive at the conclusion that

$$\begin{aligned} F(\Delta_\Theta) &\geq \frac{\kappa}{2} H(\Delta_\Theta) - \frac{3\lambda_{\mathbb{M}}}{\sqrt{2}} \sqrt{s + (K+1)h} \sqrt{H(\Delta_\Theta)} \\ &= \frac{1}{2} \sqrt{H(\Delta_\Theta)} \left\{ \kappa \sqrt{H(\Delta_\Theta)} - 3\sqrt{2}\lambda_{\mathbb{M}} \sqrt{s + (K+1)h} \right\} > 0, \end{aligned}$$

where the last relation arises from the assumption

$$H(\Delta_\Theta) > \frac{18(s + (K+1)h)\lambda_{\mathbb{M}}^2}{\kappa^2}. \quad (2.60)$$

This finishes the proof.

2.9.3 Proof of Theorem 5

Note that for any matrix $A \in \mathbb{R}^{d \times d}$, we have $\|A\|_{\text{F}} \leq \sqrt{d} \|A\|_2$. Since by our assumptions that $\|\Omega^{(k)}\|_2 = O(1)$ and $\|\check{\Omega}^{(k)}\|_2 = O(1)$, we thus have

$$\|\Omega^{(k)}\|_{\text{F}} = O\left(d^{\frac{1}{2}}\right) \quad \text{and} \quad \|\check{\Omega}^{(k)}\|_{\text{F}} = O\left(d^{\frac{1}{2}}\right).$$

For $\delta \in (0, 1]$, let

$$\lambda_{\mathbb{M}} \geq C_1 C_3 \sqrt{\frac{\log(2(K+2)d^2/\delta)}{2N}},$$

where $C_1 = 160$ and $C_3 = M_\Sigma$. Besides, by the proof of Theorem 1, when

$$\frac{\log(2(K+2)d^2/\delta)}{\min_{0 \leq k \leq K} n_k} \leq \frac{1}{2},$$

then by (2.57), we have $\mathbb{P}\{\mathbb{G}(\lambda_M)\} \geq 1 - \delta$; or equivalently, we have $\mathbb{P}\{\overline{\mathbb{G}}(\lambda_M)\} \leq \delta$, where $\overline{\mathbb{G}}(\lambda_M)$ denotes the event that $\mathbb{G}(\lambda_M)$ does not hold.

Recall that by assumption we have $2(K+2) \leq d^{\tau_1}$. Let $\delta = d^{-\tau'}$, where τ' will be specified later, then by letting

$$\lambda_M = C_1 C_3 \sqrt{\frac{(\tau' + \tau_1 + 2) \log d}{2N}}, \quad (2.61)$$

$$\frac{(\tau' + \tau_1 + 2) \log d}{\min_{0 \leq k \leq K} n_k} \leq \frac{1}{2}, \quad (2.62)$$

we have $\mathbb{P}\{\overline{\mathbb{G}}(\lambda_M)\} \leq d^{-\tau'}$. Besides, by Section 2.9.2, when (2.61)–(2.62) are true, $\mathbb{G}(\lambda_M)$ then implies that

$$\begin{aligned} \sum_{k=0}^K \alpha_k \left\| \check{\Omega}^{(k)} - \Omega^{(k)} \right\|_F^2 &\leq \frac{9(s + (K+1)h) \lambda_M^2}{4\kappa^2} \\ &\leq \frac{9C_1^2 C_3^2 (\tau' + \tau_1 + 2)}{8} \left(\frac{s}{N} + \frac{h}{n} \right) \log d. \end{aligned}$$

Note that

$$\begin{aligned} &\mathbb{E} \left[\sum_{k=0}^K \alpha_k \left\| \check{\Omega}^{(k)} - \Omega^{(k)} \right\|_F^2 \right] \\ &= \mathbb{P}\{\mathbb{G}(\lambda_M)\} \mathbb{E} \left[\sum_{k=0}^K \alpha_k \left\| \check{\Omega}^{(k)} - \Omega^{(k)} \right\|_F^2 \middle| \mathbb{G}(\lambda_M) \right] \\ &\quad + \mathbb{P}\{\overline{\mathbb{G}}(\lambda_M)\} \mathbb{E} \left[\sum_{k=0}^K \alpha_k \left\| \check{\Omega}^{(k)} - \Omega^{(k)} \right\|_F^2 \middle| \overline{\mathbb{G}}(\lambda_M) \right]. \end{aligned} \quad (2.63)$$

Given (2.61)–(2.62) are true, we have

$$\mathbb{E} \left[\sum_{k=0}^K \alpha_k \left\| \check{\Omega}^{(k)} - \Omega^{(k)} \right\|_{\text{F}}^2 \middle| \mathbb{G}(\lambda_{\text{M}}) \right] \leq \frac{9C_1^2 C_3^2 (\tau' + \tau_1 + 2)}{8} \left(\frac{s}{N} + \frac{h}{n} \right) \log d. \quad (2.64)$$

Besides, since $\|\check{\Omega}^{(k)}\|_{\text{F}}, \|\Omega^{(k)}\|_{\text{F}} = O(d^{\frac{1}{2}})$ for all $0 \leq k \leq K$, we have

$$\sum_{k=0}^K \alpha_k \left\| \check{\Omega}^{(k)} - \Omega^{(k)} \right\|_{\text{F}}^2 \leq C' d,$$

for some constant $C' > 0$, and thus we have

$$\mathbb{P} \{ \overline{\mathbb{G}}(\lambda_{\text{M}}) \} \mathbb{E} \left[\sum_{k=0}^K \alpha_k \left\| \check{\Omega}^{(k)} - \Omega^{(k)} \right\|_{\text{F}}^2 \middle| \overline{\mathbb{G}}(\lambda_{\text{M}}) \right] \leq C' d \cdot d^{-\tau'} = C' d^{-(\tau'-1)}.$$

By the assumption that we have $N \leq d^{\tau_2}$ where $\tau_2 > 0$, then when we choose τ' such that

$$\tau' \geq \tau_2 + 1 + \frac{\log(8C'/9C_1^2 C_3^2 (\tau' + \tau_1 + 2))}{\log d},$$

we then have

$$\begin{aligned} C' d^{-(\tau'-1)} &\leq \frac{9C_1^2 C_3^2 (\tau' + \tau_1 + 2)}{8} \cdot \frac{1}{N} \\ &\leq \frac{9C_1^2 C_3^2 (\tau' + \tau_1 + 2)}{8} \left(\frac{s}{N} + \frac{h}{n} \right) \log d, \end{aligned}$$

which then implies that

$$\mathbb{P} \{ \overline{\mathbb{G}}(\lambda_{\text{M}}) \} \mathbb{E} \left[\sum_{k=0}^K \alpha_k \left\| \check{\Omega}^{(k)} - \Omega^{(k)} \right\|_{\text{F}}^2 \middle| \overline{\mathbb{G}}(\lambda_{\text{M}}) \right] \leq \mathbb{E} \left[\sum_{k=0}^K \alpha_k \left\| \check{\Omega}^{(k)} - \Omega^{(k)} \right\|_{\text{F}}^2 \middle| \mathbb{G}(\lambda_{\text{M}}) \right].$$

Combine the above inequality with (2.63) and (2.64), we finally have

$$\begin{aligned}
\mathbb{E} \left[\sum_{k=0}^K \alpha_k \left\| \check{\Omega}^{(k)} - \Omega^{(k)} \right\|_{\text{F}}^2 \right] &\leq 2 \mathbb{E} \left[\sum_{k=0}^K \alpha_k \left\| \check{\Omega}^{(k)} - \Omega^{(k)} \right\|_{\text{F}}^2 \middle| \mathbb{G}(\lambda_{\text{M}}) \right] \\
&\leq \frac{18C_1^2 C_3^2 (\tau' + \tau_1 + 2)}{8} \left(\frac{s}{N} + \frac{h}{n} \right) \log d \\
&\lesssim \left(\frac{s}{N} + \frac{h}{n} \right) \log d.
\end{aligned}$$

2.9.4 Proof of Lemma 9

Let $R = \lfloor s/d \rfloor$ and let $r = s - Rd$. We follow the construction of the hard-case collection of precision matrices in Step 2 of Proof of Theorem 4.1 in Cai et al. [2016] to get $\mathcal{F}_\star$. We set $M_{n,p}$ and $c_{n,p}$ therein as $M_{n,p} = \text{some constant}$ and $c_{n,p} = R$. By (4.13) and (4.15) in Cai et al. [2016], for any $\Omega \in \mathcal{F}_\star$, we have

$$0 < c_1 \leq \gamma_{\min}(\Omega) \leq \gamma_{\max}(\Omega) \leq c_2 < \infty.$$

Furthermore, we have $|\Omega|_0 \leq c_{n,p}d \leq s$. Thus, $\mathcal{F}_\star \subseteq \mathcal{G}_1$. Following Theorem 6.1 of Cai et al. [2016], we have

$$\begin{aligned}
&\inf_{\hat{\Omega}} \sup_{\Omega \in \mathcal{G}_1} \mathbb{E} \left[\left\| \hat{\Omega} - \Omega \right\|_{\text{F}}^2 \right] \\
&\geq \inf_{\hat{\Omega}} \sup_{\Omega \in \mathcal{F}_\star} \mathbb{E} \left[\left\| \hat{\Omega} - \Omega \right\|_{\text{F}}^2 \right] \gtrsim dc_{n,p} \frac{\log d}{n} = \frac{dR}{s} \cdot \frac{s \log d}{n} \geq \frac{1}{2} \frac{s \log d}{n},
\end{aligned}$$

which completes the proof.

2.9.5 Proof of Lemma 10

Without loss of generality, we assume that d and h are even positive numbers.

Let $m = d/2$. By Lemma 21, there exist $\{b^0, b^1, \dots, b^M\} \subseteq \mathbb{R}^{m^2}$ such that

- (i) $\|b^j\|_0 \leq \frac{h}{2}$ and $\|b^j\|_2 \leq 1$ for all $0 \leq j \leq M$;
- (ii) $b^0 = 0$;
- (iii) $\|b^j - b^k\|_2 \geq \frac{1}{4}$ for all $0 \leq j \neq k \leq M$;
- (iv) $\log(M+1) \geq \frac{h}{4} \log \left(\frac{d^2/4 - h/2}{h/2} \right)$.

For $j = 0, \dots, M$, let $B^j \in \mathbb{R}^{m \times m}$ be such that $\text{vec}(B^j) = \delta \cdot b^j$, where δ is a positive number that depends on n, d, h and will be specified later. By (2.35), $d \geq 4h$ and

$$\log(M+1) \geq \frac{h}{4} \log \left(\frac{d^2/4 - h/2}{h/2} \right) = \frac{h}{4} \log \left(\frac{d^2}{2h} - 1 \right) \geq \frac{h}{4} \log \left(\frac{d^2}{4h} \right) \geq \frac{h}{4} \log d.$$

Furthermore, we have

$$\begin{aligned} |B^j|_0 &\leq \frac{h}{2}, \quad \|B^j\|_F \leq \delta, \quad \text{for } 0 \leq j \leq M, \\ B^0 &= 0, \quad \|B^j - B^k\|_F \geq \frac{\delta}{4}, \quad \text{for } 0 \leq j \neq k \leq M. \end{aligned}$$

For $j = 0, \dots, M$, let

$$\Omega^j = \begin{bmatrix} I_m & B^j \\ (B^j)^\top & I_m \end{bmatrix}.$$

We next verify that $\Omega^j \in \mathcal{G}_2$ when

$$\delta \leq \min \left\{ \frac{C_\Gamma}{\sqrt{2h}}, \frac{1 - c_1}{2}, \frac{c_2 - 1}{2} \right\}, \quad (2.65)$$

where $\mathcal{G}_2$ is defined in (2.34). First note that $\text{diag}(\Omega^j) = I_d$, $|\Omega^j - I_d|_0 = 2|B^j|_0 \leq h$, and

$$|\Omega^j - I_d|_1 = 2|B^j|_1 \leq 2\sqrt{\frac{h}{2}} \|B^j\|_F \leq \sqrt{2h}\delta \leq C_\Gamma,$$

using the choice of δ in (2.65). Furthermore, we have

$$\begin{aligned}
\gamma_{\min}(\Omega^j) &= \gamma_{\min} \left(\begin{bmatrix} I_m & 0 \\ 0 & I_m \end{bmatrix} + \begin{bmatrix} 0 & B^j \\ (B^j)^\top & 0 \end{bmatrix} \right) \\
&\geq 1 - \left\| \begin{bmatrix} 0 & B^j \\ (B^j)^\top & 0 \end{bmatrix} \right\|_2 \\
&\geq 1 - \left\| \begin{bmatrix} 0 & B^j \\ (B^j)^\top & 0 \end{bmatrix} \right\|_F \\
&= 1 - 2 \|B^j\|_F \\
&\geq c_1,
\end{aligned}$$

again using the choice of δ in (2.65). Similarly, we have $\gamma_{\max}(\Omega^j) \leq c_2$. Therefore, when (2.65) holds, we have that $\Omega^j \in \mathcal{G}_2$.

Let $\mathbb{P}_j$ denote the probability measure of $N(0, \{\Omega^j\}^{-1})$ and let $\mathbb{P}_j^n$ denote the product probability measure of $(X_1, \dots, X_n)$ where $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} \mathbb{P}_j$. Note that $\mathbb{P}_0 = N(0, I_d)$. When $\delta \leq 1/4$, Lemma 22 gives us

$$D_{\text{KL}}(\mathbb{P}_j^n \parallel \mathbb{P}_0^n) = n D_{\text{KL}}(\mathbb{P}_j \parallel \mathbb{P}_0) \leq \frac{16}{15} n \|B^j\|_F^2 \leq \frac{16}{15} n \delta^2,$$

which implies that

$$\frac{1}{M} \sum_{j=1}^M D_{\text{KL}}(\mathbb{P}_j^n \parallel \mathbb{P}_0^n) \leq \frac{16}{15} n \delta^2. \quad (2.66)$$

From condition (2.35), we have $h \log d \geq 8 \log 3$ and thus $\log(M+1) \geq \log 9$, which implies that $M \geq 8$. Furthermore, we have

$$\log M = \frac{\log M}{\log(M+1)} \log(M+1) \geq \frac{\log 8}{\log 9} \log(M+1) \geq \frac{4}{5} \log(M+1) \geq \frac{h}{5} \log d.$$

We set

$$\delta = \frac{1}{4\sqrt{2}} \sqrt{\frac{h \log d}{n}}.$$

By (2.35), we have that (2.65) holds and that $\delta \leq 1/4$. Thus (2.66) holds, implying

$$\frac{1}{M} \sum_{j=1}^M D_{\text{KL}} \left(\mathbb{P}_j^n \parallel \mathbb{P}_0^n \right) \leq \frac{1}{30} h \log d \leq \frac{1}{6} \log M. \quad (2.67)$$

In addition, we have

$$\left\| B^j - B^k \right\|_{\text{F}} \geq 2 \cdot \frac{1}{32\sqrt{2}} \sqrt{\frac{h \log d}{n}} \quad \text{for all } 0 \leq j \neq k \leq M. \quad (2.68)$$

Note that

$$\frac{\log(M+1) - \log 2}{\log M} - \frac{1}{6} \geq \frac{\log M - \log 2}{\log M} - \frac{1}{6} \geq \frac{\log 8 - \log 2}{\log 8} - \frac{1}{6} = \frac{1}{2}. \quad (2.69)$$

Following Section 2.2 and Corollary 2.6 in Tsybakov [2008], combined with (2.67)–(2.69), we have

$$\min_{\hat{\Omega}} \sup_{\Omega \in \mathcal{G}_2} \mathbb{E} \left[\left\| \hat{\Omega} - \Omega \right\|_{\text{F}}^2 \right] \geq \frac{1}{2} \left(\frac{1}{32\sqrt{2}} \sqrt{\frac{h \log d}{n}} \right)^2 = \frac{1}{4096} \cdot \frac{h \log d}{n},$$

which completes the proof.

2.9.6 Proof of Theorem 11

To prove the lower bound, we assume that $\mathcal{P}_k = N \left(0, \left(\Omega^{(k)} \right)^{-1} \right)$.

First, when $h = 0$, we have N i.i.d. samples from $\Omega^{(0)}$ where

$$\Omega^{(0)} \in \left\{ \Omega \in \mathbb{R}^{d \times d} : \Omega \succ 0, |\Omega|_0 \leq s, \|\Omega\|_2 \leq M_\Omega, \left| \Omega^{-1} \right|_\infty \leq M_\Sigma \right\} := \tilde{\mathcal{G}}'$$

for some positive universal constants M_Ω and M_Σ . Let

$$\mathcal{G}' = \left\{ \Omega \in \mathbb{R}^{d \times d} : \Omega \succ 0, |\Omega|_0 \leq s, 0 < \frac{1}{M_\Sigma} \leq \gamma_{\min}(\Omega) \leq \gamma_{\max}(\Omega) \leq M_\Omega < \infty \right\}.$$

Since $|\Omega^{-1}|_\infty \leq \|\Omega^{-1}\|_2 = \{\gamma_{\min}(\Omega)\}^{-1} \leq M_\Sigma$ for any $\Omega \in \mathcal{G}'$, we have $\mathcal{G}' \subseteq \tilde{\mathcal{G}}'$. Thus, when $s \geq d \geq c' N^\beta$, for some universal constants $c' > 0$ and $\beta > 1$ and

$$\lceil s/d \rceil = o\left(\frac{N}{(\log d)^{\frac{3}{2}}}\right)$$

it follows by Lemma 9 that

$$\inf_{\hat{\Omega}} \sup_{\Omega^{(0)} \in \tilde{\mathcal{G}}'} \mathbb{E} \left[\left\| \hat{\Omega} - \Omega^{(0)} \right\|_{\text{F}}^2 \right] \geq \inf_{\hat{\Omega}} \sup_{\Omega^{(0)} \in \mathcal{G}'} \mathbb{E} \left[\left\| \hat{\Omega} - \Omega^{(0)} \right\|_{\text{F}}^2 \right] \gtrsim \frac{s \log d}{N}. \quad (2.70)$$

If $\Omega^{(k)} = \Omega^\star = I_d$ and $\Gamma^{(k)\star} = 0$ for every $k \in [K]$, then samples from source distributions cannot be used to estimate $\Omega^{(0)}$. Therefore, we must depend solely on samples from the target distribution to estimate $\Omega^{(0)}$. Note that now we have

$$\begin{aligned} \Omega^{(0)} \in \left\{ \Omega \in \mathbb{S}^{d \times d} : \Omega \succ 0, \Omega = I_d + \Delta, \Delta_{jj} = 0 \text{ for all } 1 \leq j \leq d, \right. \\ \left. |\Delta|_0 \leq h, |\Delta|_1 \leq M_\Gamma, \|\Omega\|_2 \leq M_\Omega, |\Omega^{-1}|_\infty \leq M_\Sigma \right\} := \tilde{\mathcal{G}}''. \end{aligned}$$

Let

$$\begin{aligned} \mathcal{G}'' = \left\{ \Omega \in \mathbb{S}^{d \times d} : \Omega \succ 0, \Omega = I_d + \Delta, \Delta_{jj} = 0 \text{ for all } 1 \leq j \leq d, \right. \\ \left. |\Delta|_0 \leq h, |\Delta|_1 \leq M_\Gamma, 0 < \frac{1}{M_\Sigma} \leq \gamma_{\min}(\Omega) \leq \gamma_{\max}(\Omega) \leq M_\Omega < \infty \right\}. \end{aligned}$$

Given that for any $\Omega \in \mathcal{G}''$, it holds that $|\Omega^{-1}|_\infty \leq \|\Omega^{-1}\|_2 = \{\gamma_{\min}(\Omega)\}^{-1} \leq M_\Sigma$, we can deduce that $\mathcal{G}'' \subseteq \tilde{\mathcal{G}}''$.

By Lemma 10, when $M_\Omega, M_\Sigma > 1$ and

$$d \geq 4h, h \log d \geq 8 \log 3,$$

$$\frac{h \log d}{n_0} \leq \min \left\{ 2, 8 \left(1 - \frac{1}{M_\Sigma} \right)^2, 8(1 - M_\Omega)^2 \right\}, h \sqrt{\frac{\log d}{n_0}} \leq 4M_\Gamma,$$

we have

$$\inf_{\hat{\Omega}} \sup_{\Omega^{(0)} \in \tilde{\mathcal{G}}''} \mathbb{E} \left[\left\| \hat{\Omega} - \Omega^{(0)} \right\|_{\text{F}}^2 \right] \geq \inf_{\hat{\Omega}} \sup_{\Omega^{(0)} \in \mathcal{G}''} \mathbb{E} \left[\left\| \hat{\Omega} - \Omega^{(0)} \right\|_{\text{F}}^2 \right] \gtrsim \frac{h \log d}{n_0}. \quad (2.71)$$

The final result follows from (2.70) and (2.71).

2.9.7 Corollary 28 and Its Proof

The following corollary of Theorem 10 in Zhao et al. [2022] provides a high-probability error bound for the D-Trace Loss estimator.

Corollary 28. *Suppose that Assumption 1 and Assumption 2 hold. Let $C_\gamma = M_\Omega^{-1}$. Furthermore, suppose that for a given $\delta \in (0, 1]$ and all $k \in [K]$ we have*

$$\frac{\log(2(K+1)d^2/\delta)}{\min\{n_k, n_0\}} \leq \min \left\{ 1, \frac{\gamma_{\min}^2(\Sigma^{(k)})\gamma_{\min}^2(\Sigma^{(0)})}{C'h^2M_\Sigma^4} \right\}. \quad (2.72)$$

If

$$\lambda_\Gamma^{(k)} \geq 2C_\gamma [(C_\gamma + 2)M_\Sigma M_\Gamma + 2] M_\Sigma \sqrt{\frac{\log(2(K+1)d^2/\delta)}{2 \min\{n_k, n_0\}}}$$

for all $k \in [K]$, then

$$\left\| \hat{\Psi}^{(k)} - \Psi^{(k)} \right\|_{\text{F}} \leq \frac{C'' \sqrt{h} \lambda_\Gamma^{(k)}}{\gamma_{\min}(\Sigma^{(k)})\gamma_{\min}(\Sigma^{(0)})} \quad \text{for all } 0 \leq k \leq K$$

hold simultaneously with probability at least $1 - \delta$, where C', C'' are universal constants that depend on C_γ .

Proof. The D-Trace Loss estimator is a special case of the FuDGE estimator proposed in Zhao et al. [2019, 2022]. We adapt the proof of Theorem 10 in Zhao et al. [2022] here.

Given $\delta \in (0, 1]$, we define the event

$$\mathcal{E}^1(\delta) := \left\{ \left| \widehat{\Sigma}^{(k)} - \Sigma^{(k)} \right|_{\infty} \leq C_{\gamma} M_{\Sigma} \cdot \max \left\{ \sqrt{\frac{\log(2(K+1)d^2/\delta)}{2n_k}}, \frac{\log(2(K+1)d^2/\delta)}{n_k} \right\} \right. \\ \left. \text{for all } 0 \leq k \leq K \right\},$$

where C_{γ} is a universal constant. By Lemma 19 and the union bound, we have $\mathbb{P}(\mathcal{E}^1(\delta)) \geq 1 - \delta$. In the following, we work on the event $\mathcal{E}^1(\delta)$.

Let

$$\psi_k = C_{\gamma} M_{\Sigma} \sqrt{\frac{\log(2(K+1)d^2/\delta)}{2 \min\{n_k, n_0\}}}.$$

On the event $\mathcal{E}^1$ and when the condition (2.72) holds with $C' = 4096C_{\gamma}^2(C_{\gamma} + 2)^2$, we have

$$\max \left\{ \left| \widehat{\Sigma}^{(k)} - \Sigma^{(k)} \right|_{\infty}, \left| \widehat{\Sigma}^{(0)} - \Sigma^{(0)} \right|_{\infty} \right\} \leq \psi_k.$$

When (2.72) holds, we have

$$\psi_k \leq C_{\gamma} \cdot \max_{0 \leq k \leq K} \left| \Sigma^{(k)} \right|_{\infty} \quad \text{and} \quad \psi_k \leq \frac{\gamma_{\min}(\Sigma^{(k)}) \gamma_{\min}(\Sigma^{(0)})}{64(C_{\gamma} + 2)hM_{\Sigma}}.$$

Thus, we have

$$\kappa_{\mathcal{L}}^{(k)} := \frac{1}{2} \gamma_{\min}(\Sigma^{(k)}) \gamma_{\min}(\Sigma^{(0)}) - 16h \left(\psi_k^2 + 2\psi_k M_{\Sigma} \right) \geq \frac{1}{4} \gamma_{\min}(\Sigma^{(k)}) \gamma_{\min}(\Sigma^{(0)}),$$

where $\kappa_{\mathcal{L}}^{(k)}$ is the restricted convexity parameter defined in the proof of Theorem 10 of Zhao et al. [2022]. Set

$$\lambda_{\Gamma}^{(k)} \geq 4 \left[(C_{\gamma} + 2) M_{\Sigma} M_{\Gamma} + 1 \right] \psi_k.$$

Following the proof of Theorem 10 of Zhao et al. [2022], we have

$$\left\| \widehat{\Psi}^{(k)} - \Psi^{(k)} \right\|_{\text{F}}^2 \leq \frac{18h}{\frac{1}{16} \gamma_{\min}^2(\Sigma^{(k)}) \gamma_{\min}^2(\Sigma^{(0)})} \left(\lambda_{\Gamma}^{(k)} \right)^2,$$

which completes the proof. □

CHAPTER 3

SMART: A SPECTRAL TRANSFER APPROACH TO MULTI-TASK LEARNING

3.1 Introduction

Many modern scientific applications involve the simultaneous modeling of multiple interrelated regression tasks. A prominent example arises in genomics, where researchers aim to uncover gene-protein association networks across diverse biological contexts such as tissues or cell types [Uematsu et al., 2019, Tang et al., 2023]. In these studies, each sample corresponds to a single cell, and recent advances in single-cell multi-modal profiling technologies enable the joint measurement of both gene expression levels (e.g., mRNA abundance) and protein expression levels (e.g., surface protein markers) from the same cell [Luecken et al., 2021]. The resulting dataset consists of two high-dimensional feature matrices—one for genes and one for proteins—aligned across a common set of cells. The scientific goal is to infer the underlying regulatory relationships between genes and proteins, namely the gene-protein networks, by identifying low-dimensional structures that capture shared variation across the two modalities. Due to the modular organization of biological systems—where genes and proteins interact through a limited number of coordinated functional programs—these relationships often admit a low-rank representation in the regression coefficient matrix, reflecting a small number of latent gene-protein modules.

While multi-task learning methods have been extensively studied to improve estimation by borrowing strength across multiple related regression tasks, challenges arise when the sample sizes for certain tasks are limited. This scenario frequently occurs in biomedical studies where certain cell types or experimental conditions are rare or difficult to obtain, resulting in small-sample regimes that hinder accurate estimation. For instance, in gene-protein interaction studies across distinct cell types, some rare or disease-specific cell types

may only yield limited data, whereas others may have abundant measurements. In such contexts, transfer learning approaches offer a natural framework to leverage source knowledge from data-rich tasks to assist estimation in data-poor target tasks [Pan and Yang, 2009].

Existing transfer learning methods for high-dimensional regression predominantly rely on *parameter similarity assumptions*, wherein the coefficient vectors or matrices of the source and target tasks are presumed to be close under some norm [Li et al., 2022, 2023b]. For example, Li et al. [2022] formalizes transfer learning in high-dimensional linear regression by assuming that the parameter difference between the source and target tasks is sparse or small in ℓ_q norm. While such assumptions are intuitive in some applications, they may be overly restrictive in complex biological systems, where mechanistic differences can induce substantial shifts in individual coefficients while still preserving higher-level shared structures. For example, although specific gene-protein interactions may vary across cell types, the underlying functional pathways—comprising co-regulated genes and protein complexes—often remain largely conserved.

Motivated by these biological insights, this chapter introduces a *spectral similarity assumption* that operates at the level of the low-rank structure of the regression matrices. Specifically, we posit that the singular subspaces of the target regression matrix are contained within those of the source task. This assumption reflects the premise that the key gene-protein modules active in the target cell type are drawn from a broader repertoire of modules present in the source cell type, a phenomenon frequently observed in comparative genomics. Furthermore, we incorporate a *sparse alignment condition*, which stipulates that each target singular vector aligns with only a small number of directions from the source basis, thereby capturing the modular reuse of biological pathways across contexts. In contrast to parameter similarity assumptions, the proposed spectral transfer condition offers greater flexibility by accommodating heterogeneous shifts in individual coefficients while preserving the shared latent subspaces that govern the multi-task relationships.

Building upon these insights, we develop a new transfer learning framework for multi-task regression with low-rank structure, termed **SMART** (Spectral Multi-task Regression Transfer). The SMART estimator integrates spectral information extracted from the source task to guide estimation for the target task via structured regularization. Despite the inherent nonconvexity of the formulation, we develop an efficient optimization algorithm based on the alternating direction method of multipliers (ADMM), which leverages recent advances in manifold optimization to handle the orthogonality constraints. From a statistical perspective, we establish non-asymptotic error bounds for SMART under mild regularity conditions. The technical analysis is particularly challenging due to the interaction between nonconvexity, spectral alignment, and high-dimensionality. Empirically, we validate the proposed method through extensive simulation studies and demonstrate its practical utility by applying it to real-world gene-protein interaction data, highlighting its ability to extract biologically meaningful structures in small-sample regimes.

3.1.1 Related Work

Multi-Task Learning. Multi-task learning has a long history in computer science [Caruana, 1997], particularly in areas such as neural networks and representation learning, where shared structures are leveraged to improve generalization across tasks [Zhang and Yang, 2021]. In recent years, the statistical community has developed a rich theory and methodology for multi-task learning, especially under the multivariate linear regression framework. Reduced-rank regression (RRR), which assumes a low-rank coefficient matrix to enhance statistical efficiency and interpretability, is a central focus of this literature. Foundational works such as Anderson [1951] and Izenman [1975] introduced the key idea of modeling dependencies among multiple responses via low-rank structure, inspiring modern extensions that incorporate sparsity, adaptivity, and high-dimensional scalability. Yuan et al. [2007] and Chen et al. [2013] developed nuclear norm penalization methods that enable adaptive dimension reduction in high-dimensional settings. The dirty model of Jalali et al. [2010]

offered a complementary perspective by modeling both shared and task-specific components across tasks. Bunea et al. [2011, 2012] advanced the field by proposing optimal selection criteria for reduced-rank estimators and integrating joint variable and rank selection. Chen and Huang [2012] and Ma et al. [2020] further developed sparsity-inducing formulations that simultaneously select relevant predictors and latent factors. Uematsu et al. [2019] introduced SOFAR, a sparse orthogonal factor regression framework designed for large-scale association learning via structured SVD. These contributions collectively form a rich and versatile toolkit for modeling multi-task relationships, blending low-rank structure with sparsity to address the challenges of modern high-dimensional applications.

Transfer learning. Originally rooted in computer science, transfer learning has a long-standing presence in machine learning [Pan and Yang, 2009]. In recent years, it has begun to attract increasing interest from the statistics community, where researchers aim to investigate it within principled statistical frameworks. Notable developments include applications to high-dimensional linear regression [Li et al., 2022, He et al., 2024b] and generalized linear models [Li et al., 2023b, Tian and Feng, 2023]. In a different direction, Li et al. [2023a] proposed a transfer learning method for Gaussian graphical models. Other works address transfer under covariate shift [Pathak et al., 2022, Ma et al., 2023, Wang, 2023, Ge et al., 2024], nonparametric classification [Cai and Wei, 2021], and sequential decision-making settings such as multi-armed bandits [Cai et al., 2024a]. Transfer learning for functional data has also been studied [Lin and Reimherr, 2022, Cai et al., 2024b].

Most existing transfer learning methods are grounded in the *parameter similarity assumption*, which posits that the coefficient vectors or matrices of the source and target tasks are close under some norm. However, a few recent studies depart from this paradigm. For instance, Gu et al. [2024] proposed an angle-based transfer method for high-dimensional linear regression, while Qiao et al. [2025] developed a rank-based approach tailored for survival analysis. Closely related to our work, He et al. [2024a] introduced a spectral transfer method

based on singular value decomposition, applied to matrix completion. Although both our method and theirs exploit spectral similarity, the tasks differ fundamentally: their focus is on matrix completion, whereas our method is designed for multi-task learning. Consequently, the methodological frameworks and theoretical analyses are substantially different.

Spectral Methods. Our proposed method leverages spectral methods to facilitate transfer learning. Broadly speaking, spectral methods refer to a class of techniques that operate on the singular values (or eigenvalues) and the associated singular vectors (or eigenvectors) of carefully constructed matrices derived from data [Chen et al., 2021]. Over the past few decades, spectral methods have become foundational tools across a wide array of domains, including community detection [Newman, 2006], clustering [Ng et al., 2001], ranking [Chen et al., 2019], and econometrics and finance [Fan et al., 2021]. These methods are particularly valued for their ability to extract low-dimensional latent structures and provide interpretable representations of complex data. For a recent comprehensive overview of spectral techniques and their applications in modern data science, we refer the reader to Chen et al. [2021].

3.1.2 Organization and Notation

Notation. In this chapter, we adopt the following notational conventions. For a vector $\mathbf{v} \in \mathbb{R}^d$, we denote its ℓ_p -norm by $\|\mathbf{v}\|_p$, defined as $\|\mathbf{v}\|_p = \left(\sum_{i=1}^d |v_i|^p\right)^{1/p}$ for $1 \leq p < \infty$, and $\|\mathbf{v}\|_\infty = \max_i |v_i|$ for $p = \infty$.

For a matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$, we use $|\cdot|$ to denote elementwise norms and $\|\cdot\|$ to denote matrix norms. Specifically, $|\mathbf{A}|_1 = \sum_{i=1}^d \sum_{j=1}^d |\mathbf{A}_{ij}|$, $|\mathbf{A}|_0 = \sum_{i=1}^d \sum_{j=1}^d \mathbb{1}\{\mathbf{A}_{ij} \neq 0\}$, and $|\mathbf{A}|_\infty = \max_{1 \leq i, j \leq d} |\mathbf{A}_{ij}|$. The operator norms include $\|\mathbf{A}\|_1 = \max_{1 \leq j \leq d} \sum_{i=1}^d |\mathbf{A}_{ij}|$, $\|\mathbf{A}\|_\infty = \max_{1 \leq i \leq d} \sum_{j=1}^d |\mathbf{A}_{ij}|$, and $\|\mathbf{A}\|_2$, which denotes the largest singular value of $\mathbf{A}$. The Frobenius norm is given by $\|\mathbf{A}\|_F = \left(\sum_{i=1}^d \sum_{j=1}^d |\mathbf{A}_{ij}|^2\right)^{1/2}$.

We define the inner product between two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d \times d}$ as $\langle \mathbf{A}, \mathbf{B} \rangle = \text{tr}(\mathbf{A}^\top \mathbf{B}) = \sum_{i,j} \mathbf{A}_{ij} \mathbf{B}_{ij}$. The operator $\text{vec}(\mathbf{A})$ denotes the vectorization of $\mathbf{A}$ by stacking its columns.

When $\mathbf{A}$ is symmetric, we let $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ denote its smallest and largest eigenvalues, respectively.

For matrices $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$ and $\mathbf{B} \in \mathbb{R}^{m_1 \times m_2}$, we denote their Kronecker product as $\mathbf{A} \otimes \mathbf{B} \in \mathbb{R}^{n_1 m_1 \times n_2 m_2}$, with entries defined as $[\mathbf{A}_{ij} \mathbf{B}_{lm}]_{i,j,l,m}$. Given an index set $\mathcal{S} \subseteq [m] \times [n]$ and a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, we define $\mathbf{A}_{\mathcal{S}}$ to be the matrix with entries $\mathbf{A}_{\mathcal{S},ij} = \mathbf{A}_{ij} \mathbf{1}\{(i,j) \in \mathcal{S}\}$. We denote by $\mathbb{S}^{d \times d}$ the set of symmetric $d \times d$ matrices.

Universal constants may vary from line to line without explicit mention.

For two positive sequences $\{f(n)\}$ and $\{g(n)\}$, we use standard asymptotic notation: $f(n) = O(g(n))$ or $f(n) \lesssim g(n)$ means there exists a constant $c > 0$ such that $f(n) \leq cg(n)$ for sufficiently large n ; $f(n) = \Omega(g(n))$ or $f(n) \gtrsim g(n)$ means $f(n) \geq cg(n)$ for large n ; $f(n) = \Theta(g(n))$ or $f(n) \asymp g(n)$ means $c_1 g(n) \leq f(n) \leq c_2 g(n)$ for constants $c_1, c_2 > 0$ and large n ; and $f(n) = o(g(n))$ indicates that $f(n)/g(n) \rightarrow 0$ as $n \rightarrow \infty$.

Organization. The remainder of the paper is structured as follows. Section 3.2 introduces the problem setup and motivates the spectral transfer framework. Section 3.3 details the proposed SMART algorithm, including its ADMM-based optimization procedure and hyperparameter selection strategy. Theoretical guarantees, including technical assumptions and finite-sample error bounds, are established in Section 3.4. Section 3.5 presents comprehensive simulation studies to evaluate the empirical performance of SMART under a variety of structural conditions. Section 3.6 demonstrates the practical utility of our method through an application to a real-world multi-modal single-cell dataset. Finally, Section 3.7 concludes the paper and outlines directions for future research. Technical proofs are deferred to Section 3.8.

3.2 Problem Setup

Let $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ be n independent and identically distributed (i.i.d.) samples, where each $\mathbf{x}_i \in \mathbb{R}^p$ denotes the covariate vector and $\mathbf{y}_i \in \mathbb{R}^q$ denotes the corresponding response vector.

We consider the following multi-task linear regression model:

$$\mathbf{Y} = \mathbf{X}\mathbf{C}^* + \mathbf{E}, \quad (3.1)$$

where $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_n]^\top \in \mathbb{R}^{n \times q}$ is the response matrix, $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]^\top \in \mathbb{R}^{n \times p}$ is the design matrix, and $\mathbf{E} = [\mathbf{e}_1, \dots, \mathbf{e}_n]^\top \in \mathbb{R}^{n \times q}$ denotes the matrix of random errors. We assume that $\mathbf{X}$ and $\mathbf{E}$ are independent. Here, each column of $\mathbf{Y}$ corresponds to the observed responses for a single task across all samples, and each row corresponds to the response vector for one observation. The unknown regression coefficient matrix $\mathbf{C}^* \in \mathbb{R}^{p \times q}$ characterizes the linear relationship between the covariates and the multi-task responses.

To capture the shared structure and dependencies across multiple tasks, we impose a low-rank constraint on the regression coefficient matrix $\mathbf{C}^*$. Specifically, we assume that $\mathbf{C}^*$ admits a reduced-rank decomposition through its singular value decomposition (SVD):

$$\mathbf{C}^* = \mathbf{U}^* \mathbf{D}^* \mathbf{V}^{*\top} = \sum_{j=1}^r d_j^* \mathbf{u}_j^* \mathbf{v}_j^{*\top},$$

where $r = \text{rank}(\mathbf{C}^*) \leq \min(p, q)$ is the rank of the matrix, $\mathbf{D}^* = \text{diag}(d_1^*, \dots, d_r^*)$ is a diagonal matrix of nonzero singular values ordered such that $d_1^* \geq \dots \geq d_r^* > 0$, and $\mathbf{U}^* = [\mathbf{u}_1^*, \dots, \mathbf{u}_r^*] \in \mathbb{R}^{p \times r}$ and $\mathbf{V}^* = [\mathbf{v}_1^*, \dots, \mathbf{v}_r^*] \in \mathbb{R}^{q \times r}$ are matrices of left and right singular vectors, respectively. The orthonormal columns of $\mathbf{U}^*$ span a low-dimensional subspace of the covariates that is relevant for the tasks, while the columns of $\mathbf{V}^*$ characterize shared patterns across tasks. This low-rank structure enables information sharing among tasks and facilitates joint modeling in high-dimensional settings.

Throughout this chapter, except in the sections on numerical experiments and practical implementations, we assume for simplicity that the true rank r is known. In practice, r can be selected using a variety of model selection techniques, such as the rank selection criterion proposed by Bunea et al. [2011].

In the context of gene-protein networks, the low-rank structure of $\mathbf{C}^*$ reflects the biological intuition that a limited number of latent regulatory programs or signaling pathways govern the coordinated expression of genes and proteins. Each latent component in the decomposition corresponds to a gene-protein module—a shared mechanism by which subsets of genes exert influence over subsets of proteins. By modeling $\mathbf{C}^*$ as a low-rank matrix, we capture these cross-modal interactions in a parsimonious and interpretable manner, effectively reducing the dimensionality of the problem without sacrificing key structural information. This low-rank representation not only facilitates the discovery of biologically meaningful patterns, but also improves statistical efficiency and robustness in high-dimensional settings typical of single-cell multi-omics data.

In addition to the target dataset, we assume access to a source regression matrix $\mathbf{C}^{(0)} \in \mathbb{R}^{p \times q}$ derived from a related biological context—such as a different tissue type, cell population, or experimental condition—within the same gene-protein measurement framework. For example, in single-cell multi-omics studies, one may have previously estimated gene-protein association patterns in a well-characterized reference tissue or cell type, and now seeks to leverage this information to improve estimation in a new but related context. We assume that $\mathbf{C}^{(0)}$ admits a singular value decomposition (SVD) of the form

$$\mathbf{C}^{(0)} = \mathbf{U}^{(0)} \mathbf{D}^{(0)} \mathbf{V}^{(0)\top} = \sum_{j=1}^{r_0} d_j^{(0)} \mathbf{u}_j^{(0)} \mathbf{v}_j^{(0)\top},$$

where $1 \leq r_0 \leq \min(p, q)$ is the rank of the source matrix, $\mathbf{U}^{(0)} = [\mathbf{u}_1^{(0)}, \dots, \mathbf{u}_{r_0}^{(0)}] \in \mathbb{R}^{p \times r_0}$ and $\mathbf{V}^{(0)} = [\mathbf{v}_1^{(0)}, \dots, \mathbf{v}_{r_0}^{(0)}] \in \mathbb{R}^{q \times r_0}$ are orthonormal matrices of left and right singular vectors, respectively, and $\mathbf{D}^{(0)} = \text{diag}(d_1^{(0)}, \dots, d_{r_0}^{(0)})$ contains non-increasing positive singular values: $d_1^{(0)} \geq \dots \geq d_{r_0}^{(0)} > 0$.

To improve the estimation of the target coefficient matrix $\mathbf{C}^*$ by leveraging auxiliary information from a related source matrix $\mathbf{C}^{(0)}$, we introduce a pair of structural assumptions

that formalize the similarity between $\mathbf{C}^*$ and $\mathbf{C}^{(0)}$. Previous approaches to transfer learning often rely on norm-based proximity assumptions, which assert that the difference $\mathbf{C}^* - \mathbf{C}^{(0)}$ is small under some matrix norm. However, such assumptions can be overly restrictive in practice—particularly when the overall scale of the regression coefficients, referred to as expression strength, differs substantially between the source and target tasks.

Instead, we adopt a more flexible framework grounded in the spectral structure of the regression matrices. Specifically, we assume that the singular subspaces of the target matrix $\mathbf{C}^*$ are well aligned with those of the source matrix $\mathbf{C}^{(0)}$. This alignment is captured by the following condition:

Assumption 3 (Spectral Containment Condition). *We assume that $r_0 \geq r$ and that the singular subspaces of the target matrix $\mathbf{C}^*$ are contained within those of the source matrix $\mathbf{C}^{(0)}$; that is,*

$$\text{col}(\mathbf{U}^*) \subseteq \text{col}(\mathbf{U}^{(0)}) \quad \text{and} \quad \text{col}(\mathbf{V}^*) \subseteq \text{col}(\mathbf{V}^{(0)}).$$

This assumption is biologically plausible in many structured domains. For instance, in our real-data application, the target cell type (ILC1) exhibits a more specialized and functionally constrained regulatory program compared to the source cell type (NK), which circulates systemically and engages a broader array of immune pathways. As a result, the latent gene–protein modules active in ILC1 are naturally spanned by those in NK, satisfying the containment condition [Vivier et al., 2018].

Assumption 3 ensures that the dominant directions governing the low-rank structure of $\mathbf{C}^*$ are fully represented within the spectral span of $\mathbf{C}^{(0)}$. While this condition enables information transfer from source to target, it may be too permissive when r_0 is large, as the subspaces $\text{col}(\mathbf{U}^{(0)})$ and $\text{col}(\mathbf{V}^{(0)})$ may include many irrelevant directions. To further refine the similarity structure and encourage statistical efficiency, we introduce a second assumption that enforces sparsity in how the target singular vectors align with the source

basis. We define

$$\begin{aligned} s_u^* &:= \left| \mathbf{U}^{*\top} \mathbf{U}^{(0)} \right|_0 = \sum_{i=1}^r \# \left\{ 1 \leq j \leq r_0 : \left\langle \mathbf{u}_i^*, \mathbf{u}_j^{(0)} \right\rangle \neq 0 \right\}, \\ s_v^* &:= \left| \mathbf{V}^{*\top} \mathbf{V}^{(0)} \right|_0 = \sum_{i=1}^r \# \left\{ 1 \leq j \leq r_0 : \left\langle \mathbf{v}_i^*, \mathbf{v}_j^{(0)} \right\rangle \neq 0 \right\}. \end{aligned} \tag{3.2}$$

We then make the following assumption.

Assumption 4 (Sparse Alignment Condition). *We assume that $s_u^* \ll rp$ and $s_v^* \ll rq$.*

Together, Assumptions 3 and 4 define our structural similarity framework. The Spectral Containment Condition guarantees that the target subspaces are supported within the source subspaces, while the Sparse Alignment Condition ensures that this containment occurs in a parsimonious fashion. By definition, $s_u^*, s_v^* \leq rr_0$, so when $r_0 \ll \min(p, q)$, the sparsity condition is automatically satisfied. When r_0 is large, the Sparse Alignment Condition imposes that each target singular vector aligns with only a small number of source directions.

This sparsity assumption is particularly well justified in biological applications. For example, in our real-data analysis, the source matrix $\mathbf{C}^{(0)}$ corresponds to NK cells, which engage a wide range of regulatory programs related to cytotoxicity, migration, and systemic immune defense. In contrast, the target matrix $\mathbf{C}^*$ models ILC1s, which rely on a limited subset of these pathways—primarily those involved in tissue-resident immune surveillance and IFN- γ signaling. As a result, each latent component in the target can be expressed using only a few directions from the source spectral basis. The Sparse Alignment Condition captures this selective reuse of regulatory modules, reflecting biological modularity and supporting efficient transfer across structured cell types.

In practice, the clean source matrix $\mathbf{C}^{(0)}$ is typically unobserved. Instead, we only have access to a noisy approximation:

$$\tilde{\mathbf{C}}^{(0)} = \mathbf{C}^{(0)} + \mathbf{E}^{(0)},$$

where $\mathbf{E}^{(0)} \in \mathbb{R}^{p \times q}$ denotes an additive source noise matrix. Accordingly, we work with the singular value decomposition (SVD) of the observed noisy matrix:

$$\tilde{\mathbf{C}}^{(0)} = \tilde{\mathbf{U}}^{(0)} \tilde{\mathbf{D}}^{(0)} \tilde{\mathbf{V}}^{(0)\top} = \sum_{j=1}^{\tilde{r}_0} \tilde{d}_j^{(0)} \tilde{\mathbf{u}}_j^{(0)} \tilde{\mathbf{v}}_j^{(0)\top},$$

where the singular values are ordered as $\tilde{d}_1^{(0)} > \dots > \tilde{d}_{\tilde{r}_0}^{(0)} > 0$.

Note that $\tilde{\mathbf{C}}^{(0)}$ is not assumed to be low-rank and may even be full-rank, i.e., $\tilde{r}_0 = \min(p, q)$. As is standard, we assume the singular values are strictly decreasing to guarantee the unique ordering of the singular vectors. This assumption is mild and typically satisfied in practice, particularly when the presence of noise or numerical perturbations breaks any potential degeneracy in the spectrum.

3.3 SMART Algorithm

In this section, we introduce our proposed methodology, **SMART** (**S**pectral **M**ulti-task **A**ddaptive **R**egression **T**ransfer), which integrates low-rank regression with structured regularization guided by spectral information derived from a related source task.

We begin by extending the singular vectors $\tilde{\mathbf{U}}^{(0)}$ and $\tilde{\mathbf{V}}^{(0)}$ of the noisy source matrix $\tilde{\mathbf{C}}^{(0)}$ into full orthonormal bases of $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively. This is achieved by appending the vectors $\tilde{\mathbf{u}}_{\tilde{r}_0+1}^{(0)}, \dots, \tilde{\mathbf{u}}_p^{(0)}$ and $\tilde{\mathbf{v}}_{\tilde{r}_0+1}^{(0)}, \dots, \tilde{\mathbf{v}}_q^{(0)}$. For any pair (r_u, r_v) with $0 \leq r_u \leq p$ and $0 \leq r_v \leq q$, we define the following subspaces:

$$\tilde{\mathbf{U}}^s(r_u) := [\tilde{\mathbf{u}}_1^{(0)}, \dots, \tilde{\mathbf{u}}_{r_u}^{(0)}], \quad \tilde{\mathbf{U}}_\perp^s(r_u) := [\tilde{\mathbf{u}}_{r_u+1}^{(0)}, \dots, \tilde{\mathbf{u}}_p^{(0)}],$$

and similarly for the right singular vectors:

$$\tilde{\mathbf{V}}^s(r_v) := [\tilde{\mathbf{v}}_1^{(0)}, \dots, \tilde{\mathbf{v}}_{r_v}^{(0)}], \quad \tilde{\mathbf{V}}_\perp^s(r_v) := [\tilde{\mathbf{v}}_{r_v+1}^{(0)}, \dots, \tilde{\mathbf{v}}_q^{(0)}].$$

Given structural parameters (r_u, r_v) , we estimate the low-rank components $(\hat{\mathbf{U}}, \hat{\mathbf{D}}, \hat{\mathbf{V}})$ by solving the following structured optimization problem:

$$\begin{aligned} \hat{\mathbf{U}}, \hat{\mathbf{D}}, \hat{\mathbf{V}} \in \arg \min_{\mathbf{U}, \mathbf{D}, \mathbf{V}} & \left\{ \frac{1}{2n} \left\| \mathbf{Y} - \mathbf{XUDV}^\top \right\|_{\text{F}}^2 + \lambda_u \left| \tilde{\mathbf{U}}_{\perp}^\top(r_u) \mathbf{UD} \right|_1 + \lambda_v \left| \tilde{\mathbf{V}}_{\perp}^{\text{s}\top}(r_v) \mathbf{VD} \right|_1 \right\} \\ \text{subject to} & \quad \mathbf{U}^\top \mathbf{U} = \mathbf{I}_r, \quad \mathbf{V}^\top \mathbf{V} = \mathbf{I}_r, \end{aligned} \quad (3.3)$$

where $\mathbf{U} \in \mathbb{R}^{p \times r}$, $\mathbf{V} \in \mathbb{R}^{q \times r}$, and $\mathbf{D} = \text{diag}(d_1, \dots, d_r)$ with $d_j \geq 0$. The final estimator for the regression matrix is then given by

$$\hat{\mathbf{C}} := \hat{\mathbf{U}} \hat{\mathbf{D}} \hat{\mathbf{V}}^\top.$$

The objective in (3.3) consists of a data-fitting term and two structured regularization terms. The first term captures the empirical loss under a low-rank model. The second and third terms impose ℓ_1 penalties on the projections of the estimated coefficient matrix onto the orthogonal complements of the top- r_u and top- r_v singular vector subspaces derived from the source matrix. These penalties do not constrain alignment with the leading r_u and r_v directions, allowing the estimator to freely align with them. In contrast, alignment with the remaining directions is penalized, encouraging sparse and selective use of the less dominant source components. This mechanism enables the estimator to adaptively extract information from only the most informative source directions, consistent with the Sparse Alignment Condition (Assumption 4).

As we will show in Section 3.4, the structural ranks (r_u, r_v) govern a fundamental trade-off between approximation error—due to source noise $\mathbf{E}^{(0)}$ —and statistical estimation error. When source noise is low, aggressive regularization (e.g., $r_u = r_v = 0$) allows the estimator to fully leverage source spectral information and gain statistical efficiency. However, as the spectral norm of $\mathbf{E}^{(0)}$ increases, the estimated singular vectors $\tilde{\mathbf{U}}^{(0)}$ and $\tilde{\mathbf{V}}^{(0)}$ become less reliable. In such cases, increasing (r_u, r_v) relaxes alignment constraints and expands

the unpenalized subspace, thereby enhancing robustness by reducing sensitivity to noise-contaminated directions.

This formulation enables effective transfer by dynamically balancing reliance on prior knowledge and data-driven adaptation. It promotes borrowing strength from relevant source components while guarding against overfitting to irrelevant or noisy ones.

3.3.1 Optimization Algorithms

In this section, we present the optimization algorithm used to solve (3.3). The problem is nonconvex due to the presence of orthogonality constraints and the bilinear structure of the objective. While finding the global optimum is generally intractable for such nonconvex problems, we can efficiently obtain a high-quality local solution using an ADMM-based solver. The algorithm alternates updates over $\mathbf{U}$, $\mathbf{D}$, and $\mathbf{V}$, enforcing orthogonality through projections onto the Stiefel manifold. In practice, we initialize SMART using the truncated SVD of a warm-start estimator—such as ridge or lasso—which helps accelerate convergence and improves numerical stability. As demonstrated in Section 3.5 and Section 3.6, this optimization procedure yields solutions with strong empirical statistical accuracy performance.

To streamline notation, we write $\tilde{\mathbf{U}}_{\perp}^{\mathbf{s}}(r_u)$ as $\tilde{\mathbf{U}}_{\perp}^{\mathbf{s}}$ and $\tilde{\mathbf{V}}_{\perp}^{\mathbf{s}}(r_v)$ as $\tilde{\mathbf{V}}_{\perp}^{\mathbf{s}}$ throughout this section. We begin by rewriting (3.3) as the following equivalent constrained optimization problem:

$$\begin{aligned} (\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Omega}}) \in \arg \min_{\boldsymbol{\Theta}, \boldsymbol{\Omega}} & \left\{ \frac{1}{2n} \left\| \mathbf{Y} - \mathbf{XUDV}^{\top} \right\|_{\text{F}}^2 + \lambda_u \|\boldsymbol{\Omega}_u\|_1 + \lambda_v \|\boldsymbol{\Omega}_v\|_1 \right\}, \\ \text{subject to} \quad & \mathbf{U}^{\top} \mathbf{U} = \mathbf{I}_r, \quad \mathbf{V}^{\top} \mathbf{V} = \mathbf{I}_r, \quad \mathbf{D} \geq 0, \\ & \tilde{\mathbf{U}}_{\perp}^{\mathbf{s}\top} \mathbf{U} \mathbf{D} = \boldsymbol{\Omega}_u, \quad \tilde{\mathbf{V}}_{\perp}^{\mathbf{s}\top} \mathbf{V} \mathbf{D} = \boldsymbol{\Omega}_v, \end{aligned} \tag{3.4}$$

where $\boldsymbol{\Theta} = (\mathbf{U}, \mathbf{D}, \mathbf{V})$ and $\boldsymbol{\Omega} = (\boldsymbol{\Omega}_u, \boldsymbol{\Omega}_v)$. The constraint $\mathbf{D} \geq 0$ is interpreted elementwise.

To solve this problem, we adopt the augmented Lagrangian approach:

$$\begin{aligned}
L_\rho(\Theta, \Omega, \Gamma) = & \frac{1}{2n} \left\| \mathbf{Y} - \mathbf{XUDV}^\top \right\|_F^2 + \lambda_u \|\Omega_u\|_1 + \lambda_v \|\Omega_v\|_1 \\
& + \left\langle \Gamma_u, \tilde{\mathbf{U}}_\perp^{\text{s}\top} \mathbf{UD} - \Omega_u \right\rangle + \left\langle \Gamma_v, \tilde{\mathbf{V}}_\perp^{\text{s}\top} \mathbf{VD} - \Omega_v \right\rangle \\
& + \frac{\rho}{2} \left\| \tilde{\mathbf{U}}_\perp^{\text{s}\top} \mathbf{UD} - \Omega_u \right\|_F^2 + \frac{\rho}{2} \left\| \tilde{\mathbf{V}}_\perp^{\text{s}\top} \mathbf{VD} - \Omega_v \right\|_F^2,
\end{aligned}$$

where $\Gamma = (\Gamma_u, \Gamma_v)$ are the dual variables and $\rho > 0$ is a penalty parameter. Following the ADMM framework, we alternate updates of (Θ, Ω) and Γ in an iterative manner. The full procedure is outlined in Algorithm 4.

We now describe how to solve the subproblems in Algorithm 4.

First, the $\mathbf{U}$ -update step is equivalent to solving the following optimization problem:

$$\begin{aligned}
\mathbf{U}^{(t+1)} \leftarrow \arg \min_{\mathbf{U}} & \left\{ \frac{1}{2n} \left\| \mathbf{XUD}^{(t)} \right\|_F^2 + \frac{\rho^{(t)}}{2} \left\| \tilde{\mathbf{U}}_\perp^{\text{s}\top} \mathbf{UD}^{(t)} \right\|_F^2 + \text{tr} \left(\mathbf{U}^\top \mathbf{B}_u^{(t)} \right) \right\} \\
\text{s.t. } & \mathbf{U}^\top \mathbf{U} = \mathbf{I}_r,
\end{aligned}$$

where

$$\mathbf{B}_u^{(t)} = \tilde{\mathbf{U}}_\perp^{\text{s}} \Gamma_u^{(t)} \mathbf{D}^{(t)} - \frac{1}{n} \mathbf{X}^\top \mathbf{Y} \mathbf{V}^{(t)} \mathbf{D}^{(t)} - \rho^{(t)} \tilde{\mathbf{U}}_\perp^{\text{s}} \Omega_u^{(t)} \mathbf{D}^{(t)}.$$

This is a quadratic problem over the Stiefel manifold. We solve it using the Riemannian conjugate gradient method implemented in the `Pymanopt` package [Townsend et al., 2016].

Similarly, the $\mathbf{V}$ -update step corresponds to the following optimization problem:

$$\mathbf{V}^{(t+1)} \leftarrow \arg \min_{\mathbf{V}} \left\{ \frac{\rho^{(t)}}{2} \left\| \tilde{\mathbf{V}}_\perp^{\text{s}\top} \mathbf{VD}^{(t)} \right\|_F^2 + \text{tr} \left(\mathbf{V}^\top \mathbf{B}_v^{(t)} \right) \right\} \quad \text{s.t. } \mathbf{V}^\top \mathbf{V} = \mathbf{I}_r,$$

where

$$\mathbf{B}_v^{(t)} = \tilde{\mathbf{V}}_\perp^{\text{s}} \Gamma_v^{(t)} \mathbf{D}^{(t)} - \frac{1}{n} \mathbf{Y}^\top \mathbf{X} \mathbf{U}^{(t+1)} \mathbf{D}^{(t)} - \rho^{(t)} \tilde{\mathbf{V}}_\perp^{\text{s}} \Omega_v^{(t)} \mathbf{D}^{(t)}.$$

This problem is also solved using `Pymanopt`.

Algorithm 4 SMART Algorithm

- 1: **Input:** target rank r ; structural ranks r_u, r_v ; regularization parameters λ_u, λ_v ; source singular vectors $\tilde{\mathbf{U}}^s, \tilde{\mathbf{V}}^s$; penalty scaling factor $\gamma > 1$; maximum number of iterations $t_{\max}$.
- 2: **Initialize:** primal variables $\mathbf{U}^{(0)}, \mathbf{D}^{(0)}, \mathbf{V}^{(0)}$; auxiliary variables $\boldsymbol{\Omega}_u^{(0)}, \boldsymbol{\Omega}_v^{(0)}$; dual variables $\boldsymbol{\Gamma}_u^{(0)}, \boldsymbol{\Gamma}_v^{(0)}$; penalty parameter $\rho^{(0)}$; iteration counter $t \leftarrow 0$.
- 3: **while** stopping criteria not met and $t < t_{\max}$ **do**
- 4: Update $\mathbf{U}$:

$$\begin{aligned} \mathbf{U}^{(t+1)} \leftarrow \arg \min_{\mathbf{U}: \mathbf{U}^\top \mathbf{U} = \mathbf{I}_r} & \left\{ \frac{1}{2n} \left\| \mathbf{Y} - \mathbf{X} \mathbf{U} \mathbf{D}^{(t)} \mathbf{V}^{(t)\top} \right\|_{\text{F}}^2 \right. \\ & \left. + \left\langle \boldsymbol{\Gamma}_u^{(t)}, \tilde{\mathbf{U}}_\perp^{\text{s}\top} \mathbf{U} \mathbf{D}^{(t)} - \boldsymbol{\Omega}_u^{(t)} \right\rangle + \frac{\rho^{(t)}}{2} \left\| \tilde{\mathbf{U}}_\perp^{\text{s}\top} \mathbf{U} \mathbf{D}^{(t)} - \boldsymbol{\Omega}_u^{(t)} \right\|_{\text{F}}^2 \right\}. \end{aligned}$$

- 5: Update $\mathbf{V}$:

$$\begin{aligned} \mathbf{V}^{(t+1)} \leftarrow \arg \min_{\mathbf{V}: \mathbf{V}^\top \mathbf{V} = \mathbf{I}_r} & \left\{ \frac{1}{2n} \left\| \mathbf{Y} - \mathbf{X} \mathbf{U}^{(t+1)} \mathbf{D}^{(t)} \mathbf{V}^\top \right\|_{\text{F}}^2 \right. \\ & \left. + \left\langle \boldsymbol{\Gamma}_v^{(t)}, \tilde{\mathbf{V}}_\perp^{\text{s}\top} \mathbf{V} \mathbf{D}^{(t)} - \boldsymbol{\Omega}_v^{(t)} \right\rangle + \frac{\rho^{(t)}}{2} \left\| \tilde{\mathbf{V}}_\perp^{\text{s}\top} \mathbf{V} \mathbf{D}^{(t)} - \boldsymbol{\Omega}_v^{(t)} \right\|_{\text{F}}^2 \right\}. \end{aligned}$$

- 6: Update $\mathbf{D}$:

$$\begin{aligned} \mathbf{D}^{(t+1)} \leftarrow \arg \min_{\mathbf{D}} & \left\{ \frac{1}{2n} \left\| \mathbf{Y} - \mathbf{X} \mathbf{U}^{(t+1)} \mathbf{D} \mathbf{V}^{(t+1)\top} \right\|_{\text{F}}^2 \right. \\ & + \left\langle \boldsymbol{\Gamma}_u^{(t)}, \tilde{\mathbf{U}}_\perp^{\text{s}\top} \mathbf{U}^{(t+1)} \mathbf{D} - \boldsymbol{\Omega}_u^{(t)} \right\rangle + \left\langle \boldsymbol{\Gamma}_v^{(t)}, \tilde{\mathbf{V}}_\perp^{\text{s}\top} \mathbf{V}^{(t+1)} \mathbf{D} - \boldsymbol{\Omega}_v^{(t)} \right\rangle \\ & \left. + \frac{\rho^{(t)}}{2} \left\| \tilde{\mathbf{U}}_\perp^{\text{s}\top} \mathbf{U}^{(t+1)} \mathbf{D} - \boldsymbol{\Omega}_u^{(t)} \right\|_{\text{F}}^2 + \frac{\rho^{(t)}}{2} \left\| \tilde{\mathbf{V}}_\perp^{\text{s}\top} \mathbf{V}^{(t+1)} \mathbf{D} - \boldsymbol{\Omega}_v^{(t)} \right\|_{\text{F}}^2 \right\}. \end{aligned}$$

- 7: Update $\boldsymbol{\Omega}_u$:

$$\boldsymbol{\Omega}_u^{(t+1)} \leftarrow \arg \min_{\boldsymbol{\Omega}_u} \left\{ \frac{\rho^{(t)}}{2} \left\| \boldsymbol{\Omega}_u - \tilde{\mathbf{U}}_\perp^{\text{s}\top} \mathbf{U}^{(t+1)} \mathbf{D}^{(t+1)} \right\|_{\text{F}}^2 - \left\langle \boldsymbol{\Gamma}_u^{(t)}, \boldsymbol{\Omega}_u \right\rangle + \lambda_u \|\boldsymbol{\Omega}_u\|_1 \right\}.$$

- 8: Update $\boldsymbol{\Omega}_v$:

$$\boldsymbol{\Omega}_v^{(t+1)} \leftarrow \arg \min_{\boldsymbol{\Omega}_v} \left\{ \frac{\rho^{(t)}}{2} \left\| \boldsymbol{\Omega}_v - \tilde{\mathbf{V}}_\perp^{\text{s}\top} \mathbf{V}^{(t+1)} \mathbf{D}^{(t+1)} \right\|_{\text{F}}^2 - \left\langle \boldsymbol{\Gamma}_v^{(t)}, \boldsymbol{\Omega}_v \right\rangle + \lambda_v \|\boldsymbol{\Omega}_v\|_1 \right\}.$$

- 9: Update dual variables:

$$\begin{aligned} \boldsymbol{\Gamma}_u^{(t+1)} & \leftarrow \boldsymbol{\Gamma}_u^{(t)} + \rho^{(t)} \left(\tilde{\mathbf{U}}_\perp^{\text{s}\top} \mathbf{U}^{(t+1)} \mathbf{D}^{(t+1)} - \boldsymbol{\Omega}_u^{(t+1)} \right), \\ \boldsymbol{\Gamma}_v^{(t+1)} & \leftarrow \boldsymbol{\Gamma}_v^{(t)} + \rho^{(t)} \left(\tilde{\mathbf{V}}_\perp^{\text{s}\top} \mathbf{V}^{(t+1)} \mathbf{D}^{(t+1)} - \boldsymbol{\Omega}_v^{(t+1)} \right). \end{aligned}$$

- 10: Update penalty parameter: $\rho^{(t+1)} \leftarrow \gamma \rho^{(t)}$.

- 11: Increment iteration counter: $t \leftarrow t + 1$.

- 12: **end while**

- 13: **Output:**

$$\begin{aligned} \hat{\mathbf{U}} & \leftarrow \mathbf{U}^{(t)}, \quad \hat{\mathbf{V}} \leftarrow \mathbf{V}^{(t)}, \quad \hat{\mathbf{D}} \leftarrow \mathbf{D}^{(t)}, \quad \hat{\boldsymbol{\Omega}}_u \leftarrow \boldsymbol{\Omega}_u^{(t)}, \quad \hat{\boldsymbol{\Omega}}_v \leftarrow \boldsymbol{\Omega}_v^{(t)}, \\ \hat{\boldsymbol{\Gamma}}_u & \leftarrow \boldsymbol{\Gamma}_u^{(t)}, \quad \hat{\boldsymbol{\Gamma}}_v \leftarrow \boldsymbol{\Gamma}_v^{(t)}. \end{aligned}$$

The $\mathbf{D}$ -update step simplifies to:

$$\begin{aligned} \mathbf{D}^{(t+1)} \leftarrow \arg \min_{\mathbf{D}} \Big\{ & \frac{1}{2n} \left\| \mathbf{Y} - \mathbf{X}\mathbf{U}^{(t+1)}\mathbf{D}\mathbf{V}^{(t+1)\top} \right\|_{\text{F}}^2 \\ & + \frac{\rho^{(t)}}{2} \left\| \tilde{\mathbf{U}}_{\perp}^{\text{s}\top} \mathbf{U}^{(t+1)}\mathbf{D} - \boldsymbol{\Omega}_u^{(t)} + \rho^{(t)-1} \boldsymbol{\Gamma}_u^{(t)} \right\|_{\text{F}}^2 \\ & + \frac{\rho^{(t)}}{2} \left\| \tilde{\mathbf{V}}_{\perp}^{\text{s}\top} \mathbf{V}^{(t+1)}\mathbf{D} - \boldsymbol{\Omega}_v^{(t)} + \rho^{(t)-1} \boldsymbol{\Gamma}_v^{(t)} \right\|_{\text{F}}^2 \Big\}. \end{aligned}$$

Since $\mathbf{D} = \text{diag}(d_1, \dots, d_r)$ is diagonal, the objective decouples across its r entries. Define the shorthand:

$$\begin{aligned} \mathbf{P} &:= \mathbf{X}\mathbf{U}^{(t+1)} = [\mathbf{p}_1, \dots, \mathbf{p}_r], & \mathbf{Q} &:= \mathbf{V}^{(t+1)} = [\mathbf{q}_1, \dots, \mathbf{q}_r], \\ \mathbf{A}_u &:= \tilde{\mathbf{U}}_{\perp}^{\text{s}\top} \mathbf{U}^{(t+1)} = [\mathbf{a}_{u,1}, \dots, \mathbf{a}_{u,r}], & \mathbf{A}_v &:= \tilde{\mathbf{V}}_{\perp}^{\text{s}\top} \mathbf{V}^{(t+1)} = [\mathbf{a}_{v,1}, \dots, \mathbf{a}_{v,r}], \\ \tilde{\boldsymbol{\Omega}}_u &:= \boldsymbol{\Omega}_u^{(t)} - \rho^{(t)-1} \boldsymbol{\Gamma}_u^{(t)} = [\tilde{\boldsymbol{\omega}}_{u,1}, \dots, \tilde{\boldsymbol{\omega}}_{u,r}], & \tilde{\boldsymbol{\Omega}}_v &:= \boldsymbol{\Omega}_v^{(t)} - \rho^{(t)-1} \boldsymbol{\Gamma}_v^{(t)} = [\tilde{\boldsymbol{\omega}}_{v,1}, \dots, \tilde{\boldsymbol{\omega}}_{v,r}]. \end{aligned}$$

For $k = 1, \dots, r$, define:

$$\begin{aligned} \alpha_k &:= \frac{1}{n} \|\mathbf{p}_k\|_2^2, & \beta_k &:= \frac{1}{n} \mathbf{p}_k^{\top} \mathbf{Y} \mathbf{q}_k, \\ \gamma_{u,k} &:= \|\mathbf{a}_{u,k}\|_2^2, & \delta_{u,k} &:= \mathbf{a}_{u,k}^{\top} \tilde{\boldsymbol{\omega}}_{u,k}, \\ \gamma_{v,k} &:= \|\mathbf{a}_{v,k}\|_2^2, & \delta_{v,k} &:= \mathbf{a}_{v,k}^{\top} \tilde{\boldsymbol{\omega}}_{v,k}. \end{aligned}$$

Set:

$$c_k := \alpha_k + \rho^{(t)}(\gamma_{u,k} + \gamma_{v,k}), \quad b_k := \beta_k + \rho^{(t)}(\delta_{u,k} + \delta_{v,k}).$$

Then each d_k is updated by the closed-form solution:

$$d_k^{(t+1)} = \max \left\{ 0, \frac{b_k}{c_k} \right\}, \quad \text{for } k = 1, \dots, r.$$

Next, the $\mathbf{\Omega}_u$ -update solves:

$$\mathbf{\Omega}_u^{(t+1)} \leftarrow \arg \min_{\mathbf{\Omega}_u} \left\{ \frac{\rho^{(t)}}{2} \left\| \mathbf{\Omega}_u - \tilde{\mathbf{U}}_{\perp}^{\text{s}\top} \mathbf{U}^{(t+1)} \mathbf{D}^{(t+1)} - \rho^{(t)-1} \mathbf{\Gamma}_u^{(t)} \right\|_{\text{F}}^2 + \lambda_u \|\mathbf{\Omega}_u\|_1 \right\}.$$

This is a standard proximal update, whose closed-form solution is given entrywise by soft-thresholding:

$$\mathbf{\Omega}_{u,ij}^{(t+1)} = \mathcal{ST}_{\lambda_u/\rho^{(t)}} \left(\left[\tilde{\mathbf{U}}_{\perp}^{\text{s}\top} \mathbf{U}^{(t+1)} \mathbf{D}^{(t+1)} + \rho^{(t)-1} \mathbf{\Gamma}_u^{(t)} \right]_{ij} \right), \quad \forall i, j,$$

where the soft-thresholding operator $\mathcal{S}$ is defined as:

$$\mathcal{ST}_{\kappa}(a) = \begin{cases} a - \kappa, & \text{if } a > \kappa, \\ 0, & \text{if } |a| \leq \kappa, \\ a + \kappa, & \text{if } a < -\kappa. \end{cases}$$

The $\mathbf{\Omega}_v$ -update follows the same form:

$$\mathbf{\Omega}_{v,ij}^{(t+1)} = \mathcal{ST}_{\lambda_v/\rho^{(t)}} \left(\left[\tilde{\mathbf{V}}_{\perp}^{\text{s}\top} \mathbf{V}^{(t+1)} \mathbf{D}^{(t+1)} + \rho^{(t)-1} \mathbf{\Gamma}_v^{(t)} \right]_{ij} \right), \quad \forall i, j.$$

Stopping Criterion. Following [Boyd et al., 2011, Section 3.3], the optimality of the ADMM formulation in (3.4) requires both primal and dual feasibility to be satisfied. In practice, we monitor a pair of residuals—primal and dual—and terminate the algorithm once both fall below a prescribed tolerance.

We begin with the primal residual. Define

$$\mathbf{R}_{\text{p,u}}^{(t+1)} := \tilde{\mathbf{U}}_{\perp}^{\text{s}\top} \mathbf{U}^{(t+1)} \mathbf{D}^{(t+1)} - \mathbf{\Omega}_u^{(t+1)}, \quad \mathbf{R}_{\text{p,v}}^{(t+1)} := \tilde{\mathbf{V}}_{\perp}^{\text{s}\top} \mathbf{V}^{(t+1)} \mathbf{D}^{(t+1)} - \mathbf{\Omega}_v^{(t+1)}.$$

We then define the normalized primal residual as

$$r_{\text{primal}}^{(t+1)} = 0.5 \times \frac{\|\mathbf{R}_{\text{p,u}}^{(t+1)}\|_{\text{F}}}{\sqrt{(p-r_u)r}} + 0.5 \times \frac{\|\mathbf{R}_{\text{p,v}}^{(t+1)}\|_{\text{F}}}{\sqrt{(q-r_v)r}}.$$

Next, we consider the dual residual. Due to the optimality conditions for $\mathbf{D}^{(t+1)}$, $\boldsymbol{\Omega}_u^{(t+1)}$, and $\boldsymbol{\Omega}_v^{(t+1)}$, and their associated dual variable updates, dual feasibility is automatically satisfied for these blocks. Thus, it suffices to verify dual feasibility for the $\mathbf{U}$ and $\mathbf{V}$ updates.

To this end, define the surrogate objective functions:

$$f(\mathbf{U}) = \frac{1}{2n} \|\mathbf{XUD}^{(t+1)}\|_{\text{F}}^2 + \frac{\rho^{(t+1)}}{2} \|\tilde{\mathbf{U}}_{\perp}^{\text{s}\top} \mathbf{UD}^{(t+1)}\|_{\text{F}}^2 + \text{tr}(\mathbf{U}^{\top} \mathbf{B}_u^{(t+1)}),$$

$$g(\mathbf{V}) = \frac{1}{2n} \|\mathbf{XU}^{(t+1)} \mathbf{D}^{(t+1)} \mathbf{V}\|_{\text{F}}^2 + \frac{\rho^{(t+1)}}{2} \|\tilde{\mathbf{V}}_{\perp}^{\text{s}\top} \mathbf{VD}^{(t+1)}\|_{\text{F}}^2 + \text{tr}(\mathbf{V}^{\top} \mathbf{B}_v^{(t+0.5)}),$$

where

$$\mathbf{B}_u^{(t+1)} = \tilde{\mathbf{U}}_{\perp}^{\text{s}} \boldsymbol{\Gamma}_u^{(t+1)} \mathbf{D}^{(t+1)} - \frac{1}{n} \mathbf{X}^{\top} \mathbf{Y} \mathbf{V}^{(t+1)} \mathbf{D}^{(t+1)} - \rho^{(t+1)} \tilde{\mathbf{U}}_{\perp}^{\text{s}} \boldsymbol{\Omega}_u^{(t+1)} \mathbf{D}^{(t+1)},$$

$$\mathbf{B}_v^{(t+0.5)} = \tilde{\mathbf{V}}_{\perp}^{\text{s}} \boldsymbol{\Gamma}_v^{(t+1)} \mathbf{D}^{(t+1)} - \frac{1}{n} \mathbf{Y}^{\top} \mathbf{X} \mathbf{U}^{(t+1)} \mathbf{D}^{(t+1)} - \rho^{(t+1)} \tilde{\mathbf{V}}_{\perp}^{\text{s}} \boldsymbol{\Omega}_v^{(t+1)} \mathbf{D}^{(t+1)}.$$

Let $\mathcal{M}_{p,r} = \{\mathbf{U} \in \mathbb{R}^{p \times r} : \mathbf{U}^{\top} \mathbf{U} = \mathbf{I}_r\}$ denote the Stiefel manifold, and define $\mathcal{M}_{q,r}$ similarly for $\mathbf{V}$. Let $\nabla_{\mathcal{M}_{p,r}} f(\mathbf{U}^{(t+1)})$ denote the Riemannian gradient of f at $\mathbf{U}^{(t+1)}$, and analogously for $\nabla_{\mathcal{M}_{q,r}} g(\mathbf{V}^{(t+1)})$. Then the dual residual is defined as

$$r_{\text{dual}}^{(t+1)} = 0.5 \times \frac{\|\nabla_{\mathcal{M}_{p,r}} f(\mathbf{U}^{(t+1)})\|_{\text{F}}}{\sqrt{pr}} + 0.5 \times \frac{\|\nabla_{\mathcal{M}_{q,r}} g(\mathbf{V}^{(t+1)})\|_{\text{F}}}{\sqrt{qr}}.$$

Finally, for a user-specified tolerance $\varepsilon > 0$, the algorithm terminates when

$$r_{\text{primal}}^{(t+1)} \leq \varepsilon \quad \text{and} \quad r_{\text{dual}}^{(t+1)} \leq \varepsilon.$$

3.3.2 Hyperparameter Selection

In this section, we outline our strategy for selecting the hyperparameters of the SMART algorithm, namely $(\lambda_u, \lambda_v, r, r_u, r_v)$.

Tuning λ_u and λ_v . For fixed values of (r, r_u, r_v) , we select the sparsity-inducing regularization parameters λ_u and λ_v by minimizing the Bayesian Information Criterion (BIC), defined as:

$$\text{BIC}(\lambda_u, \lambda_v) := nq \log \left\{ \frac{1}{nq} \left\| \mathbf{Y} - \mathbf{X} \hat{\mathbf{C}}(\lambda_u, \lambda_v) \right\|_{\text{F}}^2 \right\} + \left[\left| \hat{\boldsymbol{\Omega}}_u(\lambda_u, \lambda_v) \right|_0 + \left| \hat{\boldsymbol{\Omega}}_v(\lambda_u, \lambda_v) \right|_0 \right] \log(nq),$$

where $\hat{\mathbf{C}} = \hat{\mathbf{U}} \hat{\mathbf{D}} \hat{\mathbf{V}}^\top$ is the estimated coefficient matrix, and $|\cdot|_0$ denotes the number of nonzero entries. The BIC balances the goodness-of-fit with model complexity by penalizing the effective degrees of freedom.

Choosing r , r_u , and r_v . We select the target rank r using the adaptive rank selection procedure proposed by Bunea et al. [2011], applied to the target response matrix $\mathbf{Y}$ and design matrix $\mathbf{X}$. For the source subspace dimensions r_u and r_v , we perform a grid search with K -fold cross-validation (typically $K = 5$), selecting the pair (r_u, r_v) that yields the lowest prediction error on held-out target samples.

In practice, we first estimate r from the data, then conduct cross-validation over a moderate grid of (r_u, r_v) values, and for each candidate pair, optimize (λ_u, λ_v) using BIC. This procedure strikes a balance between computational efficiency and model accuracy.

3.4 Theoretical Analysis

In this section, we present theoretical guarantees for the SMART estimator. We begin by introducing some notations that will facilitate our analysis.

We extend the source singular vectors $\mathbf{U}^{(0)}$ and $\mathbf{V}^{(0)}$ to full orthonormal bases of $\mathbb{R}^p$ and

$\mathbb{R}^q$, respectively, by appending orthogonal complements $\mathbf{U}_{\perp}^{(0)}$ and $\mathbf{V}_{\perp}^{(0)}$. Specifically,

$$\text{col}(\mathbf{U}^{(0)})^{\perp} = \text{span} \left\{ \mathbf{u}_{r_0+1}^{(0)}, \dots, \mathbf{u}_p^{(0)} \right\}, \quad \text{col}(\mathbf{V}^{(0)})^{\perp} = \text{span} \left\{ \mathbf{v}_{r_0+1}^{(0)}, \dots, \mathbf{v}_q^{(0)} \right\}.$$

Define the full orthonormal systems:

$$\mathbf{U}_{\text{full}}^{(0)} := [\mathbf{U}^{(0)} \ \mathbf{U}_{\perp}^{(0)}], \quad \mathbf{V}_{\text{full}}^{(0)} := [\mathbf{V}^{(0)} \ \mathbf{V}_{\perp}^{(0)}].$$

For $0 \leq r_u \leq p$ and $0 \leq r_v \leq q$, define the truncated signal subspaces and their complements as

$$\begin{aligned} \mathbf{U}^{\text{S}}(r_u) &:= [\mathbf{u}_1^{(0)}, \dots, \mathbf{u}_{r_u}^{(0)}], & \mathbf{U}_{\perp}^{\text{S}}(r_u) &:= [\mathbf{u}_{r_u+1}^{(0)}, \dots, \mathbf{u}_p^{(0)}], \\ \mathbf{V}^{\text{S}}(r_v) &:= [\mathbf{v}_1^{(0)}, \dots, \mathbf{v}_{r_v}^{(0)}], & \mathbf{V}_{\perp}^{\text{S}}(r_v) &:= [\mathbf{v}_{r_v+1}^{(0)}, \dots, \mathbf{v}_q^{(0)}], \end{aligned}$$

with the conventions $\mathbf{U}^{\text{S}}(0) = \mathbf{0}$ and $\mathbf{U}_{\perp}^{\text{S}}(p) = \mathbf{0}$, and likewise for $\mathbf{V}$.

We define the minimal truncation levels r_u^* and r_v^* such that the singular subspaces of the target matrix $\mathbf{C}^*$ are contained within the truncated source subspaces:

$$r_u^* := \inf \{ r_u : \text{col}(\mathbf{U}^*) \subseteq \text{col}(\mathbf{U}^{\text{S}}(r_u)) \}, \quad r_v^* := \inf \{ r_v : \text{col}(\mathbf{V}^*) \subseteq \text{col}(\mathbf{V}^{\text{S}}(r_v)) \}. \quad (3.5)$$

By Assumption 3, we have $r_u^*, r_v^* \leq r_0$.

To evaluate how well the target singular vectors align with the leading components of the source subspaces, we define the following sparsity-based overlap measures. For truncation levels r_u and r_v , let

$$s_u(r_u) := \left| \mathbf{U}^{*\top} \mathbf{U}_{\perp}^{\text{S}}(r_u) \right|_0, \quad s_v(r_v) := \left| \mathbf{V}^{*\top} \mathbf{V}_{\perp}^{\text{S}}(r_v) \right|_0, \quad (3.6)$$

where $|\cdot|_0$ denotes the number of nonzero entries. Here, $\mathbf{U}_{\perp}^{\text{S}}(r_u)$ and $\mathbf{V}_{\perp}^{\text{S}}(r_v)$ are the

orthogonal complements of the top- r_u and top- r_v source subspaces, respectively. Intuitively, $s_u(r_u)$ and $s_v(r_v)$ count the number of directions in the penalized complement subspaces that have nontrivial overlap with the target singular vectors. Smaller values indicate stronger alignment between the target subspaces and the leading source directions retained by the truncation.

By definition, we have $s_u(0) = s_u^*$ and $s_v(0) = s_v^*$, where s_u^* and s_v^* are as introduced in (3.2). Moreover, the functions $s_u(r_u)$ and $s_v(r_v)$ are non-increasing in r_u and r_v , respectively, and satisfy

$$s_u(r_u) = 0 \quad \text{for all } r_u \geq r_u^*, \quad s_v(r_v) = 0 \quad \text{for all } r_v \geq r_v^*.$$

Importantly, the definitions of r_u^* , r_v^* , $s_u(r_u)$, and $s_v(r_v)$ are independent of the specific orthonormal completion of the source basis. This follows from Assumption 3, which ensures that $\mathbf{U}^{*\top} \mathbf{u}_j^{(0)} = 0$ and $\mathbf{V}^{*\top} \mathbf{v}_j^{(0)} = 0$ for all $j > r_0$.

3.4.1 Technical Conditions

We now state the technical assumptions required for our theoretical analysis. Throughout Section 3.4, we treat the design matrix $\mathbf{X}$ as fixed and impose deterministic conditions on it. These assumptions typically hold with high probability when the rows of $\mathbf{X}$ are sampled i.i.d. from sub-Gaussian distributions.

We begin with a restricted eigenvalue condition, which ensures stable recovery in high-dimensional regimes.

Assumption 5 (Restricted Eigenvalue Condition). *There exist universal constants $\rho_1, \rho_2 > 0$ such that*

$$\frac{1}{2n} \|\mathbf{X}\boldsymbol{\beta}\|_2^2 \geq \frac{\rho_2}{2} \|\boldsymbol{\beta}\|_2^2 - \rho_1 \frac{\log p}{n} \|\boldsymbol{\beta}\|_1^2, \quad \text{for all } \boldsymbol{\beta} \in \mathbb{R}^p.$$

This condition holds with high probability when the design rows $\mathbf{x}_1, \dots, \mathbf{x}_n$ are i.i.d.

Gaussian with a non-degenerate covariance matrix; see Wainwright [2019, Theorem 7.16].

Next, we impose boundedness and non-degeneracy assumptions on the design matrix, the true coefficient matrix, and the source coefficient matrix.

Assumption 6 (Constrained Singular Values). *The following conditions hold for some universal constants $\nu > 0$ and $d_{\min}, d_{\min}^{(0)} > 0$:*

- *The design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ satisfies*

$$\sup_{\mathbf{u} \in \mathbb{R}^p: \|\mathbf{u}\|_2 \leq 1} \|\mathbf{X}\mathbf{u}\|_2^2 \leq \nu n.$$

- *The smallest nonzero singular values of the target and source coefficient matrices satisfy $d_r^* \geq d_{\min}$ and $d_{r_0}^{(0)} \geq d_{\min}^{(0)}$, respectively.*

To ensure identifiability and perturbation stability of the singular subspaces, we also impose a spectral gap condition on both the target matrix $\mathbf{C}^*$ and the source matrix $\mathbf{C}^{(0)}$.

Assumption 7 (Spectral Gap Condition). *There exists a universal constant $0 < \tau < 1$ such that the squared singular values exhibit a relative spectral-gap condition:*

$$d_{j-1}^{*2} - d_j^{*2} \geq \tau d_{j-1}^{*2}, \quad \text{for all } 2 \leq j \leq r, \quad (3.7)$$

$$\left(d_{j-1}^{(0)}\right)^2 - \left(d_j^{(0)}\right)^2 \geq \tau \left(d_{j-1}^{(0)}\right)^2, \quad \text{for all } 2 \leq j \leq r_0. \quad (3.8)$$

This assumption is standard in the low-rank estimation literature; see, for example, Uematsu et al. [2019, Condition 5].

Finally, we specify the distributional assumption on the noise matrix.

Assumption 8 (Gaussian Noise). *Let $\mathbf{E} = [\mathbf{e}_1, \dots, \mathbf{e}_n]^\top \in \mathbb{R}^{n \times q}$ denote the noise matrix. We assume that the rows $\mathbf{e}_1, \dots, \mathbf{e}_n$ are i.i.d. Gaussian vectors with mean zero and covariance*

matrix Σ . Moreover, the covariance matrix satisfies

$$\|\Sigma\|_2 \leq \alpha_{\max},$$

for some universal constant $\alpha_{\max} > 0$.

The Gaussian assumption facilitates concentration-based analysis. Our results can be extended to sub-Gaussian noise with suitable adjustments to constants in the bounds.

3.4.2 Theoretical Guarantees

In this section, we establish theoretical guarantees for the SMART estimator. The objective function in (3.3) is inherently nonconvex, posing substantial challenges for directly analyzing the statistical properties of its solution. To address this issue, we adopt a two-stage strategy inspired by the convexity-assisted nonconvex optimization (CANO) framework of Uematsu et al. [2019]. This approach enables us to derive non-asymptotic error bounds for the SMART estimator within a principled theoretical framework.

The CANO paradigm tackles nonconvex optimization problems through a structured two-stage procedure. In the first stage, we compute an initial estimator, typically a convex estimator, which serves as a coarse but consistent estimate of the true parameter. In the second stage, we refine this estimate by solving the original nonconvex SMART optimization problem, with the solution restricted to a shrinking neighborhood centered at the initializer. This localized constraint ensures that the optimization is performed in a region where the objective function exhibits more favorable geometric and statistical properties, thereby mitigating the difficulties associated with nonconvexity.

Crucially, our analysis centers on the global optimizer of the SMART objective subject to these neighborhood constraints. By incorporating structural information through spectral regularization and restricting the solution space to a statistically controlled region, the SMART estimator achieves substantially faster convergence rates than the initial estimator.

This refinement yields sharper estimation guarantees and facilitates more efficient learning.

Note that the CANO framework is used here as a theoretical device and is not intended for direct implementation. For practical algorithms, we refer the reader to Section 3.3.1.

Stage 1: Initialization. We begin by constructing an initial estimator $\hat{\mathbf{C}}^{\text{init}}$ of the true coefficient matrix $\mathbf{C}^*$ using any consistent procedure. For theoretical analysis, we treat $\hat{\mathbf{C}}^{\text{init}}$ as a black-box estimator that satisfies the following non-asymptotic error bounds:

$$\left\| \hat{\mathbf{C}}^{\text{init}} - \mathbf{C}^* \right\|_{\text{F}} \leq R_{2,n}, \quad \left| \hat{\mathbf{C}}^{\text{init}} - \mathbf{C}^* \right|_1 \leq R_{1,n}. \quad (3.9)$$

A concrete example is the Lasso estimator:

$$\hat{\mathbf{C}}^{\text{lasso}} \in \arg \min_{\mathbf{C} \in \mathbb{R}^{p \times q}} \left\{ \frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\mathbf{C}\|_{\text{F}}^2 + \lambda_0 \|\mathbf{C}\|_1 \right\},$$

with a suitable regularization parameter $\lambda_0 \geq 0$. The following theorem provides non-asymptotic guarantees for this estimator.

Theorem 29 (Error Bounds for Lasso Estimator). *Let $s_0 = |\mathbf{C}^*|_0$ denote the sparsity level of $\mathbf{C}^*$. Suppose Assumption 5 holds. Then, with probability at least $1 - 2(pq)^{1-c_0^2/2}$, the Lasso estimator satisfies*

$$\left\| \hat{\mathbf{C}}^{\text{lasso}} - \mathbf{C}^* \right\|_{\text{F}} \leq c_0 \sqrt{\frac{s_0 \log(pq)}{n}}, \quad \left| \hat{\mathbf{C}}^{\text{lasso}} - \mathbf{C}^* \right|_1 \leq c_0 s_0 \sqrt{\frac{\log(pq)}{n}},$$

for some universal constant $c_0 > \sqrt{2}$.

The proof follows directly from [Wainwright, 2019, Corollary 9.26]. Alternative initializers include ordinary least squares, ridge regression, and reduced-rank regression, provided they satisfy (3.9).

Stage 2: Refined Estimation. Next, we solve the SMART optimization problem (3.3) subject to the following constraints:

$$\left\| \widehat{\mathbf{C}} - \widehat{\mathbf{C}}^{\text{init}} \right\|_{\text{F}} \leq 2R_{2,n}, \quad \left| \widehat{\mathbf{C}} - \widehat{\mathbf{C}}^{\text{init}} \right|_1 \leq 2R_{1,n}. \quad (3.10)$$

Let $\widehat{\mathbf{C}}$ be the global optimum under these constraints, and define the estimation error $\widehat{\Delta} = \widehat{\mathbf{C}} - \mathbf{C}^*$. We introduce the spectral conditioning factor

$$\eta_r := 1 + \frac{1}{\tau} \left(\sum_{j=1}^r \left(\frac{d_1^*}{d_j^*} \right)^2 \right)^{1/2},$$

where $\tau > 0$ is the spectral gap constant from Assumption 7.

Additionally, let

$$\varepsilon^{(0)} := \left\| \widetilde{\mathbf{C}}^{(0)} - \mathbf{C}^{(0)} \right\|_2$$

denote the spectral approximation error of the source matrix. This quantity characterizes the quality of the noisy source matrix $\widetilde{\mathbf{C}}^{(0)}$, with smaller values indicating higher fidelity to the true source matrix $\mathbf{C}^{(0)}$.

Theorem 30 (Error Bound for SMART Estimator). *Suppose Assumptions 3–7 and the initialization condition (3.9) hold. Fix an error tolerance level $0 < \delta < 1$. Assume the sample size n is sufficiently large such that*

$$R_{2,n} \leq \frac{\tau}{18\sqrt{5}} \cdot \left(\frac{d_r^*}{\sqrt{r}} \right) \cdot \left(\frac{d_r^*}{d_1^*} \right)^{3/2}, \quad (3.11)$$

$$R_{1,n} \lesssim \eta_r^2 [r(r_u + r_v) + s_u(r_u) + s_v(r_v)] \cdot \frac{\log(12pq/\delta)}{\log p}, \quad (3.12)$$

and

$$\frac{9c^2\eta_r^2 R_{2,n}^2}{2\alpha_{\max} d_{\min}^2} \left[\text{tr}(\mathbf{\Sigma}) + c_1 \max \left\{ \sqrt{\text{tr}(\mathbf{\Sigma}^2) \log(6p/\delta) / c_2}, \alpha_{\max} \log(6p/\delta) / c_2 \right\} \right] \leq \log(12pr/\delta), \quad (3.13)$$

where c, c_1, c_2 are universal positive constants.

Choose regularization parameters

$$\lambda_u = 8\sqrt{2} \cdot \sqrt{\frac{\nu\alpha_{\max} \log(12pr/\delta)}{n}}, \quad \lambda_v = 2\sqrt{2} \cdot \sqrt{\frac{\nu\alpha_{\max} \log(6qr/\delta)}{n}}. \quad (3.14)$$

Then, with probability at least $1 - \delta$, the estimation error satisfies

$$\begin{aligned} \|\widehat{\Delta}\|_{\text{F}} &\lesssim \eta_r \cdot \sqrt{\frac{(r(r_u + r_v + 1) + s_u(r_u) + s_v(r_v)) \cdot \log(12r \max(p, q)/\delta)}{n}} \\ &\quad + \varepsilon_{\text{source}} \cdot \left(\frac{\log(12r \max(p, q)/\delta)}{n} \right)^{1/4}, \end{aligned} \quad (3.15)$$

where the source error $\varepsilon_{\text{source}}$ is given by

$$\begin{aligned} \varepsilon_{\text{source}} &= \left[r \{ (r_u^* - r_u)_+ + (r_v^* - r_v)_+ \} - (s_u(r_u) + s_v(r_v)) \right]^{1/2} \\ &\quad \times \left(\varepsilon^{(0)} \right)^{1/2} \left(2d_1^{(0)} + \varepsilon^{(0)} \right)^{1/2} \\ &\quad + \sqrt{r} \cdot \left[(r_u^*)^{1/4} \cdot \sqrt{p - \max(r_u^*, r_u)} + (r_v^*)^{1/4} \cdot \sqrt{q - \max(r_v^*, r_v)} \right]^{1/2} \\ &\quad \times \left(\varepsilon^{(0)} \right)^{1/4} \left(2d_1^{(0)} + \varepsilon^{(0)} \right)^{1/4}, \end{aligned} \quad (3.16)$$

and $(t)_+ = \max\{t, 0\}$ denotes the positive part.

See proof in Section 3.8.2. The estimation error on the right-hand side of (3.15) consists of two components. The first term, which follows a parametric rate, reflects the estimation error under the ideal scenario where the source matrix $\mathbf{C}^{(0)}$ is perfectly clean, i.e., $\varepsilon^{(0)} = 0$.

The quantity $r(r_u + r_v + 1) + s_u(r_u) + s_v(r_v)$ in the numerator represents the effective degrees of freedom in the model. To see this, consider the decomposition

$$\mathbf{U}^* = \underbrace{\mathbf{U}^s(r_u) \left[(\mathbf{U}^s(r_u))^\top \mathbf{U}^* \right]}_{\mathbf{P}_1^u} + \underbrace{\mathbf{U}_\perp^s(r_u) \left[(\mathbf{U}_\perp^s(r_u))^\top \mathbf{U}^* \right]}_{\mathbf{P}_2^u},$$

which implies that $\mathbf{U}^*$ can be reconstructed by estimating the projection components $\mathbf{P}_1^u$ and $\mathbf{P}_2^u$, assuming the source subspaces $\mathbf{U}^s(r_u)$ and $\mathbf{U}_\perp^s(r_u)$ are known. The matrix $\mathbf{P}_1^u \in \mathbb{R}^{r_u \times r}$ requires estimating rr_u parameters. For $\mathbf{P}_2^u$, the ℓ_1 penalty in (3.3) encourages sparsity, so we only need to estimate its nonzero entries, whose count is captured by $s_u(r_u)$ as defined in (3.6). Therefore, the total number of parameters to estimate for recovering $\mathbf{U}^*$ is $rr_u + s_u(r_u)$. A similar argument applies to $\mathbf{V}^*$, yielding $rr_v + s_v(r_v)$ parameters. Finally, estimating the diagonal matrix $\mathbf{D}^*$ contributes an additional r degrees of freedom. Altogether, the total number of parameters to estimate is $r(r_u + r_v + 1) + s_u(r_u) + s_v(r_v)$.

The second term in (3.15) accounts for the approximation error induced by noise in the source matrix $\tilde{\mathbf{C}}^{(0)}$. When the source is noiseless, i.e., $\varepsilon^{(0)} = 0$, the approximation error $\varepsilon_{\text{source}}$ vanishes, and the entire second term disappears. In contrast, when the source matrix is heavily contaminated—reflected by a large $\varepsilon^{(0)}$ —the value of $\varepsilon_{\text{source}}$ increases, indicating that the noisy source directions may misalign with the target signal. Moreover, when $r_u \geq r_u^*$ and $r_v \geq r_v^*$, where r_u^* and r_v^* denote the minimal truncation levels defined in (3.5), the first term of $\varepsilon_{\text{source}}$ in (3.16) vanishes. Since the second term in (3.16) grows more slowly with $\varepsilon^{(0)}$, increasing r_u and r_v beyond these thresholds can significantly reduce the overall approximation error.

The structural parameters r_u and r_v govern the trade-off between statistical estimation error and approximation error due to source noise. By construction, both $r_u + s_u(r_u)$ and $r_v + s_v(r_v)$ are non-decreasing functions of r_u and r_v , reflecting that increasing these structural ranks expands the unpenalized subspace. As a result, the estimation error grows with (r_u, r_v)

due to increased model complexity, while the approximation error decreases as the estimator gains more flexibility to avoid noisy or misaligned source directions.

When the source matrix is clean and the estimated singular vectors are reliable, the optimal choice is often $r_u = r_v = 0$, enabling the estimator to fully leverage the source spectral structure and minimize estimation error. However, in the presence of moderate source noise, selecting intermediate values for (r_u, r_v) helps strike a balance—preserving useful information while discarding unreliable directions. In the extreme case where the source matrix is heavily contaminated, setting $r_u = p$ and $r_v = q$ removes the regularization terms entirely. This effectively ignores the source information and reduces the estimator to a target-only model, thereby avoiding potential degradation from misleading auxiliary data.

3.5 Simulation

In this section, we evaluate the empirical performance of the proposed SMART algorithm through a series of simulation studies. We compare its estimation accuracy with several baseline methods under varying conditions, including different sample sizes, ranks, and levels of source quality.

Recall that n denotes the number of samples, p the number of input features, and q the number of output responses. We use r to represent the rank of the true target coefficient matrix $\mathbf{C}^*$, and r_0 the rank of the true source coefficient matrix $\mathbf{C}^{(0)}$, where $r \leq r_0$. The rank of the estimated coefficient matrix $\hat{\mathbf{C}}$ is denoted by $\hat{r}$.

To assess estimation quality, we report the average Frobenius norm error of the estimated coefficient matrix $\hat{\mathbf{C}} = \hat{\mathbf{U}}\hat{\mathbf{D}}\hat{\mathbf{V}}^\top$ as

$$\frac{1}{\sqrt{pq}} \left\| \hat{\mathbf{C}} - \mathbf{C}^* \right\|_{\text{F}}.$$

3.5.1 Data Generation Process

We simulate data according to the following procedure:

1. **Source Regression Coefficient Matrix:** We begin by generating two random matrices $\mathbf{U}' \in \mathbb{R}^{p \times r_0}$ and $\mathbf{V}' \in \mathbb{R}^{q \times r_0}$, whose entries are independently drawn from the standard normal distribution. The columns of $\mathbf{U}'$ and $\mathbf{V}'$ are orthonormalized using QR decomposition to obtain $\mathbf{U}^{(0)} \in \mathbb{R}^{p \times r_0}$ and $\mathbf{V}^{(0)} \in \mathbb{R}^{q \times r_0}$, respectively. We construct a diagonal matrix $\mathbf{D}^{(0)} \in \mathbb{R}^{r_0 \times r_0}$ with singular values linearly spaced between 1.0 and 10.0. The true source regression coefficient matrix is then defined as

$$\mathbf{C}^{(0)} = \mathbf{U}^{(0)} \mathbf{D}^{(0)} \mathbf{V}^{(0)\top}.$$

To model realistic scenarios with imperfect source knowledge, we add Gaussian noise:

$$\tilde{\mathbf{C}}^{(0)} = \mathbf{C}^{(0)} + \mathbf{E}^{(0)}, \quad \text{where } \mathbf{E}_{ij}^{(0)} \sim \mathcal{N}(0, \sigma_0^2) \text{ i.i.d.}$$

The noise level σ_0 controls the quality of the source matrix, with smaller values corresponding to higher fidelity. The case $\sigma_0 = 0$ represents perfect access to the true source matrix.

2. **Feature Matrix:** The input feature matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ consists of i.i.d. rows drawn from a multivariate normal distribution $\mathcal{N}_p(0, \mathbf{\Sigma}_x)$, where the covariance matrix $\mathbf{\Sigma}_x \in \mathbb{R}^{p \times p}$ has entries defined by $\Sigma_{x,ij} = 0.5^{|i-j|}$.
3. **Target Regression Coefficient Matrix:** To construct the target coefficient matrix $\mathbf{C}^*$, we randomly select r columns from both $\mathbf{U}^{(0)}$ and $\mathbf{V}^{(0)}$ (without replacement), forming $\mathbf{U}^* \in \mathbb{R}^{p \times r}$ and $\mathbf{V}^* \in \mathbb{R}^{q \times r}$. We then generate a diagonal matrix $\mathbf{D}^* \in \mathbb{R}^{r \times r}$ with values linearly spaced between 3.0 and 5.0. The target regression coefficient

matrix is given by

$$\mathbf{C}^* = \mathbf{U}^* \mathbf{D}^* \mathbf{V}^{*\top}.$$

4. **Response Matrix:** Finally, the response matrix $\mathbf{Y} \in \mathbb{R}^{n \times q}$ is generated according to the standard multivariate linear model:

$$\mathbf{Y} = \mathbf{X} \mathbf{C}^* + \mathbf{E}, \quad \text{where } \mathbf{E}_{ij} \sim \mathcal{N}(0, \sigma^2) \text{ i.i.d.}$$

The noise standard deviation is fixed at $\sigma = 0.5$ for all experiments.

3.5.2 *Experimental Design*

We evaluate the performance of SMART across multiple simulation settings with varying problem dimensions. Specifically, we consider the following three configurations:

- **Model I:** $p = 100, q = 50$
- **Model II:** $p = 150, q = 100$
- **Model III:** $p = 300, q = 200$

To assess SMART under different conditions, we conduct four experiments. In each experiment, one key parameter is varied while the others are held fixed at their default values. The default settings are: sample size $n = 200$, estimated rank $\hat{r} = 5$, source subspace ranks $r_u = r_v = 10$, and source noise level $\sigma_0 = 0.01$. For all models, we fix the true rank of the source coefficient matrix at $r_0 = 10$ and the true rank of the target coefficient matrix at $r = 5$. Each configuration is repeated 100 times to compute reliable averages and standard errors.

1. **Experiment 1:** Vary the sample size n , while keeping $\hat{r}, r_u, r_v$, and σ_0 fixed.
2. **Experiment 2:** Vary the estimated rank $\hat{r}$, with n, r_u, r_v , and σ_0 held constant.

3. **Experiment 3:** Let $r_u = r_v = r_s$, and vary r_s , keeping n , $\hat{r}$, and σ_0 fixed.

4. **Experiment 4:** Vary the source noise level σ_0 , with $\hat{r}$, r_u , and r_v fixed.

In the above experiments, the hyperparameters $(\hat{r}, r_u, r_v)$ for SMART are either pre-specified or selected in a fully data-driven manner. For the fixed-rank version, we specify $(\hat{r}, r_u, r_v)$ manually as above, and tune the regularization parameters (λ_u, λ_v) using the BIC criterion described in Section 3.3.2; we denote this method as **SMART (fixed ranks)**.

We also implement a fully automated variant of SMART that uses the hyperparameter selection strategy from Section 3.3.2 to jointly select $(\hat{r}, r_u, r_v)$ along with (λ_u, λ_v) ; we denote this method simply as **SMART**. Since SMART does not rely on manually specified ranks, its performance remains constant across experiments that vary $\hat{r}$ or r_u, r_v . Therefore, we display its results in Experiments 2 and 3 as horizontal reference lines corresponding to the case $n = 200$.

3.5.3 Baseline Methods

We compare the proposed SMART algorithm against the following state-of-the-art low-rank regression methods:

- **RRR (Reduced-Rank Regression)**. An adaptive low-rank regression method that extends traditional RRR by employing adaptive nuclear norm penalization to promote low-rank structure with improved estimation accuracy [Chen et al., 2013].
- **SRRR (Sparse Reduced-Rank Regression)**. An extension of classical RRR that incorporates sparsity-inducing group lasso penalties, enabling simultaneous dimension reduction and variable selection [Chen and Huang, 2012].
- **SOFAR (Sparse Orthogonal Factor Regression)**. A structured estimation method that jointly enforces sparsity and orthogonality by solving a regularized optimization

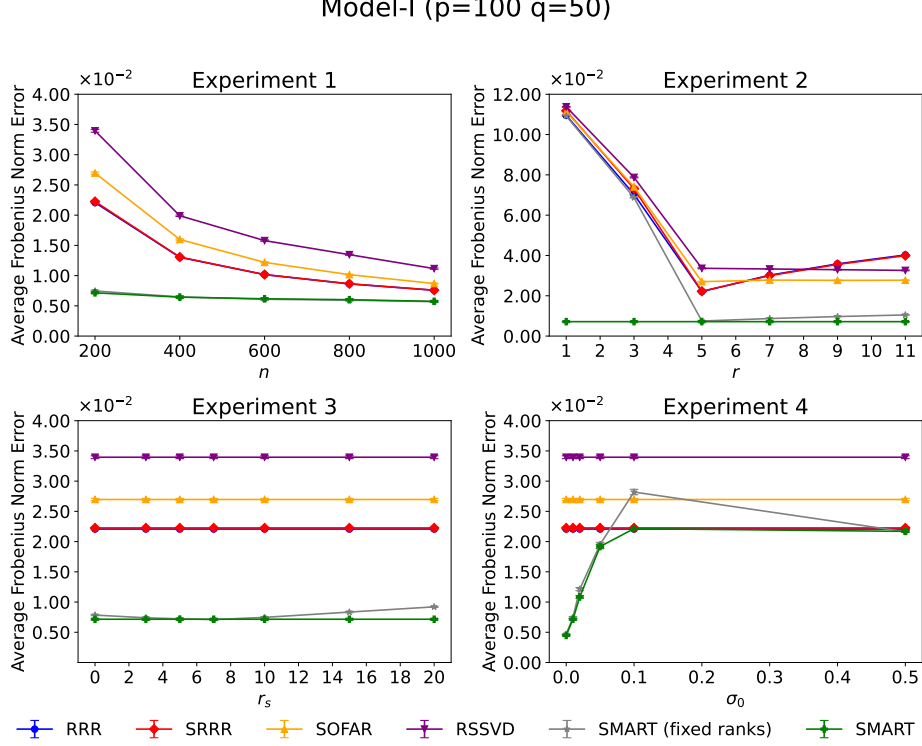


Figure 3.1: Simulation results of SMART for Model I.

problem over sparse singular vectors subject to orthogonality constraints [Uematsu et al., 2019].

- **RSSVD (Reduced-Rank Regression with Sparse SVD).** A layer-wise estimation method that applies adaptive penalties to individual rank-one components, inducing sparsity in the singular vectors of the coefficient matrix [Mukherjee and Zhu, 2011, Mukherjee et al., 2015].

All baseline methods rely solely on the target data for estimation and do not incorporate any source information. Therefore, comparisons with these methods allow us to quantify the benefit of transfer learning enabled by SMART.

We adopt the implementations provided in the R package `rrpack` [Chen and Wang, 2022]. Except in Experiment 2, we set the estimator rank to the ground truth value $r = 5$. Since the baseline methods do not leverage source information, their performance remains unchanged

in Experiments 3 and 4. Accordingly, we display their results as horizontal reference lines, corresponding to the outcomes obtained at $n = 200$ in Experiment 1.

3.5.4 Results

The results for Models I–III are presented in Figures 3.1–3.3. Across all experiments and settings, the SMART algorithm consistently outperforms the baseline methods.

In Experiments 2 and 3, we observe that SMART with pre-specified ranks exhibits strong robustness to the choice of hyperparameters $(\hat{r}, r_u, r_v)$. Furthermore, results from Experiment 4 show that the fully data-driven version of SMART, which selects $(\hat{r}, r_u, r_v)$ via the strategy described in Section 3.3.2, never underperforms relative to the target-only baselines.

More specifically, when the source matrix is accurate (i.e., low noise), SMART achieves substantial improvements over the baselines. Conversely, when the source matrix is highly contaminated, SMART automatically downweights or discards the unreliable source information and performs comparably to target-only methods. These results demonstrate that SMART is highly robust to the quality of the source data and is capable of effectively avoiding negative transfer.

3.6 Real-World Data Analysis

In this section, we evaluate the performance of the proposed SMART algorithm on a real-world dataset. The dataset we use is a single-cell multi-omics dataset collected from bone marrow mononuclear cells of 12 healthy human donors ¹. Leveraging recent advances in multi-modal single-cell technologies, the dataset provides simultaneous measurements of gene expression (GEX) and protein expression (ADT) for each cell. Specifically, we treat each row of $\mathbf{Y}$ as the protein expression levels of an individual cell, and each corresponding row of $\mathbf{X}$ as the gene expression levels for that same cell. Our objective is to build a linear model

1. The dataset was released for the Multimodal Single-Cell Data Integration Challenge at NeurIPS 2021 and is publicly available at <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE194122>

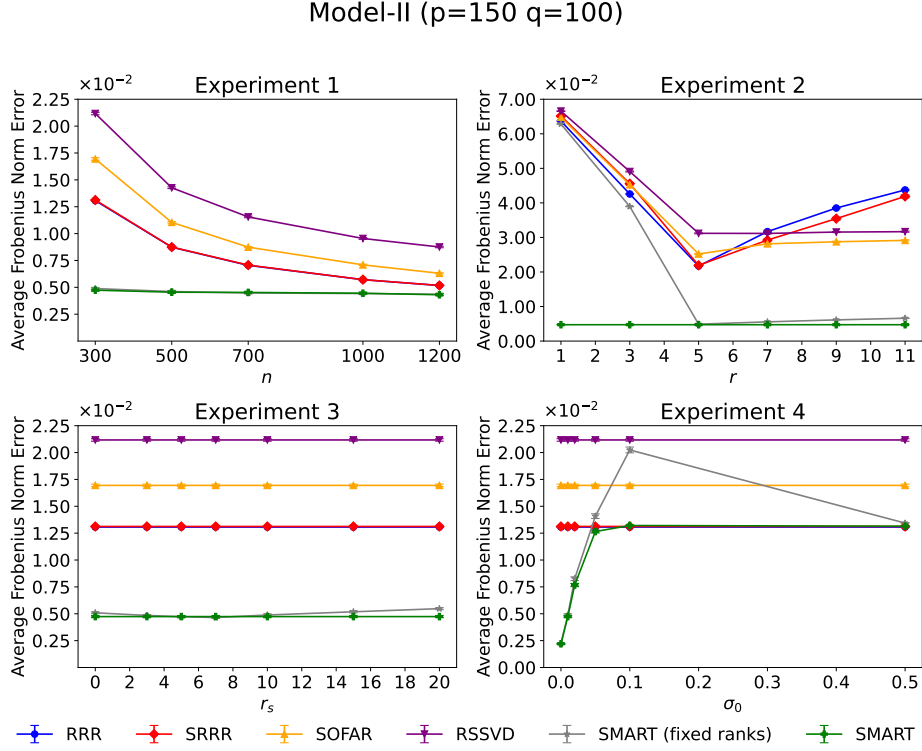


Figure 3.2: Simulation results of SMART for Model II.

that predicts protein expression levels from gene expression profiles.

In addition to protein and gene expression data, the dataset includes cell-type annotations. Notably, the number of cells varies considerably across cell types: while some types are well-represented, others have only limited samples. This imbalance poses challenges for accurate model estimation using only the data from a rare cell type. Figure 3.4 displays the distribution of cell counts across different cell types.

To address the data scarcity challenge for underrepresented cell types, we leverage information from biologically related and well-sampled source populations. In this case study, we designate ILC1 (tissue-resident innate lymphoid cells) as the target cell type and NK cells (circulating cytotoxic natural killer cells) as the source. This choice is motivated not only by sample size considerations—ILC1 has limited representation ($n = 552$) compared to NK cells ($n = 5434$)—but also by the asymmetry embedded in the spectral containment

Model-III (p=300 q=200)

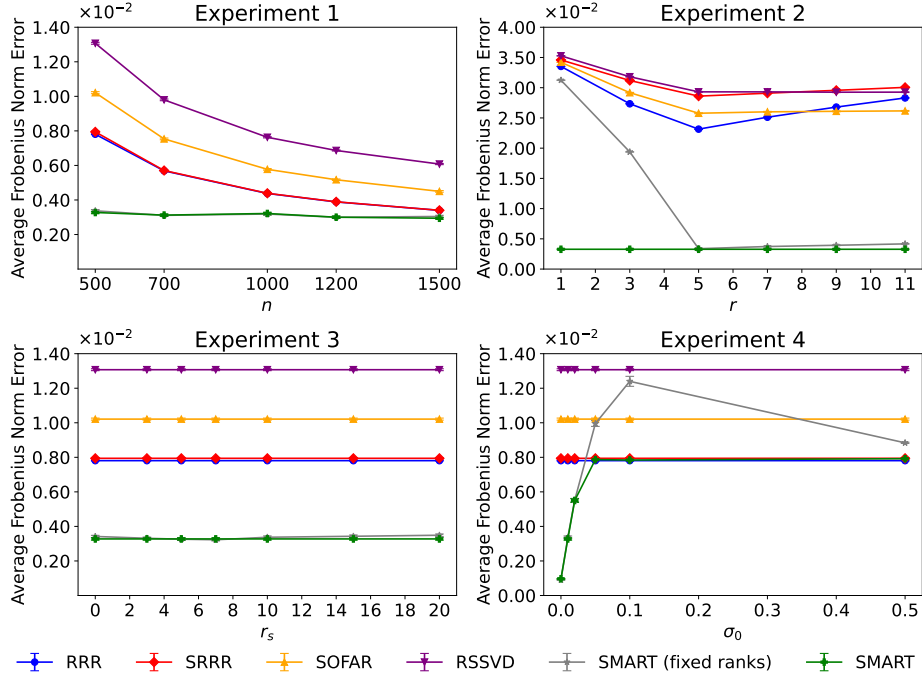


Figure 3.3: Simulation results of SMART for Model III.

assumption. Biologically, NK cells and ILC1s are both members of the group 1 innate lymphoid cell family and share key transcription factors (e.g., T-bet) and effector cytokines (e.g., IFN- γ). However, NK cells are developmentally and functionally broader: they circulate systemically, exhibit cytotoxic activity, and operate in diverse immunological environments. In contrast, ILC1s are tissue-resident, more specialized in local immune regulation, and functionally constrained. This developmental hierarchy suggests that the regulatory programs active in ILC1s are likely a specialized subset of those in NK cells [Vivier et al., 2018]. Consequently, the gene-protein association patterns (i.e., the left and right singular subspaces of the regression matrix) for ILC1 are expected to be contained within those of NK cells. This containment aligns with the spectral containment assumption, justifying the use of NK cells as the source and ILC1 as the target in our transfer learning framework.

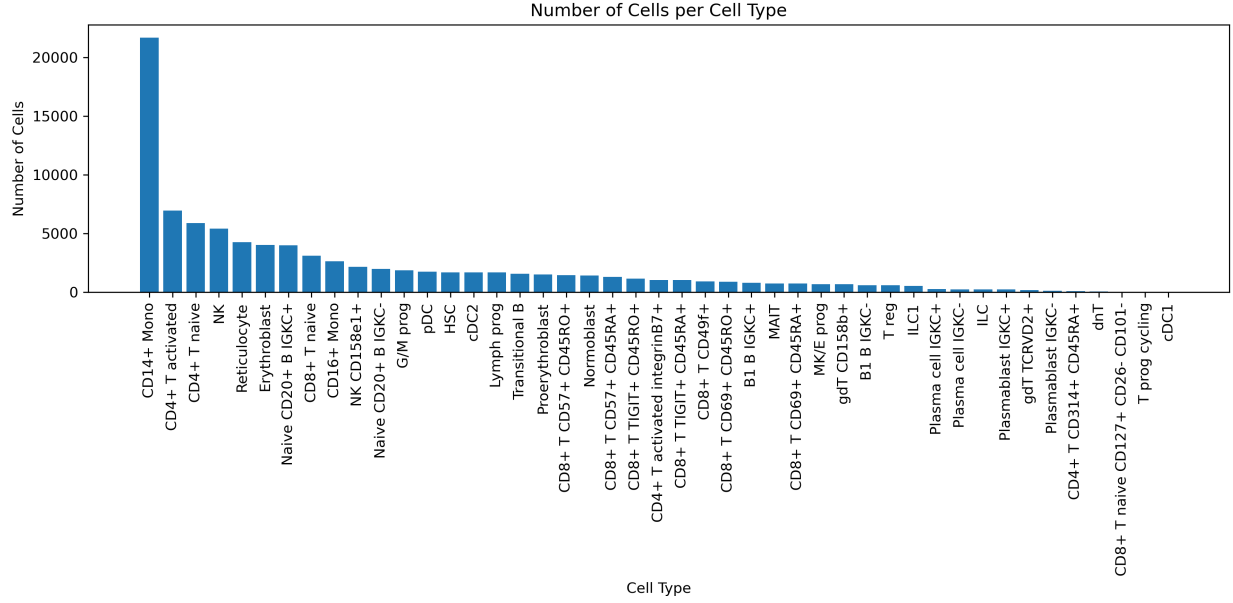


Figure 3.4: Cell counts for each annotated cell type in the single-cell multi-modal dataset.

Data Preprocessing. We applied a series of preprocessing steps to prepare the gene and protein expression matrices for modeling. Starting from the raw dataset, we extracted the gene expression (GEX) and protein expression (ADT) matrices, comprising 13,953 genes and 134 proteins measured across 90,261 cells. The GEX matrix was processed via library size normalization, log-transformation, selection of the top 3,000 highly variable genes (HVGs), and standardization. For the ADT matrix, we applied centered log-ratio (CLR) normalization to address compositional effects. We then filtered the dataset to retain only NK and ILC1 cells, resulting in 5,434 source cells and 552 target cells. To ensure modeling stability, we removed genes with zero variance in either group, retaining 2,998 GEX features and all 134 ADT markers. Finally, we performed marginal screening on both tasks and took the union of selected features, yielding 248 shared gene predictors.

Implementation. To evaluate the SMART algorithm, we first construct the noisy source matrix $\tilde{\mathbf{C}}^{(0)}$ by applying various baseline estimators to the NK (source) data. Specifically, we consider seven methods: Lasso, Ridge, OLS, RRR, SRRR, RSSVD, and SOFAR. For each method, the estimated matrix $\tilde{\mathbf{C}}^{(0)}$ is used as input to SMART to produce a target

regression estimate $\hat{\mathbf{C}}$. The model is trained on a random 70% subset of ILC1 (target) cells and evaluated on the remaining 30% held-out data. Prediction accuracy is measured using the Frobenius norm of the test error:

$$\frac{1}{\sqrt{n_{\text{test}}q}} \left\| \mathbf{X}_{\text{test}} \hat{\mathbf{C}} - \mathbf{Y}_{\text{test}} \right\|_{\text{F}},$$

where n_{test} is the number of test samples and q is the number of output variables. To ensure robustness, we repeat this procedure 100 times using different train/test splits generated from a fixed list of random seeds. For each method, we report the mean, standard deviation, median, and interquartile range (Q1 and Q3) of the prediction errors across repetitions.

Baseline Methods. To contextualize the performance of SMART, we compare it against a set of baseline methods grouped into two categories: *source-only* and *target-only*. The source-only methods apply the regression matrix estimated from NK data directly to the ILC1 test set without any retraining. In contrast, the target-only methods apply the same set of estimators—Lasso, Ridge, OLS, RRR, SRRR, RSSVD, and SOFAR—solely to the ILC1 data, without access to the source. For consistency, we evaluate each method using the same 70%/30% train/test splits and error metric as in SMART. The source-only methods are evaluated directly on the test set, while the target-only methods are trained on the target training set. Each experiment is repeated over 100 random splits, and summary statistics are computed across repetitions.

Results. Table 3.1 summarizes the prediction error statistics for all methods on the single-cell multi-modal dataset. Among the methods evaluated, SMART consistently achieves the lowest prediction errors, with mean values near 0.20 and minimal variability across different source initializations. In contrast, target-only methods exhibit significantly higher errors, with means ranging from 0.33 (Lasso, Ridge) to above 0.46 (RSSVD), highlighting the difficulty of fitting accurate models using limited ILC1 data alone. Source-only methods

Method	Mean	SE	Median	Q1	Q3
<i>Target Only Methods</i>					
Target Only (RRR)	0.410	2.4×10^{-2}	0.307	0.259	0.478
Target Only (SRRR)	0.449	3.4×10^{-2}	0.292	0.227	0.533
Target Only (RSSVD)	0.464	3.4×10^{-2}	0.309	0.239	0.489
Target Only (SOFAR)	0.360	1.8×10^{-2}	0.297	0.267	0.355
Target Only (Lasso)	0.331	3.0×10^{-4}	0.331	0.330	0.333
Target Only (OLS)	0.432	2.0×10^{-3}	0.428	0.416	0.442
Target Only (Ridge)	0.336	4.0×10^{-4}	0.336	0.332	0.338
<i>Source Only Methods</i>					
Source Only (RRR)	0.331	4.0×10^{-4}	0.331	0.328	0.334
Source Only (SRRR)	0.332	4.0×10^{-4}	0.332	0.329	0.334
Source Only (RSSVD)	0.358	6.0×10^{-4}	0.358	0.355	0.362
Source Only (SOFAR)	0.382	4.0×10^{-4}	0.383	0.380	0.385
Source Only (Lasso)	0.365	3.0×10^{-4}	0.365	0.363	0.367
Source Only (OLS)	0.358	2.0×10^{-4}	0.358	0.357	0.360
Source Only (Ridge)	0.361	3.0×10^{-4}	0.361	0.360	0.363
<i>SMART Methods</i>					
SMART (RRR)	0.200	3.0×10^{-4}	0.200	0.198	0.202
SMART (SRRR)	0.200	3.0×10^{-4}	0.200	0.198	0.202
SMART (RSSVD)	0.203	5.0×10^{-4}	0.202	0.200	0.204
SMART (SOFAR)	0.211	2.1×10^{-3}	0.204	0.201	0.215
SMART (Ridge)	0.203	4.0×10^{-4}	0.203	0.201	0.205
SMART (Lasso)	0.201	5.0×10^{-4}	0.201	0.199	0.203
SMART (OLS)	0.200	4.0×10^{-4}	0.200	0.198	0.202

Table 3.1: Summary statistics of prediction errors across 100 repeated experiments on the single-cell multi-modal dataset.

demonstrate stable but suboptimal performance, underscoring the importance of task-specific adaptation. Notably, SMART not only improves upon both baselines in terms of average accuracy but also yields the narrowest interquartile ranges, indicating strong generalization and robustness across data splits.

3.7 Conclusion

We propose SMART, a novel spectral transfer learning framework for multi-task linear regression that enhances estimation in low-sample target settings by leveraging structural knowledge from a related source task. SMART imposes structured regularization that aligns the target coefficient matrix with the singular vectors of the source, enabling effective information transfer. The resulting nonconvex optimization problem is efficiently solved via an ADMM-based algorithm with provable convergence guarantees. Theoretically, SMART achieves non-asymptotic error bounds under mild spectral conditions. Empirically, it consistently outperforms baseline estimators across a wide range of synthetic and real-world experiments, while exhibiting robustness to source misspecification and effectively mitigating negative transfer.

There are several promising avenues for future research. First, extending the spectral transfer framework to more general model classes—such as generalized linear models and nonlinear regressions—could broaden its applicability and impact. In particular, incorporating spectral alignment into neural network architectures may offer new perspectives and practical benefits in deep transfer learning. Second, developing formal inference procedures to quantify uncertainty would enhance interpretability and enable more reliable decision-making in high-stakes settings. Lastly, designing adaptive methods for selecting structural ranks and tuning hyperparameters—while maintaining both computational efficiency and theoretical guarantees—remains an important and challenging direction for future work.

3.8 Technical Details

This section presents the technical proofs for the main theorems in this chapter. To facilitate the statement, we define several notational conventions.

We denote the structured factor matrices and their estimators as follows:

$$\mathbf{A}^* = \mathbf{U}^* \mathbf{D}^*, \quad \mathbf{B}^* = \mathbf{V}^* \mathbf{D}^*, \quad \widehat{\mathbf{A}} = \widehat{\mathbf{U}} \widehat{\mathbf{D}}, \quad \text{and} \quad \widehat{\mathbf{B}} = \widehat{\mathbf{V}} \widehat{\mathbf{D}}.$$

We define the estimation error of the regression matrix and its structured components as:

$$\widehat{\Delta} = \widehat{\mathbf{C}} - \mathbf{C}^*, \quad \widehat{\Delta}^a = \widehat{\mathbf{A}} - \mathbf{A}^*, \quad \widehat{\Delta}^b = \widehat{\mathbf{B}} - \mathbf{B}^*, \quad \widehat{\Delta}^d = \widehat{\mathbf{D}} - \mathbf{D}^*, \quad (3.17)$$

and we let $\widehat{\Delta}^{\text{init}} = \widehat{\mathbf{C}}^{\text{init}} - \mathbf{C}^*$ denote the error of the initial estimator.

We also define the generalized inverse diagonal matrix

$$\widehat{\mathbf{D}}^- = \text{diag}(\widehat{d}_1^-, \dots, \widehat{d}_r^-), \quad \text{where} \quad \widehat{d}_j^- = \widehat{d}_j^{-1} \cdot \mathbb{1}\{\widehat{d}_j > 0\}, \quad \text{for all } 1 \leq j \leq r.$$

3.8.1 Preliminary Lemmas

In this section, we present and prove a set of preliminary lemmas that are instrumental for establishing our main theoretical results. We begin with a perturbation bound that ensures sign consistency between estimated and true singular vectors.

Lemma 31 (Sign Consistency). *Let $\mathbf{C} = \mathbf{U} \mathbf{D} \mathbf{V}^\top$ and $\mathbf{C}^* = \mathbf{U}^* \mathbf{D}^* \mathbf{V}^{*\top}$ be two matrices with $\text{rank}(\mathbf{C}^*) = r$. Suppose the squared singular values of $\mathbf{C}^*$ satisfy the separation condition*

$$d_{j-1}^{*2} - d_j^{*2} \geq \tau d_{j-1}^{*2} \quad \text{for all } 2 \leq j \leq r,$$

for some universal constant $0 < \tau < 1$.

Fix an index $1 \leq j \leq r$. If the perturbation satisfies

$$\|\mathbf{C} - \mathbf{C}^*\|_2 \leq \frac{\tau}{6\sqrt{5}} \cdot \left(\frac{d_j^*}{\sqrt{r}} \right) \cdot \left(\frac{d_j^*}{d_1^*} \right)^{3/2}, \quad (3.18)$$

then the j -th pair of singular vectors has aligned sign:

$$\left(\mathbf{u}_j^\top \mathbf{u}_j^* \right) \cdot \left(\mathbf{v}_j^\top \mathbf{v}_j^* \right) > 0. \quad (3.19)$$

Consequently, if condition (3.18) holds for $j = r$, then the sign alignment (3.19) holds uniformly for all $j = 1, \dots, r$.

Proof. We proceed by contradiction. Suppose

$$\left(\mathbf{u}_j^\top \mathbf{u}_j^* \right) \cdot \left(\mathbf{v}_j^\top \mathbf{v}_j^* \right) \leq 0.$$

Then,

$$\begin{aligned} \mathbf{u}_j^{*\top} (\mathbf{C}^* - \mathbf{C}) \mathbf{v}_j^* &= d_j^* - d_j \left(\mathbf{u}_j^\top \mathbf{u}_j^* \right) \left(\mathbf{v}_j^\top \mathbf{v}_j^* \right) - \sum_{k \neq j} d_k \left(\mathbf{u}_k^\top \mathbf{u}_j^* \right) \left(\mathbf{v}_k^\top \mathbf{v}_j^* \right) \\ &\geq d_j^* - \sum_{k \neq j} d_k \left(\mathbf{u}_k^\top \mathbf{u}_j^* \right) \left(\mathbf{v}_k^\top \mathbf{v}_j^* \right). \end{aligned} \quad (3.20)$$

By Theorem 3 of Yu et al. [2015], there exists a sign $s_j^u \in \{-1, +1\}$ such that

$$\left\| s_j^u \mathbf{u}_j - \mathbf{u}_j^* \right\|_2 \leq \frac{2^{3/2} (2d_1^* + \|\mathbf{C} - \mathbf{C}^*\|_2) \|\mathbf{C} - \mathbf{C}^*\|_2}{\min \left(d_{j-1}^{*2} - d_j^{*2}, d_j^{*2} - d_{j+1}^{*2} \right)},$$

where we define $d_0^{*2} = \infty$ and $d_{r+1}^{*2} = 0$. Using the bound from (3.18) and the eigenvalue

separation condition, we have

$$\left\| s_j^u \mathbf{u}_j - \mathbf{u}_j^* \right\|_2 \leq \frac{3 \cdot 2^{3/2} d_1^* \|\mathbf{C} - \mathbf{C}^*\|_2}{\tau d_j^{*2}}. \quad (3.21)$$

Similarly, there exists $s_j^v \in \{-1, +1\}$ such that

$$\left\| s_j^v \mathbf{v}_j - \mathbf{v}_j^* \right\|_2 \leq \frac{3 \cdot 2^{3/2} d_1^* \|\mathbf{C} - \mathbf{C}^*\|_2}{\tau d_j^{*2}}. \quad (3.22)$$

Using the orthogonality of $\mathbf{u}_k$ and $\mathbf{v}_k$ for $k \neq j$, we rewrite:

$$\begin{aligned} \sum_{k \neq j} d_k \left(\mathbf{u}_k^\top \mathbf{u}_j^* \right) \left(\mathbf{v}_k^\top \mathbf{v}_j^* \right) &= \sum_{k \neq j} d_k \left(\mathbf{u}_k^\top (\mathbf{u}_j^* - s_j^u \mathbf{u}_j) \right) \left(\mathbf{v}_k^\top (\mathbf{v}_j^* - s_j^v \mathbf{v}_j) \right) \\ &\leq \left\| \mathbf{u}_j^* - s_j^u \mathbf{u}_j \right\|_2 \cdot \left\| \mathbf{v}_j^* - s_j^v \mathbf{v}_j \right\|_2 \cdot \sum_{k=1}^r d_k. \end{aligned} \quad (3.23)$$

Combining (3.21), (3.22), and (3.23), we obtain:

$$\sum_{k \neq j} d_k \left(\mathbf{u}_k^\top \mathbf{u}_j^* \right) \left(\mathbf{v}_k^\top \mathbf{v}_j^* \right) \leq \frac{72 d_1^{*2} \|\mathbf{C} - \mathbf{C}^*\|_2^2}{\tau^2 d_j^{*4}} \cdot \sum_{k=1}^r d_k \leq \frac{72 r d_1^{*2} d_1 \|\mathbf{C} - \mathbf{C}^*\|_2^2}{\tau^2 d_j^{*4}}. \quad (3.24)$$

Since (3.18) implies

$$\|\mathbf{C} - \mathbf{C}^*\|_2 \leq \frac{\tau}{6\sqrt{5}} \cdot \left(\frac{d_j^*}{\sqrt{r}} \right) \cdot \left(\frac{d_j^*}{d_1^*} \right)^{3/2} \leq \frac{1}{6\sqrt{5}} d_j^* \leq \frac{1}{4} d_j^*, \quad (3.25)$$

we can bound the top singular value of $\mathbf{C}$:

$$d_1 = \|\mathbf{C}\|_2 \leq \|\mathbf{C}^*\|_2 + \|\mathbf{C} - \mathbf{C}^*\|_2 \leq d_1^* + \frac{1}{4} d_r^* \leq \frac{5}{4} d_1^*. \quad (3.26)$$

Substituting (3.25) and (3.26) into (3.24) yields

$$\sum_{k \neq j} d_k \left(\mathbf{u}_k^\top \mathbf{u}_j^* \right) \left(\mathbf{v}_k^\top \mathbf{v}_j^* \right) \leq \frac{90 r d_1^{*3} \|\mathbf{C} - \mathbf{C}^*\|_2^2}{\tau^2 d_j^{*4}}. \quad (3.27)$$

Furthermore, from (3.18), we know

$$\|\mathbf{C} - \mathbf{C}^*\|_2^2 \leq \frac{\tau^2 d_j^{*5}}{180 r d_1^{*3}}.$$

Plugging this into (3.27) gives

$$\sum_{k \neq j} d_k \left(\mathbf{u}_k^\top \mathbf{u}_j^* \right) \left(\mathbf{v}_k^\top \mathbf{v}_j^* \right) \leq \frac{1}{2} d_j^*.$$

Finally, combining this with (3.20), we obtain

$$\|\mathbf{C} - \mathbf{C}^*\|_2 \geq \mathbf{u}_j^{*\top} (\mathbf{C}^* - \mathbf{C}) \mathbf{v}_j^* \geq \frac{1}{2} d_j^*, \quad (3.28)$$

which contradicts the bound in (3.25). Therefore,

$$\left(\mathbf{u}_j^\top \mathbf{u}_j^* \right) \cdot \left(\mathbf{v}_j^\top \mathbf{v}_j^* \right) > 0.$$

□

The following lemma provides Frobenius norm bounds on the deviations of singular values and singular vectors under perturbation, assuming an eigen-gap condition and sign alignment.

Lemma 32 (Perturbation Bound under Singular Value Separation). *Let $\mathbf{C} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ and*

$\mathbf{C}^* = \mathbf{U}^* \mathbf{D}^* \mathbf{V}^{*\top}$ be two matrices with $\text{rank}(\mathbf{C}^*) = r$. Suppose the following conditions hold:

$$\|\mathbf{C} - \mathbf{C}^*\|_2 \leq d_1^*, \quad \left(\mathbf{u}_j^\top \mathbf{u}_j^* \right) \cdot \left(\mathbf{v}_j^\top \mathbf{v}_j^* \right) \geq 0 \quad \text{for all } 1 \leq j \leq r, \quad (3.29)$$

and the squared singular values of $\mathbf{C}^*$ satisfy the separation condition

$$d_{j-1}^{*2} - d_j^{*2} \geq \tau d_{j-1}^{*2} \quad \text{for all } 2 \leq j \leq r,$$

for some universal constant $0 < \tau < 1$. Then the following bounds hold:

$$\|\mathbf{D} - \mathbf{D}^*\|_F \leq \|\mathbf{C} - \mathbf{C}^*\|_F, \quad (3.30)$$

$$\|\mathbf{U}\mathbf{D} - \mathbf{U}^*\mathbf{D}^*\|_F \leq c\eta_r \|\mathbf{C} - \mathbf{C}^*\|_F, \quad (3.31)$$

$$\|\mathbf{V}\mathbf{D} - \mathbf{V}^*\mathbf{D}^*\|_F \leq c\eta_r \|\mathbf{C} - \mathbf{C}^*\|_F, \quad (3.32)$$

where

$$\eta_r = 1 + \frac{1}{\tau} \left(\sum_{j=1}^r \left(\frac{d_1^*}{d_j^*} \right)^2 \right)^{1/2},$$

and $c > 0$ is a universal constant.

Proof. We prove the result by correcting a technical oversight in the proof of Lemma 3 from the supplementary material of Uematsu et al. [2019].

Note that the pair of singular vectors $(\mathbf{u}_j^*, \mathbf{v}_j^*)$ and $(-\mathbf{u}_j^*, -\mathbf{v}_j^*)$ yield the same matrix $\mathbf{C}^*$, since the signs cancel in the outer product. Therefore, without loss of generality, we may choose the signs of the singular vectors so that

$$\left(\mathbf{u}_j^\top \mathbf{u}_j^* \right) \geq 0 \quad \text{and} \quad \left(\mathbf{v}_j^\top \mathbf{v}_j^* \right) \geq 0.$$

This choice is justified under the condition (3.29). With this sign alignment, the remain-

der of the proof follows directly from Lemma 3 in the supplementary material of Uematsu et al. [2019]. $\square$

Next, we present a lemma that provides an upper bound on the source error when the source matrix $\mathbf{C}^{(0)}$ is observed with additive noise. Recall that we observe a perturbed version of the source matrix given by

$$\tilde{\mathbf{C}}^{(0)} = \mathbf{C}^{(0)} + \mathbf{E}^{(0)},$$

where $\mathbf{E}^{(0)}$ denotes the noise matrix. Let the singular value decomposition (SVD) of $\tilde{\mathbf{C}}^{(0)}$ be

$$\tilde{\mathbf{C}}^{(0)} = \tilde{\mathbf{U}}^{(0)} \tilde{\mathbf{D}}^{(0)} \tilde{\mathbf{V}}^{(0)\top} = \sum_{j=1}^{\tilde{r}_0} \tilde{d}_j^{(0)} \tilde{\mathbf{u}}_j^{(0)} \tilde{\mathbf{v}}_j^{(0)\top}.$$

We extend $\tilde{\mathbf{U}}^{(0)}$ and $\tilde{\mathbf{V}}^{(0)}$ to full orthonormal bases of $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively, by appending the vectors $\tilde{\mathbf{u}}_{\tilde{r}_0+1}^{(0)}, \dots, \tilde{\mathbf{u}}_p^{(0)}$ and $\tilde{\mathbf{v}}_{\tilde{r}_0+1}^{(0)}, \dots, \tilde{\mathbf{v}}_q^{(0)}$.

For any integers $0 \leq r_u \leq p$ and $0 \leq r_v \leq q$, we define the truncated and orthogonal complement subspaces of the left and right singular vectors of $\tilde{\mathbf{C}}^{(0)}$ as

$$\tilde{\mathbf{U}}^s(r_u) := [\tilde{\mathbf{u}}_1^{(0)}, \dots, \tilde{\mathbf{u}}_{r_u}^{(0)}], \quad \tilde{\mathbf{U}}_{\perp}^s(r_u) := [\tilde{\mathbf{u}}_{r_u+1}^{(0)}, \dots, \tilde{\mathbf{u}}_p^{(0)}],$$

and analogously,

$$\tilde{\mathbf{V}}^s(r_v) := [\tilde{\mathbf{v}}_1^{(0)}, \dots, \tilde{\mathbf{v}}_{r_v}^{(0)}], \quad \tilde{\mathbf{V}}_{\perp}^s(r_v) := [\tilde{\mathbf{v}}_{r_v+1}^{(0)}, \dots, \tilde{\mathbf{v}}_q^{(0)}].$$

We also define the full source-augmented orthogonal basis matrices:

$$\tilde{\mathbf{U}}_{\text{full}} := \begin{bmatrix} \tilde{\mathbf{U}}^s & \tilde{\mathbf{U}}_{\perp}^s \end{bmatrix} \in \mathbb{R}^{p \times p}, \quad \tilde{\mathbf{V}}_{\text{full}} := \begin{bmatrix} \tilde{\mathbf{V}}^s & \tilde{\mathbf{V}}_{\perp}^s \end{bmatrix} \in \mathbb{R}^{q \times q}.$$

Let $\mathcal{S}_u$, $\mathcal{S}_u^c$, $\mathcal{S}_v$, and $\mathcal{S}_v^c$ be defined as in (3.61) and (3.62), and recall the quantities $s_u(r_u)$

and $s_v(r_v)$ introduced in (3.6). Additionally, let $\varepsilon^{(0)}$ denote the spectral error of the source matrix, defined as

$$\varepsilon^{(0)} := \left\| \mathbf{E}^{(0)} \right\|_2.$$

Throughout, we also adopt the standard notation $x_+ := \max(x, 0)$ for any real number $x \in \mathbb{R}$.

We are now ready to state the following lemma, which quantifies the approximation error incurred by using the noisy singular vectors of the source matrix.

Lemma 33 (Approximation Error from Noisy Source Singular Vectors). *Suppose Condition (3.8) in Assumption 7 holds. Then, for all $0 \leq r_u \leq p$, we have*

$$\begin{aligned} \left| \left[\tilde{\mathbf{U}}_{\perp}^{\text{s}\top}(r_u) \mathbf{U}^* \right]_{\mathcal{S}_u^c} \right|_1 &\leq \frac{2^{3/2} [r(r_u^* - r_u)_+ - s_u(r_u)]}{\tau(d_{r_u^*}^{(0)})^2} (2d_1^{(0)} + \varepsilon^{(0)}) \varepsilon^{(0)} \\ &\quad + \frac{2^{5/4} r r_u^{*1/4} \sqrt{p - \max(r_u^*, r_u)}}{d_{r_u^*}^{(0)}} (\varepsilon^{(0)})^{1/2} (\varepsilon^{(0)} + 2d_1^{(0)})^{1/2}. \end{aligned}$$

Similarly, for all $0 \leq r_v \leq q$, we have

$$\begin{aligned} \left| \left[\tilde{\mathbf{V}}_{\perp}^{\text{s}\top}(r_v) \mathbf{V}^* \right]_{\mathcal{S}_v^c} \right|_1 &\leq \frac{2^{3/2} [r(r_v^* - r_v)_+ - s_v(r_v)]}{\tau(d_{r_v^*}^{(0)})^2} (2d_1^{(0)} + \varepsilon^{(0)}) \varepsilon^{(0)} \\ &\quad + \frac{2^{5/4} r r_v^{*1/4} \sqrt{q - \max(r_v^*, r_v)}}{d_{r_v^*}^{(0)}} (\varepsilon^{(0)})^{1/2} (\varepsilon^{(0)} + 2d_1^{(0)})^{1/2}. \end{aligned}$$

Proof. To streamline the notation, we abbreviate $\mathbf{U}^{\text{s}}(r_u)$ and $\mathbf{U}_{\perp}^{\text{s}}(r_u)$ as $\mathbf{U}^{\text{s}}$ and $\mathbf{U}_{\perp}^{\text{s}}$, respectively, and similarly write $\mathbf{V}^{\text{s}}(r_v)$ and $\mathbf{V}_{\perp}^{\text{s}}(r_v)$ as $\mathbf{V}^{\text{s}}$ and $\mathbf{V}_{\perp}^{\text{s}}$. The notations $\tilde{\mathbf{U}}^{\text{s}}$, $\tilde{\mathbf{U}}_{\perp}^{\text{s}}$, $\tilde{\mathbf{V}}^{\text{s}}$, and $\tilde{\mathbf{V}}_{\perp}^{\text{s}}$ are defined analogously. In addition, we simplify $s_u(r_u)$ and $s_v(r_v)$ to s_u and s_v for the remainder of the proof.

Recall the definition of r_u^* in (3.5). We first consider the case $r_u \leq r_u^*$. Define the

intermediate and residual components of the singular vector basis as

$$\mathbf{U}_{\text{gap}} = [\mathbf{u}_{r_u+1}^{(0)}, \dots, \mathbf{u}_{r_u^*}^{(0)}], \quad \mathbf{U}_{\text{tail}} = [\mathbf{u}_{r_u^*+1}^{(0)}, \dots, \mathbf{u}_p^{(0)}],$$

and analogously,

$$\tilde{\mathbf{U}}_{\text{gap}} = [\tilde{\mathbf{u}}_{r_u+1}^{(0)}, \dots, \tilde{\mathbf{u}}_{r_u^*}^{(0)}], \quad \tilde{\mathbf{U}}_{\text{tail}} = [\tilde{\mathbf{u}}_{r_u^*+1}^{(0)}, \dots, \tilde{\mathbf{u}}_p^{(0)}].$$

In the degenerate case $r_u = r_u^*$, both $\mathbf{U}_{\text{gap}}$ and $\tilde{\mathbf{U}}_{\text{gap}}$ are interpreted as $p \times 0$ empty matrices.

It then follows immediately that

$$\mathbf{U}_{\perp}^{\text{s}} = [\mathbf{U}_{\text{gap}} \ \mathbf{U}_{\text{tail}}], \quad \tilde{\mathbf{U}}_{\perp}^{\text{s}} = [\tilde{\mathbf{U}}_{\text{gap}} \ \tilde{\mathbf{U}}_{\text{tail}}].$$

Next, we define the support sets:

$$\begin{aligned} \mathcal{M}_u &:= \left\{ (i, j) \in [r] \times [r_u^* - r_u] : [\mathbf{U}^{*\top} \mathbf{U}_{\text{gap}}]_{ij} \neq 0 \right\}, \\ \mathcal{M}_u^c &:= \left\{ (i, j) \in [r] \times [r_u^* - r_u] : [\mathbf{U}^{*\top} \mathbf{U}_{\text{gap}}]_{ij} = 0 \right\}. \end{aligned}$$

Since $\mathbf{U}_{\text{tail}}^{\top} \mathbf{U}^* = 0$, the definition of $\mathcal{S}_u$ in (3.61) implies that

$$\left| [\tilde{\mathbf{U}}_{\perp}^{\text{s}\top} \mathbf{U}^*]_{\mathcal{S}_u^c} \right|_1 = \left| [\mathbf{U}^{*\top} \tilde{\mathbf{U}}_{\text{gap}}]_{\mathcal{M}_u^c} \right|_1 + \left| \mathbf{U}^{*\top} \tilde{\mathbf{U}}_{\text{tail}} \right|_1. \quad (3.33)$$

Therefore, it suffices to derive upper bounds for the two terms on the right-hand side of (3.33).

We begin by bounding the term $\left| [\mathbf{U}^{*\top} \tilde{\mathbf{U}}_{\text{gap}}]_{\mathcal{M}_u^c} \right|_1$. For any $(i, j) \in \mathcal{M}_u^c$, we have $\langle \mathbf{u}_i^*, \mathbf{u}_j^{(0)} \rangle = 0$ by construction. Hence, for any such pair and any sign $s_j^u \in \{-1, +1\}$, it

follows that

$$\begin{aligned}
\left| \left[\mathbf{U}^{*\top} \tilde{\mathbf{U}}_{\text{gap}} \right]_{ij} \right| &= \left| \left\langle \mathbf{u}_i^*, \tilde{\mathbf{u}}_j^{(0)} \right\rangle \right| = \left| \left\langle \mathbf{u}_i^*, s_j^u \tilde{\mathbf{u}}_j^{(0)} \right\rangle \right| \\
&= \left| \left\langle \mathbf{u}_i^*, s_j^u \tilde{\mathbf{u}}_j^{(0)} - \mathbf{u}_j^{(0)} \right\rangle \right| \leq \left\| s_j^u \tilde{\mathbf{u}}_j^{(0)} - \mathbf{u}_j^{(0)} \right\|_2.
\end{aligned} \tag{3.34}$$

By Theorem 3 of Yu et al. [2015], there exists a sign $\tilde{s}_j^u \in \{-1, +1\}$ such that

$$\left\| \tilde{s}_j^u \tilde{\mathbf{u}}_j^{(0)} - \mathbf{u}_j^{(0)} \right\|_2 \leq \frac{2^{3/2} \left(2d_1^{(0)} + \left\| \mathbf{E}^{(0)} \right\|_2 \right) \left\| \mathbf{E}^{(0)} \right\|_2}{\min \left\{ \left(d_{j-1}^{(0)} \right)^2 - \left(d_j^{(0)} \right)^2, \left(d_j^{(0)} \right)^2 - \left(d_{j+1}^{(0)} \right)^2 \right\}}.$$

Under Condition (3.8) in Assumption 7, the spectral gap is bounded below by $\tau \left(d_j^{(0)} \right)^2$, so the above reduces to

$$\left\| \tilde{s}_j^u \tilde{\mathbf{u}}_j^{(0)} - \mathbf{u}_j^{(0)} \right\|_2 \leq \frac{2^{3/2}}{\tau \left(d_j^{(0)} \right)^2} \left(2d_1^{(0)} + \left\| \mathbf{E}^{(0)} \right\|_2 \right) \left\| \mathbf{E}^{(0)} \right\|_2. \tag{3.35}$$

Combining inequalities (3.34) and (3.35) yields the elementwise bound

$$\left| \left[\mathbf{U}^{*\top} \tilde{\mathbf{U}}_{\text{gap}} \right]_{ij} \right| \leq \frac{2^{3/2}}{\tau \left(d_j^{(0)} \right)^2} \left(2d_1^{(0)} + \left\| \mathbf{E}^{(0)} \right\|_2 \right) \left\| \mathbf{E}^{(0)} \right\|_2.$$

Since $d_j^{(0)} \geq d_{r_u^*}^{(0)}$ for all $j \in [r_u + 1, r_u^*]$, we may uniformly bound the maximum as

$$\max_{(i,j) \in \mathcal{M}_u^c} \left| \left[\mathbf{U}^{*\top} \tilde{\mathbf{U}}_{\text{gap}} \right]_{ij} \right| \leq \frac{2^{3/2}}{\tau \left(d_{r_u^*}^{(0)} \right)^2} \left(2d_1^{(0)} + \left\| \mathbf{E}^{(0)} \right\|_2 \right) \left\| \mathbf{E}^{(0)} \right\|_2.$$

Finally, observe that $|\mathcal{M}_u^c| = r(r_u^* - r_u) - s_u$, since $|\mathcal{M}_u| = s_u$ by the definition of $\mathcal{S}_u$.

Therefore,

$$\begin{aligned}
\left| \left[\mathbf{U}^{*\top} \tilde{\mathbf{U}}_{\text{gap}} \right]_{\mathcal{M}_u^c} \right|_1 &\leq |\mathcal{M}_u^c| \cdot \max_{(i,j) \in \mathcal{M}_u^c} \left| \left[\mathbf{U}^{*\top} \tilde{\mathbf{U}}_{\text{gap}} \right]_{ij} \right| \\
&\leq \frac{2^{3/2} (r(r_u^* - r_u) - s_u)}{\tau (d_{r_u^*}^{(0)})^2} \left(2d_1^{(0)} + \left\| \mathbf{E}^{(0)} \right\|_2 \right) \left\| \mathbf{E}^{(0)} \right\|_2.
\end{aligned} \tag{3.36}$$

We next bound the second term $\left| \mathbf{U}^{*\top} \tilde{\mathbf{U}}_{\text{tail}} \right|_1$ in (3.33). Define the matrices

$$\tilde{\mathbf{U}}_{\text{oracle}} := \begin{bmatrix} \tilde{\mathbf{U}}^s & \tilde{\mathbf{U}}_{\text{gap}} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{u}}_1^{(0)}, \dots, \tilde{\mathbf{u}}_{r_u^*}^{(0)} \end{bmatrix}, \quad \mathbf{U}_{\text{oracle}} := \begin{bmatrix} \mathbf{U}^s & \mathbf{U}_{\text{gap}} \end{bmatrix} = \begin{bmatrix} \mathbf{u}_1^{(0)}, \dots, \mathbf{u}_{r_u^*}^{(0)} \end{bmatrix}.$$

By orthogonality of the singular vectors, we have the identity

$$\left\| \mathbf{U}^{*\top} \tilde{\mathbf{U}}_{\text{tail}} \right\|_{\text{F}}^2 + \left\| \mathbf{U}^{*\top} \tilde{\mathbf{U}}_{\text{oracle}} \right\|_{\text{F}}^2 = \left\| \mathbf{U}^* \right\|_{\text{F}}^2 = r. \tag{3.37}$$

It follows that

$$\begin{aligned}
\left| \mathbf{U}^{*\top} \tilde{\mathbf{U}}_{\text{tail}} \right|_1^2 &\leq r(p - r_u^*) \left\| \mathbf{U}^{*\top} \tilde{\mathbf{U}}_{\text{tail}} \right\|_{\text{F}}^2 \\
&= r(p - r_u^*) \left\{ r - \left\| \mathbf{U}^{*\top} \tilde{\mathbf{U}}_{\text{oracle}} \right\|_{\text{F}}^2 \right\} \\
&= r(p - r_u^*) \left(\sqrt{r} + \left\| \mathbf{U}^{*\top} \tilde{\mathbf{U}}_{\text{oracle}} \right\|_{\text{F}} \right) \left(\sqrt{r} - \left\| \mathbf{U}^{*\top} \tilde{\mathbf{U}}_{\text{oracle}} \right\|_{\text{F}} \right),
\end{aligned}$$

which implies

$$\left| \mathbf{U}^{*\top} \tilde{\mathbf{U}}_{\text{tail}} \right|_1^2 \leq 2r^{3/2}(p - r_u^*) \left(\sqrt{r} - \left\| \mathbf{U}^{*\top} \tilde{\mathbf{U}}_{\text{oracle}} \right\|_{\text{F}} \right). \tag{3.38}$$

To control the residual term, observe that for any orthonormal matrix $\mathbf{O}_{\text{oracle}} \in \mathbb{R}^{r_u^* \times r_u^*}$,

$$\begin{aligned} \left\| \mathbf{U}^{*\top} \tilde{\mathbf{U}}_{\text{oracle}} \right\|_{\text{F}} &= \left\| \mathbf{U}^{*\top} \tilde{\mathbf{U}}_{\text{oracle}} \mathbf{O}_{\text{oracle}} \right\|_{\text{F}} \\ &= \left\| \mathbf{U}^{*\top} \left(\tilde{\mathbf{U}}_{\text{oracle}} \mathbf{O}_{\text{oracle}} - \mathbf{U}_{\text{oracle}} \right) + \mathbf{U}^{*\top} \mathbf{U}_{\text{oracle}} \right\|_{\text{F}} \\ &\geq \left\| \mathbf{U}^{*\top} \mathbf{U}_{\text{oracle}} \right\|_{\text{F}} - \left\| \mathbf{U}^{*\top} \left(\tilde{\mathbf{U}}_{\text{oracle}} \mathbf{O}_{\text{oracle}} - \mathbf{U}_{\text{oracle}} \right) \right\|_{\text{F}}. \end{aligned}$$

By Assumption 3, we have $\mathbf{U}^{*\top} \mathbf{U}_{\text{tail}} = 0$, which implies $\left\| \mathbf{U}^{*\top} \mathbf{U}_{\text{oracle}} \right\|_{\text{F}} = \sqrt{r}$. Therefore,

$$\begin{aligned} \sqrt{r} - \left\| \mathbf{U}^{*\top} \tilde{\mathbf{U}}_{\text{oracle}} \right\|_{\text{F}} &\leq \left\| \mathbf{U}^* \right\|_{\text{F}} \left\| \tilde{\mathbf{U}}_{\text{oracle}} \mathbf{O}_{\text{oracle}} - \mathbf{U}_{\text{oracle}} \right\|_{\text{F}} \\ &= \sqrt{r} \left\| \tilde{\mathbf{U}}_{\text{oracle}} \mathbf{O}_{\text{oracle}} - \mathbf{U}_{\text{oracle}} \right\|_{\text{F}}. \end{aligned} \tag{3.39}$$

Substituting this into (3.38), we obtain

$$\left| \mathbf{U}^{*\top} \tilde{\mathbf{U}}_{\text{tail}} \right|_1^2 \leq 2r^2(p - r_u^*) \left\| \tilde{\mathbf{U}}_{\text{oracle}} \mathbf{O}_{\text{oracle}} - \mathbf{U}_{\text{oracle}} \right\|_{\text{F}}$$

for all orthonormal $\mathbf{O}_{\text{oracle}} \in \mathbb{R}^{r_u^* \times r_u^*}$. Finally, by Theorem 3 of Yu et al. [2015], there exists an orthonormal matrix $\tilde{\mathbf{O}}_{\text{oracle}}$ such that

$$\left\| \tilde{\mathbf{U}}_{\text{oracle}} \tilde{\mathbf{O}}_{\text{oracle}} - \mathbf{U}_{\text{oracle}} \right\|_{\text{F}} \leq \frac{2\sqrt{2r_u^*}}{\left(d_{r_u^*}^{(0)}\right)^2} \left\| \mathbf{E}^{(0)} \right\|_2 \left(\left\| \mathbf{E}^{(0)} \right\|_2 + 2d_1^{(0)} \right). \tag{3.40}$$

Plugging this into the previous inequality yields

$$\left| \mathbf{U}^{*\top} \tilde{\mathbf{U}}_{\text{tail}} \right|_1 \leq \frac{2^{5/4} r \sqrt{p - r_u^*} r_u^{*1/4}}{d_{r_u^*}^{(0)}} \left\| \mathbf{E}^{(0)} \right\|_2^{1/2} \left(\left\| \mathbf{E}^{(0)} \right\|_2 + 2d_1^{(0)} \right)^{1/2}. \tag{3.41}$$

Combining the decomposition in (3.33), the bound for the gap component in (3.36), and

the bound for the tail component in (3.41), we conclude that

$$\begin{aligned} \left| \left[\tilde{\mathbf{U}}_{\perp}^{\text{s}\top} \mathbf{U}^* \right]_{\mathcal{S}_u^c} \right|_1 &\leq \frac{2^{3/2} (r(r_u^* - r_u) - s_u)}{\tau \left(d_{r_u^*}^{(0)} \right)^2} \left(2d_1^{(0)} + \left\| \mathbf{E}^{(0)} \right\|_2 \right) \left\| \mathbf{E}^{(0)} \right\|_2 \\ &\quad + \frac{2^{5/4} r r_u^{*1/4} \sqrt{p - r_u^*}}{d_{r_u^*}^{(0)}} \left\| \mathbf{E}^{(0)} \right\|_2^{1/2} \left(\left\| \mathbf{E}^{(0)} \right\|_2 + 2d_1^{(0)} \right)^{1/2}. \end{aligned} \quad (3.42)$$

We now consider the case $r_u > r_u^*$. In this setting, define

$$\tilde{\mathbf{U}}_{\text{tail}} = \left[\tilde{\mathbf{u}}_{r_u+1}^{(0)}, \dots, \tilde{\mathbf{u}}_p^{(0)} \right], \quad \mathbf{U}_{\text{tail}} = \left[\mathbf{u}_{r_u+1}^{(0)}, \dots, \mathbf{u}_p^{(0)} \right].$$

By definition, we have $\tilde{\mathbf{U}}_{\perp}^{\text{s}} = \tilde{\mathbf{U}}_{\text{tail}}$. Moreover, due to the structure of the support set $\mathcal{S}_u$, which is empty now, we have $\mathcal{S}_u^c$ containing all the rows associated with $\tilde{\mathbf{U}}_{\text{tail}}$. We thus have

$$\left| \left[\tilde{\mathbf{U}}_{\perp}^{\text{s}\top} \mathbf{U}^* \right]_{\mathcal{S}_u^c} \right|_1 = \left| \mathbf{U}^{*\top} \tilde{\mathbf{U}}_{\text{tail}} \right|_1. \quad (3.43)$$

This equality critically depends on the definition of $\mathcal{S}_u^c$, which ensures that only indices corresponding to the tail subspace are included. Applying Hölder's inequality, we further obtain

$$\left| \mathbf{U}^{*\top} \tilde{\mathbf{U}}_{\text{tail}} \right|_1 \leq \sqrt{r(p - r_u)} \left\| \mathbf{U}^{*\top} \tilde{\mathbf{U}}_{\text{tail}} \right\|_{\text{F}}.$$

To facilitate analysis, we define the intermediate matrices:

$$\tilde{\mathbf{U}}_{\text{gap}} := \left[\tilde{\mathbf{u}}_{r_u^*+1}^{(0)}, \dots, \tilde{\mathbf{u}}_{r_u}^{(0)} \right], \quad \mathbf{U}_{\text{gap}} := \left[\mathbf{u}_{r_u^*+1}^{(0)}, \dots, \mathbf{u}_{r_u}^{(0)} \right],$$

and reuse the oracle basis components:

$$\tilde{\mathbf{U}}_{\text{oracle}} := \left[\tilde{\mathbf{u}}_1^{(0)}, \dots, \tilde{\mathbf{u}}_{r_u^*}^{(0)} \right], \quad \mathbf{U}_{\text{oracle}} := \left[\mathbf{u}_1^{(0)}, \dots, \mathbf{u}_{r_u^*}^{(0)} \right].$$

Since these form an orthonormal basis of $\mathbb{R}^p$, we have the decomposition

$$\left\| \mathbf{U}^{*\top} \tilde{\mathbf{U}}_{\text{oracle}} \right\|_{\text{F}}^2 + \left\| \mathbf{U}^{*\top} \tilde{\mathbf{U}}_{\text{gap}} \right\|_{\text{F}}^2 + \left\| \mathbf{U}^{*\top} \tilde{\mathbf{U}}_{\text{tail}} \right\|_{\text{F}}^2 = \left\| \mathbf{U}^* \right\|_{\text{F}}^2 = r.$$

Therefore,

$$\begin{aligned} \left\| \mathbf{U}^{*\top} \tilde{\mathbf{U}}_{\text{tail}} \right\|_{\text{F}}^2 &\leq r - \left\| \mathbf{U}^{*\top} \tilde{\mathbf{U}}_{\text{oracle}} \right\|_{\text{F}}^2 \\ &= \left(\sqrt{r} + \left\| \mathbf{U}^{*\top} \tilde{\mathbf{U}}_{\text{oracle}} \right\|_{\text{F}} \right) \left(\sqrt{r} - \left\| \mathbf{U}^{*\top} \tilde{\mathbf{U}}_{\text{oracle}} \right\|_{\text{F}} \right) \\ &\leq 2\sqrt{r} \left(\sqrt{r} - \left\| \mathbf{U}^{*\top} \tilde{\mathbf{U}}_{\text{oracle}} \right\|_{\text{F}} \right). \end{aligned} \quad (3.44)$$

By the bounds in (3.39) and (3.40), we have

$$\sqrt{r} - \left\| \mathbf{U}^{*\top} \tilde{\mathbf{U}}_{\text{oracle}} \right\|_{\text{F}} \leq \frac{2\sqrt{2rr_u^*}}{\left(d_{r_u^*}^{(0)}\right)^2} \left\| \mathbf{E}^{(0)} \right\|_2 \left(\left\| \mathbf{E}^{(0)} \right\|_2 + 2d_1^{(0)} \right).$$

Substituting this into (3.43) and (3.44), we obtain

$$\left| \left[\tilde{\mathbf{U}}_{\perp}^{\text{s}\top} \mathbf{U}^* \right]_{\mathcal{S}_u^c} \right|_1 \leq \frac{2^{5/4} r r_u^{*1/4} \sqrt{p - r_u}}{d_{r_u^*}^{(0)}} \left\| \mathbf{E}^{(0)} \right\|_2^{1/2} \left(\left\| \mathbf{E}^{(0)} \right\|_2 + 2d_1^{(0)} \right)^{1/2}. \quad (3.45)$$

Combining the bounds derived in (3.42) and (3.45), we obtain a unified inequality that holds for all $0 \leq r_u \leq p$:

$$\begin{aligned} \left| \left[\tilde{\mathbf{U}}_{\perp}^{\text{s}\top} \mathbf{U}^* \right]_{\mathcal{S}_u^c} \right|_1 &\leq \frac{2^{3/2} (r(r_u^* - r_u)_+ - s_u)}{\tau \left(d_{r_u^*}^{(0)}\right)^2} \left(2d_1^{(0)} + \left\| \mathbf{E}^{(0)} \right\|_2 \right) \left\| \mathbf{E}^{(0)} \right\|_2 \\ &\quad + \frac{2^{5/4} r r_u^{*1/4} \sqrt{p - \max(r_u, r_u^*)}}{d_{r_u^*}^{(0)}} \left\| \mathbf{E}^{(0)} \right\|_2^{1/2} \left(\left\| \mathbf{E}^{(0)} \right\|_2 + 2d_1^{(0)} \right)^{1/2}, \end{aligned}$$

where $x_+ = \max(x, 0)$ denotes the positive part.

A symmetric argument yields the analogous bound for the right singular vectors:

$$\begin{aligned} \left| \left[\tilde{\mathbf{V}}_{\perp}^{\top} \mathbf{V}^* \right]_{\mathcal{S}_v^c} \right|_1 &\leq \frac{2^{3/2} (r(r_v^* - r_v)_+ - s_v)}{\tau \left(d_{r_v^*}^{(0)} \right)^2} \left(2d_1^{(0)} + \left\| \mathbf{E}^{(0)} \right\|_2 \right) \left\| \mathbf{E}^{(0)} \right\|_2 \\ &\quad + \frac{2^{5/4} r r_v^{*1/4} \sqrt{p - \max(r_v, r_v^*)}}{d_{r_v^*}^{(0)}} \left\| \mathbf{E}^{(0)} \right\|_2^{1/2} \left(\left\| \mathbf{E}^{(0)} \right\|_2 + 2d_1^{(0)} \right)^{1/2}, \end{aligned}$$

for all $0 \leq r_v \leq p$.

□

The following lemma identifies a collection of desirable “good” events and establishes that they occur with high probability. In subsequent analysis, we condition on these events to derive the main theoretical results.

Lemma 34 (High-Probability Good Events). *Suppose Assumptions 5–8, Conditions (3.9)–(3.11), and Condition (3.13) hold. Fix any $0 < \delta < 1$, and define the regularization parameters λ_u and λ_v as in (3.14). Then, with probability at least $1 - \delta$, the following inequalities hold simultaneously:*

$$\left| \frac{1}{n} \tilde{\mathbf{U}}_{\text{full}}^{\top} \mathbf{X}^{\top} \mathbf{E} \hat{\mathbf{B}} \hat{\mathbf{D}} \right|_{\infty} \leq \lambda_u/2, \quad (3.46)$$

$$\left| \frac{1}{n} (\mathbf{X} \mathbf{U}^*)^{\top} \mathbf{E} \hat{\mathbf{B}} \hat{\mathbf{D}} \right|_{\infty} \leq \lambda_u/2, \quad (3.47)$$

$$\left| \frac{1}{n} (\mathbf{X} \mathbf{U}^*)^{\top} \mathbf{E} \tilde{\mathbf{V}}_{\text{full}} \right|_{\infty} \leq \lambda_v/2. \quad (3.48)$$

Proof. It suffices to show that each of the inequalities (3.46), (3.47), and (3.48) holds with probability at least $1 - \delta/3$. The final result then follows by the Bonferroni inequality.

Define the shorthand notations:

$$\tilde{\mathbf{X}} := \mathbf{X} \tilde{\mathbf{U}}_{\text{full}}, \quad \tilde{\mathbf{E}} := \mathbf{E} \tilde{\mathbf{V}}_{\text{full}}, \quad \mathbf{X}_* := \mathbf{X} \mathbf{U}^*, \quad \mathbf{E}_* := \mathbf{E} \mathbf{V}^*.$$

Let $\tilde{\mathbf{x}}^{(j)}$ and $\tilde{\mathbf{e}}^{(j)}$ denote the j -th columns of $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{E}}$, respectively. Similarly, let $\mathbf{x}_*^{(j)}$ and $\mathbf{e}_*^{(j)}$ denote the j -th columns of $\mathbf{X}_*$ and $\mathbf{E}_*$.

Under Condition (3.9) and the constraints of (3.10), we have the bound

$$\|\hat{\Delta}\|_{\text{F}} \leq 3R_{2,n}. \quad (3.49)$$

Then, by the sample size condition (3.11), it follows that

$$\|\hat{\Delta}\|_2 \leq \|\hat{\Delta}\|_{\text{F}} \leq 3R_{2,n} \leq \frac{\tau}{6\sqrt{5}} \cdot \left(\frac{d_r^*}{\sqrt{r}}\right) \cdot \left(\frac{d_r^*}{d_1^*}\right)^{3/2}. \quad (3.50)$$

Applying Lemma 31, Lemma 32, and Assumption 7, we further obtain the following bounds:

$$\|\hat{\Delta}^a\|_{\text{F}} \leq c\eta_r \|\hat{\Delta}\|_{\text{F}}, \quad \|\hat{\Delta}^b\|_{\text{F}} \leq c\eta_r \|\hat{\Delta}\|_{\text{F}}, \quad \|\hat{\Delta}^d\|_{\text{F}} \leq \|\hat{\Delta}\|_{\text{F}}, \quad (3.51)$$

where η_r is defined as

$$\eta_r = 1 + \frac{1}{\tau} \left(\sum_{j=1}^r \left(\frac{d_1^*}{d_j^*} \right)^2 \right)^{1/2},$$

with $\tau \in (0, 1)$ and $c > 0$ denoting universal constants.

We begin by establishing that inequality (3.46) holds with probability at least $1 - \delta/3$.

Observe that

$$\hat{\mathbf{B}}\hat{\mathbf{D}}^- = (\hat{\mathbf{B}} - \mathbf{B}^* + \mathbf{B}^*)\hat{\mathbf{D}}^- = \mathbf{B}^*\hat{\mathbf{D}}^- + \hat{\Delta}^b\hat{\mathbf{D}}^-.$$

It then follows that

$$\left| \frac{1}{n} \tilde{\mathbf{U}}_{\text{full}}^\top \mathbf{X}^\top \mathbf{E} \hat{\mathbf{B}} \hat{\mathbf{D}}^- \right|_{\infty} = \left| \frac{1}{n} \tilde{\mathbf{X}}^\top \mathbf{E} \hat{\mathbf{B}} \hat{\mathbf{D}}^- \right|_{\infty} \leq \left| \frac{1}{n} \tilde{\mathbf{X}}^\top \mathbf{E} \mathbf{B}^* \hat{\mathbf{D}}^- \right|_{\infty} + \left| \frac{1}{n} \tilde{\mathbf{X}}^\top \mathbf{E} \hat{\Delta}^b \hat{\mathbf{D}}^- \right|_{\infty}.$$

Therefore, by the union bound, to ensure that (3.46) holds with probability at least $1 - \delta/3$, it suffices to show that the following two inequalities hold with probability at least $1 - \delta/6$

each:

$$\left| \frac{1}{n} \tilde{\mathbf{X}}^\top \mathbf{E} \mathbf{B}^* \hat{\mathbf{D}}^- \right|_\infty \leq \lambda_u/4, \quad (3.52)$$

$$\left| \frac{1}{n} \tilde{\mathbf{X}}^\top \mathbf{E} \hat{\Delta}^b \hat{\mathbf{D}}^- \right|_\infty \leq \lambda_u/4. \quad (3.53)$$

We now show that inequality (3.52) holds with probability at least $1 - \delta/6$.

By the bound (3.50) and the bound (3.51), we have

$$\left| \hat{d}_j - d_j^* \right| \leq \left\| \hat{\Delta}^d \right\|_F \leq \left\| \hat{\Delta} \right\|_F \leq \frac{1}{2} d_j^* \quad \text{for all } 1 \leq j \leq r.$$

This implies that

$$\hat{d}_j \geq d_j^* - \left| \hat{d}_j - d_j^* \right| \geq \frac{1}{2} d_j^*,$$

and hence,

$$\hat{\mathbf{D}}_{jj}^- = \hat{d}_j^{-1} \leq 2d_j^{*-1} \quad \text{and} \quad \left(\mathbf{D}^* \hat{\mathbf{D}}^- \right)_{jj} = d_j^* \cdot \hat{d}_j^{-1} \leq 2 \quad \text{for all } 1 \leq j \leq r. \quad (3.54)$$

Therefore, we can bound the original term as follows:

$$\begin{aligned} \left| \frac{1}{n} \tilde{\mathbf{X}}^\top \mathbf{E} \mathbf{B}^* \hat{\mathbf{D}}^- \right|_\infty &= \left| \frac{1}{n} \tilde{\mathbf{X}}^\top \mathbf{E} \mathbf{V}^* \mathbf{D}^* \hat{\mathbf{D}}^- \right|_\infty \\ &= \left| \frac{1}{n} \tilde{\mathbf{X}}^\top \mathbf{E}_* \mathbf{D}^* \hat{\mathbf{D}}^- \right|_\infty \\ &\leq \frac{1}{n} \left| \tilde{\mathbf{X}}^\top \mathbf{E}_* \right|_\infty \cdot \left| \mathbf{D}^* \hat{\mathbf{D}}^- \right|_\infty \\ &\leq \frac{2}{n} \left| \tilde{\mathbf{X}}^\top \mathbf{E}_* \right|_\infty, \end{aligned}$$

where the final bound follows from (3.54).

Note that

$$\left| \tilde{\mathbf{X}}^\top \mathbf{E}_* \right|_\infty = \max_{1 \leq i \leq p} \max_{1 \leq j \leq r} \left| \tilde{\mathbf{x}}^{(i)\top} \mathbf{e}_*^{(j)} \right|.$$

Therefore, by the union bound (Bonferroni's inequality), we have

$$\mathbb{P} \left\{ \left| \frac{1}{n} \tilde{\mathbf{X}}^\top \mathbf{E} \mathbf{B}^* \hat{\mathbf{D}}^- \right|_\infty > \frac{\lambda_u}{4} \right\} \leq \sum_{i=1}^p \sum_{j=1}^r \mathbb{P} \left\{ \left| \tilde{\mathbf{x}}^{(i)\top} \mathbf{e}_*^{(j)} \right| > \frac{n\lambda_u}{8} \right\}. \quad (3.55)$$

Since

$$\mathbf{e}_*^{(j)} = \mathbf{E} \mathbf{v}_j^* \sim \mathcal{N} \left(0, \left(\mathbf{v}_j^{*\top} \boldsymbol{\Sigma} \mathbf{v}_j^* \right) \mathbf{I}_n \right),$$

it follows that each product $\tilde{\mathbf{x}}^{(i)\top} \mathbf{e}_*^{(j)}$ is Gaussian with zero mean and variance

$$\begin{aligned} \mathbb{E} \left[\left(\tilde{\mathbf{x}}^{(i)\top} \mathbf{e}_*^{(j)} \right)^2 \right] &= \tilde{\mathbf{x}}^{(i)\top} \mathbb{E} \left[\mathbf{e}_*^{(j)} \mathbf{e}_*^{(j)\top} \right] \tilde{\mathbf{x}}^{(i)} \\ &= \left(\mathbf{v}_j^{*\top} \boldsymbol{\Sigma} \mathbf{v}_j^* \right) \cdot \left\| \tilde{\mathbf{x}}^{(i)} \right\|_2^2 \\ &\leq \alpha_{\max} \left\| \mathbf{X} \tilde{\mathbf{u}}_i^{(0)} \right\|_2^2 \leq \nu \alpha_{\max} n, \end{aligned}$$

where the final two bounds follow from Assumptions 6 and 8.

By the Gaussian tail bound (see, e.g., [Wainwright, 2019, Proposition 2.5]), we obtain

$$\mathbb{P} \left\{ \left| \tilde{\mathbf{x}}^{(i)\top} \mathbf{e}_*^{(j)} \right| > \frac{n\lambda_u}{8} \right\} \leq 2 \exp \left(-\frac{n^2 \lambda_u^2 / 64}{2 \cdot \nu \alpha_{\max} n} \right) = 2 \exp \left(-\frac{n \lambda_u^2}{128 \nu \alpha_{\max}} \right).$$

Combining this with (3.55), we obtain

$$\mathbb{P} \left\{ \left| \frac{1}{n} \tilde{\mathbf{X}}^\top \mathbf{E} \mathbf{B}^* \hat{\mathbf{D}}^- \right|_\infty > \frac{\lambda_u}{4} \right\} \leq 2pr \exp \left(-\frac{n \lambda_u^2}{128 \nu \alpha_{\max}} \right). \quad (3.56)$$

Therefore, if λ_u is chosen as in (3.14), inequality (3.52) holds with probability at least $1 - \delta/6$.

We now turn to proving that inequality (3.53) holds with probability at least $1 - \delta/6$.

By inequality (3.54) and Assumption 6, we have

$$\left| \frac{1}{n} \tilde{\mathbf{X}}^\top \mathbf{E} \hat{\Delta}^b \hat{\mathbf{D}}^- \right|_\infty \leq \left| \frac{1}{n} \tilde{\mathbf{X}}^\top \mathbf{E} \hat{\Delta}^b \right|_\infty \cdot \left| \hat{\mathbf{D}}^- \right|_\infty \leq \frac{2}{nd_r^*} \left| \tilde{\mathbf{X}}^\top \mathbf{E} \hat{\Delta}^b \right|_\infty \leq \frac{2}{nd_{\min}} \left| \tilde{\mathbf{X}}^\top \mathbf{E} \hat{\Delta}^b \right|_\infty. \quad (3.57)$$

By Lemma 35, and noting that the j -th row of $\tilde{\mathbf{X}}^\top \mathbf{E}$ is $\tilde{\mathbf{x}}^{(j)\top} \mathbf{E}$, we obtain

$$\begin{aligned} \frac{1}{n} \left| \tilde{\mathbf{X}}^\top \mathbf{E} \hat{\Delta}^b \right|_\infty &\leq \frac{1}{n} \left\{ \max_{1 \leq j \leq p} \left\| \tilde{\mathbf{x}}^{(j)\top} \mathbf{E} \right\|_2 \right\} \cdot \left\{ \max_{1 \leq j \leq q} \left\| \hat{\Delta}_j^b \right\|_2 \right\} \\ &\leq \frac{1}{n} \left\{ \max_{1 \leq j \leq p} \left\| \tilde{\mathbf{x}}^{(j)\top} \mathbf{E} \right\|_2 \right\} \cdot \left\| \hat{\Delta}^b \right\|_F. \end{aligned}$$

Applying the bound from (3.51) and the bound from (3.49), we conclude that

$$\frac{1}{n} \left| \tilde{\mathbf{X}}^\top \mathbf{E} \hat{\Delta}^b \right|_\infty \leq \frac{3c\eta r}{n} \left\{ \max_{1 \leq j \leq p} \left\| \tilde{\mathbf{x}}^{(j)\top} \mathbf{E} \right\|_2 \right\} \cdot R_{2,n}. \quad (3.58)$$

Note that

$$\frac{1}{n} \mathbf{E}^\top \tilde{\mathbf{x}}^{(j)} = \frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{X}}_{ij} \mathbf{e}_i \sim \mathcal{N} \left(0, \frac{\beta_j}{n} \Sigma \right),$$

where $\beta_j = \left\| \frac{1}{\sqrt{n}} \tilde{\mathbf{x}}^{(j)} \right\|_2^2$. By Lemma 36, we have the concentration inequality

$$\mathbb{P} \left\{ \left\| \frac{1}{n} \mathbf{E}^\top \tilde{\mathbf{x}}^{(j)} \right\|_2^2 - \frac{\beta_j}{n} \text{tr}(\Sigma) \geq t \right\} \leq \exp \left\{ -c_2 \min \left(\frac{n^2 t^2}{c_1^2 \beta_j^2 \text{tr}(\Sigma^2)}, \frac{nt}{c_1 \beta_j \lambda_{\max}(\Sigma)} \right) \right\} \quad (3.59)$$

for all $t \geq 0$, where $c_1, c_2 > 0$ are universal constants.

Let $\mathbf{1}_j$ denote the j -th canonical basis vector. By Assumption 6, we have

$$\beta_j = \left\| \frac{1}{\sqrt{n}} \tilde{\mathbf{x}}^{(j)} \right\|_2^2 = \left\| \frac{1}{\sqrt{n}} \mathbf{X} \tilde{\mathbf{U}}_{\text{full}} \mathbf{1}_j \right\|_2^2 \leq \nu.$$

Moreover, Assumption 8 ensures that $\lambda_{\max}(\Sigma) \leq \alpha_{\max}$.

Substituting these bounds into inequality (3.59), we obtain

$$\mathbb{P} \left\{ \left\| \frac{1}{n} \mathbf{E}^\top \tilde{\mathbf{x}}^{(j)} \right\|_2^2 \geq \frac{\nu}{n} \text{tr}(\Sigma) + t \right\} \leq \exp \left\{ -c_2 \min \left(\frac{n^2 t^2}{c_1^2 \nu^2 \text{tr}(\Sigma^2)}, \frac{nt}{c_1 \nu \alpha_{\max}} \right) \right\}.$$

Applying the union bound over $j = 1, \dots, p$, we have

$$\mathbb{P} \left\{ \max_{1 \leq j \leq p} \left\| \frac{1}{n} \mathbf{E}^\top \tilde{\mathbf{x}}^{(j)} \right\|_2^2 \geq \frac{\nu}{n} \text{tr}(\mathbf{\Sigma}) + t \right\} \leq p \exp \left\{ -c_2 \min \left(\frac{n^2 t^2}{c_1^2 \nu^2 \text{tr}(\mathbf{\Sigma}^2)}, \frac{nt}{c_1 \nu \alpha_{\max}} \right) \right\}.$$

Letting the right-hand side equal $\delta/6$ and solving for t gives

$$t = \frac{c_1 \nu}{n} \max \left\{ \sqrt{\frac{\text{tr}(\mathbf{\Sigma}^2) \log(6p/\delta)}{c_2}}, \frac{\alpha_{\max} \log(6p/\delta)}{c_2} \right\}.$$

Therefore, with probability at least $1 - \delta/6$, we have

$$\max_{1 \leq j \leq p} \left\| \frac{1}{n} \mathbf{E}^\top \tilde{\mathbf{x}}^{(j)} \right\|_2^2 \leq \frac{\nu}{n} \text{tr}(\mathbf{\Sigma}) + \frac{c_1 \nu}{n} \max \left\{ \sqrt{\frac{\text{tr}(\mathbf{\Sigma}^2) \log(6p/\delta)}{c_2}}, \frac{\alpha_{\max} \log(6p/\delta)}{c_2} \right\}.$$

Combining the above concentration bound with (3.58), we obtain

$$\begin{aligned} & \frac{1}{n^2} \left| \tilde{\mathbf{X}}^\top \mathbf{E} \hat{\mathbf{\Delta}}^b \right|_\infty^2 \\ & \leq 9c^2 \eta_r^2 R_{2,n}^2 \nu \left[\frac{\text{tr}(\mathbf{\Sigma})}{n} + \frac{c_1}{n} \max \left\{ \sqrt{\frac{\text{tr}(\mathbf{\Sigma}^2) \log(6p/\delta)}{c_2}}, \frac{\alpha_{\max} \log(6p/\delta)}{c_2} \right\} \right] \end{aligned}$$

with probability at least $1 - \delta/6$.

Furthermore, applying the bound from (3.57), we conclude that

$$\begin{aligned} & \left| \frac{1}{n} \tilde{\mathbf{X}}^\top \mathbf{E} \hat{\mathbf{\Delta}}^b \hat{\mathbf{D}} \right|_\infty^2 \leq \frac{4}{d_{\min}^2} \left| \frac{1}{n} \tilde{\mathbf{X}}^\top \mathbf{E} \hat{\mathbf{\Delta}}^b \right|_\infty^2 \\ & \leq \frac{36c^2 \eta_r^2 R_{2,n}^2 \nu}{d_{\min}^2} \left[\frac{\text{tr}(\mathbf{\Sigma})}{n} + \frac{c_1}{n} \max \left\{ \sqrt{\frac{\text{tr}(\mathbf{\Sigma}^2) \log(6p/\delta)}{c_2}}, \frac{\alpha_{\max} \log(6p/\delta)}{c_2} \right\} \right] \end{aligned}$$

also holds with probability at least $1 - \delta/6$.

To ensure that (3.53) holds with this probability, it suffices that the right-hand side of the above inequality is bounded by $\lambda_u^2/16$, where λ_u is specified in (3.14). This is equivalent

to requiring

$$\frac{36c^2\eta_r^2 R_{2,n}^2 \nu}{d_{\min}^2} \left[\frac{\text{tr}(\mathbf{\Sigma})}{n} + \frac{c_1}{n} \max \left\{ \sqrt{\frac{\text{tr}(\mathbf{\Sigma}^2) \log(6p/\delta)}{c_2}}, \frac{\alpha_{\max} \log(6p/\delta)}{c_2} \right\} \right] \leq 8\nu\alpha_{\max} \log(12pr/\delta),$$

which is guaranteed by Condition (3.13).

Thus, we have shown that inequality (3.46) holds with probability at least $1 - \delta/3$.

A similar argument can be applied to establish that (3.47) also holds with probability at least $1 - \delta/3$.

Finally, we turn to showing that (3.48) holds with probability at least $1 - \delta/3$. Note that

$$\left| \frac{1}{n} (\mathbf{X}\mathbf{U}^*)^\top \mathbf{E}\tilde{\mathbf{V}}_{\text{full}} \right|_\infty = \max_{1 \leq i \leq r} \max_{1 \leq j \leq q} \left| \frac{1}{n} \mathbf{x}_*^{(i)\top} \tilde{\mathbf{e}}^{(j)} \right|.$$

By the union bound (Bonferroni's inequality), we have

$$\mathbb{P} \left\{ \left| \frac{1}{n} (\mathbf{X}\mathbf{U}^*)^\top \mathbf{E}\tilde{\mathbf{V}}_{\text{full}} \right|_\infty > \frac{\lambda_v}{2} \right\} \leq \sum_{i=1}^r \sum_{j=1}^q \mathbb{P} \left\{ \left| \mathbf{x}_*^{(i)\top} \tilde{\mathbf{e}}^{(j)} \right| > \frac{n\lambda_v}{2} \right\}. \quad (3.60)$$

Since each column $\tilde{\mathbf{e}}^{(j)} = \mathbf{E}\tilde{\mathbf{v}}_j$ satisfies

$$\tilde{\mathbf{e}}^{(j)} \sim \mathcal{N} \left(0, \left(\tilde{\mathbf{v}}_j^\top \mathbf{\Sigma} \tilde{\mathbf{v}}_j \right) \cdot \mathbf{I}_n \right),$$

we conclude that $\mathbf{x}_*^{(i)\top} \tilde{\mathbf{e}}^{(j)}$ is Gaussian with mean zero and variance

$$\begin{aligned} \mathbb{E} \left[\left(\mathbf{x}_*^{(i)\top} \tilde{\mathbf{e}}^{(j)} \right)^2 \right] &= \mathbf{x}_*^{(i)\top} \mathbb{E} \left[\tilde{\mathbf{e}}^{(j)} \tilde{\mathbf{e}}^{(j)\top} \right] \mathbf{x}_*^{(i)} \\ &= \left(\tilde{\mathbf{v}}_j^\top \mathbf{\Sigma} \tilde{\mathbf{v}}_j \right) \cdot \left\| \mathbf{x}_*^{(i)} \right\|_2^2 \\ &\leq \alpha_{\max} \left\| \mathbf{X}\mathbf{u}_i^* \right\|_2^2 \leq \nu\alpha_{\max} n, \end{aligned}$$

where the final two bounds follow from Assumptions 6 and 8.

Applying the Gaussian tail bound [Wainwright, 2019, Proposition 2.5], we obtain

$$\mathbb{P} \left\{ \left| \mathbf{x}_*^{(i)\top} \tilde{\mathbf{e}}^{(j)} \right| > \frac{n\lambda_v}{2} \right\} \leq 2 \exp \left(-\frac{n^2 \lambda_v^2 / 4}{2 \cdot \nu \alpha_{\max} n} \right) = 2 \exp \left(-\frac{n\lambda_v^2}{8\nu\alpha_{\max}} \right).$$

Combining this with (3.60), we obtain

$$\mathbb{P} \left\{ \left| \frac{1}{n} (\mathbf{X}\mathbf{U}^*)^\top \mathbf{E} \tilde{\mathbf{V}}_{\text{full}} \right|_\infty > \frac{\lambda_v}{2} \right\} \leq 2qr \exp \left(-\frac{n\lambda_v^2}{8\nu\alpha_{\max}} \right).$$

Therefore, if λ_v is chosen as in (3.14), the right-hand side of the inequality is $\delta/3$, and thus (3.48) holds with probability at least $1 - \delta/3$.

□

3.8.2 Proof of Theorem 30

To simplify notation, we denote $\mathbf{U}^s(r_u)$ and $\mathbf{U}_\perp^s(r_u)$ by $\mathbf{U}^s$ and $\mathbf{U}_\perp^s$, respectively. Similarly, we write $\mathbf{V}^s(r_v)$ and $\mathbf{V}_\perp^s(r_v)$ as $\mathbf{V}^s$ and $\mathbf{V}_\perp^s$.

In addition, we define $\tilde{\mathbf{U}}^s$, $\tilde{\mathbf{U}}_\perp^s$, $\tilde{\mathbf{V}}^s$, and $\tilde{\mathbf{V}}_\perp^s$ analogously. Throughout the remainder of the proof, we also abbreviate $s_u(r_u)$ and $s_v(r_v)$ as s_u and s_v , respectively.

We define the following support sets for the source-truncated orthogonal complements:

$$\begin{aligned} \mathcal{S}_u &:= \left\{ (i, j) : \left[\mathbf{U}_\perp^{s\top} \mathbf{U}^* \right]_{ij} \neq 0 \right\} \subseteq [p - r_u] \times [r], \\ \mathcal{S}_u^c &:= \left\{ (i, j) : \left[\mathbf{U}_\perp^{s\top} \mathbf{U}^* \right]_{ij} = 0 \right\} \subseteq [p - r_u] \times [r]. \end{aligned} \tag{3.61}$$

By the definition of s_u , we have $|\mathcal{S}_u| = s_u$ and $|\mathcal{S}_u^c| = r(p - r_u) - s_u$.

Similarly, for the v -direction, we define:

$$\begin{aligned}\mathcal{S}_v &:= \left\{ (i, j) : \left[\mathbf{V}_\perp^{\text{s}\top} \mathbf{V}^* \right]_{ij} \neq 0 \right\} \subseteq [q - r_v] \times [r], \\ \mathcal{S}_v^c &:= \left\{ (i, j) : \left[\mathbf{V}_\perp^{\text{s}\top} \mathbf{V}^* \right]_{ij} = 0 \right\} \subseteq [q - r_v] \times [r],\end{aligned}\tag{3.62}$$

so that $|\mathcal{S}_v| = s_v$ and $|\mathcal{S}_v^c| = r(q - r_v) - s_v$.

For convenience, we also define the full source-augmented orthogonal basis matrices:

$$\tilde{\mathbf{U}}_{\text{full}} := \begin{bmatrix} \tilde{\mathbf{U}}^{\text{s}} & \tilde{\mathbf{U}}_\perp^{\text{s}} \end{bmatrix} \in \mathbb{R}^{p \times p}, \quad \tilde{\mathbf{V}}_{\text{full}} := \begin{bmatrix} \tilde{\mathbf{V}}^{\text{s}} & \tilde{\mathbf{V}}_\perp^{\text{s}} \end{bmatrix} \in \mathbb{R}^{q \times q}.$$

Under Condition (3.9) and constraint (3.10), we have the following bound on the estimation error:

$$\left\| \hat{\Delta} \right\|_{\text{F}} \leq 3R_{2,n}, \quad \text{and} \quad \left| \hat{\Delta} \right|_1 \leq 3R_{1,n}.\tag{3.63}$$

Given the regularization parameters

$$\lambda_u = 8\sqrt{2} \cdot \sqrt{\frac{\nu\alpha_{\max} \log(12pr/\delta)}{n}}, \quad \lambda_v = 2\sqrt{2} \cdot \sqrt{\frac{\nu\alpha_{\max} \log(6qr/\delta)}{n}},$$

Lemma 34 together with Conditions (3.9)–(3.11) and Condition (3.13) imply that the following concentration inequalities hold with probability at least $1 - \delta$:

$$\left| \frac{1}{n} \tilde{\mathbf{U}}_{\text{full}}^\top \mathbf{X}^\top \mathbf{E} \hat{\mathbf{B}} \hat{\mathbf{D}} \right|_\infty \leq \frac{\lambda_u}{2},\tag{3.64}$$

$$\left| \frac{1}{n} (\mathbf{X} \mathbf{U}^*)^\top \mathbf{E} \hat{\mathbf{B}} \hat{\mathbf{D}} \right|_\infty \leq \frac{\lambda_u}{2},\tag{3.65}$$

$$\left| \frac{1}{n} (\mathbf{X} \mathbf{U}^*)^\top \mathbf{E} \tilde{\mathbf{V}}_{\text{full}} \right|_\infty \leq \frac{\lambda_v}{2}.\tag{3.66}$$

Therefore, to complete the proof, it suffices to establish the error bound deterministically, conditional on the inequalities (3.64)–(3.66).

By Conditions (3.9)–(3.10) and the optimality of $(\widehat{\mathbf{U}}, \widehat{\mathbf{D}}, \widehat{\mathbf{V}})$, we have

$$\begin{aligned} & \frac{1}{2n} \left\| \mathbf{Y} - \mathbf{X} \widehat{\mathbf{U}} \widehat{\mathbf{D}} \widehat{\mathbf{V}}^\top \right\|_{\text{F}}^2 + \lambda_u |\widetilde{\mathbf{U}}_\perp^{\text{s}\top} \widehat{\mathbf{U}} \widehat{\mathbf{D}}|_1 + \lambda_v |\widetilde{\mathbf{V}}_\perp^{\text{s}\top} \widehat{\mathbf{V}} \widehat{\mathbf{D}}|_1 \\ & \leq \frac{1}{2n} \left\| \mathbf{Y} - \mathbf{X} \mathbf{U}^* \mathbf{D}^* \mathbf{V}^{*\top} \right\|_{\text{F}}^2 + \lambda_u |\widetilde{\mathbf{U}}_\perp^{\text{s}\top} \mathbf{U}^* \mathbf{D}^*|_1 + \lambda_v |\widetilde{\mathbf{V}}_\perp^{\text{s}\top} \mathbf{V}^* \mathbf{D}^*|_1. \end{aligned}$$

Combined with (3.1) and the definitions in (3.17), the above inequality implies that

$$\frac{1}{2n} \|\mathbf{X} \widehat{\Delta}\|_{\text{F}}^2 \leq \left\langle \frac{1}{n} \mathbf{X}^\top \mathbf{E}, \widehat{\Delta} \right\rangle + \lambda_u \left\{ \left| \widetilde{\mathbf{U}}_\perp^{\text{s}\top} \mathbf{A}^* \right|_1 - \left| \widetilde{\mathbf{U}}_\perp^{\text{s}\top} \widehat{\mathbf{A}} \right|_1 \right\} + \lambda_v \left\{ \left| \widetilde{\mathbf{V}}_\perp^{\text{s}\top} \mathbf{B}^* \right|_1 - \left| \widetilde{\mathbf{V}}_\perp^{\text{s}\top} \widehat{\mathbf{B}} \right|_1 \right\}. \quad (3.67)$$

Note that

$$\begin{aligned} \widehat{\Delta} &= \widehat{\mathbf{U}} \widehat{\mathbf{D}} \widehat{\mathbf{V}}^\top - \mathbf{U}^* \mathbf{D}^* \mathbf{V}^{*\top} = \widehat{\mathbf{A}} \widehat{\mathbf{D}}^- \widehat{\mathbf{B}}^\top - \mathbf{U}^* \mathbf{B}^{*\top} \\ &= \widehat{\Delta}^a \left(\widehat{\mathbf{B}} \widehat{\mathbf{D}}^- \right)^\top + \mathbf{A}^* \widehat{\mathbf{D}}^- \widehat{\mathbf{B}}^\top - \mathbf{U}^* \mathbf{B}^{*\top} \\ &= \widehat{\Delta}^a \left(\widehat{\mathbf{B}} \widehat{\mathbf{D}}^- \right)^\top - \mathbf{U}^* \widehat{\Delta}^d \left(\widehat{\mathbf{B}} \widehat{\mathbf{D}}^- \right)^\top + \mathbf{U}^* \left(\widehat{\Delta}^b \right)^\top. \end{aligned}$$

Thus, we have

$$\left\langle \frac{1}{n} \mathbf{X}^\top \mathbf{E}, \widehat{\Delta} \right\rangle = \left\langle \frac{1}{n} \mathbf{X}^\top \mathbf{E} \widehat{\mathbf{B}} \widehat{\mathbf{D}}^-, \widehat{\Delta}^a \right\rangle - \left\langle \frac{1}{n} \mathbf{U}^{*\top} \mathbf{X}^\top \mathbf{E} \widehat{\mathbf{B}} \widehat{\mathbf{D}}^-, \widehat{\Delta}^d \right\rangle + \left\langle \frac{1}{n} \mathbf{E}^\top \mathbf{X} \mathbf{U}^*, \widehat{\Delta}^b \right\rangle. \quad (3.68)$$

Besides, since

$$\widehat{\Delta}^a = \widetilde{\mathbf{U}}^{\text{s}} \widetilde{\mathbf{U}}^{\text{s}\top} \widehat{\Delta}^a + \widetilde{\mathbf{U}}_\perp^{\text{s}} \widetilde{\mathbf{U}}_\perp^{\text{s}\top} \widehat{\Delta}^a,$$

we then have

$$\begin{aligned} \left\langle \frac{1}{n} \mathbf{X}^\top \mathbf{E} \widehat{\mathbf{B}} \widehat{\mathbf{D}}^-, \widehat{\Delta}^a \right\rangle &= \left\langle \frac{1}{n} \widetilde{\mathbf{U}}^{\text{s}\top} \mathbf{X}^\top \mathbf{E} \widehat{\mathbf{B}} \widehat{\mathbf{D}}^-, \widetilde{\mathbf{U}}^{\text{s}\top} \widehat{\Delta}^a \right\rangle + \left\langle \frac{1}{n} \widetilde{\mathbf{U}}_\perp^{\text{s}\top} \mathbf{X}^\top \mathbf{E} \widehat{\mathbf{B}} \widehat{\mathbf{D}}^-, \widetilde{\mathbf{U}}_\perp^{\text{s}\top} \widehat{\Delta}^a \right\rangle \\ &\leq \left| \frac{1}{n} \widetilde{\mathbf{U}}^{\text{s}\top} \mathbf{X}^\top \mathbf{E} \widehat{\mathbf{B}} \widehat{\mathbf{D}}^- \right|_\infty \left| \widetilde{\mathbf{U}}^{\text{s}\top} \widehat{\Delta}^a \right|_1 + \left| \frac{1}{n} \widetilde{\mathbf{U}}_\perp^{\text{s}\top} \mathbf{X}^\top \mathbf{E} \widehat{\mathbf{B}} \widehat{\mathbf{D}}^- \right|_\infty \left| \widetilde{\mathbf{U}}_\perp^{\text{s}\top} \widehat{\Delta}^a \right|_1. \end{aligned}$$

Note that

$$\max \left\{ \left| \frac{1}{n} \tilde{\mathbf{U}}^{\text{s}\top} \mathbf{X}^\top \mathbf{E} \hat{\mathbf{B}} \hat{\mathbf{D}}^- \right|_\infty, \left| \frac{1}{n} \tilde{\mathbf{U}}_\perp^{\text{s}\top} \mathbf{X}^\top \mathbf{E} \hat{\mathbf{B}} \hat{\mathbf{D}}^- \right|_\infty \right\} = \left| \frac{1}{n} \tilde{\mathbf{U}}_{\text{full}}^\top \mathbf{X}^\top \mathbf{E} \hat{\mathbf{B}} \hat{\mathbf{D}}^- \right|_\infty,$$

combined with (3.64), we then have

$$\begin{aligned} \left\langle \frac{1}{n} \mathbf{X}^\top \mathbf{E} \hat{\mathbf{B}} \hat{\mathbf{D}}^-, \hat{\Delta}^a \right\rangle &\leq \frac{\lambda_u}{2} \left| \tilde{\mathbf{U}}^{\text{s}\top} \hat{\Delta}^a \right|_1 + \frac{\lambda_u}{2} \left| \tilde{\mathbf{U}}_\perp^{\text{s}\top} \hat{\Delta}^a \right|_1 \\ &\leq \frac{\lambda_u}{2} \left| \tilde{\mathbf{U}}^{\text{s}\top} \hat{\Delta}^a \right|_1 + \frac{\lambda_u}{2} \left| \left[\tilde{\mathbf{U}}_\perp^{\text{s}\top} \hat{\Delta}^a \right]_{\mathcal{S}_u} \right|_1 + \frac{\lambda_u}{2} \left| \left[\tilde{\mathbf{U}}_\perp^{\text{s}\top} \hat{\Delta}^a \right]_{\mathcal{S}_u^c} \right|_1. \end{aligned}$$

By the fact that

$$\left| \left[\tilde{\mathbf{U}}_\perp^{\text{s}\top} \hat{\Delta}^a \right]_{\mathcal{S}_u^c} \right|_1 \leq \left| \left[\tilde{\mathbf{U}}_\perp^{\text{s}\top} \hat{\mathbf{A}} \right]_{\mathcal{S}_u^c} \right|_1 + \left| \left[\tilde{\mathbf{U}}_\perp^{\text{s}\top} \mathbf{A}^* \right]_{\mathcal{S}_u^c} \right|_1,$$

we have

$$\begin{aligned} \left\langle \frac{1}{n} \mathbf{X}^\top \mathbf{E} \hat{\mathbf{B}} \hat{\mathbf{D}}^-, \hat{\Delta}^a \right\rangle &\leq \frac{\lambda_u}{2} \left| \tilde{\mathbf{U}}^{\text{s}\top} \hat{\Delta}^a \right|_1 + \frac{\lambda_u}{2} \left| \left[\tilde{\mathbf{U}}_\perp^{\text{s}\top} \hat{\Delta}^a \right]_{\mathcal{S}_u} \right|_1 \\ &\quad + \frac{\lambda_u}{2} \left| \left[\tilde{\mathbf{U}}_\perp^{\text{s}\top} \hat{\mathbf{A}} \right]_{\mathcal{S}_u^c} \right|_1 + \frac{\lambda_u}{2} \left| \left[\tilde{\mathbf{U}}_\perp^{\text{s}\top} \mathbf{A}^* \right]_{\mathcal{S}_u^c} \right|_1. \quad (3.69) \end{aligned}$$

Besides, we also have

$$\begin{aligned} \left| \tilde{\mathbf{U}}_\perp^{\text{s}\top} \mathbf{A}^* \right|_1 - \left| \tilde{\mathbf{U}}_\perp^{\text{s}\top} \hat{\mathbf{A}} \right|_1 &= \left| \left[\tilde{\mathbf{U}}_\perp^{\text{s}\top} \mathbf{A}^* \right]_{\mathcal{S}_u} \right|_1 - \left| \left[\tilde{\mathbf{U}}_\perp^{\text{s}\top} \hat{\mathbf{A}} \right]_{\mathcal{S}_u} \right|_1 + \left| \left[\tilde{\mathbf{U}}_\perp^{\text{s}\top} \mathbf{A}^* \right]_{\mathcal{S}_u^c} \right|_1 - \left| \left[\tilde{\mathbf{U}}_\perp^{\text{s}\top} \hat{\mathbf{A}} \right]_{\mathcal{S}_u^c} \right|_1 \\ &\leq \left| \left[\tilde{\mathbf{U}}_\perp^{\text{s}\top} \mathbf{A}^* \right]_{\mathcal{S}_u} \right|_1 - \left| \left[\tilde{\mathbf{U}}_\perp^{\text{s}\top} \hat{\mathbf{A}} \right]_{\mathcal{S}_u} \right|_1 + \left| \left[\tilde{\mathbf{U}}_\perp^{\text{s}\top} \mathbf{A}^* \right]_{\mathcal{S}_u^c} \right|_1 - \left| \left[\tilde{\mathbf{U}}_\perp^{\text{s}\top} \hat{\mathbf{A}} \right]_{\mathcal{S}_u^c} \right|_1 \\ &= \left| \left[\tilde{\mathbf{U}}_\perp^{\text{s}\top} \hat{\Delta}^a \right]_{\mathcal{S}_u} \right|_1 + \left| \left[\tilde{\mathbf{U}}_\perp^{\text{s}\top} \mathbf{A}^* \right]_{\mathcal{S}_u^c} \right|_1 - \left| \left[\tilde{\mathbf{U}}_\perp^{\text{s}\top} \hat{\mathbf{A}} \right]_{\mathcal{S}_u^c} \right|_1. \quad (3.70) \end{aligned}$$

Combining (3.69) and (3.70), we have

$$\begin{aligned}
& \left\langle \frac{1}{n} \mathbf{X}^\top \mathbf{E} \widehat{\mathbf{B}} \widehat{\mathbf{D}}^-, \widehat{\Delta}^a \right\rangle + \lambda_u \left\{ \left| \widetilde{\mathbf{U}}_\perp^{\text{s}\top} \mathbf{A}^* \right|_1 - \left| \widetilde{\mathbf{U}}_\perp^{\text{s}\top} \widehat{\mathbf{A}} \right|_1 \right\} \\
& \leq \frac{\lambda_u}{2} \left| \widetilde{\mathbf{U}}^{\text{s}\top} \widehat{\Delta}^a \right|_1 + \frac{3\lambda_u}{2} \left| \left[\widetilde{\mathbf{U}}_\perp^{\text{s}\top} \widehat{\Delta}^a \right]_{\mathcal{S}_u} \right|_1 + \frac{3\lambda_u}{2} \left| \left[\widetilde{\mathbf{U}}_\perp^{\text{s}\top} \mathbf{A}^* \right]_{\mathcal{S}_u^c} \right|_1 - \frac{\lambda_u}{2} \left| \left[\widetilde{\mathbf{U}}_\perp^{\text{s}\top} \widehat{\mathbf{A}} \right]_{\mathcal{S}_u^c} \right|_1 \\
& \leq \frac{3\lambda_u}{2} \left| \widetilde{\mathbf{U}}^{\text{s}\top} \widehat{\Delta}^a \right|_1 + \frac{3\lambda_u}{2} \left| \left[\widetilde{\mathbf{U}}_\perp^{\text{s}\top} \widehat{\Delta}^a \right]_{\mathcal{S}_u} \right|_1 + \frac{3\lambda_u}{2} \left| \left[\widetilde{\mathbf{U}}_\perp^{\text{s}\top} \mathbf{A}^* \right]_{\mathcal{S}_u^c} \right|_1. \tag{3.71}
\end{aligned}$$

In addition, note that $\left\| \widehat{\Delta}^a \right\|_{\text{F}}^2 = \left\| \widetilde{\mathbf{U}}^{\text{s}\top} \widehat{\Delta}^a \right\|_{\text{F}}^2 + \left\| \widetilde{\mathbf{U}}_\perp^{\text{s}\top} \widehat{\Delta}^a \right\|_{\text{F}}^2$, we then have

$$\left| \widetilde{\mathbf{U}}^{\text{s}\top} \widehat{\Delta}^a \right|_1 \leq \sqrt{r_u r} \left\| \widetilde{\mathbf{U}}^{\text{s}\top} \widehat{\Delta}^a \right\|_{\text{F}} \leq \sqrt{r_u r} \left\| \widehat{\Delta}^a \right\|_{\text{F}}$$

and

$$\left| \left[\widetilde{\mathbf{U}}_\perp^{\text{s}\top} \widehat{\Delta}^a \right]_{\mathcal{S}_u} \right|_1 \leq \sqrt{s_u} \left\| \left[\widetilde{\mathbf{U}}_\perp^{\text{s}\top} \widehat{\Delta}^a \right]_{\mathcal{S}_u} \right\|_{\text{F}} \leq \sqrt{s_u} \left\| \widetilde{\mathbf{U}}_\perp^{\text{s}\top} \widehat{\Delta}^a \right\|_{\text{F}} \leq \sqrt{s_u} \left\| \widehat{\Delta}^a \right\|_{\text{F}}.$$

Thus, we have

$$\left| \widetilde{\mathbf{U}}^{\text{s}\top} \widehat{\Delta}^a \right|_1 + \left| \left[\widetilde{\mathbf{U}}_\perp^{\text{s}\top} \widehat{\Delta}^a \right]_{\mathcal{S}_u} \right|_1 \leq (\sqrt{r_u r} + \sqrt{s_u}) \left\| \widehat{\Delta}^a \right\|_{\text{F}} \leq \sqrt{2(rr_u + s_u)} \left\| \widehat{\Delta}^a \right\|_{\text{F}},$$

where the last inequality follows from the fact that $\sqrt{a} + \sqrt{b} \leq \sqrt{2(a+b)}$ for all $a, b \geq 0$.

Combining the above inequality with (3.71), we have

$$\begin{aligned}
& \left\langle \frac{1}{n} \mathbf{X}^\top \mathbf{E} \widehat{\mathbf{B}} \widehat{\mathbf{D}}^-, \widehat{\Delta}^a \right\rangle + \lambda_u \left\{ \left| \widetilde{\mathbf{U}}_\perp^{\text{s}\top} \mathbf{A}^* \right|_1 - \left| \widetilde{\mathbf{U}}_\perp^{\text{s}\top} \widehat{\mathbf{A}} \right|_1 \right\} \\
& \leq \frac{3\lambda_u}{2} \sqrt{2(rr_u + s_u)} \left\| \widehat{\Delta}^a \right\|_{\text{F}} + \frac{3\lambda_u}{2} \left| \left[\widetilde{\mathbf{U}}_\perp^{\text{s}\top} \mathbf{A}^* \right]_{\mathcal{S}_u^c} \right|_1. \tag{3.72}
\end{aligned}$$

Similarly, given (3.66) holds, we can show that

$$\begin{aligned}
& \left\langle \frac{1}{n} \mathbf{E}^\top \mathbf{X} \mathbf{U}^*, \widehat{\Delta}^b \right\rangle + \lambda_v \left\{ \left| \widetilde{\mathbf{V}}_\perp^{\text{s}\top} \mathbf{B}^* \right|_1 - \left| \widetilde{\mathbf{V}}_\perp^{\text{s}\top} \widehat{\mathbf{B}} \right|_1 \right\} \\
& \leq \frac{3\lambda_v}{2} \sqrt{2(rr_v + s_v)} \left\| \widehat{\Delta}^b \right\|_{\text{F}} + \frac{3\lambda_v}{2} \left| \left[\widetilde{\mathbf{V}}_\perp^{\text{s}\top} \mathbf{B}^* \right]_{\mathcal{S}_v^c} \right|_1. \tag{3.73}
\end{aligned}$$

And given (3.65), we have

$$\left\langle \frac{1}{n} \mathbf{U}^{*\top} \mathbf{X}^\top \mathbf{E} \widehat{\mathbf{B}} \widehat{\mathbf{D}}^-, \widehat{\Delta}^d \right\rangle \leq \frac{\lambda_u}{2} \left| \widehat{\Delta}^d \right|_1 \leq \frac{\lambda_u}{2} \sqrt{r} \left\| \widehat{\Delta}^d \right\|_{\text{F}}. \quad (3.74)$$

Define

$$\lambda_{\max} := \max \{ \lambda_u, \lambda_v \}.$$

Then, by definition of λ_u and λ_v , we have the upper bound

$$\lambda_{\max} \leq c_\lambda \cdot \sqrt{\frac{\log(12 \max(p, q) \cdot r / \delta)}{n}},$$

where $c_\lambda = 8\sqrt{2\nu\alpha_{\max}}$ is a universal constant.

Combining equations (3.67), (3.68), and (3.72)–(3.74), we obtain the following deterministic bound:

$$\begin{aligned} \frac{1}{2n} \left\| \mathbf{X} \widehat{\Delta} \right\|_{\text{F}}^2 &\leq \lambda_{\max} \left\{ \frac{3}{2} \sqrt{2(rr_u + s_u)} \left\| \widehat{\Delta}^a \right\|_{\text{F}} + \frac{3}{2} \sqrt{2(rr_v + s_v)} \left\| \widehat{\Delta}^b \right\|_{\text{F}} + \frac{1}{2} \sqrt{r} \left\| \widehat{\Delta}^d \right\|_{\text{F}} \right\} \\ &\quad + \frac{3\lambda_{\max}}{2} \left| \left[\widetilde{\mathbf{U}}_{\perp}^{\text{s}\top} \mathbf{A}^* \right]_{\mathcal{S}_u^c} \right|_1 + \frac{3\lambda_{\max}}{2} \left| \left[\widetilde{\mathbf{V}}_{\perp}^{\text{s}\top} \mathbf{B}^* \right]_{\mathcal{S}_v^c} \right|_1. \end{aligned} \quad (3.75)$$

By (3.63) and Condition (3.11), we have

$$\left\| \widehat{\Delta} \right\|_2 \leq \left\| \widehat{\Delta} \right\|_{\text{F}} \leq \frac{\tau}{6\sqrt{5}} \cdot \left(\frac{d_r^*}{\sqrt{r}} \right) \cdot \left(\frac{d_r^*}{d_1^*} \right)^{3/2}.$$

Then by Lemma 31, Lemma 32 and Assumption 7, we have

$$\left\| \widehat{\Delta}^a \right\|_{\text{F}} \leq c\eta_r \left\| \widehat{\Delta} \right\|_{\text{F}}, \quad \left\| \widehat{\Delta}^b \right\|_{\text{F}} \leq c\eta_r \left\| \widehat{\Delta} \right\|_{\text{F}} \quad \text{and} \quad \left\| \widehat{\Delta}^d \right\|_{\text{F}} \leq \left\| \widehat{\Delta} \right\|_{\text{F}},$$

where $\eta_r = 1 + (1/\tau) \cdot \left(\sum_{j=1}^r (d_1^*/d_j^*)^2 \right)^{1/2}$, $0 < \tau < 1$ and $c > 0$ are universal constants.

Substituting these bounds into equation (3.75), we obtain:

$$\begin{aligned}
\frac{1}{2n} \left\| \mathbf{X} \hat{\Delta} \right\|_{\text{F}}^2 &\leq \lambda_{\max} \left\{ \frac{3\sqrt{2}}{2} c \eta_r \sqrt{r r_u + s_u} + \frac{3\sqrt{2}}{2} c \eta_r \sqrt{r r_v + s_v} + \frac{1}{2} \sqrt{r} \right\} \cdot \left\| \hat{\Delta} \right\|_{\text{F}} \\
&\quad + \frac{3\lambda_{\max}}{2} \left| \left[\tilde{\mathbf{U}}_{\perp}^{\text{s}\top} \mathbf{A}^* \right]_{\mathcal{S}_u^c} \right|_1 + \frac{3\lambda_{\max}}{2} \left| \left[\tilde{\mathbf{V}}_{\perp}^{\text{s}\top} \mathbf{B}^* \right]_{\mathcal{S}_v^c} \right|_1 \\
&\leq c' \lambda_{\max} \eta_r \sqrt{r(r_u + r_v + 1) + s_u + s_v} \cdot \left\| \hat{\Delta} \right\|_{\text{F}} \\
&\quad + \frac{3\lambda_{\max}}{2} \left| \left[\tilde{\mathbf{U}}_{\perp}^{\text{s}\top} \mathbf{A}^* \right]_{\mathcal{S}_u^c} \right|_1 + \frac{3\lambda_{\max}}{2} \left| \left[\tilde{\mathbf{V}}_{\perp}^{\text{s}\top} \mathbf{B}^* \right]_{\mathcal{S}_v^c} \right|_1, \tag{3.76}
\end{aligned}$$

where $c' > 0$ is a universal constant.

By Assumption 5, the following restricted strong convexity (RSC) inequality holds:

$$\frac{1}{2n} \left\| \mathbf{X} \hat{\Delta} \right\|_{\text{F}}^2 \geq \frac{\rho_2}{2} \left\| \hat{\Delta} \right\|_{\text{F}}^2 - \rho_1 \cdot \frac{\log p}{n} \left| \hat{\Delta} \right|_1 \geq \frac{\rho_2}{2} \left\| \hat{\Delta} \right\|_{\text{F}}^2 - 3\rho_1 \cdot \frac{\log p}{n} R_{1,n},$$

where the final inequality follows from the bound in (3.63). Combining this with the upper bound in (3.76), we deduce:

$$\begin{aligned}
\left\| \hat{\Delta} \right\|_{\text{F}}^2 &\leq \frac{2c'}{\rho_2} \lambda_{\max} \eta_r \sqrt{r(r_u + r_v + 1) + s_u + s_v} \cdot \left\| \hat{\Delta} \right\|_{\text{F}} \\
&\quad + \frac{3\lambda_{\max}}{\rho_2} \left| \left[\tilde{\mathbf{U}}_{\perp}^{\text{s}\top} \mathbf{A}^* \right]_{\mathcal{S}_u^c} \right|_1 + \frac{3\lambda_{\max}}{\rho_2} \left| \left[\tilde{\mathbf{V}}_{\perp}^{\text{s}\top} \mathbf{B}^* \right]_{\mathcal{S}_v^c} \right|_1 + \frac{6\rho_1}{\rho_2} \cdot \frac{\log p}{n} R_{1,n} \\
&= \frac{2c' c_{\lambda}}{\rho_2} \eta_r \sqrt{\frac{[r(r_u + r_v + 1) + s_u + s_v] \cdot \log(12r \max(p, q)/\delta)}{n}} \cdot \left\| \hat{\Delta} \right\|_{\text{F}} \\
&\quad + \frac{3c_{\lambda}}{\rho_2} \left\{ \left| \left[\tilde{\mathbf{U}}_{\perp}^{\text{s}\top} \mathbf{A}^* \right]_{\mathcal{S}_u^c} \right|_1 + \left| \left[\tilde{\mathbf{V}}_{\perp}^{\text{s}\top} \mathbf{B}^* \right]_{\mathcal{S}_v^c} \right|_1 \right\} \cdot \sqrt{\frac{\log(12r \max(p, q)/\delta)}{n}} \\
&\quad + \frac{6\rho_1}{\rho_2} \cdot \frac{\log p}{n} R_{1,n}, \tag{3.77}
\end{aligned}$$

For any $t \in \mathbb{R}$ satisfying $t^2 - bt - c \leq 0$ with $b, c > 0$, we have the inequality $t \leq b + \sqrt{c}$.

Applying this result to (3.77), we conclude that:

$$\begin{aligned}
\|\widehat{\Delta}\|_{\text{F}} &\leq \frac{2c'c_{\lambda}}{\rho_2}\eta_r\sqrt{\frac{[r(r_u+r_v+1)+s_u+s_v]\cdot\log(12r\max(p,q)/\delta)}{n}} \\
&\quad + \sqrt{\frac{3c_{\lambda}}{\rho_2}}\left\{\left|\left[\widetilde{\mathbf{U}}_{\perp}^{\text{s}\top}\mathbf{A}^*\right]_{\mathcal{S}_u^c}\right|_1 + \left|\left[\widetilde{\mathbf{V}}_{\perp}^{\text{s}\top}\mathbf{B}^*\right]_{\mathcal{S}_v^c}\right|_1\right\}^{1/2}\left(\frac{\log(12r\max(p,q)/\delta)}{n}\right)^{1/4} \\
&\quad + \left(\frac{6\rho_1}{\rho_2}\cdot\frac{\log p}{n}R_{1,n}\right)^{1/2}.
\end{aligned} \tag{3.78}$$

By Lemma 33, we have

$$\begin{aligned}
&\left|\left[\widetilde{\mathbf{U}}_{\perp}^{\text{s}\top}\mathbf{A}^*\right]_{\mathcal{S}_u^c}\right|_1 + \left|\left[\widetilde{\mathbf{V}}_{\perp}^{\text{s}\top}\mathbf{B}^*\right]_{\mathcal{S}_v^c}\right|_1 \\
&\leq 2^{3/2}\left[r\{(r_u^*-r_u)_+ + (r_v^*-r_v)_+\} - (s_u(r_u) + s_v(r_v))\right]\frac{(2d_1^{(0)} + \varepsilon^{(0)})\varepsilon^{(0)}}{\tau\left(\min(d_{r_u^*}^{(0)}, d_{r_v^*}^{(0)})\right)^2} \\
&\quad + 2^{5/4}r\left[r_u^{*1/4}\sqrt{p - \max(r_u^*, r_u)} + r_v^{*1/4}\sqrt{p - \max(r_v^*, r_v)}\right]\frac{(\varepsilon^{(0)})^{1/2}(\varepsilon^{(0)} + 2d_1^{(0)})^{1/2}}{\min(d_{r_u^*}^{(0)}, d_{r_v^*}^{(0)})}.
\end{aligned}$$

Combine the above inequality with (3.78), we have

$$\begin{aligned}
\|\widehat{\Delta}\|_{\text{F}} &\leq \frac{2c'c_{\lambda}}{\rho_2}\eta_r\sqrt{\frac{(r(r_u+r_v+1)+s_u+s_v)\log(12r\max(p,q)/\delta)}{n}} \\
&\quad + \frac{2}{\tau}\sqrt{\frac{6c_{\lambda}}{\rho_2}}\widetilde{\varepsilon}_{\text{source}}\left(\frac{\log(12r\max(p,q)/\delta)}{n}\right)^{\frac{1}{4}} + \left(\frac{6\rho_1}{\rho_2}\cdot\frac{\log p}{n}R_{1,n}\right)^{\frac{1}{2}},
\end{aligned} \tag{3.79}$$

where

$$\begin{aligned}
\tilde{\varepsilon}_{\text{source}} &= \left[r \left\{ (r_u^* - r_u)_+ + (r_v^* - r_v)_+ \right\} - (s_u(r_u) + s_v(r_v)) \right]^{1/2} \\
&\quad \times \frac{\left(\varepsilon^{(0)} \right)^{1/2} \left(2d_1^{(0)} + \varepsilon^{(0)} \right)^{1/2}}{\min \left(d_{r_u^*}^{(0)}, d_{r_v^*}^{(0)} \right)} \\
&\quad + \sqrt{r} \cdot \left[(r_u^*)^{1/4} \cdot \sqrt{p - \max(r_u^*, r_u)} + (r_v^*)^{1/4} \cdot \sqrt{q - \max(r_v^*, r_v)} \right]^{1/2} \\
&\quad \times \frac{\left(\varepsilon^{(0)} \right)^{1/4} \left(2d_1^{(0)} + \varepsilon^{(0)} \right)^{1/4}}{\sqrt{\min \left(d_{r_u^*}^{(0)}, d_{r_v^*}^{(0)} \right)}}.
\end{aligned}$$

By Assumption 6, we have

$$\min \left(d_{r_u^*}^{(0)}, d_{r_v^*}^{(0)} \right) \geq d_{\min}^{(0)} = \Omega(1).$$

Thus, we have

$$\begin{aligned}
\left\| \hat{\Delta} \right\|_{\text{F}} &\lesssim \eta_r \sqrt{\frac{(r(r_u + r_v + 1) + s_u + s_v) \log(12r \max(p, q)/\delta)}{n}} \\
&\quad + \varepsilon_{\text{source}} \left(\frac{\log(12r \max(p, q)/\delta)}{n} \right)^{\frac{1}{4}} + R_{1,n}^{\frac{1}{2}} \sqrt{\frac{\log p}{n}},
\end{aligned}$$

where $\varepsilon_{\text{source}}$ is defined in (3.16).

Moreover, combined with Condition (3.12), we have

$$\begin{aligned}
\left\| \hat{\Delta} \right\|_{\text{F}} &\lesssim \eta_r \sqrt{\frac{(r(r_u + r_v + 1) + s_u + s_v) \log(12r \max(p, q)/\delta)}{n}} \\
&\quad + \varepsilon_{\text{source}} \left(\frac{\log(12r \max(p, q)/\delta)}{n} \right)^{\frac{1}{4}}.
\end{aligned}$$

This completes the proof.

3.8.3 Additional Useful Lemmas

This section presents auxiliary lemmas that facilitate the theoretical analysis.

Lemma 35. *Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{B} \in \mathbb{R}^{n \times q}$ be arbitrary matrices. Denote by $\mathbf{A}_{i:}$ the i -th row of $\mathbf{A}$ and by $\mathbf{B}_{:,j}$ the j -th column of $\mathbf{B}$. Then,*

$$\|\mathbf{AB}\|_\infty \leq \left(\max_{1 \leq i \leq m} \|\mathbf{A}_{i:}\|_2 \right) \left(\max_{1 \leq j \leq q} \|\mathbf{B}_{:,j}\|_2 \right).$$

Proof. For any $1 \leq i \leq m$ and $1 \leq j \leq q$, the (i, j) -th entry of $\mathbf{AB}$ satisfies

$$[\mathbf{AB}]_{ij} = \sum_{k=1}^n \mathbf{A}_{ik} \mathbf{B}_{kj} = \langle \mathbf{A}_{i:}, \mathbf{B}_{:,j} \rangle \leq \|\mathbf{A}_{i:}\|_2 \|\mathbf{B}_{:,j}\|_2,$$

where the inequality follows from the Cauchy–Schwarz inequality. Taking the maximum over all i and j completes the proof. $\square$

Lemma 36 (Gaussian Vector Norm Concentration). *Let $\mathbf{g} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ be a d -dimensional Gaussian random vector. Then, for every $t \geq 0$, we have*

$$\mathbb{P} \left\{ \|\mathbf{g}\|_2^2 - \text{tr}(\boldsymbol{\Sigma}) \geq t \right\} \leq \exp \left\{ -c_2 \min \left(\frac{t^2}{c_1^2 \text{tr}(\boldsymbol{\Sigma}^2)}, \frac{t}{c_1 \lambda_{\max}(\boldsymbol{\Sigma})} \right) \right\},$$

where $c_1, c_2 > 0$ are universal constants.

Proof. Let $\boldsymbol{\Sigma} = \mathbf{U} \boldsymbol{\Lambda} \mathbf{U}^\top$ be the eigenvalue decomposition of $\boldsymbol{\Sigma}$, where $\mathbf{U} \in \mathbb{R}^{d \times d}$ is orthonormal and $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_d)$ with $\lambda_j \geq 0$. Define $\mathbf{z} := \boldsymbol{\Lambda}^{-1/2} \mathbf{U}^\top \mathbf{g}$, so that $\mathbf{g} = \mathbf{U} \boldsymbol{\Lambda}^{1/2} \mathbf{z}$. By construction, $\mathbf{z} = (Z_1, \dots, Z_d)^\top$ consists of i.i.d. standard normal variables.

We compute

$$\|\mathbf{g}\|_2^2 = \left\| \boldsymbol{\Lambda}^{1/2} \mathbf{z} \right\|_2^2 = \sum_{j=1}^d \lambda_j Z_j^2,$$

and thus,

$$\|\mathbf{g}\|_2^2 - \text{tr}(\mathbf{\Sigma}) = \sum_{j=1}^d \lambda_j (Z_j^2 - 1). \quad (3.80)$$

By Lemma 2.7.4 of Vershynin [2018], each Z_j^2 is sub-exponential, and hence $Z_j^2 - 1$ are i.i.d. mean-zero sub-exponential variables. Let $c_1 = \|Z_j^2 - 1\|_{\psi_1}$ denote the sub-exponential norm. Then by Theorem 2.8.3 of the same reference, for all $t \geq 0$,

$$\mathbb{P} \left\{ \sum_{j=1}^d \lambda_j (Z_j^2 - 1) \geq t \right\} \leq \exp \left\{ -c_2 \min \left(\frac{t^2}{c_1^2 \sum_{j=1}^d \lambda_j^2}, \frac{t}{c_1 \max_j \lambda_j} \right) \right\}, \quad (3.81)$$

where $c_2 > 0$ is a universal constant.

Substituting $\text{tr}(\mathbf{\Sigma}^2) = \sum_{j=1}^d \lambda_j^2$ and $\lambda_{\max}(\mathbf{\Sigma}) = \max_j \lambda_j$ into (3.81) and combining with (3.80) yields the result. $\square$

CHAPTER 4

ADAPTIVE CLIENT SAMPLING IN FEDERATED LEARNING VIA ONLINE LEARNING WITH BANDIT FEEDBACK

4.1 Introduction

Modern edge devices, such as personal mobile phones, wearable devices, and sensor systems in vehicles, collect large amounts of data that are valuable for training machine learning models. If each device only uses its local data to train a model, the resulting generalization performance will be limited due to the number of available samples on each device. Traditional approaches where data are transferred to a central server, which trains a model based on all available data, have fallen out of fashion due to privacy concerns and high communication costs. Federated Learning (FL) has emerged as a paradigm that allows for collaboration between different devices (clients) to train a global model while keeping data locally and only exchanging model updates [McMahan et al., 2017].

In a typical FL process, we have clients that contain data and a central server that orchestrates the training process [Kairouz et al., 2021]. The following process is repeated until the model is trained: (i) the server selects a subset of available clients; (ii) the server broadcasts the current model parameters and sometimes also a training program (e.g., a Tensorflow graph [Abadi et al., 2016]); (iii) the selected clients make updates to the model parameters based on their local data; (iv) the local model updates are uploaded to the server; (v) the server aggregates the local updates and makes a global update of the shared model. In this chapter, we focus on the first step and develop a practical strategy for selecting clients with provable guarantees.

To train a machine learning model in a FL setting with M clients, we would like to

minimize the following objective ¹:

$$\min_w F(w) := \sum_{m \in [M]} \lambda_m \phi(w; \mathcal{D}_m), \quad (4.1)$$

where $\phi(w; \mathcal{D}_m)$ is the loss function used to assess the quality of a machine learning model parameterized by the vector w based on the local data $\mathcal{D}_m$ on the client $m \in [M]$. The parameter λ_m denotes the weight for client m . Typically, we have $\lambda_m = n_m/n$, where $n_m = |\mathcal{D}_m|$ is the number of samples on the client m , and the total number of samples is $n = \sum_{m=1}^M n_m$. At the beginning of the t -th communication round, the server uses the sampling distribution $p^t = (p_1^t, \dots, p_M^t)^\top$ to choose K clients by sampling with replacement from $[M]$ ². Let $S^t \subseteq [M]$ denote the set of chosen clients with $|S^t| = K$. The server transmits the current model parameter vector w^t to each client $m \in S^t$. The client m computes the local update g_m^t ³ and sends it back to the server ⁴. After receiving local updates from clients in S^t , the server constructs a stochastic estimate of the global gradient as

$$g^t = \frac{1}{K} \sum_{m \in S^t} \frac{\lambda_m}{p_m^t} g_m^t, \quad (4.2)$$

and makes the global update of the parameter w^t using g^t . For example, $w^{t+1} = w^t - \mu^t g^t$,

1. We use $[M]$ to denote the set $\{1, \dots, M\}$.

2. In this chapter, we assume that all clients are available in each round and the purpose of client sampling is to reduce the communication cost, which is also the case considered by some previous research [Chen et al., 2022]. However, in practice, it is possible that only a subset of clients are available at the beginning of each round due to physical constraint. In Section 4.9.2, we discuss how to extend our proposed methods to deal with such situations. Analyzing such an extension is highly non-trivial and we leave it for further study. See detailed discussion in Section 4.9.2.

3. Here by model update, we actually mean the negative direction of model update. For example, when applying gradient descent, we refer the gradient as the model update, while the model parameter makes an update at the direction of the negative gradient. We stick with the term model update since it is more commonly used.

4. Throughout the chapter, except in Section 4.3, we do not specify how g_m^t is obtained. One possibility that the reader could keep in mind for concreteness is the LocalUpdate algorithm [Charles and Konečný, 2020], which covers well-known algorithms such as mini-batch SGD and FedAvg [McMahan et al., 2017].

if the server is using stochastic gradient descent (SGD) with the stepsize sequence $\{\mu^t\}_{t \geq 1}$ [Bottou et al., 2018]. However, the global update can be obtained using other procedures as well.

The sampling distribution in FL is typically uniform over clients, that is, $p^t = p^{\text{unif}} = (1/M, \dots, 1/M)^\top$. However, nonuniform sampling (also called importance sampling) can lead to faster convergence, both in theory and practice, as has been illustrated in stochastic optimization [Zhao and Zhang, 2015, Needell et al., 2016]. While the sampling distribution can be designed based on prior knowledge [Zhao and Zhang, 2015, Johnson and Guestrin, 2018, Needell et al., 2016, Stich et al., 2017], we cast the problem of choosing the sampling distribution as an online learning task and need no prior knowledge about (4.1).

Existing approaches to designing a sampling distribution using online learning take a stationary online learning framework and focus on matching the best sampling distribution that does not change over the training process. As a result, the obtained algorithms update the sampling distribution by treating all history information equally. However, as the training proceeds, the best sampling distribution changes with iterations. To address this problem, we take a non-stationary online learning framework and use an online stochastic mirror descent (OSMD) algorithm that puts more emphasis on recent feedback and learns to ‘forget’ the history. Consequently, our method shows empirical advantages over the previous methods. Besides, we derive a dynamic regret upper bound that allows the comparators to change with iterations, which generalizes the theoretical results on static regret of previous research. Moreover, we provably show how our sampling method improves the convergence guarantee of federated optimization methods over uniform sampling by reducing the dependency on the heterogeneity of the problem, which is also new to the best of our knowledge.

4.1.1 *Related Work*

Client Sampling. Chen et al. [2022] proposed a theoretically optimal client sampling scheme. However, their approach requires all clients to compute local updates in every

round, which is impractical in the presence of stragglers. Ribero and Vikalo [2020] modeled the evolution of model parameters during training using an Ornstein–Uhlenbeck process and derived an optimal sampling strategy from this formulation. Cho et al. [2020b] developed a biased client selection method and analyzed its convergence behavior. However, the algorithm incurs a non-vanishing bias and is not guaranteed to converge to the true optimum. Moreover, it requires participation from more clients per round than our method, leading to higher communication and computational costs. Other works such as Kim et al. [2020], Cho et al. [2020a], Yang et al. [2021] cast client selection as a multi-armed bandit problem, but provide only limited theoretical guarantees. Wang et al. [2020] applied reinforcement learning to balance accuracy and communication efficiency in client selection.

Importance Sampling. Our approach is closely related to importance sampling techniques in stochastic optimization. Zhao and Zhang [2015] and Needell et al. [2016] demonstrated that sampling data points from a non-uniform distribution—constructed using upper bounds on gradient norms—can accelerate convergence in gradient-based methods. Csiba and Richtárik [2018] extended importance sampling to mini-batch settings. Adaptive sampling strategies were developed by Stich et al. [2017], Johnson and Guestrin [2018], Gopal [2016], where the sampling distribution evolves over time. Importance sampling has also been explored in stochastic coordinate descent by Nesterov [2012], Perekrestenko et al. [2017], Zhu et al. [2016], Salehi et al. [2018].

Several works designed sampling distributions using online learning with bandit feedback. Namkoong et al. [2017] and Salehi et al. [2017] applied the EXP3 algorithm [Lattimore and Szepesvári, 2020, Chapter 11], while Borsos et al. [2018] used the follow-the-regularized-leader (FTRL) algorithm [Lattimore and Szepesvári, 2020, Chapter 28] for online convex optimization. Borsos et al. [2019] proposed restricting the sampling distribution to a linear combination of distributions from a predefined set and used online Newton steps to update the mixture weights. These methods target a stationary distribution, while in practice, the

optimal sampling strategy evolves dynamically across iterations. As a result, they may suffer from suboptimal empirical performance and provide only regret bounds with respect to a fixed comparator in hindsight.

To overcome this limitation, Hanchi and Stephens [2020] adopted a non-stationary framework that maintains recent information for each client and used a decaying step size to establish regret bounds. In contrast, our approach establishes regret bounds relative to a *dynamic comparator sequence* of sampling distributions. Our regret analysis does not require step size decay and explicitly captures the dependence on the total variation of the comparator sequence, which reflects the intrinsic non-stationarity of the problem.

Online Convex Optimization. Our work also contributes to the literature on online convex optimization (OCO). We formulate the client sampling problem as an online learning task [Hazan, 2016] and adopt techniques from dynamic OCO to solve it. Hall and Willett [2015], Yang et al. [2016], Daniely et al. [2015] proposed algorithms that achieve sublinear dynamic regret with respect to time-varying comparator sequences. Notably, Hall and Willett [2015] developed a dynamic mirror descent algorithm whose regret scales with the total variation in the comparator sequence.

Compared to these studies, our setting presents two key challenges. First, we only observe bandit feedback, rather than full knowledge of the loss function. Second, the loss functions are unbounded, violating standard assumptions in OCO. To address the first issue, we construct unbiased estimators for the loss and its gradient, which are then used to update the sampling distribution. For the second challenge, we first derive regret bounds under a bounded-loss assumption on a restricted region of the simplex and then quantify the additional regret induced by projecting comparator distributions into this region.

4.1.2 Organization and Notation

The remainder of this chapter is structured as follows. In Section 4.2, the motivation for employing importance sampling in federated learning (FL) is discussed, and an adaptive client sampling algorithm is introduced. Section 4.3 establishes optimization guarantees for mini-batch SGD and FedAvg under the proposed sampling framework. A regret analysis, which constitutes a fundamental component of the optimization analysis, is presented in Section 4.4. Section 4.5 extends the sampling algorithm to address scenarios with unknown problem parameters, while Section 4.6 further generalizes the approach to accommodate a varying number of rounds and sampling without replacement. Section 4.7 and Section 4.8 report experimental results on simulated and real-world datasets, respectively. The chapter concludes in Section 4.9 with a summary of the main findings and a discussion of potential directions for future research. All technical proofs are provided in Section 4.10.

Notation. In this chapter, we adopt the following notations and conventions. Let $\mathbb{R}_+^M := [0, \infty)^M$ and $\mathbb{R}_{++}^M := (0, \infty)^M$. For $M \in \mathbb{N}^+$, the $(M - 1)$ -dimensional probability simplex is defined as

$$\mathcal{P}_{M-1} := \left\{ x \in \mathbb{R}_+^M : \sum_{i=1}^M x_i = 1 \right\}.$$

We use $p = (p_1, \dots, p_M)^\top$ to denote a sampling distribution supported on $[M] := \{1, \dots, M\}$, and write $p^{1:T}$ for a sequence of sampling distributions $\{p^t\}_{t=1}^T$.

For any vector $x \in \mathbb{R}^n$ and $p \in [1, \infty]$, the ℓ_p -norm is defined as

$$\|x\|_p := \begin{cases} (\sum_{i=1}^n |x_i|^p)^{1/p}, & \text{if } 1 \leq p < \infty, \\ \max_{1 \leq i \leq n} |x_i|, & \text{if } p = \infty. \end{cases}$$

The dual norm associated with any norm $\|\cdot\|$ is defined as $\|z\|_\star := \sup\{z^\top x : \|x\| \leq 1\}$. For any finite set $\mathcal{B}$, we denote its cardinality by $|\mathcal{B}|$.

Let $\Phi : \mathcal{D} \subseteq \mathbb{R}^M \rightarrow \mathbb{R}$ be a differentiable convex function defined on a convex open set $\mathcal{D}$. The closure of $\mathcal{D}$ is denoted by $\bar{\mathcal{D}}$. The Bregman divergence between any $x, y \in \mathcal{D}$ with respect to Φ is defined as

$$D_\Phi(x \| y) := \Phi(x) - \Phi(y) - \langle \nabla \Phi(y), x - y \rangle.$$

In particular, we consider the unnormalized negative entropy given by

$$\Phi(x) := \sum_{m=1}^M x_m \log x_m - \sum_{m=1}^M x_m, \quad x = (x_1, \dots, x_M)^\top \in \mathcal{D} = \mathbb{R}_+^M,$$

where $0 \log 0$ is defined as zero by convention.

For two positive sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n = O(b_n)$ if there exists $C > 0$ such that $a_n/b_n \leq C$ for all sufficiently large n . Similarly, $a_n = \Omega(b_n)$ if there exists $c > 0$ such that $a_n/b_n \geq c$ for all sufficiently large n . We write $a_n = \Theta(b_n)$ if both $a_n = O(b_n)$ and $a_n = \Omega(b_n)$ hold. We write $a_n = o(b_n)$ if $\lim_{n \rightarrow \infty} a_n/b_n = 0$. In addition, $a_n = \tilde{O}(b_n)$ if $a_n = O(b_n \log^k b_n)$ for some $k \geq 0$.

4.2 Adaptive Client Sampling

In this section, we show how to cast the client sampling problem as an online learning task. Subsequently, we solve the online learning problem using the OSMD algorithm.

4.2.1 Client Sampling as an Online Learning Problem

Recall that at the beginning of the t -th communication round, the server uses a sampling distribution p^t to choose a set of clients S^t , by sampling with replacement K clients from $[M]$, to update the parameter vector w^t . For a chosen client $m \in S^t$, the local update is denoted as g_m^t . For example, the local update $g_m^t = \nabla \phi(w^t; \mathcal{D}_m)$ may be the full gradient; when mini-batch SGD/FedSGD is used, then $g_m^t = (1/B) \sum_{b=1}^B \nabla \phi(w^t; \xi_m^{t,b})$, where $\xi_m^{t,b} \stackrel{i.i.d.}{\sim} \mathcal{D}_m$ and B is the batch size; when FedAvg [McMahan et al., 2017] is used, then $g_m^t = w^t - w_m^{t,B}$,

where $w_m^{t,b} = w_m^{t,b-1} - \mu_l^t \nabla f(w_m^{t,b-1}; \xi_m^{t,b-1})$, $b = 0, \dots, B-1$, $w_m^{t,0} = w^t$, $\xi_m^{t,b} \stackrel{i.i.d.}{\sim} \mathcal{D}_m$, and μ_l^t is the local stepsize at t -th communication round.

The randomness comes from two sampling processes. The first sampling happens on the client level, and the second sampling happens locally when computing local updates. Client sampling is dealing with the first randomness. Since the two sampling processes are independent, we may treat g_m^t as deterministic in this section to ease the understanding. We will include the second randomness when analyzing regret and specific optimization algorithms in the following sections.

We define the aggregated oracle update at the t -th communication round as

$$J^t = \sum_{m=1}^M \lambda_m g_m^t.$$

The oracle update J^t is constructed only for theoretical purposes and is not computed in practice. The stochastic estimate g^t , defined in (4.2), is an unbiased estimate of J^t , that is, $\mathbb{E}_{S^t}[g^t] = J^t$. The variance of g^t is

$$\mathbb{V}_{S^t}[g^t] = \frac{1}{K} \left(\sum_{m=1}^M \frac{\lambda_m^2 \|g_m^t\|_2^2}{p_m^t} - \|J^t\|_2^2 \right). \quad (4.3)$$

Our goal is to design the sampling distribution p^t , used to sample S^t , to minimize the variance in (4.3). In doing so, we can ignore the second term, as it is independent of p^t . Minimizing variance is our goal in designing the sampling distribution because we require g^t to be an unbiased estimate of J^t . Allowing g^t to be biased, as in the biased client selection literature [Cho et al., 2020b, Qu et al., 2022, Ribero and Vikalo, 2020], may render minimizing variance ineffective. Our focus is on unbiased client selection, leaving biased client selection for future research.

Let $a_m^t = \lambda_m^2 \|g_m^t\|_2^2$. For any sampling distribution $q = (q_1, \dots, q_M)^\top$, the *variance*

reduction loss⁵ is defined as

$$l_t(q) = \frac{1}{K} \sum_{m=1}^M \frac{a_m^t}{q_m}. \quad (4.4)$$

Then for S^t sampled via q , we have

$$\mathbb{V}_{S^t}[g^t] = l_t(q) - \frac{1}{K} \|J^t\|_2^2.$$

Given a sequence of sampling distributions $q^{1:T}$, the cumulative variance reduction loss is defined as $L(q^{1:T}) := \sum_{t=1}^T l_t(q^t)$. When the choice of $q^{1:T}$ is random, the expected cumulative variance reduction loss is defined as $\bar{L}(q^{1:T}) := \mathbb{E} [L(q^{1:T})]$.

The variance reduction loss appears in the bound on the sub-optimality of a stochastic optimization algorithm. As a motivating example, suppose $F(\cdot)$ in (4.1) is ν -strongly convex. Furthermore, suppose the local update $g_m^t = \nabla \phi(w^t; \mathcal{D}_m)$ is the full gradient of the local loss and the global update is made by SGD with stepsize $\mu^t = 2/(\nu t)$. Theorem 3 of Salehi et al. [2017] then states that for any $T \geq 1$:

$$\mathbb{E} \left[F \left(\frac{2}{T(T+1)} \sum_{t=1}^T t \cdot w^t \right) \right] - F(w^*) \leq \frac{2}{\nu T(T+1)} \bar{L}(p^{1:T}), \quad (4.5)$$

where w^* is the minimizer of the objective in (4.1). Therefore, by choosing the sequence of sampling distributions $p^{1:T}$ to make the $\bar{L}(p^{1:T})$ small, one can achieve faster convergence. This observation holds in other stochastic optimization problems as well. We develop an algorithm that creates a sequence of sampling distributions $p^{1:T}$ to minimize $\bar{L}(p^{1:T})$ using only the norm of local updates, and without imposing assumptions on the loss functions or how the local and global updates are made. As a result, the proposed sampling algorithm is agnostic to both optimization algorithms and optimization problems. In Section 4.3, we

5. The variance reduction loss $l_t(\cdot)$ should be distinguished from the training loss $\phi(\cdot)$. While the former is always convex, $\phi(\cdot)$ can be non-convex.

show how our sampling method improves over uniform sampling by providing tighter upper bounds on mini-batch SGD and FedAvg with non-convex objective $\phi(\cdot)$.

Suppose that at the beginning of the t -th communication round we know all $\{a_m^t\}_{m=1}^M$. Then the optimal sampling distribution

$$p_\star^t = (p_{\star,1}^t, \dots, p_{\star,M}^t)^\top = \arg \min_{p \in \mathcal{P}_{M-1}} l_t(p)$$

is obtained as $p_{\star,m}^t = \sqrt{a_m^t} / (\sum_{m=1}^M \sqrt{a_m^t})$. Computing the distribution $p_\star^t$ is impractical as it requires local updates of all clients, which eradicates the need for client sampling. From the form of $p_\star^t$, we observe that clients with a large a_m^t are more “important” and should have a higher probability of being selected. Since we do not know $\{a_m^t\}_{m=1}^M$, we will need to explore the environment to learn about the importance of clients before we can exploit the best strategy. Finally, we note that the relative importance of clients will change over time, which makes the environment dynamic and challenging.

Based on the above discussion, we cast the problem of creating a sequence of sampling distributions as an online learning task with bandit feedback, where a game is played between the server and environment. Let p^1 be the initial sampling distribution. At the beginning of iteration t , the server samples with replacement K clients from $[M]$, denoted by S^t , using p^t . The environment reveals $\{a_m^t\}_{m \in S^t}$ to the server, where $a_m^t = \lambda_m^2 \|g_m^t\|_2^2$. The environment also computes $l_t(p^t)$; however, this loss is not revealed to the server. The server then updates p^{t+1} based on the feedback $\{\{a_m^u\}_{m \in S^u}\}_{u=1}^t$ and sampling distributions $\{p^u\}_{u=1}^t$. Note that in this game, the server only gets information about the chosen clients and, based on this partial information, or bandit feedback, needs to update the sampling distribution. On the other hand, we would like to be competitive with an oracle that can calculate the cumulative variance reduction loss. We will design p^t in a way that is agnostic to the generation mechanism of $\{a^t\}_{t \geq 1}$, and will treat the environment as deterministic, with randomness coming only from $\{S^t\}_{t \geq 1}$ when designing p^t . We describe an OSMD-based

approach to solve this online learning problem.

4.2.2 OSMD Sampler

Note that the variance-reduction loss function l_t is a convex function on $\mathcal{P}_{M-1}$ and

$$\nabla l_t(q) = -\frac{1}{K} \left(\frac{a_1^t}{(q_1)^2}, \dots, \frac{a_M^t}{(q_M)^2} \right)^\top \in \mathbb{R}^M \quad \text{for all } q = (q_1, \dots, q_M)^\top \in \mathbb{R}_{++}^M.$$

Since we do not observe a^t , we cannot compute $l_t(\cdot)$ or $\nabla l_t(\cdot)$. Instead, we can construct unbiased estimates of them. For any $q \in \mathcal{P}_{M-1}$, let $\hat{l}_t(q; p^t)$ be an estimate of $l_t(q)$ defined as

$$\hat{l}_t(q; p^t) = \frac{1}{K^2} \sum_{m=1}^M \frac{a_m^t}{q_m p_m^t} \mathcal{N}\{m \in S^t\}, \quad (4.6)$$

and $\nabla \hat{l}_t(q; p^t) \in \mathbb{R}^M$ has the m -th entry defined as

$$\left[\nabla \hat{l}_t(q; p^t) \right]_m = -\frac{1}{K^2} \cdot \frac{a_m^t}{q_m^2 p_m^t} \mathcal{N}\{m \in S^t\}. \quad (4.7)$$

The set S^t is sampled with replacement from $[M]$ using p^t and $\mathcal{N}\{m \in S^t\}$ denotes the number of times that a client m is chosen in S^t . Thus, $0 \leq \mathcal{N}\{m \in S^t\} \leq K$. Given q and p^t , $\hat{l}_t(q; p^t)$ and $\nabla \hat{l}_t(q; p^t)$ are random variables in $\mathbb{R}$ and $\mathbb{R}^M$ that satisfy

$$\mathbb{E}_{S^t} \left[\hat{l}_t(q; p^t) \mid p^t \right] = l_t(q), \quad \mathbb{E}_{S^t} \left[\nabla \hat{l}_t(q; p^t) \mid p^t \right] = \nabla l_t(q).$$

When S^t and $p^t \in \mathbb{R}_{++}^M$ are given, $\hat{l}_t(q; p^t)$ is a convex function with respect to q on $\mathbb{R}_{++}^M$ and satisfies $\hat{l}_t(q; p^t) - \hat{l}_t(q'; p^t) \leq \langle \nabla \hat{l}_t(q; p^t), q - q' \rangle$, for $q, q' \in \mathbb{R}_{++}^M$. The constructed estimates $\hat{l}_t(q; p^t)$ and $\nabla \hat{l}_t(q; p^t)$ are crucial for designing updates to the sampling distribution.

OSMD Sampler is an online stochastic mirror descent algorithm for updating the sampling distribution, detailed in Algorithm 5. The sampling distribution is restricted to lie in the space $\mathcal{A} = \mathcal{P}_{M-1} \cap [\alpha/M, \infty)^M$, $\alpha \in (0, 1]$, to prevent the server from assigning

Algorithm 5 OSMD Sampler

- 1: **Input:** Learning rate η , parameter $\alpha \in (0, 1]$, and number of iterations T .
 - 2: **Output:** $\hat{p}^{1:T}$.
 - 3: **Initialize:** $\hat{p}^1 = p^{\text{unif}}$.
 - 4: **for** $t = 1, 2, \dots, T - 1$ **do**
 - 5: Sample S^t by $\hat{p}^t$.
 - 6: Compute $\nabla \hat{l}_t(\hat{p}^t; \hat{p}^t)$ via (4.7).
 - 7: $\hat{p}^{t+1} = \arg \min_{p \in \mathcal{A}} \eta \langle p, \nabla \hat{l}_t(\hat{p}^t; \hat{p}^t) \rangle + D_\Phi(p \parallel \hat{p}^t)$.
 - 8: **end for**
-

too small probabilities to devices. The learning rates $\{\eta_t\}_{t \geq 1}$ are positive ⁶. $\Phi(x) = \sum_{m=1}^M x_m \log x_m - \sum_{m=1}^M x_m$, $x = (x_1, \dots, x_M)^\top \in \mathcal{D} = \mathbb{R}_+^M$, with $0 \log 0$ defined as 0, is the unnormalized negative entropy. The Bregman divergence between any $x, y \in \mathcal{D}$ with respect to the function Φ is given as $D_\Phi(x \parallel y) = \Phi(x) - \Phi(y) - \langle \nabla \Phi(y), x - y \rangle$. Line 7 of Algorithm 5 provides an update to the sampling distribution using the mirror descent update. The available feedback is used to construct an estimate of the loss, while the Bregman divergence between the current and next sampling distribution is used as a regularizer, ensuring that the updated sampling distribution does not change too much. The update only uses the most recent information, while forgetting the history, which results in nonstationarity of the sequence of sampling distributions.

An efficient algorithm to solve the mirror descent update in Line 7 of Algorithm 5 is shown in Algorithm 6, justified by the following proposition.

Proposition 37. *Let*

$$\tilde{p}_m^{t+1} = p_m^t \exp \left\{ \mathcal{N} \left\{ m \in S^t \right\} \eta_t a_m^t / (K^2 (p_m^t)^3) \right\}, \quad m \in [M].$$

Let $\pi : [M] \mapsto [M]$ be a permutation such that $\tilde{p}_{\pi(1)}^{t+1} \leq \tilde{p}_{\pi(2)}^{t+1} \leq \dots \leq \tilde{p}_{\pi(M)}^{t+1}$. Let $m_\star^t$ be the

6. We use the term *learning rate* when discussing an online algorithm that learns a sampling distribution, while the term *stepsize* is used in the context of an optimization algorithm.

Algorithm 6 Solver of Step 7 of Algorithm 5

- 1: **Input:** $\hat{p}^t$, S^t , $\{a_m^t\}_{m \in S^t}$, and $\mathcal{A} = \mathcal{P}_{M-1} \cap [\alpha/M, \infty)^M$.
 - 2: **Output:** $\hat{p}^{t+1}$.
 - 3: Let $\tilde{p}_m^{t+1} = p_m^t \exp \{ \mathcal{N} \{ m \in S^t \} \eta_t a_m^t / (K^2 (p_m^t)^3) \}$ for $m \in [M]$.
 - 4: Sort $\{\tilde{p}_m^{t+1}\}_{m=1}^M$ in a non-decreasing order: $\tilde{p}_{\pi(1)}^{t+1} \leq \tilde{p}_{\pi(2)}^{t+1} \leq \dots \leq \tilde{p}_{\pi(M)}^{t+1}$.
 - 5: Let $v_m = \tilde{p}_{\pi(m)}^{t+1} \left(1 - \frac{m-1}{M} \alpha \right)$ for $m \in [M]$.
 - 6: Let $u_m = \frac{\alpha}{M} \sum_{j=m}^M \tilde{p}_{\pi(j)}^{t+1}$ for $m \in [M]$.
 - 7: Find the smallest m such that $v_m > u_m$, denoted as $m_\star^t$.
 - 8: Let $\hat{p}_m^{t+1} = \begin{cases} \alpha/M & \text{if } \pi(m) < m_\star^t \\ ((1 - ((m_\star^t - 1)/M)\alpha)\tilde{p}_m^{t+1}) / \left(\sum_{j=m_\star^t}^M \tilde{p}_{\pi(j)}^{t+1} \right) & \text{otherwise.} \end{cases}$
-

smallest integer m such that

$$\tilde{p}_{\pi(m)}^{t+1} \left(1 - \frac{m-1}{M} \alpha \right) > \frac{\alpha}{M} \sum_{j=m}^M \tilde{p}_{\pi(j)}^{t+1}.$$

Then

$$\hat{p}_m^{t+1} = \begin{cases} \alpha/M & \text{if } \pi(m) < m_\star^t \\ ((1 - ((m_\star^t - 1)/M)\alpha)\tilde{p}_m^{t+1}) / \left(\sum_{j=m_\star^t}^M \tilde{p}_{\pi(j)}^{t+1} \right) & \text{otherwise.} \end{cases}$$

See proof in Section 4.10.2. The main cost comes from sorting the sequence $\{\tilde{p}_m^{t+1}\}_{m=1}^M$, which can be done with the computational complexity of $O(M \log M)$. However, note that we only update a few entries of $\hat{p}^t$ to get $\tilde{p}^{t+1}$ and $\hat{p}^t$ is sorted. Therefore, most entries of $\tilde{p}^{t+1}$ are also sorted. Using this observation, we can usually achieve a much faster running time, for example, by using an adaptive sorting algorithm [Estivill-Castro and Wood, 1992].

4.3 Application of OSMD Sampler on Federated Optimization

Algorithms

We illustrate how OSMD Sampler can be used to provably improve the convergence rates of federated optimization algorithms by reducing the heterogeneity. We choose two algorithms

that are most commonly used in federated learning as our illustrative examples, namely the SGD mini-batch and FedAvg [McMahan et al., 2017]. We use these two algorithms as motivational examples to show how adaptive sampling improves the convergence guarantees of optimization algorithms. However, the analysis here could be generalized to other optimization algorithms as well.

To simplify the notation, we denote $F_m(w) := \phi(w; \mathcal{D}_m)$ and let $\lambda_m = 1/M$ for all $m \in [M]$ in problem (4.1). We assume that the client objectives are differentiable and L -smooth functions.

Assumption 9. *For all $m \in [M]$, $F_m(\cdot)$ is differentiable and L -smooth, that is,*

$$\|\nabla F_m(x) - \nabla F_m(y)\|_2 \leq L\|x - y\|_2, \quad \text{for all } x, y \in \mathbb{R}^d.$$

Note that we allow $F_m(\cdot)$ to be non-convex. We also assume that the objective function $F(\cdot)$ is lower-bounded.

Assumption 10. *We assume that $\inf_w F(w) > -\infty$. We then denote $F^* := \inf_w F(w)$.*

In addition, we make the following assumption about the local stochastic gradient.

Assumption 11. *We assume that*

$$\mathbb{E}_{\xi \sim \mathcal{D}_m} [\nabla \phi(w; \xi)] = \nabla F_m(w) \quad \text{and} \quad \mathbb{E}_{\xi \sim \mathcal{D}_m} [\|\nabla \phi(w; \xi) - \nabla F_m(w)\|_2^2] \leq \sigma^2$$

for all w and $m \in [M]$; besides, $\|\nabla \phi(w; \xi)\|_2 \leq G$ for all w and ξ .

While bounded gradient variance is often assumed in federated learning literature [Patel et al., 2022], bounded gradient norm is less common. We use this assumption to simplify regret analysis, making the loss in (4.4) bounded. Removing it is feasible but complex, diverging from the main focus of this chapter. From a practical viewpoint, many loss functions,

like logistic regression loss, naturally satisfy this assumption. For others, one can project the gradient into a bounded norm subspace. If any minimizer $w^\star$ has a bounded gradient norm $\|\nabla\phi(w^\star; \xi)\|_2$ and is within this subspace, the projection will not increase the distance to the minimizer. However, while the projection step can provide a slightly stronger theoretical guarantee, it brings few practical benefits and makes the algorithm harder to follow. Therefore, we adopt a stronger assumption to simplify the presentation.

We start our analysis by building the connection between the heterogeneity and client sampling. We first introduce quantities that characterize the heterogeneity of the optimization problem. Specifically, heterogeneity characterizes how the objective functions of different clients differ from each other. In a federated learning problem, heterogeneity can be large and it is important to understand its effect on the convergence of algorithms. Let

$$\zeta_{\text{unif}}^2 := \sup_w \frac{1}{M} \sum_{m=1}^M \|\nabla F_m(w) - \nabla F(w)\|_2^2 = \sup_w \left\{ \frac{1}{M} \sum_{m=1}^M \|\nabla F_m(w)\|_2^2 - \|\nabla F(w)\|_2^2 \right\}. \quad (4.8)$$

The quantity ζ_{unif} has been commonly used to quantify the first-order heterogeneity in the literature [Karimireddy et al., 2020a,b]. To understand the relationship between heterogeneity and client sampling, let $\tilde{m}$ be a random index drawn from $[M]$, with $\mathbb{P}\{\tilde{m} = m\} = p_m$ for all $m \in [M]$. Then a natural unbiased estimator of $\nabla F(w)$ is $\nabla F_{\tilde{m}}(w)/(M\tilde{p}_m)$. We define

$$V(p, w) := \mathbb{E}_{\tilde{m}} \left[\left\| \frac{1}{M\tilde{p}_{\tilde{m}}} \nabla F_{\tilde{m}}(w) - \nabla F(w) \right\|_2^2 \right],$$

to be the variance of the estimator at parameter w , where $\mathbb{E}_{\tilde{m}}[\cdot]$ denotes the expectation taken with respect to the random index $\tilde{m}$. Note that we have

$$V(p, w) = \frac{1}{M^2} \sum_{m=1}^M \frac{1}{p_m} \|\nabla F_m(w)\|_2^2 - \|\nabla F(w)\|_2^2. \quad (4.9)$$

Thus, it is clear that $\zeta_{\text{unif}}^2 = \sup_w V(p^{\text{unif}}, w)$. In other words, the common definition of heterogeneity can be viewed as the worst-case variance of uniform client sampling.

When we use adaptive sampling to do client sampling, there are two sources of flexibility that allow us to reduce the heterogeneity: (i) we can use non-uniform sampling that may depend on parameter w ; (ii) we allow the sampling distribution to change over iterations. To reflect the consequential effect, we introduce a new concept termed *dynamic heterogeneity*. Let $\text{TV}(q^{1:T}) = \sum_{t=1}^{T-1} \|q^{t+1} - q^t\|_1$ be the total variation of any sequence of sampling distributions $q^{1:T} \in \mathcal{P}_{M-1}^T$. The dynamic heterogeneity is defined as

$$\zeta_T^2(\alpha, \beta) = \frac{1}{T} \sup_{w^1} \min_{p^1 \in \mathcal{A}} \cdots \sup_{w^T} \min_{p^T \in \mathcal{A}} \sum_{t=1}^T V(p^t, w^t) \quad \text{subject to } \text{TV}(p^{1:T}) \leq \beta,$$

where $V(p, w)$ is defined in (4.9) and $\beta \geq 0$ is the total variation budget. The dynamic heterogeneity $\zeta_T^2(\alpha, \beta)$ can be regarded as the worst-case variance of dynamic samplings in $\mathcal{A}^T$ with the total variation budget β . To see how $\zeta_T^2(\alpha, \beta)$ improves over ζ_{unif}^2 , let

$$\zeta_{\text{fix}}^2(\alpha) = \min_{p \in \mathcal{A}} \sup_w V(p, w). \quad (4.10)$$

The quantity $\zeta_{\text{fix}}^2(\alpha)$ can be regarded as the minimum heterogeneity by using the best fixed sampling distribution in $\mathcal{A}$ that does not depend on parameter w . Let p_f be the solution of p to the min-max problem (4.10), that is, $\sup_w V(p_f, w) = \zeta_{\text{fix}}^2(\alpha)$. Since $p^{\text{unif}} \in \mathcal{A}$ for all $\alpha \geq 0$, it is easy to see that

$$\zeta_{\text{fix}}^2(\alpha) = \min_{p \in \mathcal{A}} \sup_w V(p, w) \leq \sup_w V(p^{\text{unif}}, w) = \zeta_{\text{unif}}^2.$$

Note that $\zeta_T^2(\alpha, \beta)$ is a non-increasing function of β with any given α . Thus, we have

$$\zeta_T^2(\alpha, \beta) \leq \zeta_{\text{fix}}^2(\alpha) \leq \zeta_{\text{unif}}^2 \quad \forall 0 \leq \alpha \leq 1, \beta \geq 0. \quad (4.11)$$

Proof. Note that

$$\begin{aligned}
\zeta_T^2(\alpha, \beta) &= \frac{1}{T} \sup_{w^1} \min_{p^1 \in \mathcal{A}} \cdots \sup_{w^T} \min_{p^T \in \mathcal{A}} \sum_{t=1}^T V(p^t, w^t) \quad \text{subject to } \text{TV}(p^{1:T}) \leq \beta \\
&\leq \frac{1}{T} \sup_{w^1} \min_{p^1 \in \mathcal{A}} \cdots \sup_{w^T} \min_{p^T \in \mathcal{A}} \sum_{t=1}^T V(p^t, w^t) \quad \text{subject to } \text{TV}(p^{1:T}) = 0 \\
&= \frac{1}{T} \sup_{w^1} \min_{p \in \mathcal{A}} \sup_{w^2} \cdots \sup_{w^T} \sum_{t=1}^T V(p, w^t) \\
&\leq \frac{1}{T} \min_{p \in \mathcal{A}} \sup_{w^1} \cdots \sup_{w^T} \sum_{t=1}^T V(p, w^t) \\
&\leq \frac{1}{T} \sup_{w^1} \cdots \sup_{w^T} \sum_{t=1}^T V(p_f, w^t) = \sup_w V(p_f, w) = \zeta_{\text{fix}}^2(\alpha) \leq \zeta_{\text{unif}}^2.
\end{aligned}$$

□

Note that the above inequality also implies that the dynamic sampling distribution may potentially improve over a fixed sampling distribution. As we will see shortly, when $\zeta_{\text{unif}}^2 > \zeta_{\text{fix}}^2(\alpha)$, the OSMD Sampler can always improve over uniform sampling asymptotically.

When $\beta \geq 2(T-1)$ and α is small enough such that

$$p_m^\star(w) := \frac{\|\nabla F_m(w)\|_2}{\sum_{m'=1}^M \|\nabla F_{m'}(w)\|_2} \geq \frac{\alpha}{M} \quad \text{for all } w \text{ and } m \in [M],$$

we have

$$\zeta_T^2(\alpha, \beta) = \sup_w \min_{p \in \mathcal{A}} V(w, p) = \sup_w \left\{ \left(\frac{1}{M} \sum_{m=1}^M \|\nabla F_m(w)\|_2 \right)^2 - \|\nabla F(w)\|_2^2 \right\} \triangleq \zeta_{\text{min}}^2,$$

which is the smallest heterogeneity possibly achievable.

Algorithm 7 Mini-batch SGD with OSMD Sampler

- 1: **Input:** Number of communication rounds T , number of clients chosen in each round K , local batch size B , initial model parameter w^1 , stepsizes $\{\mu^t\}_{t=1}^T$, learning rate η and parameter $\alpha \in (0, 1]$.
- 2: **Output:** The final model parameter w^R .
- 3: **Initialize:** $\hat{p}^1 = p^{\text{unif}}$.
- 4: **for** $t = 1, \dots, T$ **do**
- 5: Sample S^t with replacement from $[M]$ with probability $\hat{p}^t$, such that $|S^t| = K$.
- 6: **for** $m \in S^t$ **do**
- 7: Download the current model parameter w^t .
- 8: Locally sample a mini-batch $\mathcal{B}_m^t = \{\xi_m^{t,1}, \dots, \xi_m^{t,B}\}$ i.i.d. uniformly random from $[n_m]$, where $|\mathcal{B}_m^t| = B$.
- 9: Locally compute and upload $g_m^t = (1/B) \sum_{b=1}^B \nabla \phi(w^t; \xi_m^{t,b})$ to the server.
- 10: **end for**
- 11: Server computes $a_m^t = \lambda_m^2 \|g_m^t\|^2$ for $m \in S^t$ and

$$g^t = \frac{1}{K} \sum_{m \in S^t} \frac{\lambda_m}{\hat{p}_m^t} g_m^t. \quad (4.12)$$

- 12: Server makes update of the model parameter $w^{t+1} \leftarrow w^t - \mu^t g^t$.
 - 13: Server obtains updated sampling distribution $\hat{p}^{t+1}$ by Algorithm 5.
 - 14: **end for**
-

4.3.1 Convergence Analysis of Mini-batch SGD with OSMD Sampler

We introduce the convergence analysis of mini-batch SGD with OSMD Sampler. The detailed algorithm is given in Algorithm 7. Compared to the classical mini-batch SGD, the key ingredients of Algorithm 7 are Line 13, where the server updates the sampling distribution by OSMD Sampler, and Line 5, where the server samples clients from a non-uniform sampling distribution. In (4.12), we use a weighted average to compute the global gradient.

Recall that B is the local batch size in Algorithm 7 and $K = |S^t|$. Let $D^F := F(w^1) - F^\star$. We then have the following convergence guarantee for Algorithm 7.

Theorem 38. *Assume Assumption 9—11 holds. Let $\{w^1, \dots, w^T\}$ be the sequence of iterates generated by Algorithm 7 and let w^R denote an element of that sequence chosen uniformly*

at random. Let

$$\eta = \frac{K\alpha^3}{MG^2} \sqrt{\frac{2\log M + 4\beta\log(M/\alpha)}{T}}, \quad (4.13)$$

and $\mu_t \equiv \mu$ for all $t \in [T]$, where

$$\mu = \min \left\{ \frac{1}{L}, \frac{1}{\sigma} \sqrt{\frac{D^F K B \alpha}{LT}}, \frac{1}{\zeta_T(\alpha, \beta)} \sqrt{\frac{D^F K}{LT}}, \frac{\sqrt{D^F K} \alpha^{\frac{3}{2}}}{\sqrt{LMT}^{\frac{1}{4}} G \left(\frac{1}{2} \log M + \beta \log(M/\alpha) \right)^{\frac{1}{4}}} \right\},$$

we then have

$$\begin{aligned} & \mathbb{E} \left[\left\| \nabla F(w^R) \right\|^2 \right] \\ & \lesssim \frac{D^F L}{T} + \frac{\sigma \sqrt{D^F L}}{\sqrt{TKB\alpha}} + \frac{\zeta_T(\alpha, \beta) \sqrt{D^F L}}{\sqrt{TK}} + \frac{\sqrt{D^F L} M^{\frac{1}{2}} G}{T^{\frac{3}{4}} K^{\frac{1}{2}} \alpha^{\frac{3}{2}}} \left(\frac{1}{2} \log M + \beta \log(M/\alpha) \right)^{\frac{1}{4}}. \end{aligned} \quad (4.14)$$

Proof. The key proof technique is the construction of a ghost subset $\tilde{S}^t$ in each round that is sampled from a carefully designed comparator sampling distribution. In this way, the regret can be compared with the difference between the convergence rate under OSMD Sampler and the convergence rate under the comparator sampling distribution. Note that $\tilde{S}^t$ is only constructed for theoretical analysis, and does not need to be actually computed in practice. The rest of the proof then follows the regret analysis as in Theorem 40. See detailed proof in Section 4.10.3. $\square$

To see how the convergence rate in Theorem 38 is better than the rate of uniform sampling, recall that the convergence rate of mini-batch SGD under uniform sampling [Ghadimi and Lan, 2013] is

$$R_{\text{unif}}^{\text{MB}} := \frac{D^F L}{T} + \frac{\sigma \sqrt{D^F L}}{\sqrt{TKB}} + \frac{\zeta_{\text{unif}} \sqrt{D^F L}}{\sqrt{TK}}.$$

Denote the right hand side of (4.14) as $R_{\text{osmd}}^{\text{MB}}$, then to have $R_{\text{osmd}}^{\text{MB}} \lesssim R_{\text{unif}}^{\text{MB}}$, we only need

that

$$\frac{\sqrt{D^F L} M^{\frac{1}{2}} G}{T^{\frac{3}{4}} K^{\frac{1}{2}} \alpha^{\frac{3}{2}}} \left(\frac{1}{2} \log M + \beta \log (M/\alpha) \right)^{\frac{1}{4}} \lesssim \frac{(\zeta_{\text{unif}} - \zeta_T(\alpha, \beta)) \sqrt{D^F L}}{\sqrt{TK}},$$

which is equivalent to a requirement that

$$\frac{M^{\frac{1}{2}} G}{\alpha^{\frac{3}{2}}} \left(\frac{\frac{1}{2} \log M + \beta \log (M/\alpha)}{T} \right)^{\frac{1}{4}} \lesssim \zeta_{\text{unif}} - \zeta_T(\alpha, \beta).$$

By treating all the other quantities except for T , ζ_{unif} , and $\zeta_T(\alpha, \beta)$ as constants and setting $\beta = o(T)$, we have the left hand side of the above inequality as $o(1)$. On the other hand, note that $\zeta_T(\alpha, \beta) \leq \zeta_{\text{fix}} \leq \zeta_{\text{unif}}$ for any $\beta \geq 0$ and $0 < \alpha \leq 1$, where ζ_{fix} is defined in (4.10), so whenever $\zeta_{\text{unif}} > \zeta_{\text{fix}}$, we have $\zeta_{\text{unif}} - \zeta_T(\alpha, \beta) \geq \zeta_{\text{unif}} - \zeta_{\text{fix}} = \Omega(1)$. In conclusion, when $\zeta_{\text{unif}} > \zeta_{\text{fix}}$ and by setting $\beta = o(T)$, the OSMD Sampler can achieve a better convergence rate than uniform sampling for mini-batch SGD.

4.3.2 Convergence Analysis of FedAvg with OSMD Sampler

We establish the convergence guarantee for FedAvg when combined with the OSMD Sampler. The complete algorithm is provided in Algorithm 8. Compared to the standard FedAvg algorithm in McMahan et al. [2017], the main modifications in Algorithm 8 are:

- In Line 15, the server updates the sampling distribution using the OSMD Sampler.
- In Line 5, clients are sampled from a non-uniform sampling distribution.
- The global model update (4.15) employs a weighted average.

Moreover, when defining a_m^t , we rescale $\|g_m^t\|^2$ by $(\mu_l^t)^2 B$ to ensure that $\mathbb{E} [\|g_m^t\|^2 / (\mu_l^t)^2 B] = \Theta(1)$ as $\mu_l^t \rightarrow 0$ and $B \rightarrow \infty$.

We now present the main result.

Theorem 39 (Convergence of FedAvg with OSMD Sampler). *Let B denote the local batch size in Algorithm 7, and $K = |S^t|$ be the number of sampled clients at each round. Define*

Algorithm 8 FedAvg with OSMD Sampler

- 1: **Input:** Communication rounds T , clients per round K , local steps B , initial model parameter w^1 , global stepsizes $\{\mu_l^t\}_{t=1}^T$, local stepsizes $\{\mu_l^t\}_{t=1}^T$, learning rate η , and parameter $\alpha \in (0, 1]$.
- 2: **Output:** The final model parameter w^R .
- 3: **Initialize:** $\hat{p}^1 = p^{\text{unif}}$.
- 4: **for** $t = 1, \dots, T$ **do**
- 5: Sample S^t with replacement from $[M]$ with probability $\hat{p}^t$, such that $|S^t| = K$.
- 6: **for** $m \in S^t$ **do**
- 7: Download the current model parameter w^t and let $w_m^{t,0} = w^t$.
- 8: **for** $b = 0, 1, \dots, B-1$ **do**
- 9: Sample $\xi_m^{t,b}$ from $[n_m]$ uniformly random.
- 10: Compute $w_m^{t,b+1} = w_m^{t,b} - \mu_l^t \nabla \phi(w_m^{t,b}; \xi_m^{t,b})$.
- 11: **end for**
- 12: Locally compute $g_m^t = w^t - w_m^{t,B}$ and upload it to the server.
- 13: **end for**
- 14: Server computes

$$a_m^t = \frac{\lambda_m^2 \|g_m^t\|_2^2}{(\mu_l^t)^2 B} = \frac{\lambda_m^2}{B} \left\| \sum_{b=0}^{B-1} \nabla \phi(w_m^{t,b}; \xi_m^{t,b}) \right\|_2^2$$

for $m \in S^t$ and let

$$w^{t+1} = w^t - \frac{\mu^t}{K} \sum_{m \in S^t} \frac{\lambda_m}{\hat{p}_m^t} g_m^t. \quad (4.15)$$

- 15: Server obtains updated sampling distribution $\hat{p}^{t+1}$ by Algorithm 5.
 - 16: **end for**
-

$D^F := F(w^1) - F^*$. Assume that Assumptions 9–11 hold.

Let $\{w^1, \dots, w^T\}$ be the sequence of iterates generated by Algorithm 8, and let w^R be an iterate selected uniformly at random from $\{w^1, \dots, w^T\}$.

Set the learning rate

$$\eta = \frac{K\alpha^3}{MBG^2} \sqrt{\frac{2\log M + 4\beta\log(M/\alpha)}{T}}, \quad (4.16)$$

and let $\mu^t = \mu \geq 1$ and $\mu_l^t = \mu_l$ for all $t \in [T]$, where

$$\mu_l = \min \left\{ \frac{1}{4\mu BL} \sqrt{\frac{1}{2 + 1/\alpha}}, \frac{(D^F)^{1/3}}{\left(4 + \frac{2}{\alpha}\right)^{1/3} \mu BL^{2/3} \left(\zeta_{\text{unif}}^2 + \frac{\sigma^2}{2B}\right)^{1/3} T^{1/3}}, \frac{\sqrt{2D^F}}{\mu B \sqrt{L} \sqrt{\frac{2\zeta_T^2(\alpha, \beta)}{K} + \frac{\sigma^2}{KB\alpha} + \sqrt{\frac{\frac{1}{2} \log M + \beta \log(M/\alpha)}{T}} \sqrt{T}}} \right\}.$$

Then, we have

$$\begin{aligned} & \mathbb{E} \left[\left\| \nabla F(w^R) \right\|^2 \right] \\ & \lesssim \frac{D^F L \sqrt{2 + \frac{1}{\alpha}}}{T} + \frac{\left(4 + \frac{2}{\alpha}\right)^{1/3} (D^F L)^{2/3} \zeta_{\text{unif}}^{2/3}}{T^{2/3}} + \frac{\left(4 + \frac{2}{\alpha}\right)^{1/3} (D^F L)^{2/3} \sigma^{2/3}}{B^{1/3} T^{2/3}} \\ & \quad + \frac{\sqrt{D^F L} \zeta_T(\alpha, \beta)}{\sqrt{TK}} + \frac{\sqrt{D^F L} \sigma}{\sqrt{TKB\alpha}} + \frac{\sqrt{D^F L}}{\sqrt{T}} \left(\frac{\frac{1}{2} \log M + \beta \log(M/\alpha)}{T} \right)^{1/4}. \end{aligned} \quad (4.17)$$

Proof. See Section 4.10.4. The proof follows similar steps as Theorem 38. The key technique is the construction of a ghost subset sampled from the comparator sampling distribution. The remainder of the proof proceeds by applying the regret bound from Theorem 40. $\square$

To see how OSMD Sampler improves over uniform sampling, note that the convergence rate of FedAvg under uniform sampling [Karimireddy et al., 2020b] is

$$R_{\text{unif}}^{\text{Avg}} := \frac{D^F L}{T} + \frac{(D^F L)^{\frac{2}{3}} \zeta_{\text{unif}}^{\frac{2}{3}}}{T^{\frac{2}{3}}} + \frac{(D^F L)^{\frac{2}{3}} \sigma^{\frac{2}{3}}}{B^{\frac{1}{3}} T^{\frac{2}{3}}} + \frac{\sqrt{D^F L} \zeta_{\text{unif}}}{\sqrt{TK}} + \frac{\sqrt{D^F L} \sigma}{\sqrt{TKB}}.$$

Denote the right hand side of (4.17) as $R_{\text{osmd}}^{\text{Avg}}$. By treating all the other quantities except for T , ζ_{unif} and $\zeta_T(\alpha, \beta)$ as constants, to have $R_{\text{osmd}}^{\text{Avg}} \lesssim R_{\text{unif}}^{\text{Avg}}$, we only need that

$$\sqrt{K} \left(\frac{\frac{1}{2} \log M + \beta \log(M/\alpha)}{T} \right)^{\frac{1}{4}} \lesssim \zeta_{\text{unif}} - \zeta_T(\alpha, \beta).$$

Similar to the argument in Section 4.3.1, by setting $\beta = o(T)$, we have the left hand side of the above inequality as $o(1)$. On the other hand, when $\zeta_{\text{unif}} > \zeta_{\text{fix}}$, we have $\zeta_{\text{unif}} - \zeta_T(\alpha, \beta) \geq \zeta_{\text{unif}} - \zeta_{\text{fix}} = \Omega(1)$. Thus, when $\zeta_{\text{unif}} > \zeta_{\text{fix}}$ and by setting $\beta = o(T)$, we have shown that OSMD Sampler can achieve a better convergence rate than uniform sampling for FedAvg.

4.4 Regret Analysis of OSMD Sampler

In this section, we present the regret analysis of the OSMD Sampler, which serves as a key component in the convergence analysis of Section 4.3.

We first introduce the notion of *dynamic regret*, which quantifies the performance of an online algorithm that generates a sequence of sampling distributions $\{\hat{p}^t\}_{t=1}^T$ in a non-stationary environment. For any comparator sequence $q^{1:T} \in \mathcal{P}_{M-1}^T$, the dynamic regret is defined as

$$\text{D-Regret}_T(q^{1:T}) = \bar{L}(\hat{p}^{1:T}) - \bar{L}(q^{1:T}). \quad (4.18)$$

In contrast, the *static regret* measures performance relative to the best fixed sampling distribution, i.e., when $q^1 = \dots = q^T$ [Namkoong et al., 2017, Salehi et al., 2017, Borsos et al., 2018, 2019]. In this case, we write the regret as $\text{D-Regret}_T(q)$.

The variability of a comparator sequence $q^{1:T}$ is captured by its *total variation*, defined as

$$\text{TV}(q^{1:T}) = \sum_{t=1}^{T-1} \|q^{t+1} - q^t\|_1.$$

A larger total variation indicates that the comparator sequence is more dynamic and, therefore, more difficult to track.

We also introduce auxiliary quantities that quantify the distance of q^t from the feasible

set $\mathcal{A}$. For any $q^t \in \mathcal{P}_{M-1}$ and $\alpha \in (0, 1]$, define

$$\begin{aligned} \psi(q^t, \alpha) &:= \sum_{m=1}^M \left(\frac{\alpha}{M} - q_m^t \right) \mathbb{1} \left\{ q_m^t < \frac{\alpha}{M} \right\}, \\ \omega(q^t, \alpha) &:= \frac{\sum_{m=1}^M \left(\frac{\alpha}{M} - q_m^t \right) \mathbb{1} \left\{ q_m^t < \frac{\alpha}{M} \right\}}{\sum_{m=1}^M \left(q_m^t - \frac{\alpha}{M} \right) \mathbb{1} \left\{ q_m^t \geq \frac{\alpha}{M} \right\}}, \quad \phi(q^t, \alpha) := \frac{\omega(q^t, \alpha)}{1 - \omega(q^t, \alpha) \left(1 - \frac{\alpha}{M} \right)}. \end{aligned} \quad (4.19)$$

These quantities will be used to characterize the projection error in the main result below.

Theorem 40 (Regret Bound for OSMD Sampler). *Let $\eta_t \equiv \eta$ for all t , and let $\{\hat{p}^t\}_{t=1}^T$ be the sequence generated by Algorithm 5. Then for any (possibly random) comparator sequence $q^{1:T}$, we have*

$$\begin{aligned} \text{D-Regret}_T(q^{1:T}) &\leq \underbrace{\frac{\log M}{\eta} + \frac{2 \log(M/\alpha)}{\eta} \mathbb{E} \left[\text{TV} \left(q^{1:T} \right) \right] + \frac{\eta M^6}{2K^2 \alpha^6} \sum_{t=1}^T \mathbb{E} \left[\left(a_{\max}^t \right)^2 \right]}_{\text{Intrinsic Regret}} \\ &\quad + \underbrace{\frac{8 \log(M/\alpha)}{\eta} \sum_{t=1}^T \mathbb{E} \left[\psi(q^t, \alpha) \right] + \frac{1}{K} \sum_{t=1}^T \mathbb{E} \left[\phi(q^t, \alpha) l_t(q^t) \right]}_{\text{Projection Error}}, \end{aligned}$$

where $a_{\max}^t := \max_{1 \leq m \leq M} a_m^t = \max_{1 \leq m \leq M} \lambda_m^2 \|g_m^t\|^2$.

Proof. The main technical challenge is to construct a projection of an arbitrary comparator sequence $q^{1:T}$ onto $\mathcal{A}^T$ and control the associated projection error. To the best of our knowledge, this projection error bound for dynamic comparator sequences is novel. The additional difficulty lies in relating the cumulative regret to the total variation of the comparator sequence. The detailed proof is provided in Section 4.10.5. $\square$

From Theorem 40, we see that the bound on the dynamic regret consists of two parts. The first part is the *intrinsic regret*, which quantifies the difficulty of tracking a comparator sequence within $\mathcal{A}^T$. The second part is the *projection error*, which arises from projecting an arbitrary comparator sequence onto $\mathcal{A}^T$.

As shown in Section 4.10.5, we have $0 \leq \omega(q^t, \alpha) \leq 1$ for all $\alpha \in (0, 1]$, which implies that $\phi(q^t, \alpha) \leq M/\alpha$. Furthermore, since

$$\psi(q^t, \alpha) \leq \sum_{m=1}^M \frac{\alpha}{M} \mathbb{1} \left\{ q_m^t < \frac{\alpha}{M} \right\} \leq \alpha,$$

the projection error can be bounded by

$$\frac{8T\alpha \log(M/\alpha)}{\eta} + \frac{M}{\alpha} \sum_{t=1}^T \mathbb{E} \left[l_t(q^t) \right].$$

More importantly, when $q^t \in \mathcal{A}$, we have $\psi(q^t, \alpha) = \omega(q^t, \alpha) = \phi(q^t, \alpha) = 0$. In this case, the projection error vanishes, and only the intrinsic regret remains.

As α decreases from one to zero, the intrinsic regret increases since we are enforcing a stricter constraint on the feasible set $\mathcal{A}$; however, the projection error decreases because the gap between $\mathcal{A}$ and the simplex $\mathcal{P}_{M-1}$ diminishes as $\alpha \rightarrow 0$. An optimal choice of α balances these two sources of regret.

4.5 Adaptive-OSMD Sampler

In this section, we discuss an extension of OSMD Sampler that can automatically choose the learning rate η and is agnostic to the optimization method. There are two main tuning parameters in OSMD Sampler, namely α and η . As we show empirically in Section 4.7.3, the performance of the algorithm is relatively robust to the choice of α . However, the choice of η may have a large effect on the performance of OSMD Sampler. One way to choose η is by minimizing the regret in Theorem 40, which is stated in the following corollary.

Corollary 41. *Let $\eta_t \equiv \eta$ for all t and $\hat{p}^{1:T}$ be a sequence generated by Algorithm 5. Assume that there exists $A_{\max} > 0$ such that $a_{\max}^t \leq A_{\max}$ for all t , where $a_{\max}^t = \max_{1 \leq m \leq M} a_m^t = \max_{1 \leq m \leq M} \lambda_m^2 \|g_m^t\|_2^2$. For any comparator sequence $q^{1:T}$, where q^t is allowed to be random,*

such that $q^t \in \mathcal{A}$ for all $t \in [T]$ and $\mathbb{E}[TV(q^{1:T})] \leq \beta$, let

$$\eta = \frac{K\alpha^3}{M^3 A_{\max}} \sqrt{\frac{2 \log M + 4\beta \log(M/\alpha)}{T}}, \quad (4.20)$$

then

$$D\text{-Regret}_T(q^{1:T}) \leq \frac{M^3 A_{\max}}{K\alpha^3} \sqrt{T \left[\frac{1}{2} \log M + \beta \log(M/\alpha) \right]}. \quad (4.21)$$

The proof of Corollary 41 follows directly from Theorem 40. Note that under Assumption 9–11 and when $\lambda_m = \frac{1}{M}$ for all $m \in [M]$, we have $A_{\max} = \frac{G^2}{M^2}$ for Mini-batch SGD (Algorithm 7) and $A_{\max} = \frac{BG^2}{M^2}$ for FedAvg (Algorithm 4.3.2)⁷, thus (4.20) recovers (4.13) and (4.16).

In practice, since the gradient norm is usually decreasing, we can estimate $A_{\max}$ by adding a pre-training phase where we broadcast the initial model parameter w^0 to all devices before the start of the training, and collect the returned $\|g_m^0\|_2^2$ from all responsive devices, which we denote as S^0 . Then we can estimate $A_{\max}$ by $\hat{A}_{\max} = \max_{m \in S^0} \lambda_m \|g_m^0\|_2^2$.

On the other hand, the optimal choice of β in (4.20) depends on specific problems and is hard to estimate before training starts. Thus, it is preferable to have a tuning strategy that is adaptive to any $\beta > 0$, which we describe in the following.

The main idea is to run a set of expert algorithms, each with a different learning rate for Algorithm 5. We then use a prediction-with-expert-advice algorithm to track the best performing expert algorithm.⁸ More specifically, we define the set of expert learning rates as

$$\mathcal{E} := \left\{ 2^{e-1} \cdot \frac{K\alpha^3}{M^3 A_{\max}} \sqrt{\frac{2 \log M}{T}} \mid e = 1, 2, \dots, E \right\}, \quad (4.22)$$

7. See Section 4.10.3 and Section 4.10.4 for proof.

8. We refer the reader to Cesa-Bianchi and Lugosi [2006, Chapter 2] for an overview of prediction-with-expert-advice algorithms.

Algorithm 9 Adaptive-OSMD Sampler

- 1: **Input:** Meta learning rate γ ; the set of expert learning rates $\mathcal{E} = \{\eta_1 \leq \eta_2 \leq \dots \leq \eta_E\}$ with $E = |\mathcal{E}|$; parameter $\alpha \in (0, 1]$, $\mathcal{A} = \mathcal{P}_{M-1} \cap [\alpha/M, \infty)^M$; number of iterations T ; initial distribution p^{init} .
- 2: **Output:** $\hat{p}^{1:T}$.
- 3: Set $\theta_e^1 = (1 + 1/E)/(e(e+1))$ and $\hat{p}_e^1 = p^{\text{init}}$, $\forall e \in [E]$.
- 4: **for** $t = 1, 2, \dots, T-1$ **do**
- 5: Compute $\hat{p}^t = \sum_{e=1}^E \theta_e^t \hat{p}_e^t$.
- 6: Sample S^t by $\hat{p}^t$.
- 7: **for** $e = 1, 2, \dots, E$ **do**
- 8: Compute $\hat{l}_t(\hat{p}_e^t; \hat{p}^t)$ via (4.6) and $\nabla \hat{l}_t(\hat{p}_e^t; \hat{p}^t)$ via (4.7).
- 9: Solve $\hat{p}_e^{t+1} = \arg \min_{p \in \mathcal{A}} \eta_e \langle p, \nabla \hat{l}_t(\hat{p}_e^t; \hat{p}^t) \rangle + D_\Phi(p \| \hat{p}_e^t)$ via Algorithm 6.
- 10: **end for**
- 11: Update the weight of each expert:

$$\theta_e^{t+1} = \frac{\theta_e^t \exp \left\{ -\gamma \hat{l}_t(\hat{p}_e^t; \hat{p}^t) \right\}}{\sum_{e=1}^E \theta_e^t \exp \left\{ -\gamma \hat{l}_t(\hat{p}_e^t; \hat{p}^t) \right\}}, \quad \forall e \in [E].$$

12: **end for**

where

$$E = \left\lceil \frac{1}{2} \log_2 \left(1 + \frac{4 \log(M/\alpha)}{\log M} (T-1) \right) \right\rceil + 1. \quad (4.23)$$

Then for each $\eta_e \in \mathcal{E}$, Adaptive-OSMD Sampler algorithm runs an expert algorithm to generate a sequence of sampling distributions $\hat{p}_e^{1:T}$. Meanwhile, it also runs a meta-algorithm that uses an exponentially-weighted-average strategy to aggregate $\{\hat{p}_e^{1:T}\}_{e=1}^E$ into a single output $\hat{p}^{1:T}$, which achieves performance close to the best expert.

Algorithm 9 details Adaptive-OSMD Sampler. Note that since we can compute $\hat{l}_t(\hat{p}_e^t; \hat{p}^t)$ and $\nabla \hat{l}_t(\hat{p}_e^t; \hat{p}^t)$ directly, there is no need to use a surrogate loss as in van Erven and Koolen [2016] and Zhang et al. [2018].

From the computational perspective, the major cost comes from solving step 9 of Algorithm 9, which needs to be run for a total number of $T|\mathcal{E}| = O(T \log_2 T)$ times. Compared with Algorithm 5, the computational complexity only increases by $\log_2 T$ times.

We have the following regret guarantee on Algorithm 9.

Theorem 42. Assume that there exists $A_{\max} > 0$ such that $a_{\max}^t \leq A_{\max}$ for all t , where $a_{\max}^t = \max_{1 \leq m \leq M} a_m^t = \max_{1 \leq m \leq M} \lambda_m^2 \|g_m^t\|_2^2$. Let $\hat{p}^{1:T}$ be the output of Algorithm 9 with $\gamma = \frac{\alpha}{M} \sqrt{\frac{8K}{TA_{\max}}}$, $p^{init} = p^{\text{unif}}$ and $\mathcal{E}$ as in (4.22). Then for any comparator sequence $q^{1:T}$, where q^t is allowed to be random, such that $q^t \in \mathcal{A}$ for all $t \in [T]$ and $\mathbb{E}[TV(q^{1:T})] \leq \beta$, we have

$$D\text{-Regret}_T(q^{1:T}) \leq \frac{3M^3 A_{\max}}{K\alpha^3} \sqrt{T \left[\frac{1}{2} \log M + \beta \log(M/\alpha) \right]} + \frac{M}{\alpha} \sqrt{\frac{TA_{\max}}{8K}} (1 + 2 \log E).$$

See the proof in Section 4.10.6. Since the additional regret term is $\tilde{O}((M/\alpha)\sqrt{T/K})$, which is no larger than the first term asymptotically in its dependency on T except for log terms, the bound on the regret is of the same order as in (4.21). However, we do not need to specify β to set the learning rate.

Following Theorem 42 and the proofs of Theorem 38 and Theorem 39, we then have the following optimization guarantees on the Adaptive-OSMD Sampler.

Theorem 43. Assume Assumption 9–11 holds and $\lambda_m = \frac{1}{M}$ for all $m \in [M]$.

- *Mini-batch SGD with Adaptive-OSMD Sampler.* Let $\{w^1, \dots, w^T\}$ be the sequence of iterates generated by Algorithm 7, where in Line 13, the sampling distribution is updated by Algorithm 9 with $A_{\max} = \frac{G^2}{M^2}$. Let w^R denote an element of that sequence chosen uniformly at random. Besides, let $\mu_t = \mu$ for all $t \in [T]$, where

$$\mu = \min \left\{ \frac{1}{L}, \frac{1}{\sigma} \sqrt{\frac{D^F K B \alpha}{LT}}, \frac{1}{\zeta_T(\alpha, \beta)} \sqrt{\frac{D^F K}{LT}}, \frac{\sqrt{D^F K} \alpha^{\frac{3}{2}}}{\sqrt{LMT}^{\frac{1}{4}} G \left(\frac{1}{2} \log M + \beta \log(M/\alpha) \right)^{\frac{1}{4}}}, \sqrt{\frac{\alpha D^F}{LG}} \left(\frac{K}{T} \right)^{\frac{1}{2}} \sqrt{\frac{1}{1 + 2 \log E}} \right\},$$

we then have

$$\begin{aligned}
& \mathbb{E} \left[\left\| \nabla F(w^R) \right\|^2 \right] \\
& \lesssim \frac{D^F L}{T} + \frac{\sigma \sqrt{D^F L}}{\sqrt{TKB\alpha}} + \frac{\zeta_T(\alpha, \beta) \sqrt{D^F L}}{\sqrt{TK}} + \frac{\sqrt{D^F L} M^{\frac{1}{2}} G}{T^{\frac{3}{4}} K^{\frac{1}{2}} \alpha^{\frac{3}{2}}} \left(\frac{1}{2} \log M + \beta \log(M/\alpha) \right)^{\frac{1}{4}} \\
& \quad + \sqrt{\frac{D^F L G}{\alpha}} \left(\frac{1}{K} \right)^{\frac{1}{4}} \left(\frac{1}{T} \right)^{\frac{3}{4}} \sqrt{1 + 2 \log E}.
\end{aligned} \tag{4.24}$$

- *FedAvg with Adaptive-OSMD Sampler.* Let $\{w^1, \dots, w^T\}$ be the sequence of iterates generated by Algorithm 8, where in Line 13, the sampling distribution is updated by Algorithm 9 with $A_{\max} = \frac{BG^2}{M^2}$. Let w^R denote an element of that sequence chosen uniformly at random. Besides, let $\mu^t = \mu \geq 1$ and $\mu_l^t = \mu_l$ for all $t \in [T]$, where

$$\begin{aligned}
\mu_l = \min \left\{ \frac{1}{4\mu BL} \sqrt{\frac{1}{2 + 1/\alpha}}, \frac{(D^F)^{\frac{1}{3}}}{\left(4 + \frac{2}{\alpha}\right)^{\frac{1}{3}} \mu BL^{\frac{2}{3}} \left(\zeta_{unif}^2 + \frac{\sigma^2}{2B}\right)^{\frac{1}{3}} T^{\frac{1}{3}}}, \right. \\
\frac{\sqrt{2D^F}}{\mu B \sqrt{L} \sqrt{\frac{2\zeta_T^2(\alpha, \beta)}{K} + \frac{\sigma^2}{KB\alpha} + \sqrt{\frac{\frac{1}{2} \log M + \beta \log(M/\alpha)}{T}} \sqrt{T}}}, \\
\left. \frac{1}{\mu} \sqrt{\frac{2\alpha D^F}{LG(1 + 2 \log E)}} \left(\frac{1}{B} \right)^{\frac{3}{4}} \left(\frac{8K}{T} \right)^{\frac{1}{4}} \right\},
\end{aligned}$$

we then have

$$\begin{aligned}
\mathbb{E} \left[\left\| \nabla F(w^R) \right\|_2^2 \right] & \lesssim \frac{D^F L \sqrt{2 + \frac{1}{\alpha}}}{T} + \frac{\left(4 + \frac{2}{\alpha}\right)^{\frac{1}{3}} (D^F L)^{\frac{2}{3}} \zeta_{unif}^{\frac{2}{3}}}{T^{\frac{2}{3}}} + \frac{\left(4 + \frac{2}{\alpha}\right)^{\frac{1}{3}} (D^F L)^{\frac{2}{3}} \sigma^{\frac{2}{3}}}{B^{\frac{1}{3}} T^{\frac{2}{3}}} \\
& \quad + \frac{\sqrt{D^F L} \zeta_T(\alpha, \beta)}{\sqrt{TK}} + \frac{\sqrt{D^F L} \sigma}{\sqrt{TKB\alpha}} + \frac{\sqrt{D^F L}}{\sqrt{T}} \left(\frac{\frac{1}{2} \log M + \beta \log(M/\alpha)}{T} \right)^{\frac{1}{4}} \\
& \quad + \sqrt{\frac{D^F L G}{\alpha}} \left(\frac{1}{KB} \right)^{\frac{1}{4}} \left(\frac{1}{T} \right)^{\frac{3}{4}} \sqrt{1 + 2 \log E}. \tag{4.25}
\end{aligned}$$

Compare (4.14) with (4.24) (respectively, (4.17) with (4.25)), there is an additional error term in (4.24) (respectively, (4.25)). However, this additional error term is no larger than the previous error term when considering its dependency on T . Furthermore, when we let $\beta = T^c$ with $0 < c < 1$ and treat all the other parameters except for β and T to be constants, we then have the last term dominated by the penultimate term. On the other hand, in exchange for the additional error, we do not need to specify β when setting the learning rate η .

Note that Adaptive-OSMD Sampler still requires the number of iterations T as an input; however, in some applications, T is not available before the training starts. For example, one may use some stopping criterion to determine when to stop. To deal with such cases, in Section 4.6.1, we introduce an extension of Adaptive-OSMD Sampler that does not need T as input by using the doubling trick. The extension algorithm enjoys similar regret and optimization guarantees as Adaptive-OSMD Sampler. See Section 4.6.1 for more details.

4.6 Extensions of Adaptive-OSMD Sampler

In this section, we discuss additional extensions of Adaptive-OSMD Sampler. In Section 4.6.1, we discuss how to choose η without knowing T in advance. In Section 4.6.2, we discuss how to extend Adaptive-OSMD Sampler to sample without replacement setting.

4.6.1 Adaptive-OSMD Sampler with Doubling Trick

Algorithm 9 requires the total number of iterations T as input, which is not always available in practice. In those cases, we use the doubling trick [Cesa-Bianchi and Lugosi, 2006, Section 2.3] to avoid this requirement. The basic idea is to restart the Adaptive-OSMD Sampler at exponentially increasing time points $T_b = 2^{b-1}$, $b \geq 1$. The learning rates of experts in Algorithm 9 are reset at the beginning of each time interval, and the meta-algorithm learning rate γ is chosen optimally for the interval length.

More specifically, let $A_{\max} > 0$ such that $a_{\max}^t \leq A_{\max}$ for all t , where $a_{\max}^t =$

$\max_{1 \leq m \leq M} a_m^t = \max_{1 \leq m \leq M} \lambda_m^2 \|g_m^t\|_2^2$. At the time point T_b , we let

$$\mathcal{E}_b := \left\{ 2^{e-\frac{b}{2}-\frac{1}{2}} \cdot \frac{K\alpha^3 \sqrt{2 \log M}}{M^3 A_{\max}} \mid e = 1, 2, \dots, E_b \right\}, \quad (4.26)$$

where

$$E_b = \left\lceil \frac{1}{2} \log_2 \left(1 + \frac{4 \log(M/\alpha)}{\log M} (2^{b-1} - 1) \right) \right\rceil + 1, \quad (4.27)$$

and $\gamma_b = \frac{\alpha}{M} \sqrt{\frac{8K}{2^{b-1} A_{\max}}}$, $b \geq 1$.

In a practical implementation, at the time point $t = T_b$, instead of initializing all expert algorithms using a uniform distribution, we can initialize them with the output of the meta-algorithm for $t = T_b - 1$. Besides, since the gradient norm is usually decreasing, we can estimate $A_{\max}$ by adding a pre-training phase where we broadcast the initial model parameter w^0 to all devices before the start of the training, and collect the returned $\|g_m^0\|_2^2$ from all responsive devices, which we denote as S^0 . Then we can estimate $A_{\max}$ by $\hat{A}_{\max} = \max_{m \in S^0} \lambda_m \|g_m^0\|_2^2$.

Adaptive-Doubling-OSMD Sampler is detailed in Algorithm 10. From the computational perspective, by the proof of Theorem 44, Algorithm 10 needs to run Step 9 of Algorithm 9 for a total number of $O(T|\mathcal{E}|^2) = O(T \lceil \log_2 T \rceil)$ times. Therefore, the computational complexity of Adaptive-Doubling-OSMD Sampler is asymptotically the same as that of Adaptive-OSMD Sampler, while it increases by only a $\log(T)$ factor compared to OSMD Sampler. The following theorem provides a bound on the dynamic regret for Adaptive-Doubling-OSMD Sampler.

Theorem 44. *Suppose the training is stopped after T iterations. Let $\hat{p}^{1:T}$ be the output of Algorithm 10, where p^{unif} is used in Step 6. Then for any comparator sequence $q^{1:T}$, where*

Algorithm 10 Adaptive-OSMD Sampler with Doubling Trick (Adaptive-Doubling-OSMD)

```
1: Input: Parameters  $\alpha$  and  $A_{\max}$ .
2: Output:  $\hat{p}^t$  for  $t = 1, \dots, T$ .
3: while True do
4:   Set  $\mathcal{E}_b$  as in (4.26).
5:   Let  $\gamma_b = \frac{\alpha}{M} \sqrt{\frac{8K}{2^{b-1}A_{\max}}}$ .
6:   Obtain  $\{\hat{p}^t\}_{t=2^{b-1}}^{2^b-1}$  from Algorithm 9 with parameters:  $\gamma_b$ ,  $\mathcal{E}_b$ ,  $\alpha$ , the number of iterations  $2^{b-1}$ , and the initial distribution  $p^{\text{unif}}$  or  $\hat{p}^{2^{b-1}-1}$  (when  $b > 1$ ).
7:   if Training Process is Converged then
8:     Break.
9:   end if
10:  Let  $b \leftarrow b + 1$ .
11: end while
```

q^t is allowed to be random, such that $q^t \in \mathcal{A}$ for all $t \in [T]$ and $\mathbb{E}[TV(q^{1:T})] \leq \beta$, we have

$$\begin{aligned} D\text{-Regret}_T(q^{1:T}) &\leq \frac{6M^3 A_{\max}}{(\sqrt{2}-1)K\alpha^3} \sqrt{T \left[\frac{1}{2} \log M + \beta \log(M/\alpha) \right]} \\ &\quad + \frac{2M}{(\sqrt{2}-1)\alpha} \sqrt{\frac{TA_{\max}}{8K}} (1 + 2 \log E). \end{aligned}$$

See the proof in Section 4.10.8. Comparing Theorem 44 with Theorem 42, we see that the regret bound of Adaptive-Doubling-OSMD has the same order as that of Adaptive-OSMD Sampler. However, Adaptive-Doubling-OSMD Sampler does not need to know T in advance. Mimicking the proof of Theorem 43, we can also show the optimization guarantees on Adaptive-Doubling-OSMD with mini-batch SGD and FedAvg, which are basically the same as Theorem 43, thus are omitted here.

4.6.2 Adaptive Sampling Without Replacement

In the discussion so far, we have assumed that the set S^t is obtained by sampling with replacement from p^t . When K is relatively large compared to M and p^t is far from uniform distribution, sampling without replacement can be more efficient than sampling with replacement. However, when sampling without replacement using p^t , the variance reduction

Algorithm 11 Adaptive sampling without replacement

```
1: Input:  $w^1$  and  $\hat{p}^1$ .
2: for  $t = 1, 2, \dots, T - 1$  do
3:   Let  $\hat{p}_{(1)}^t = \hat{p}^t$  and sample  $m_1^t$  from  $[M]$  by  $\hat{p}_{(1)}^t$ .
4:   for  $k = 2, \dots, K$  do
5:     /* Design the sampling distribution for sampling the  $k$ -th client
6:       in the  $t$ -th round */
7:     Construct  $\hat{p}_{(k)}^t$  by letting
      
$$\hat{p}_{(k),m}^t = \begin{cases} \left(1 - \sum_{l=1}^{k-1} \hat{p}_{m_l^t}^t\right)^{-1} \hat{p}_m^t & \text{if } m \in [M] \setminus \{m_1^t, \dots, m_{k-1}^t\} \\ 0 & \text{otherwise.} \end{cases}$$

8:     /* Sample the  $k$ -th client */
9:     Sample  $m_k^t$  from  $[M] \setminus \{m_1^t, \dots, m_{k-1}^t\}$  by  $\hat{p}_{(k)}^t$ .
10:   end for
11:   Let  $S^t = \{m_1^t, \dots, m_K^t\}$ .
12:   The server broadcasts the model parameter  $w^t$  to clients in  $S^t$ .
13:   The clients in  $S^t$  compute and upload the set of local gradients  $\{g_{m_1^t}^t, \dots, g_{m_K^t}^t\}$ .
14:   /* Construct global gradient estimate */
15:   Let  $g_{(1)}^t = \lambda_{m_1^t}^t g_{m_1^t}^t / \hat{p}_{(1),m_1^t}^t$ .
16:   for  $k = 2, \dots, K$  do
17:     Let  $g_{(k)}^t = \lambda_{m_k^t}^t g_{m_k^t}^t / \hat{p}_{(k),m_k^t}^t + \sum_{l=1}^{k-1} \lambda_{m_l^t}^t g_{m_l^t}^t$ .
18:   end for
19:   Let  $\tilde{g}^t = K^{-1} \sum_{k=1}^K g_{(k)}^t$ .
20:   /* Update the model weight based on the global gradient estimate */
21:   Obtain the updated model parameter  $w^{t+1}$  using  $w^t$  and  $\tilde{g}^t$ .
22:   /* Update sampling distribution */
23:   Let  $a_m^t = \lambda_m^2 \|g_m^t\|^2$  for  $m \in S^t$ .
24:   Input  $\{a_m^t\}_{m \in S^t}$  into Adaptive-OSMD Sampler to get  $\hat{p}^{t+1}$ .
25: end for
```

loss does not have a clean form as in (4.4). As a result, an online design of the sampling distribution is more challenging. In this section, we discuss how to use the sampling distribution obtained by Adaptive-OSMD Sampler to sample clients without replacement, following the approach taken in Hanchi and Stephens [2020].

The detailed sampling procedure is described in Algorithm 11. We still use the Adaptive-

OSMD Sampler to update the sampling distribution. However, we use the designed sampling distribution in a way that no client is chosen twice. Furthermore, Step 18 of Algorithm 11 constructs the gradient estimate with the following properties.

Proposition 45 (Proposition 3 of Hanchi and Stephens [2020]). *Let $\tilde{p}^t = p$ and let $\tilde{g}^t$ be as in Step 18 of Algorithm 11. Note that $\tilde{g}^t = \tilde{g}^t(p)$ depends on p . Recall that $J^t = \sum_{m=1}^M \lambda_m g_m^t$. We have*

$$\mathbb{E}_{S^t} [\tilde{g}^t] = J^t \quad \text{and} \quad \arg \min_{p \in \mathcal{P}_{M-1}} \mathbb{E}_{S^t} [\|\tilde{g}^t - J^t\|_2^2] = \arg \min_{p \in \mathcal{P}_{M-1}} l_t(p),$$

where $l_t(\cdot)$ is defined in (4.4) and the expectation is taken over S^t .

From Proposition 45, we see that $\tilde{g}^t$ is an unbiased stochastic gradient. Furthermore, the variance of $\tilde{g}^t$ is minimized by the same sampling distribution that minimizes the variance reduction loss in (4.4). Therefore, it is reasonable to use the sampling distribution generated by the Adaptive-OSMD Sampler to design $\tilde{g}^t$.

4.7 Simulation Experiments

In this section, we use simulated data to demonstrate the performance of Adaptive-OSMD Sampler (Algorithm 9). We compare our method against uniform sampling in Section 4.7.1 and compare against other bandit feedback online learning samplers in Section 4.7.2. In addition, we examine the robustness of Adaptive-OSMD Sampler to the choice of α in Section 4.7.3, while in Section 4.7.4, we compare Adaptive-OSMD Sampler with the Lipschitz constant-based importance sampling.

We generate data as follows. We set the number of clients as $M = 100$, and each client has $n_m = 100$ samples, $m \in [M]$. Samples on each client are generated as

$$y_{m,i} = \langle w_\star, x_{m,i} \rangle + N(0, 0.1^2), \quad i \in [n_m], \quad (4.28)$$

where the coefficient vector $w_\star \in \mathbb{R}^d$ has elements generated as i.i.d. $N(10, 3)$, and the feature vector $x_{m,i} \in \mathbb{R}^d$ is generated as $x_{m,i} \sim N(0, \Sigma_m)$, where $\Sigma_m = s_m \cdot \Sigma$, Σ is a diagonal matrix with $\Sigma_{jj} = \kappa^{(j-1)/(d-1)-1}$, $\forall j \in [d]$ and $\kappa > 0$ is the condition number of Σ . We generate $\{s_m\}_{m=1}^M$ i.i.d. from $e^{N(0, \sigma^2)}$ and rescale them as $s_m \leftarrow (s_m / \max_{m \in [M]} s_m) \times 10$ so that $s_m \leq 10$ for all $m \in [M]$. In this setting, κ controls the difficulty of each problem when solved separately, while σ controls the level of heterogeneity across clients. In all experiments, we fix $\kappa = 25$, which corresponds to a hard problem, and change σ to simulate different heterogeneity levels. We expect that uniform sampling suffers when the heterogeneity level is high. The dimension d of the problem is set as $d = 10$. The results are averaged over 10 independent runs.

We use the mean squared error loss defined as

$$L(w) = \frac{1}{M} \sum_{m=1}^M L_m(w), \quad \text{where} \quad L_m(w) = \frac{1}{2n_m} (y_{m,i} - \langle w_\star, x_{m,i} \rangle)^2.$$

We use the stochastic gradient descent to make global updates. At each round t , we choose a subset of $K = 5$ clients, denoted as S^t . For each client $m \in S^t$, we choose a mini-batch of samples, $\mathcal{B}_m^t$, of size $\bar{B} = 10$, and compute the mini-batch stochastic gradient. The parameter w is updated as

$$w^{t+1} = w^t + \frac{\mu_{\text{SGD}}}{MK\bar{B}} \sum_{m \in S^t} \frac{1}{p_m^t} \sum_{i \in \mathcal{B}_m^t} (y_{m,i} - \langle w_\star, x_{m,i} \rangle) \cdot x_{m,i},$$

where μ_{SGD} is the learning rate, set as $\mu_{\text{SGD}} = 0.1$ in simulations.

In all experiments, we set α in Adaptive-OSMD Sampler as $\alpha = 0.4$. The tuning parameters for MABS, VRB, and Avare are set as in their original papers.

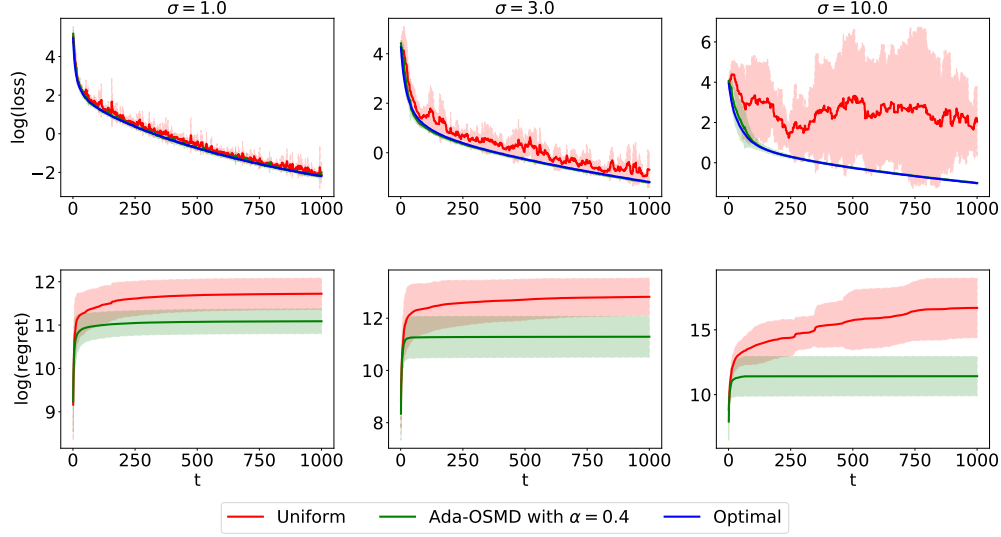


Figure 4.1: The training loss (top row) and cumulative regret (bottom row) are compared for the Adaptive-OSMD Sampler, Uniform Sampler, and Optimal Sampler, under $\sigma = 1.0$, $\sigma = 3.0$, and $\sigma = 10.0$. Solid lines represent the mean values, while the shaded regions indicate mean $\pm$ standard deviation across independent runs.

4.7.1 Adaptive-OSMD Sampler vs Uniform Sampling

The results of the training process and the cumulative regret are shown in Figure 4.1. For the training loss, we see that when the heterogeneity level is low ($\sigma = 1.0$), the uniform sampling performs as well as Adaptive-OSMD Sampler and theoretically optimal sampling; however, as the heterogeneity level increases, the performance of uniform sampling gradually suffers; when $\sigma = 10.0$, uniform sampling performs poorly. On the other hand, Adaptive-OSMD Sampler performs well across all levels of heterogeneity and is very close to the theoretically optimal sampling. Similarly, for the cumulative regret, when the heterogeneity level is low, the cumulative regret of uniform sampling is close to Adaptive-OSMD Sampler; however, when the heterogeneity level increases, the cumulative regret of uniform sampling gets much larger than Adaptive-OSMD Sampler. Based on the above results, we can conclude that while the widely used choice of uniform sampling may be reasonable when heterogeneity is low, our proposed sampling strategy is robust across different levels of heterogeneity, and thus should be considered as the default option.

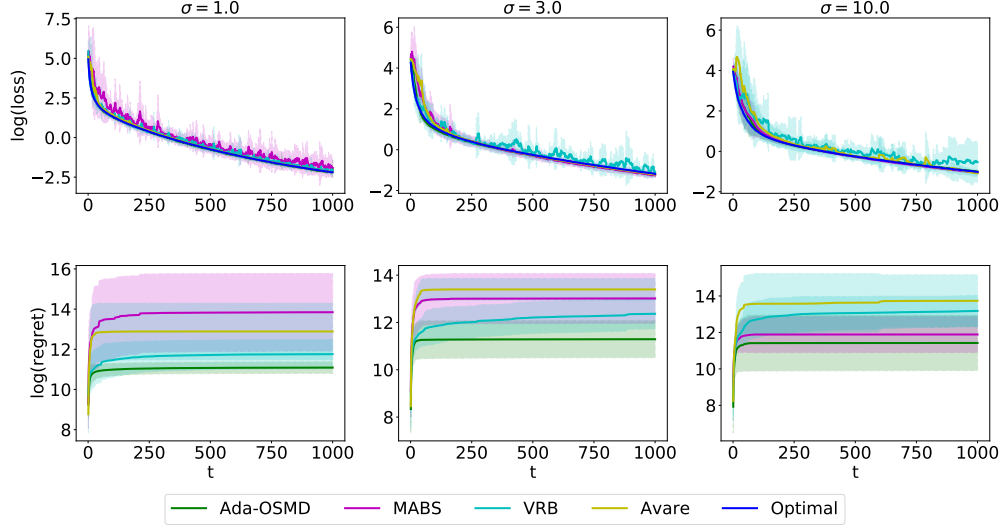


Figure 4.2: The training loss (top row) and cumulative regret (bottom row) are compared across the Adaptive-OSMD Sampler, MABS, VRB, and Avare methods for $\sigma = 1.0$, $\sigma = 3.0$, and $\sigma = 10.0$. Solid lines represent the mean values, while shaded regions indicate mean $\pm$ standard deviation across independent runs.

4.7.2 Adaptive-OSMD Sampler vs MABS vs VRB vs Avare

We compare Adaptive-OSMD Sampler to other bandit feedback online learning samplers: MABS [Salehi et al., 2017], VRB [Borsos et al., 2018] and Avare [Hanchi and Stephens, 2020]. Training loss and cumulative regret are shown in Figure 4.2. We see that while VRB and Avare perform better when the heterogeneity level is low and MABS performs better when the heterogeneity level is high, Adaptive-OSMD Sampler always achieves the best in both training loss and cumulative regret across all different levels of heterogeneity. Thus, we conclude that Adaptive-OSMD is a better choice than other online learning samplers.

4.7.3 Robustness of Adaptive-OSMD Sampler to the Choice of α

We examine the robustness of Adaptive-OSMD Sampler to the choice of α . We run Adaptive-OSMD Sampler separately for each $\alpha \in \{0.01, 0.1, 0.4, 0.7, 0.9, 1.0\}$. Note that when $\alpha = 1.0$, the Adaptive-OSMD Sampler outputs a uniform distribution. Training loss and cumulative regret are shown in Figure 4.3. We observe that Adaptive-OSMD Sampler is robust to the

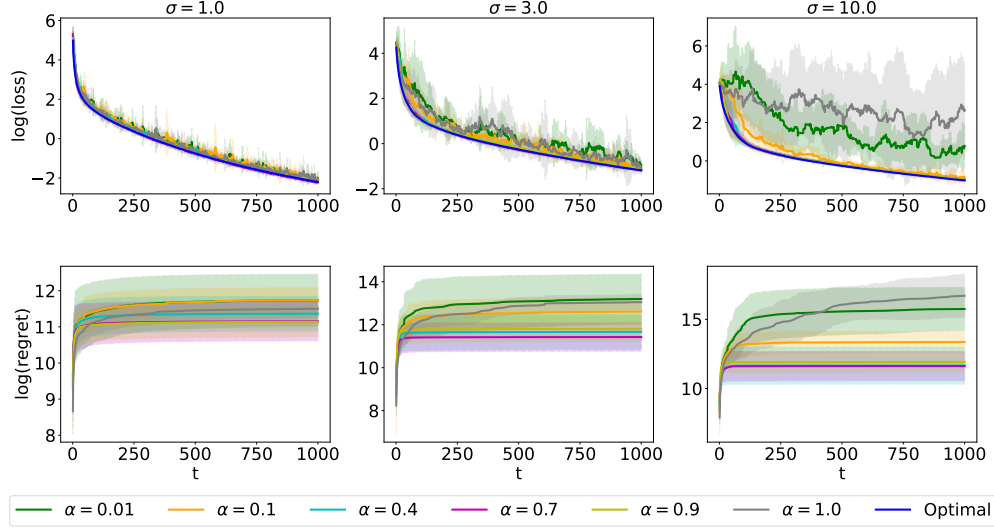


Figure 4.3: The training loss (top row) and cumulative regret (bottom row) are shown for the Adaptive-OSMD Sampler with different values of α under $\sigma = 1.0$, $\sigma = 3.0$, and $\sigma = 10.0$. Solid lines represent the mean values, while shaded regions indicate mean $\pm$ standard deviation across independent runs.

choice of α , and performs well as long as α is not too close to zero or too close to one.

4.7.4 Dynamic Sampling Distribution v.s. Fixed Sampling Distribution

In this chapter, we allow both our sampling distribution and competitor sampling distribution to change over time, while previous studies either use a fixed sampling distribution [Zhao and Zhang, 2015, Needell et al., 2016] or they compare against a fixed sampling distribution [Namkoong et al., 2017, Salehi et al., 2017, Borsos et al., 2018, 2019]. In this section, we show that under certain settings, a dynamic sampling distribution can achieve a significant advantage over a fixed sampling distribution. More specifically, we compare the Adaptive-OSMD Sampler with the Lipschitz constant-based importance sampling distribution proposed by Zhao and Zhang [2015], Needell et al. [2016], which we denote as p^{IS} .

We still use the same model as in (4.28) to generate data. but we generate $w_\star$ and $x_{m,i}$ differently. Motivated by Zhao et al. [2023b], for each $m \in [M]$, we choose uniformly at random one dimension among $\mathbb{R}^d$, denoted as $\text{supp}(m) \in [d]$, as the support of $x_{m,i}$ for all $i \in [n_m]$, while the remaining dimensions of $x_{m,i}$ are set to be zero. The nonzero dimension

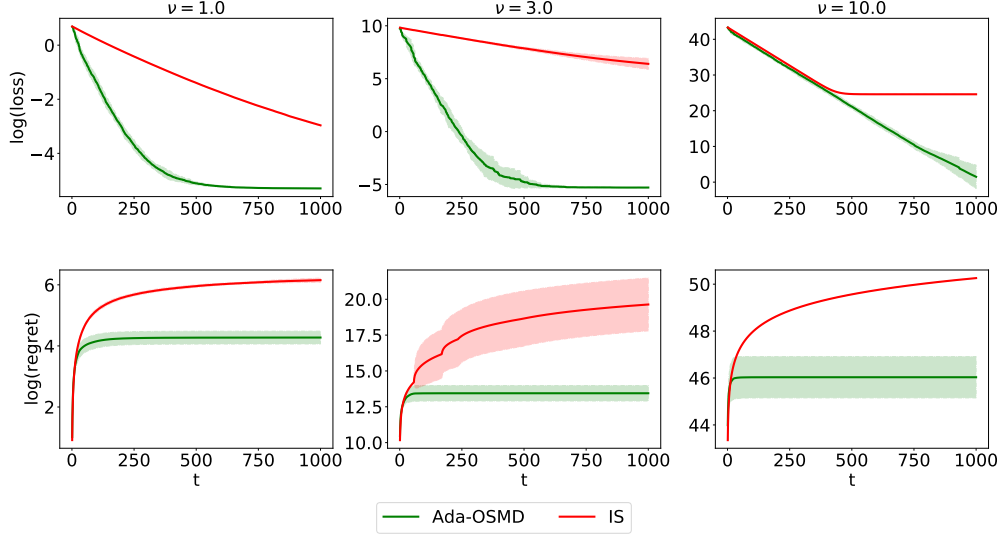


Figure 4.4: The training loss (top row) and cumulative regret (bottom row) are compared between the Adaptive-OSMD Sampler and p^{IS} for $\nu = 1.0$, $\nu = 3.0$, and $\nu = 10.0$. Solid lines represent the mean values, while shaded regions illustrate mean $\pm$ standard deviation across independent runs.

of $x_{m,i}$ is generated from $N(1.0, 0.1^2)$. The entries of $w_\star$ are generated i.i.d. from $e^{N(0, \nu^2)}$. Therefore, ν controls the variance of entries of $w_\star$.

Besides, we choose the optimal stepsize from the set $\{1.0, 0.5, 0.1, 0.05, 0.01\}$ for each method separately. The final result is shown in Figure 4.4. We see that Adaptive-OSMD Sampler performs better than p^{IS} across all levels of ν . Note that in practice, in order to implement p^{IS} , we need prior information about Lipschitz constants of $L_m(\cdot)$'s, while Adaptive-OSMD Sampler does not need prior information. This way, our proposed method does not only have better practical performance, but also requires less prior information.

4.7.5 Comparison Between Sampling with Replacement and Sampling without Replacement

Following the same simulation setup as in Section 4.7, we empirically compare sampling with replacement and sampling without replacement procedure proposed in Section 4.6.2 when used together with the Adaptive-OSMD sampler. Training loss and cumulative regret are

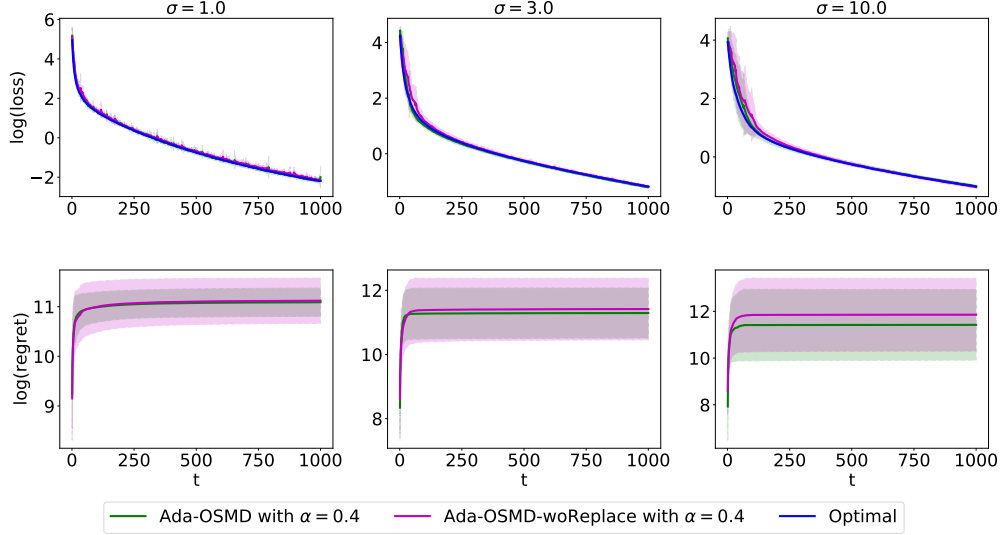


Figure 4.5: The training loss (top row) and cumulative regret (bottom row) are compared for the Adaptive-OSMD Sampler with replacement and without replacement, across $\sigma = 1.0$, $\sigma = 3.0$, and $\sigma = 10.0$. Solid lines represent the mean values, while shaded regions indicate mean $\pm$ standard deviation across independent runs.

shown in Figure 4.5. We observe that using sampling with replacement results in a slightly smaller cumulative regret and a slightly better training loss. However, these differences are not significant.

4.8 Real Data Experiment

We compare Adaptive-OSMD Sampler with uniform sampling and other online learning samplers including MABS [Salehi et al., 2017], VRB [Borsos et al., 2018] and Avare [Hanchi and Stephens, 2020] on real data. We use three commonly used computer vision data sets: MNIST [LeCun and Cortes, 2010]⁹, KMINIST [Clanuwat et al., 2018]¹⁰, and FMINIST [Xiao et al., 2017]¹¹. We set the number of devices to be $M = 500$. To better simulate the situation

9. Yann LeCun and Corinna Cortes hold the copyright of MNIST data set, which is a derivative work from original NIST data sets. MNIST data set is made available under the terms of the Creative Commons Attribution-Share Alike 3.0 license.

10. KMINIST data set is licensed under a permissive CC BY-SA 4.0 license, except where specified within some benchmark scripts.

11. FMINIST data set is under The MIT License (MIT) Copyright © [2017] Zalando SE, <https://tech.zalando.com>

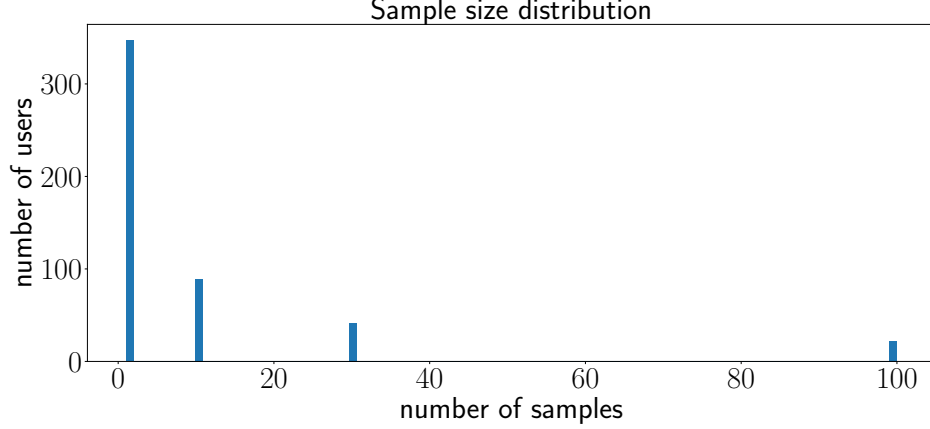


Figure 4.6: The sample size distribution in the training set across clients.

where our method brings significant convergence speed improvement, we create a highly skewed sample size distribution of the training set among clients: 65% of clients have only one training sample, 20% of clients have 5 training samples, 10% of clients have 30 training samples, and 5% of clients have 100 training samples. This setting tries to illustrate a real-life situation where most of the data come from a small fraction of users, while most of the users have only a small number of samples. The skewed sample size distribution is common in other FL data sets, such as LEAF [Caldas et al., 2018]. The sample size distribution in the training set is shown in Figure 4.6. In addition, each client has 10 validation samples used to measure the prediction accuracy of the model over the training process.

We use a multi-class logistic regression model. For a given gray scale picture with the label $y \in \{1, 2, \dots, C\}$, we unroll its pixel matrix into a vector $x \in \mathbb{R}^p$. Given a parameter matrix $W \in \mathbb{R}^{C \times p}$, the training loss function defined in (4.1) is

$$\phi(W; x, y) := l_{\text{CE}}(\varsigma(Wx); y),$$

where $\varsigma(\cdot) : \mathbb{R}^C \rightarrow \mathbb{R}^C$ is the softmax function defined as

$$[\varsigma(x)]_i = \frac{\exp(x_i)}{\sum_{j=1}^K \exp(x_j)}, \quad \text{for all } x \in \mathbb{R}^C,$$

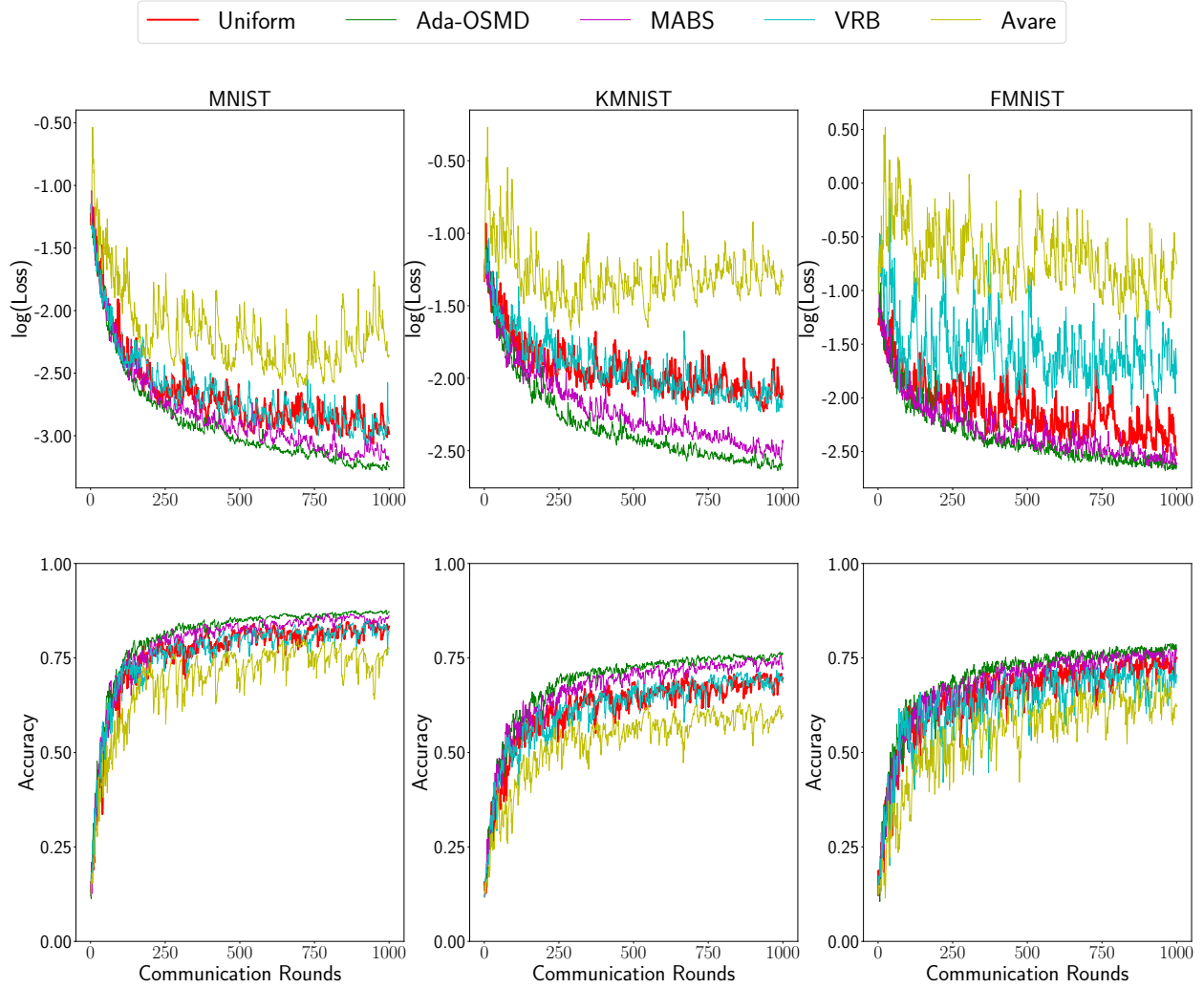


Figure 4.7: Comparison of Adaptive-OSMD Sampler, Uniform Sampler, and other online learning samplers on real datasets. Each column corresponds to a different dataset. The top row shows training loss, and the bottom row shows validation accuracy. Results are averaged over five independent runs.

and $l_{\text{CE}}(x; y) = \sum_{i=1}^C \mathbb{1}(y = i) \log x_i$, $x \in \mathbb{R}^C$, $y \in \{1, \dots, C\}$, is the cross-entropy function.

We use the same algorithms and tuning parameters as in Section 4.7. Learning rate in SGD is set to 0.075 for MNIST and KMNIST, and is set to 0.03 for FMNIST. The total number of communication rounds is to 1,000. In each round of communication, we choose $K = 10$ clients to participate (2% of the total number of clients). For a chosen client m , we compute its local mini-batch gradient with the batch size equal to $\min\{5, n_m\}$, where n_m is

the training sample size on the client m .

Figure 4.7 shows both the training loss and validation accuracy. Each figure shows the average performance over 5 independent runs. We use the same random seed for both Adaptive-OSMD Sampler and competitors, and change random seeds across different runs. The main focus is on minimizing the training loss, and the validation accuracy is only included for completeness. We observe that Adaptive-OSMD Sampler performs better than uniform sampling and other online learning samplers across all data sets.

4.9 Conclusion

We studied the client sampling problem in FL. We proposed an online learning with bandit feedback approach to tackle client sampling. We used online stochastic mirror descent to solve the online learning problem and applied the online ensemble method to choose the tuning parameters. We established an upper bound on the dynamic regret relative to any sequence of sampling distributions. Besides, we provide optimization guarantees for our sampling method when used with mini-batch SGD and FedAvg. Extensive numerical experiments demonstrated the benefits of our approach over both widely used uniform sampling and other competitors.

In this chapter, we have focused on sampling with replacement. However, sampling without replacement would ideally be a more efficient approach. In Section 4.6.2, we discussed a natural extension of the Adaptive-OSMD Sampler to a setting where sampling without replacement is used. However, this approach does not directly minimize the variance of the gradient g^t . When sampling without replacement is used, the variance function becomes more complicated, and the design of an algorithm to directly minimize the variance is an interesting future direction.

Besides, in federated learning, privacy is a major concern. In this chapter, the non-uniform sampling distribution may make the protection of clients' privacy more challenging than uniform sampling. One possible solution is to add noise to the gradient feedback

and protect the clients' privacy under the Differential Privacy (DP) concept [Dwork, 2008]. However, the added noise may hurt the performance of our sampling design and increase the regret. Studying the trade-off between privacy protection and regret is an important direction for addressing societal concerns in real-world applications.

Other fruitful future directions include the design of sampling algorithms for minimizing personalized FL objectives and sampling with physical constraints in the FL system, which we discuss in the following.

4.9.1 *Client Sampling with Personalized FL Objective*

Data distributions across clients are often heterogeneous. Personalized FL has emerged as one effective way to handle such heterogeneity [Kulkarni et al., 2020]. Hanzely et al. [2023] illustrated how many existing approaches to personalization can be studied through a unified framework, and, in this section, we discuss a natural extension of Adaptive-OSMD Sampler to this personalized objective. Specifically, we study the following optimization problem

$$\min_{w, \beta} F(w, \beta) := \sum_{m=1}^M \lambda_m \phi(w, \beta_m; \mathcal{D}_m), \quad (4.29)$$

where $w \in \mathbb{R}^{d_0}$ corresponds to the shared parameter and $\beta = (\beta_1, \dots, \beta_M)$ with $\beta_m \in \mathbb{R}^{d_m}$ corresponds to the local parameters. The objective in (4.29) covers a wide range of personalized federated learning problems [Hanzely et al., 2023]. We further generalize the approach and study the following bilevel optimization problem:

$$\begin{aligned} \min_w h(w) &:= \sum_{m=1}^M \lambda_m F_m(w, \hat{\beta}_m(w)) := \sum_{m=1}^M \lambda_m \phi(w, \hat{\beta}_m(w); \mathcal{D}_m) \\ \text{subject to } \hat{\beta}_m(w) &= \arg \min_{\beta_m} G_m(w, \beta_m) := \phi(w, \beta_m; \bar{\mathcal{D}}_m). \end{aligned} \quad (4.30)$$

When $\mathcal{D}_m = \bar{\mathcal{D}}_m$, then (4.30) recovers (4.29). When $\bar{\mathcal{D}}_m \neq \mathcal{D}_m$, we then optimize the shared and local parameters on different data sets, which may prevent overfitting. The formulation in (4.30) is closely related to the implicit MAML [Rajeswaran et al., 2019].

In the following, we use ∇_w to denote a partial derivative with respect to w with β_m fixed, ∇_{β_m} to denote a partial derivative with respect to β_m with w fixed, and ∇ to denote a derivative with respect to w where $\beta_m(w)$ is treated as a function of w . Let $\nabla_{\beta_m \beta_m^\top}^2 G_m(w, \beta_m) \in \mathbb{R}^{d_m \times d_m}$ be the Hessian matrix of G_m with respect to β_m where w is fixed, and $\nabla_{w \beta_m^\top}^2 G_m(w, \beta_m) \in \mathbb{R}^{d_0 \times d_m}$ be the Hessian matrix of G_m with respect to w and β_m , that is,

$$\begin{aligned} \left[\nabla_{\beta_m \beta_m^\top}^2 G_m(w, \beta_m) \right]_{i,j} &= \frac{\partial G_m(w, \beta_m)}{\partial \beta_{m,i} \beta_{m,j}} \quad \text{for all } i, j = 1, 2, \dots, d_m, \\ \left[\nabla_{w \beta_m^\top}^2 G_m(w, \beta_m) \right]_{i,j} &= \frac{\partial G_m(w, \beta_m)}{\partial w_i \beta_{m,j}} \quad \text{for all } i = 1, 2, \dots, d_0, j = 1, 2, \dots, d_m. \end{aligned}$$

By the implicit function theorem, we have

$$\nabla h(w) = \underbrace{\frac{1}{M} \sum_{m=1}^M \lambda_m \nabla_1 F_m(w, \hat{\beta}_m(w))}_{\nabla_1 h(w)} + \underbrace{\frac{1}{M} \sum_{m=1}^M \lambda_m \nabla_2 F_m(w, \hat{\beta}_m(w))}_{\nabla_2 h(w)} \quad (4.31)$$

where

$$\begin{aligned} \nabla_1 F_m(w, \hat{\beta}_m(w)) &:= \nabla_w F_m(w, \hat{\beta}_m(w)), \\ \nabla_2 F_m(w, \hat{\beta}_m(w)) &:= -\nabla_{w \beta_m^\top}^2 G_m(w, \hat{\beta}_m(w)) \left[\nabla_{\beta_m \beta_m^\top}^2 G_m(w, \hat{\beta}_m(w)) \right]^{-1} \nabla_{\beta_m} F_m(w, \hat{\beta}_m(w)). \end{aligned}$$

There are two parts to $\nabla h(w^t)$ and, therefore, instead of choosing a single subset of clients for computing both parts, we decouple S^t into two subsets S_1^t and S_2^t , $S^t = S_1^t \cup S_2^t$. We use clients in S_1^t to compute local updates of the first part, and clients in S_2^t to compute the local updates of the second part. To get an estimate of $\nabla h(w)$, we can estimate $\nabla_1 h(w)$ and

$\nabla_2 h(w)$ separably and then combine. Assume that $g_{1,m}^t$ is an estimate of $\nabla_1 F_m(w, \hat{\beta}_m(w))$ and $g_{2,m}^t$ is an estimate of $\nabla_2 F_m(w, \hat{\beta}_m(w))$, we can then construct estimates of $\nabla_1 h(w)$ and $\nabla_2 h(w)$ as

$$g_1^t = \frac{1}{K_1} \sum_{m \in S_1^t} \lambda_m \frac{g_{1,m}^t}{p_{1,m}^t}, \quad g_2^t = \frac{1}{K_2} \sum_{m \in S_2^t} \lambda_m \frac{g_{2,m}^t}{p_{2,m}^t},$$

where $K_1 = |S_1^t|$ and $K_2 = |S_2^t|$. Then $g^t = g_1^t + g_2^t$ is an estimate of $\nabla h(w)$.

We design p_1^t and p_2^t to choose S_1^t and S_2^t by minimizing the variance of the gradients.

Note that

$$\begin{aligned} \min_{p_1^t} \mathbb{E}_1 \left[\mathbb{E}_{S_1^t} \left[\left\| g_1^t - \nabla_1 h(w^t) \right\|^2 \right] \right] &+ \min_{p_2^t} \mathbb{E}_2 \left[\mathbb{E}_{S_2^t} \left[\left\| g_2^t - \nabla_2 h(w^t) \right\|^2 \right] \right] \\ &\leq \min_{p_1^t = p_2^t = p^t} \mathbb{E} \left[\mathbb{E}_{S^t} \left[\left\| g_1^t - \nabla_1 h(w^t) + g_2^t - \nabla_2 h(w^t) \right\|^2 \right] \right], \end{aligned}$$

so that the decomposition allows us to better minimize the variance. We term this approach as *doubly variance reduction for personalized Federated Learning*. The first part minimizes the variance of updates to the shared global parameter, when the best local parameters are fixed; and the second part minimizes the variance of updates to local parameters, when the global part is fixed. While these two parts are related, any given machine will have different contributions to these two tasks.

Adaptive-OSMD Sampler can be used to minimize the variance for both parts of the gradient. We note that this is a heuristic approach to solving the client sampling problem when minimizing a personalized FL objective. Personalized FL objectives have additional structures that should be used to design more efficient sampling strategies. Furthermore, designing sampling strategies that improve the statistical performance of trained models, rather than improving computational speed, is important in the heterogeneous setting. Addressing these questions is an important area for future research.

4.9.2 Sampling with Physical Constraint in FL System

In this chapter, we assume that all clients are available in each round. However, in practical FL applications, a subset of the clients may be inactive due to physical constraints; thus, we have to assign zero probabilities to them. In this section, we propose a simple extension of our proposed sampling method to such cases.

Specifically, denote the subset of clients that are active at the beginning of round t as $I^t \subseteq [M]$. If we have $|I^t| \leq K$, we can then use all clients in I^t to make updates in round t ; otherwise, we would like to choose a smaller subset $S^t \subseteq I^t$ to participate. This can be achieved by rescaling the output sampling distribution of any of our proposed methods, which we denote as $\hat{p}^t$. We let $\tilde{p}_m^t = \hat{p}_m^t / (\sum_{i \in I^t} \hat{p}_i^t)$ and $\tilde{p}_m^t = 0$ for all $m \notin I^t$. We can then use $\tilde{p}_m^t$ to choose S^t from I^t .

However, analyzing such a method in terms of convergence and regret guarantee is highly non-trivial. Typically, for general active clients sequence $\{I^t\}_{t=1}^T$, the optimization algorithms are not guaranteed to converge even if we involve all clients in I^t in each round. This can happen, for example, if a client is active only once in the whole training process. Thus, to ensure convergence, we need additional assumptions about $\{I^t\}_{t=1}^T$. Moreover, deriving a regret bound is also very challenging, as assigning zero probability to any client will make the variance-reduction loss unbounded and thus, the regret can be arbitrarily large. To achieve such a theoretical result, one may need to appropriately redefine the regret concept. Such an analysis is beyond the scope of this chapter and we leave it for future research.

4.10 Technical Details

In this section, we present the technical proofs of the theorems stated in the preceding sections of this chapter.

4.10.1 Useful Lemmas

In this section, we present several auxiliary lemmas that will be used in the proofs of the main theorems.

Lemma 46. *Suppose that f is a differentiable convex function defined on $\text{dom} f$, and $\mathcal{X} \subseteq \text{dom} f$ is a closed convex set. Then x is the minimizer of f on $\mathcal{X}$ if and only if*

$$\nabla f(x)^\top (y - x) \geq 0 \quad \text{for all } y \in \mathcal{X}.$$

Proof. See Section 4.2.3 of Boyd et al. [2004]. □

Lemma 47 (Smoothness Upper Bound). *Let $f : \mathcal{W} \rightarrow \mathbb{R}$ be defined on a convex set $\mathcal{W} \subseteq \mathbb{R}^d$. Suppose that f is continuously differentiable and L -smooth, i.e.,*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \text{for all } x, y \in \mathcal{W}.$$

Then, for any $x, y \in \mathcal{W}$, we have

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2} \|y - x\|^2.$$

Proof. The proof follows Theorem 2.1.5 of Nesterov et al. [2018]. By the fundamental theorem of calculus, we have

$$\begin{aligned} f(y) - f(x) &= \int_0^1 \langle \nabla f(x + t(y - x)), y - x \rangle dt, \\ \implies f(y) - f(x) - \langle \nabla f(x), y - x \rangle &= \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle dt. \end{aligned}$$

Taking absolute values and applying Cauchy–Schwarz inequality gives

$$\begin{aligned} |f(y) - f(x) - \langle \nabla f(x), y - x \rangle| &\leq \int_0^1 \|\nabla f(x + t(y - x)) - \nabla f(x)\| \cdot \|y - x\| dt \\ &\leq \int_0^1 Lt\|y - x\|^2 dt = \frac{L}{2}\|y - x\|^2, \end{aligned}$$

where we used the L -smoothness assumption. This completes the proof. $\square$

Lemma 48 (Relaxed Triangle Inequality). *This is Lemma 3 of Karimireddy et al. [2020b].*

Let $\{v_1, \dots, v_\tau\}$ be τ vectors in $\mathbb{R}^d$. Then the following are true:

$$\begin{aligned} \|v_i + v_j\|^2 &\leq (1 + a)\|v_i\|^2 + \left(1 + \frac{1}{a}\right)\|v_j\|^2 \quad \text{for any } a > 0, \\ \left\| \sum_{i=1}^{\tau} v_i \right\|_2^2 &\leq \tau \sum_{i=1}^{\tau} \|v_i\|_2^2. \end{aligned}$$

Proof. See Lemma 3 of Karimireddy et al. [2020b]. $\square$

Lemma 49. *For $q \in \mathcal{P}_{M-1}$ we have $D_\Phi(q \| p^{\text{unif}}) \leq \log M$, where Φ is the unnormalized negative entropy.*

Proof. Since $\Phi(q) = \sum_{m=1}^M q_m(\log q_m - 1) \leq 0$, $\Phi(p^{\text{unif}}) = -\log M$, and

$$\langle \nabla \Phi(p^{\text{unif}}), q - p^{\text{unif}} \rangle = \sum_{m=1}^M (q_m - \frac{1}{M}) \log \frac{1}{M} = 0,$$

we have $D_\Phi(q \| p) \leq \log M$. $\square$

Lemma 50 (Hoeffding’s Inequality). *Let X be a random variable with $a \leq X \leq b$ for $a, b \in \mathbb{R}$. Then for all $s \in \mathbb{R}$, we have*

$$\log \mathbb{E} \left[e^{sX} \right] \leq s\mathbb{E}[X] + \frac{s^2(b - a)^2}{8}.$$

Proof. See Section 2 of Wainwright [2019]. $\square$

Lemma 51 (Based on Exercise 26.12 of Lattimore and Szepesvári [2020]). *Let $\alpha \in [0, 1]$, $\mathcal{A} = \mathcal{P}_{M-1} \cap [\alpha/M, 1]^M$, $\mathcal{D} = [0, \infty)^M$, and Φ is the unnormalised entropy on $\mathcal{D}$. For $y \in [0, \infty)^M$, let $x = \arg \min_{v \in \mathcal{A}} D_\Phi(v \| y)$. Suppose $y_1 \leq y_2 \leq \dots \leq y_M$. Let $m^\star$ be the smallest value such that*

$$y_{m^\star} \left(1 - \frac{m^\star - 1}{M} \alpha \right) > \frac{\alpha}{M} \sum_{m=m^\star}^M y_m.$$

Then

$$x_m = \begin{cases} \frac{\alpha}{M} & \text{if } m < m^\star \\ \frac{(1 - \frac{m^\star - 1}{M} \alpha) y_m}{\sum_{n=m^\star}^M y_n} & \text{otherwise.} \end{cases}$$

Proof. Consider the following constrained optimization problem:

$$\begin{aligned} \min_{u \in [0, \infty)^M} \quad & \sum_{m=1}^M u_m \log \frac{u_m}{y_m}, \\ \text{s.t.} \quad & \sum_{m=1}^M u_m = 1, \\ & u_m \geq \frac{\alpha}{M}, \quad \forall m \in [M]. \end{aligned}$$

Since x is the solution to this problem, by the Karush-Kuhn-Tucker (KKT) conditions, there exist multipliers $\lambda \in \mathbb{R}$ and $\nu_1, \dots, \nu_M \in \mathbb{R}$ such that

$$\log \frac{x_m}{y_m} + 1 - \lambda - \nu_m = 0, \quad \forall m \in [M], \tag{4.32}$$

$$\sum_{m=1}^M x_m = 1, \tag{4.33}$$

$$x_m - \frac{\alpha}{M} \geq 0, \quad \forall m \in [M], \tag{4.34}$$

$$\nu_m \geq 0, \quad \forall m \in [M], \tag{4.35}$$

$$\nu_m \left(x_m - \frac{\alpha}{M} \right) = 0, \quad \forall m \in [M]. \tag{4.36}$$

From (4.32), we have $x_m = y_m \exp(-1 + \lambda + \nu_m)$. By complementary slackness (4.36), we distinguish two cases:

- If $x_m = \frac{\alpha}{M}$, then $\nu_m \geq 0$ and

$$x_m = y_m \exp(-1 + \lambda + \nu_m) \geq y_m \exp(-1 + \lambda).$$

- If $x_m > \frac{\alpha}{M}$, then $\nu_m = 0$, and thus $x_m = y_m \exp(-1 + \lambda)$.

Without loss of generality, assume the variables are ordered so that

$$x_1 = \cdots = x_{m^*-1} = \frac{\alpha}{M} < x_{m^*} \leq \cdots \leq x_M.$$

Then the normalization constraint (4.33) gives

$$1 = \sum_{m=1}^M x_m = (m^* - 1) \frac{\alpha}{M} + \exp(-1 + \lambda) \sum_{m=m^*}^M y_m,$$

which implies

$$\exp(-1 + \lambda) = \frac{1 - (m^* - 1) \frac{\alpha}{M}}{\sum_{m=m^*}^M y_m}. \quad (4.37)$$

Substituting into x_{m^*} , we obtain

$$x_{m^*} = y_{m^*} \exp(-1 + \lambda) = y_{m^*} \cdot \frac{1 - (m^* - 1) \frac{\alpha}{M}}{\sum_{m=m^*}^M y_m} > \frac{\alpha}{M}.$$

Thus,

$$y_{m^*} \left(1 - \frac{m^* - 1}{M} \alpha \right) > \frac{\alpha}{M} \sum_{m=m^*}^M y_m. \quad (4.38)$$

It remains to show that for any $1 \leq m' \leq m^* - 1$,

$$y_{m'} \left(1 - \frac{m' - 1}{M} \alpha \right) \leq \frac{\alpha}{M} \sum_{m=m'}^M y_m. \quad (4.39)$$

This, together with (4.38), completes the proof.

To verify (4.39), note that for $m' \leq m^* - 1$, we have $x_{m'} = \frac{\alpha}{M}$ and from (4.32),

$$\frac{\alpha}{M} = y_{m'} \exp(-1 + \lambda + \nu_{m'}).$$

Since $y_1 \leq \dots \leq y_M$, we have $\nu_1 \geq \dots \geq \nu_{m^*-1}$.

Now, for $m' \leq m^* - 1$,

$$\begin{aligned} (m^* - m') \frac{\alpha}{M} &= \sum_{m=m'}^{m^*-1} y_m \exp(-1 + \lambda + \nu_m) \\ &\leq \exp(-1 + \lambda + \nu_{m'}) \sum_{m=m'}^{m^*-1} y_m. \end{aligned} \quad (4.40)$$

Also, from (4.37),

$$1 - (m^* - 1) \frac{\alpha}{M} = \exp(-1 + \lambda) \sum_{m=m^*}^M y_m \leq \exp(-1 + \lambda + \nu_{m'}) \sum_{m=m^*}^M y_m. \quad (4.41)$$

Combining (4.40) and (4.41), we obtain

$$\begin{aligned} \frac{1 - (m' - 1) \frac{\alpha}{M}}{\sum_{m=m'}^M y_m} &= \frac{1 - (m^* - 1) \frac{\alpha}{M} + (m^* - m') \frac{\alpha}{M}}{\sum_{m=m'}^M y_m} \\ &\leq \frac{\exp(-1 + \lambda + \nu_{m'}) \left(\sum_{m=m'}^{m^*-1} y_m + \sum_{m=m^*}^M y_m \right)}{\sum_{m=m'}^M y_m} \\ &= \exp(-1 + \lambda + \nu_{m'}) = \frac{\alpha}{M y_{m'}}, \end{aligned}$$

which rearranges to give (4.39). □

4.10.2 Proof of Proposition 37

First, we show that the solution $\hat{p}^{t+1}$ in Step 7 of Algorithm 5 can be found as

$$\begin{aligned}\tilde{p}^{t+1} &= \arg \min_{p \in \mathcal{D}} \eta_t \langle p, \nabla \hat{l}_t(\hat{p}^t; \hat{p}^t) \rangle + D_{\Phi}(p \| \hat{p}^t), \\ \hat{p}^{t+1} &= \arg \min_{p \in \mathcal{A}} D_{\Phi}(p \| \tilde{p}^{t+1}).\end{aligned}$$

The optimality condition for $\tilde{p}^{t+1}$ implies that

$$\eta_t \nabla \hat{l}_t(\hat{p}^t; \hat{p}^t) + \nabla \Phi(\tilde{p}^{t+1}) - \nabla \Phi(\hat{p}^t) = 0. \quad (4.42)$$

By Lemma 46, the optimality condition for $\hat{p}^{t+1}$ implies that

$$\langle p - \hat{p}^{t+1}, \nabla \Phi(\hat{p}^{t+1}) - \nabla \Phi(\tilde{p}^{t+1}) \rangle \geq 0, \quad \text{for all } p \in \mathcal{A}.$$

Combining the last two displays, we have

$$\langle p - \hat{p}^{t+1}, \eta_t \nabla \hat{l}_t(\hat{p}^t; \hat{p}^t) + \nabla \Phi(\hat{p}^{t+1}) - \nabla \Phi(\hat{p}^t) \rangle \geq 0, \quad \text{for all } p \in \mathcal{A}.$$

By Lemma 46, this is the optimality condition for $\hat{p}^{t+1}$ to be the solution in Step 7 of Algorithm 5.

Note that (4.42) implies that

$$-\frac{\eta_t}{K^2} \cdot \frac{a_m^t}{(\hat{p}_m^t)^3} \mathcal{N}\{m \in S^t\} + \log(\tilde{p}_m^{t+1}) - \log(\hat{p}_m^t) = 0, \quad m \in [M].$$

Therefore,

$$\tilde{p}_m^{t+1} = \hat{p}_m^t \exp \left(\frac{\eta_t a_m^t}{K^2 (\hat{p}_m^t)^3} \mathcal{N}\{m \in S^t\} \right), \quad m \in [M],$$

and the final result follows from Lemma 51.

4.10.3 Proof of Theorem 38

Our proof builds on techniques from the proof of Theorem 2.1 in Ghadimi and Lan [2013], with a key innovation: we introduce a *ghost subset*, which is sampled from $[M]$ according to the comparator's sampling distribution. This ghost subset is a conceptual tool used purely for theoretical analysis—it does not need to be computed or observed in practice.

Let $\{w^t\}_{t=1}^T \triangleq w^{1:T}$ be an arbitrary sequence of iterates. Define $q_s^{1:T}$ as the solution to the following optimization problem:

$$\begin{aligned} \tilde{\zeta}(w^{1:T}) &= \frac{1}{T} \min_{p^1} \cdots \min_{p^T} \sum_{t=1}^T V(p^t, w^t), \\ \text{subject to } \text{TV}(p^{1:T}) &\leq \beta \quad \text{and} \quad p^t \in \sigma(w^1, \dots, w^t), \end{aligned} \tag{4.43}$$

where $\sigma(w^1, \dots, w^t)$ denotes the σ -algebra generated by the filtration $\{w^1, \dots, w^t\}$.

By definition, this implies

$$\tilde{\zeta}(w^{1:T}) = \frac{1}{T} \sum_{t=1}^T V(q_s^t(w^t), w^t), \quad \text{with} \quad \text{TV}(q_s^{1:T}) \leq \beta.$$

Recalling the definition of $\zeta_T^2(\alpha, \beta)$, we obtain the identity:

$$\frac{1}{T} \sup_{w^1} \cdots \sup_{w^T} \sum_{t=1}^T V(q_s^t, w^t) = \zeta_T^2(\alpha, \beta). \tag{4.44}$$

Let $\delta^t = g^t - \nabla F(w^t)$ denote the stochastic gradient noise at iteration t . Under Assump-

tion 9, we can apply Lemma 47 to obtain the following bound:

$$\begin{aligned}
& F(w^{t+1}) \\
& \leq F(w^t) + \langle \nabla F(w^t), w^{t+1} - w^t \rangle + \frac{L}{2} \mu^2 \|g^t\|_2^2 \\
& = F(w^t) - \mu \langle \nabla F(w^t), g^t \rangle + \frac{L}{2} \mu^2 \|g^t\|_2^2 \\
& = F(w^t) - \mu \|\nabla F(w^t)\|_2^2 - \mu \langle \nabla F(w^t), \delta^t \rangle \\
& \quad + \frac{L}{2} \mu^2 \left[\|\nabla F(w^t)\|_2^2 + 2 \langle \nabla F(w^t), \delta^t \rangle + \|\delta^t\|_2^2 \right] \\
& = F(w^t) - \left(\mu - \frac{L}{2} \mu^2 \right) \|\nabla F(w^t)\|_2^2 - \left(\mu - L\mu^2 \right) \langle \nabla F(w^t), \delta^t \rangle + \frac{L}{2} \mu^2 \|\delta^t\|_2^2. \quad (4.45)
\end{aligned}$$

We use the notation $\mathbb{E}_{S^t}[\cdot]$ to denote expectation taken with respect to the mini-batch S^t . Since $\mathbb{E}_{S^t}[g^t \mid w^t, \hat{p}^t] = \nabla F(w^t)$ by construction, it follows that

$$\mathbb{E}[\delta^t \mid w^t, \hat{p}^t] = 0.$$

Hence, taking the full expectation, we obtain

$$\mathbb{E}[\langle \nabla F(w^t), \delta^t \rangle] = \mathbb{E}[\mathbb{E}[\langle \nabla F(w^t), \delta^t \rangle \mid w^t, \hat{p}^t]] = 0. \quad (4.46)$$

Conditioned on $w^1, \dots, w^t$, the distribution q_s^t is deterministic. We can therefore introduce a *ghost subset* of clients $\tilde{S}^t$, of size $|\tilde{S}^t| = K$, sampled from $[M]$ according to q_s^t .

Recall that the full-batch average gradient is defined as

$$J^t := \frac{1}{M} \sum_{m=1}^M g_m^t.$$

We also define the ghost mini-batch gradient as

$$\tilde{g}^t := \frac{1}{MK} \sum_{m \in \tilde{S}^t} \frac{g_m^t}{q_{s,m}^t}.$$

We now analyze the conditional variance of the stochastic gradient estimator:

$$\begin{aligned} & \mathbb{E}_{S^t} \left[\left\| \delta^t \right\|_2^2 \mid w^1, \dots, w^t, \hat{p}^t \right] \\ &= \mathbb{E}_{S^t} \left[\left\| g^t - J^t + J^t - \nabla F(w^t) \right\|_2^2 \mid w^1, \dots, w^t, \hat{p}^t \right] \\ &= \mathbb{E}_{S^t} \left[\left\| g^t - J^t \right\|_2^2 \mid w^1, \dots, w^t, \hat{p}^t \right] + \left\| J^t - \nabla F(w^t) \right\|_2^2 \\ &= \ell_t(\hat{p}^t) - \frac{1}{K} \left\| J^t \right\|_2^2 + \left\| J^t - \nabla F(w^t) \right\|_2^2 \\ &= \ell_t(q_s^t) - \frac{1}{K} \left\| J^t \right\|_2^2 + \left\| J^t - \nabla F(w^t) \right\|_2^2 + \left(\ell_t(\hat{p}^t) - \ell_t(q_s^t) \right) \\ &= \mathbb{E}_{\tilde{S}^t} \left[\left\| \tilde{g}^t - \nabla F(w^t) \right\|_2^2 \mid w^1, \dots, w^t, \hat{p}^t \right] + \left(\ell_t(\hat{p}^t) - \ell_t(q_s^t) \right). \end{aligned}$$

Next, we analyze the first term:

$$\begin{aligned} & \mathbb{E} \left[\left\| \tilde{g}^t - \nabla F(w^t) \right\|_2^2 \mid w^1, \dots, w^t, \hat{p}^t \right] \\ &= \mathbb{E} \left[\left\| \frac{1}{MK} \sum_{m \in \tilde{S}^t} \frac{g_m^t}{q_{s,m}^t} - \nabla F(w^t) \right\|_2^2 \mid w^1, \dots, w^t, \hat{p}^t \right] \\ &\leq \frac{2}{M^2 K^2} \mathbb{E} \left[\sum_{m \in \tilde{S}^t} \mathbb{E} \left[\frac{\left\| g_m^t - \nabla F_m(w^t) \right\|_2^2}{(q_{s,m}^t)^2} \right] \mid w^1, \dots, w^t, \hat{p}^t \right] \\ &\quad + \frac{2}{K} \left(\frac{1}{M^2} \sum_{m=1}^M \frac{\left\| \nabla F_m(w^t) \right\|_2^2}{q_{s,m}^t} - \left\| \nabla F(w^t) \right\|_2^2 \right) \\ &= \frac{2\sigma^2}{M^2 K B} \sum_{m=1}^M \frac{1}{q_{s,m}^t} + \frac{2}{K} V(q_s^t, w^t) \\ &\leq \frac{2\sigma^2}{K B \alpha} + \frac{2}{K} V(q_s^t, w^t), \end{aligned}$$

where the third line uses the variance bound $\mathbb{E} \left[\|g_m^t - \nabla F_m(w^t)\|_2^2 \right] \leq \sigma^2/B$ and the definition of $V(p, w)$. The final inequality uses the assumption $q_{s,m}^t \geq \alpha/M$ for all m , since $q_s^t \in \mathcal{A}$.

Putting this back into the earlier expression, we obtain:

$$\mathbb{E} \left[\|\delta^t\|_2^2 \mid w^1, \dots, w^t, \hat{p}^t \right] \leq \frac{2\sigma^2}{KB\alpha} + \frac{2}{K} V(q_s^t, w^t) + \ell_t(\hat{p}^t) - \ell_t(q_s^t),$$

which implies

$$\begin{aligned} \sum_{t=1}^T \mathbb{E} \left[\|\delta^t\|_2^2 \right] &\leq \frac{2T\sigma^2}{KB\alpha} + \frac{2}{K} \mathbb{E} \left[\sum_{t=1}^T V(q_s^t, w^t) \right] + \mathbb{E} \left[\sum_{t=0}^{T-1} \ell_t(\hat{p}^t) - \ell_t(q_s^t) \right] \\ &= \frac{2T\sigma^2}{KB\alpha} + \frac{2}{K} \mathbb{E} \left[\sum_{t=1}^T V(q_s^t, w^t) \right] + \text{D-Regret}_T(q_s^{1:T}) \\ &\leq \frac{2T\sigma^2}{KB\alpha} + \frac{2T\zeta_T^2(\alpha, \beta)}{K} + \text{D-Regret}_T(q_s^{1:T}), \end{aligned} \tag{4.47}$$

where the final inequality follows from

$$\begin{aligned} \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T V(q_s^t, w^t) \right] &= \frac{1}{T} \mathbb{E}_{w^1, \dots, w^T} \left[\sum_{t=1}^T V(q_s^t, w^t) \right] \\ &\leq \frac{1}{T} \sup_{w^1, \dots, w^T} \sum_{t=1}^T V(q_s^t, w^t) = \zeta_T^2(\alpha, \beta), \end{aligned}$$

where the last equality uses (4.44).

Combining inequalities (4.45), (4.46), and (4.47), we obtain:

$$\begin{aligned}
& \left(\mu - \frac{L}{2}\mu^2 \right) \sum_{t=1}^T \mathbb{E} \left[\left\| \nabla F(w^t) \right\|_2^2 \right] \\
& \leq F(w^1) - F(w^T) + \frac{L}{2}\mu^2 \left(\frac{2T\sigma^2}{KB\alpha} + \frac{2T\zeta_T^2(\alpha, \beta)}{K} + \text{D-Regret}_T(q_s^{1:T}) \right) \\
& \leq D^F + \frac{L}{2}\mu^2 \left(\frac{2T\sigma^2}{KB\alpha} + \frac{2T\zeta_T^2(\alpha, \beta)}{K} + \text{D-Regret}_T(q_s^{1:T}) \right),
\end{aligned}$$

where $D^F = F(w^1) - \inf_w F(w)$ bounds the descent range.

Since $\mu \leq 1/L$, it follows that $\mu - \frac{L}{2}\mu^2 \geq \mu/2$. Dividing both sides by T and rearranging yields:

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[\left\| \nabla F(w^t) \right\|_2^2 \right] \leq \frac{2D^F}{T\mu} + L\mu \left(\frac{2\sigma^2}{KB\alpha} + \frac{2\zeta_T^2(\alpha, \beta)}{K} + \frac{\text{D-Regret}_T(q_s^{1:T})}{T} \right). \quad (4.48)$$

Next, we bound the dynamic regret term $\text{D-Regret}_T(q_s^{1:T})$. Note that

$$a_{\max}^t = \frac{1}{M^2} \max_{1 \leq m \leq M} \left\| g_m^t \right\|_2^2 = \frac{1}{M^2} \max_{1 \leq m \leq M} \left\| \frac{1}{B} \sum_{b=1}^B \nabla \phi(w^t; \xi_m^{t,b}) \right\|_2^2 \leq \frac{G^2}{M^2},$$

since $\|\nabla \phi(w; \xi)\|_2 \leq G$ for all w and ξ .

Then, by Theorem 40, and using the facts that $q_s^t \in \mathcal{A}$ for all $t \in [T]$ and $\text{TV}(q_s^{1:T}) \leq \beta$, we obtain:

$$\begin{aligned}
\text{D-Regret}_T(q_s^{1:T}) & \leq \frac{\log M}{\eta} + \frac{2 \log(M/\alpha)}{\eta} \cdot \mathbb{E} \left[\text{TV}(q_s^{1:T}) \right] + \frac{\eta M^6}{2K^2\alpha^6} \sum_{t=1}^T \mathbb{E} \left[\left(a_{\max}^t \right)^2 \right] \\
& \leq \frac{\log M}{\eta} + \frac{2\beta \log(M/\alpha)}{\eta} + \frac{\eta T M^2 G^4}{2K^2\alpha^6}.
\end{aligned}$$

Let the learning rate be chosen as

$$\eta = \frac{K\alpha^3}{MG^2} \sqrt{\frac{2\log M + 4\beta\log(M/\alpha)}{T}},$$

which minimizes the sum of the three regret terms above. Substituting into the bound yields

$$\text{D-Regret}_T(q_s^{1:T}) \leq \frac{\sqrt{T}MG^2}{K\alpha^3} \sqrt{\frac{1}{2}\log M + \beta\log(M/\alpha)}.$$

Plugging this bound into inequality (4.48), we obtain the final convergence result:

$$\begin{aligned} & \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[\left\| \nabla F(w^t) \right\|_2^2 \right] \\ & \leq \frac{2D^F}{T\mu} + L\mu \left(\frac{2\sigma^2}{KB\alpha} + \frac{2\zeta_T^2(\alpha, \beta)}{K} + \frac{MG^2}{K\alpha^3} \sqrt{\frac{\frac{1}{2}\log M + \beta\log(M/\alpha)}{T}} \right). \end{aligned}$$

Finally, let

$$\mu = \min \left\{ \frac{1}{L}, \frac{1}{\sigma} \sqrt{\frac{D^F KB\alpha}{LT}}, \frac{1}{\zeta_T(\alpha, \beta)} \sqrt{\frac{D^F K}{LT}}, \sqrt{\frac{D^F K\alpha^3}{L\sqrt{T}MG^2 \sqrt{\frac{1}{2}\log M + \beta\log(M/\alpha)}}} \right\}, \quad (4.49)$$

then we have

$$\begin{aligned}
\mathbb{E} \left[\left\| \nabla F(w^R) \right\|_2^2 \right] &= \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[\left\| \nabla F(w^t) \right\|_2^2 \right] \\
&\lesssim \frac{D^F}{T} \max \left\{ L, \sigma \sqrt{\frac{LT}{D^F K B \alpha}}, \zeta_T(\alpha, \beta) \sqrt{\frac{TL}{D^F K}}, \sqrt{\frac{L \sqrt{T} M G^2 \sqrt{\frac{1}{2} \log M + \beta \log(M/\alpha)}}{D^F K \alpha^3}} \right\} \\
&\quad + \frac{\sigma \sqrt{D^F L}}{\sqrt{TK B \alpha}} + \frac{\zeta_T(\alpha, \beta) \sqrt{D^F L}}{\sqrt{TK}} + \frac{\sqrt{D^F L} M^{\frac{1}{2}} G}{T^{\frac{3}{4}} K^{\frac{1}{2}} \alpha^{\frac{3}{2}}} \left(\frac{1}{2} \log M + \beta \log(M/\alpha) \right)^{\frac{1}{4}} \\
&\lesssim \frac{D^F}{T} \left(L + \sigma \sqrt{\frac{LT}{D^F K B \alpha}} + \zeta_T(\alpha, \beta) \sqrt{\frac{TL}{D^F K}} + \sqrt{\frac{L \sqrt{T} M G^2 \sqrt{\frac{1}{2} \log M + \beta \log(M/\alpha)}}{D^F K \alpha^3}} \right) \\
&\quad + \frac{\sigma \sqrt{D^F L}}{\sqrt{TK B \alpha}} + \frac{\zeta_T(\alpha, \beta) \sqrt{D^F L}}{\sqrt{TK}} + \frac{\sqrt{D^F L} M^{\frac{1}{2}} G}{T^{\frac{3}{4}} K^{\frac{1}{2}} \alpha^{\frac{3}{2}}} \left(\frac{1}{2} \log M + \beta \log(M/\alpha) \right)^{\frac{1}{4}} \\
&\lesssim \frac{D^F L}{T} + \frac{\sigma \sqrt{D^F L}}{\sqrt{TK B \alpha}} + \frac{\zeta_T(\alpha, \beta) \sqrt{D^F L}}{\sqrt{TK}} + \frac{\sqrt{D^F L} M^{\frac{1}{2}} G}{T^{\frac{3}{4}} K^{\frac{1}{2}} \alpha^{\frac{3}{2}}} \left(\frac{1}{2} \log M + \beta \log(M/\alpha) \right)^{\frac{1}{4}}.
\end{aligned}$$

4.10.4 Proof of Theorem 39

As in the proof of Theorem 38, the central technique is the construction of a *ghost subset*, drawn from $[M]$ according to the comparator sampling distribution. This subset is introduced purely for theoretical analysis and does not need to be computed in practice.

Let $\{w^t\}_{t=1}^T$ be any sequence of iterates, and let $q_s^{1:T}$ denote the solution to the dynamic game defined in Equation (4.43).

By Assumption 9 and Lemma 47, the smoothness of F yields

$$F(w^{t+1}) \leq F(w^t) + \left\langle \nabla F(w^t), w^{t+1} - w^t \right\rangle + \frac{L}{2} \left\| w^{t+1} - w^t \right\|_2^2,$$

where

$$w^{t+1} - w^t = -\frac{\mu}{MK} \sum_{m \in S^t} \frac{1}{\hat{p}_m^t} g_m^t, \quad g_m^t = \mu_l \sum_{b=0}^{B-1} \nabla \phi(w_m^{t,b}; \xi_m^{t,b}).$$

Let $\mathcal{B}_m^t = \{\xi_m^{t,0}, \dots, \xi_m^{t,B-1}\}$ denote the local minibatch, and let $\mathbb{E}_{S^t}[\cdot]$ denote the conditional expectation over the sampled client set S^t . Then,

$$\begin{aligned} \mathbb{E}_{S^t} \left[F(w^{t+1}) \mid w^1, \dots, w^t, \hat{p}^t \right] &\leq F(w^t) - \mu \left\langle \nabla F(w^t), \frac{1}{M} \sum_{m=1}^M g_m^t \right\rangle \\ &\quad + \frac{L\mu^2}{2} \mathbb{E}_{S^t} \left[\left\| \frac{1}{K} \sum_{m \in S^t} \frac{1}{M\hat{p}_m^t} g_m^t \right\|_2^2 \mid w^1, \dots, w^t, \hat{p}^t \right]. \end{aligned}$$

Define $J^t := \frac{1}{M} \sum_{m=1}^M g_m^t$ as the full-batch gradient estimator. Then,

$$\begin{aligned} \mathbb{E}_{S^t} \left[F(w^{t+1}) \right] &\leq F(w^t) - \mu \left\langle \nabla F(w^t), J^t \right\rangle \\ &\quad + \frac{L\mu^2}{2} \mathbb{E}_{S^t} \left[\left\| \frac{1}{K} \sum_{m \in S^t} \frac{1}{M\hat{p}_m^t} g_m^t - J^t + J^t \right\|_2^2 \mid w^1, \dots, w^t, \hat{p}^t \right] \\ &= F(w^t) - \mu \left\langle \nabla F(w^t), J^t \right\rangle + \frac{L\mu^2}{2} \|J^t\|_2^2 \\ &\quad + \frac{L\mu^2}{2} \mathbb{E}_{S^t} \left[\left\| \frac{1}{K} \sum_{m \in S^t} \frac{1}{M\hat{p}_m^t} g_m^t - J^t \right\|_2^2 \mid w^1, \dots, w^t, \hat{p}^t \right]. \end{aligned}$$

Recall from Algorithm 8 that we define $a_m^t := \frac{\|g_m^t\|_2^2}{B(M\mu_l)^2}$. Using this definition, we compute the conditional variance term as follows:

$$\begin{aligned} \mathbb{E}_{S^t} \left[\left\| \frac{1}{K} \sum_{m \in S^t} \frac{1}{M\hat{p}_m^t} g_m^t - J^t \right\|_2^2 \mid w^1, \dots, w^t, \hat{p}^t \right] &= \frac{1}{K} \sum_{m=1}^M \frac{\|g_m^t\|_2^2}{M^2\hat{p}_m^t} - \frac{1}{K} \|J^t\|_2^2 \\ &= B\mu_l^2 \cdot \frac{1}{K} \sum_{m=1}^M \frac{a_m^t}{\hat{p}_m^t} - \frac{1}{K} \|J^t\|_2^2 \\ &= B\mu_l^2 \cdot l_t(\hat{p}^t) - \frac{1}{K} \|J^t\|_2^2, \quad (4.50) \end{aligned}$$

where $l_t(\cdot)$ is the variance reduction loss function defined in Equation (4.4).

Substituting Equation (4.50) into the earlier bound on $\mathbb{E}_{S^t}[F(w^{t+1})]$, we obtain

$$\begin{aligned} \mathbb{E}_{S^t} \left[F(w^{t+1}) \mid w^1, \dots, w^t, \hat{p}^t \right] &\leq F(w^t) - \mu \left\langle \nabla F(w^t), J^t \right\rangle + \frac{L\mu^2}{2} \|J^t\|_2^2 \\ &\quad + \frac{L\mu^2}{2} \left(B\mu_l^2 \cdot l_t(\hat{p}^t) - \frac{1}{K} \|J^t\|_2^2 \right). \end{aligned} \quad (4.51)$$

Conditioned on $w^1, \dots, w^t$, the comparator sampling distribution q_s^t is deterministic. For analytical purposes, we introduce a *ghost subset* of clients $\tilde{S}^t$, consisting of K clients sampled with replacement from $[M]$ according to q_s^t .

Define the associated gradient estimator:

$$\tilde{g}^t := \frac{1}{K} \sum_{m \in \tilde{S}^t} \frac{g_m^t}{Mq_{s,m}^t}.$$

Then, following the same derivation as in Equation (4.50), we obtain

$$\mathbb{E}_{\tilde{S}^t} \left[\|\tilde{g}^t - J^t\|_2^2 \mid w^1, \dots, w^t, \hat{p}^t \right] = B\mu_l^2 \cdot l_t(q_s^t) - \frac{1}{K} \|J^t\|_2^2.$$

Combining this with the bound in Equation (4.51), we have:

$$\begin{aligned} &\mathbb{E}_{S^t} \left[F(w^{t+1}) \mid w^{1:t}, \hat{p}^t \right] \\ &\leq F(w^t) - \mu \left\langle \nabla F(w^t), J^t \right\rangle + \frac{L\mu^2}{2} \|J^t\|_2^2 + \frac{BL\mu^2\mu_l^2}{2} \left(l_t(\hat{p}^t) - l_t(q_s^t) \right) \\ &\quad + \frac{L\mu^2}{2} \left(B\mu_l^2 \cdot l_t(q_s^t) - \frac{1}{K} \|J^t\|_2^2 \right) \\ &= F(w^t) - \mu \left\langle \nabla F(w^t), J^t \right\rangle + \frac{L\mu^2}{2} \|J^t\|_2^2 + \frac{BL\mu^2\mu_l^2}{2} \left(l_t(\hat{p}^t) - l_t(q_s^t) \right) \\ &\quad + \frac{L\mu^2}{2} \mathbb{E}_{\tilde{S}^t} \left[\|\tilde{g}^t - J^t\|_2^2 \mid w^{1:t}, \hat{p}^t \right] \\ &= F(w^t) - \mu \left\langle \nabla F(w^t), J^t \right\rangle + \frac{BL\mu^2\mu_l^2}{2} \left(l_t(\hat{p}^t) - l_t(q_s^t) \right) + \frac{L\mu^2}{2} \mathbb{E}_{\tilde{S}^t} \left[\|\tilde{g}^t\|_2^2 \mid w^{1:t}, \hat{p}^t \right]. \end{aligned}$$

Let $\mathbb{E}_t[\cdot]$ denote the conditional expectation $\mathbb{E}[\cdot \mid w^1, \dots, w^t, \hat{p}^t]$. Then we can write:

$$\begin{aligned} \mathbb{E}_t \left[F(w^{t+1}) \right] &\leq F(w^t) - \frac{\mu\mu_l}{M} \mathbb{E}_t \left[\left\langle \nabla F(w^t), \sum_{m=1}^M \sum_{b=0}^{B-1} \nabla F_m(w_m^{t,b}) \right\rangle \right] \\ &\quad + \frac{L\mu^2}{2} \mathbb{E}_t \left[\|\hat{g}^t\|_2^2 \right] + \frac{BL\mu^2\mu_l^2}{2} \mathbb{E}_t \left[l_t(\hat{p}^t) - l_t(q_s^t) \right]. \end{aligned} \quad (4.52)$$

We begin by recalling a standard inequality: for any $u, v \in \mathbb{R}^d$,

$$-\langle u, v \rangle = -\frac{1}{2}\|u\|^2 + \frac{1}{2}\|u - v\|^2 - \frac{1}{2}\|v\|^2 \leq -\frac{1}{2}\|u\|^2 + \frac{1}{2}\|u - v\|^2.$$

Applying this inequality to the inner product in the gradient term, we obtain:

$$\begin{aligned} &\mathbb{E}_t \left[-\frac{\mu\mu_l}{M} \sum_{m=1}^M \sum_{b=0}^{B-1} \left\langle \nabla F(w^t), \nabla F_m(w_m^{t,b}) \right\rangle \right] \\ &= -\mu\mu_l \sum_{b=0}^{B-1} \mathbb{E}_t \left[\left\langle \nabla F(w^t), \frac{1}{M} \sum_{m=1}^M \nabla F_m(w_m^{t,b}) \right\rangle \right] \\ &\leq -\frac{\mu\mu_l B}{2} \|\nabla F(w^t)\|^2 + \frac{\mu\mu_l}{2} \sum_{b=0}^{B-1} \mathbb{E}_t \left[\left\| \frac{1}{M} \sum_{m=1}^M (\nabla F_m(w^t) - \nabla F_m(w_m^{t,b})) \right\|^2 \right] \\ &\leq -\frac{\mu\mu_l B}{2} \|\nabla F(w^t)\|^2 + \frac{\mu\mu_l}{2M} \sum_{m=1}^M \sum_{b=0}^{B-1} \mathbb{E}_t \left[\|\nabla F_m(w^t) - \nabla F_m(w_m^{t,b})\|^2 \right] \\ &\leq -\frac{\mu\mu_l B}{2} \|\nabla F(w^t)\|^2 + \frac{\mu\mu_l L^2}{2M} \sum_{m=1}^M \sum_{b=0}^{B-1} \mathbb{E}_t \left[\|w^t - w_m^{t,b}\|^2 \right], \end{aligned} \quad (4.53)$$

where the last line uses the L -smoothness of each local objective F_m .

We now decompose the second-moment term $\mathbb{E}_t \left[\|\tilde{g}^t\|^2 \right]$ as follows:

$$\begin{aligned}
\mathbb{E}_t \left[\|\tilde{g}^t\|^2 \right] &= \mathbb{E}_t \left[\left\| \frac{1}{K} \sum_{m \in \tilde{S}^t} \frac{g_m^t}{M q_{s,m}^t} \right\|^2 \right] \\
&= \mu_l^2 \cdot \mathbb{E}_t \left[\left\| \frac{1}{K} \sum_{m \in \tilde{S}^t} \frac{1}{M q_{s,m}^t} \sum_{b=0}^{B-1} \nabla \phi(w_m^{t,b}; \xi_m^{t,b}) \right\|^2 \right] \\
&= \mu_l^2 \cdot (I_1 + I_2),
\end{aligned} \tag{4.54}$$

where the decomposition separates the stochastic and deterministic components of the gradient via:

$$\begin{aligned}
I_1 &:= \mathbb{E}_t \left[\left\| \frac{1}{K} \sum_{m \in \tilde{S}^t} \frac{1}{M q_{s,m}^t} \sum_{b=0}^{B-1} \left(\nabla \phi(w_m^{t,b}; \xi_m^{t,b}) - \nabla F_m(w_m^{t,b}) \right) \right\|^2 \right], \\
I_2 &:= \mathbb{E}_t \left[\left\| \frac{1}{K} \sum_{m \in \tilde{S}^t} \frac{1}{M q_{s,m}^t} \sum_{b=0}^{B-1} \nabla F_m(w_m^{t,b}) \right\|^2 \right].
\end{aligned}$$

To bound I_1 , observe that

$$\begin{aligned}
I_1 &= \frac{1}{K} \sum_{m=1}^M \frac{1}{M^2 q_{s,m}^t} \mathbb{E}_t \left[\left\| \sum_{b=0}^{B-1} \left(\nabla \phi(w_m^{t,b}; \xi_m^{t,b}) - \nabla F_m(w_m^{t,b}) \right) \right\|_2^2 \right] \\
&= \frac{1}{K} \sum_{m=1}^M \frac{1}{M^2 q_{s,m}^t} \sum_{b=0}^{B-1} \mathbb{E}_t \left[\left\| \nabla \phi(w_m^{t,b}; \xi_m^{t,b}) - \nabla F_m(w_m^{t,b}) \right\|_2^2 \right] \\
&\leq \frac{\sigma^2 B}{M^2 K} \sum_{m=1}^M \frac{1}{q_{s,m}^t}.
\end{aligned}$$

By the assumption $q_{s,m}^t \geq \alpha/M$ for all m , it follows that

$$I_1 \leq \frac{\sigma^2 B}{K \alpha}.$$

We now derive an upper bound for I_2 . Using the triangle inequality, Jensen's inequality, and Lemma 48, we obtain:

$$\begin{aligned}
I_2 &\leq 2\mathbb{E}_t \left[\sum_{m=1}^M \frac{1}{M^2 q_{s,m}^t} \left\| \sum_{b=0}^{B-1} \left(\nabla F_m(w_m^{t,b}) - \nabla F_m(w^t) \right) \right\|_2^2 \right] \\
&\quad + 2B^2 \mathbb{E}_t \left[\left\| \frac{1}{K} \sum_{m \in \tilde{S}^t} \frac{1}{M q_{s,m}^t} \nabla F_m(w^t) - \nabla F(w^t) \right\|_2^2 \right] + 2B^2 \left\| \nabla F(w^t) \right\|_2^2 \\
&\leq 2B \mathbb{E}_t \left[\sum_{m=1}^M \frac{1}{M^2 q_{s,m}^t} \sum_{b=0}^{B-1} \left\| \nabla F_m(w_m^{t,b}) - \nabla F_m(w^t) \right\|_2^2 \right] + 2B^2 \left\| \nabla F(w^t) \right\|_2^2 \\
&\quad + \frac{2B^2}{K} \left(\frac{1}{M^2} \sum_{m=1}^M \frac{1}{q_{s,m}^t} \left\| \nabla F_m(w^t) \right\|_2^2 - \left\| \nabla F(w^t) \right\|_2^2 \right) \\
&= 2BL^2 \mathbb{E}_t \left[\sum_{m=1}^M \frac{1}{M^2 q_{s,m}^t} \sum_{b=0}^{B-1} \left\| w_m^{t,b} - w^t \right\|_2^2 \right] + 2B^2 \left\| \nabla F(w^t) \right\|_2^2 + \frac{2B^2}{K} \mathbb{E}_t \left[V(q_s^t, w^t) \right],
\end{aligned} \tag{4.55}$$

where the last line follows from the L -smoothness of F_m and the definition of the variance proxy $V(p, w)$ in Equation (4.9).

Finally, using the lower bound $q_{s,m}^t \geq \alpha/M$, we obtain:

$$I_2 \leq \frac{2BL^2}{M\alpha} \mathbb{E}_t \left[\sum_{m=1}^M \sum_{b=0}^{B-1} \left\| w_m^{t,b} - w^t \right\|_2^2 \right] + 2B^2 \left\| \nabla F(w^t) \right\|_2^2 + \frac{2B^2}{K} \mathbb{E}_t \left[V(q_s^t, w^t) \right]. \tag{4.56}$$

Combining the bounds from Equations (4.54) through (4.56), we obtain:

$$\begin{aligned}
\frac{L\mu^2}{2} \mathbb{E}_t \left[\left\| \tilde{g}^t \right\|_2^2 \right] &\leq \frac{\mu^2 \mu_l^2 L \sigma^2 B}{2K\alpha} + \frac{\mu^2 \mu_l^2 L^3 B}{M\alpha} \mathbb{E}_t \left[\sum_{m=1}^M \sum_{b=0}^{B-1} \left\| w_m^{t,b} - w^t \right\|_2^2 \right] \\
&\quad + \mu^2 \mu_l^2 B^2 L \left\| \nabla F(w^t) \right\|_2^2 + \frac{\mu^2 \mu_l^2 L B^2}{K} \mathbb{E}_t \left[V(q_s^t, w^t) \right].
\end{aligned}$$

Substituting this into Equation (4.52), and using the bound from Equation (4.53), we obtain:

$$\begin{aligned}\mathbb{E}_t \left[F(w^{t+1}) \right] &\leq F(w^t) - \frac{\mu\mu_l B}{2} (1 - 2\mu\mu_l BL) \left\| \nabla F(w^t) \right\|_2^2 \\ &\quad + \frac{\mu\mu_l L^2}{M} \left(\frac{1}{2} + \frac{\mu\mu_l LB}{\alpha} \right) \mathbb{E}_t \left[\sum_{m=1}^M \sum_{b=0}^{B-1} \left\| w_m^{t,b} - w^t \right\|_2^2 \right] \\ &\quad + \frac{\mu^2 \mu_l^2 L \sigma^2 B}{2K\alpha} + \frac{\mu^2 \mu_l^2 LB^2}{K} \mathbb{E}_t \left[V(q_s^t, w^t) \right] + \frac{BL\mu^2 \mu_l^2}{2} \mathbb{E}_t \left[l_t(\hat{p}^t) - l_t(q_s^t) \right].\end{aligned}$$

Now define the shorthand $\tilde{\mu} := \mu\mu_l$ and assume $\tilde{\mu} \leq \frac{1}{4BL}$ so that $1 - 2\tilde{\mu}BL \geq \frac{1}{2}$. Then the above simplifies to:

$$\begin{aligned}\mathbb{E}_t \left[F(w^{t+1}) \right] &\leq F(w^t) - \frac{\tilde{\mu}B}{4} \left\| \nabla F(w^t) \right\|_2^2 \\ &\quad + \tilde{\mu}L^2 \left(\frac{1}{2} + \frac{1}{4\alpha} \right) \mathbb{E}_t \left[\frac{1}{M} \sum_{m=1}^M \sum_{b=0}^{B-1} \left\| w_m^{t,b} - w^t \right\|_2^2 \right] \\ &\quad + \frac{\tilde{\mu}^2 BL \sigma^2}{2K\alpha} + \frac{\tilde{\mu}^2 B^2 L}{K} \mathbb{E}_t \left[V(q_s^t, w^t) \right] + \frac{BL\tilde{\mu}^2}{2} \mathbb{E}_t \left[l_t(\hat{p}^t) - l_t(q_s^t) \right].\end{aligned}\tag{4.57}$$

Following Lemma 8 of Karimireddy et al. [2020b], we now establish the bound

$$\mathbb{E}_t \left[\frac{1}{M} \sum_{m=1}^M \sum_{b=0}^{B-1} \left\| w_m^{t,b} - w^t \right\|_2^2 \right] \leq 8B^3 \mu_l^2 \zeta_{\text{unif}}^2 + 8B^3 \mu_l^2 \left\| \nabla F(w^t) \right\|_2^2 + 4B^2 \mu_l^2 \sigma^2. \tag{4.58}$$

To prove this, note that for $b = 0$, we have $w_m^{t,0} = w^t$; hence

$$\mathbb{E}_t \left[\left\| w_m^{t,0} - w^t \right\|_2^2 \right] = 0.$$

For $1 \leq b \leq B-1$, using the update rule and expanding recursively, we obtain:

$$\begin{aligned}
& \mathbb{E}_t \left[\left\| w_m^{t,b} - w^t \right\|_2^2 \right] \\
&= \mathbb{E}_t \left[\left\| w_m^{t,b-1} - \mu_l \nabla \phi(w_m^{t,b-1}; \xi_m^{t,b-1}) - w^t \right\|_2^2 \right] \\
&= \mathbb{E}_t \left[\left\| w_m^{t,b-1} - \mu_l \nabla F_m(w_m^{t,b-1}) - w^t \right\|_2^2 \right] \\
&\quad + \mu_l^2 \mathbb{E}_t \left[\left\| \nabla \phi(w_m^{t,b-1}; \xi_m^{t,b-1}) - \nabla F_m(w_m^{t,b-1}) \right\|_2^2 \right] \\
&\leq \left(1 + \frac{1}{B-1} \right) \mathbb{E}_t \left[\left\| w_m^{t,b-1} - w^t \right\|_2^2 \right] + B\mu_l^2 \mathbb{E}_t \left[\left\| \nabla F_m(w_m^{t,b-1}) \right\|_2^2 \right] + \mu_l^2 \sigma^2 \\
&\leq \left(1 + \frac{1}{B-1} \right) \mathbb{E}_t \left[\left\| w_m^{t,b-1} - w^t \right\|_2^2 \right] \\
&\quad + 2B\mu_l^2 \left\| \nabla F_m(w_m^{t,b-1}) - \nabla F_m(w^t) \right\|_2^2 + 2B\mu_l^2 \left\| \nabla F_m(w^t) \right\|_2^2 + \mu_l^2 \sigma^2 \\
&\leq \left(1 + \frac{1}{B-1} + 2B\mu_l^2 L^2 \right) \mathbb{E}_t \left[\left\| w_m^{t,b-1} - w^t \right\|_2^2 \right] + 2B\mu_l^2 \left\| \nabla F_m(w^t) \right\|_2^2 + \mu_l^2 \sigma^2,
\end{aligned}$$

where the first inequality uses Assumption 11 and Lemma 48, and the second and third use Jensen's inequality and the L -smoothness assumption (Assumption 9).

Assume now that $\mu_l \leq \frac{1}{\sqrt{2BL}}$, so that $2B\mu_l^2 L^2 \leq \frac{1}{B} \leq \frac{1}{B-1}$. Then the recurrence becomes:

$$\mathbb{E}_t \left[\left\| w_m^{t,b} - w^t \right\|_2^2 \right] \leq \left(1 + \frac{2}{B-1} \right) \mathbb{E}_t \left[\left\| w_m^{t,b-1} - w^t \right\|_2^2 \right] + 2B\mu_l^2 \left\| \nabla F_m(w^t) \right\|_2^2 + \mu_l^2 \sigma^2.$$

By induction, this implies:

$$\mathbb{E}_t \left[\left\| w_m^{t,b} - w^t \right\|_2^2 \right] \leq \left(2B\mu_l^2 \left\| \nabla F_m(w^t) \right\|_2^2 + \mu_l^2 \sigma^2 \right) \sum_{\tau=0}^{b-1} \left(1 + \frac{2}{B-1} \right)^\tau.$$

The geometric sum is bounded by:

$$\sum_{\tau=0}^{b-1} \left(1 + \frac{2}{B-1}\right)^\tau = \frac{\left(1 + \frac{2}{B-1}\right)^b - 1}{\left(1 + \frac{2}{B-1}\right) - 1} \leq \frac{B-1}{2} \left(\left(1 + \frac{2}{B-1}\right)^{B-1} - 1 \right) \leq 4B,$$

which gives:

$$\mathbb{E}_t \left[\left\| w_m^{t,b} - w^t \right\|_2^2 \right] \leq 8B^2 \mu_l^2 \left\| \nabla F_m(w^t) \right\|_2^2 + 4B \mu_l^2 \sigma^2.$$

Summing over b and averaging over m , we conclude:

$$\begin{aligned} \frac{1}{M} \sum_{m=1}^M \sum_{b=0}^{B-1} \mathbb{E}_t \left[\left\| w_m^{t,b} - w^t \right\|_2^2 \right] &\leq 8B^3 \mu_l^2 \cdot \frac{1}{M} \sum_{m=1}^M \left\| \nabla F_m(w^t) \right\|_2^2 + 4B^2 \mu_l^2 \sigma^2 \\ &\leq 8B^3 \mu_l^2 \left(\zeta_{\text{unif}}^2 + \left\| \nabla F(w^t) \right\|_2^2 \right) + 4B^2 \mu_l^2 \sigma^2, \end{aligned}$$

where the final inequality follows from the definition of ζ_{unif}^2 .

This completes the proof of Equation (4.58).

We now upper bound the interpolation term in Equation (4.57) using the result from Equation (4.58):

$$\begin{aligned} &\tilde{\mu} L^2 \left(\frac{1}{2} + \frac{1}{4\alpha} \right) \mathbb{E}_t \left[\frac{1}{M} \sum_{m=1}^M \sum_{b=0}^{B-1} \left\| w_m^{t,b} - w^t \right\|_2^2 \right] \\ &\leq \left(4 + \frac{2}{\alpha} \right) \tilde{\mu} \mu_l^2 B^3 L^2 \zeta_{\text{unif}}^2 + \left(4 + \frac{2}{\alpha} \right) \tilde{\mu} \mu_l^2 B^3 L^2 \left\| \nabla F(w^t) \right\|_2^2 + \left(2 + \frac{1}{\alpha} \right) \tilde{\mu} \mu_l^2 B^2 L^2 \sigma^2. \end{aligned}$$

Substituting into (4.57), we obtain:

$$\begin{aligned} \mathbb{E}_t \left[F(w^{t+1}) \right] &\leq F(w^t) - \frac{\tilde{\mu} B}{4} \left(1 - 4 \left(4 + \frac{2}{\alpha} \right) \mu_l^2 B^2 L^2 \right) \left\| \nabla F(w^t) \right\|_2^2 \\ &\quad + \left(4 + \frac{2}{\alpha} \right) \tilde{\mu} \mu_l^2 B^2 L^2 \left(B \zeta_{\text{unif}}^2 + \frac{\sigma^2}{2} \right) \\ &\quad + \frac{\tilde{\mu}^2 B L}{K} \left(\frac{\sigma^2}{2\alpha} + B \mathbb{E}_t \left[V(q_s^t, w^t) \right] \right) + \frac{B L \tilde{\mu}^2}{2} \mathbb{E}_t \left[l_t(\hat{p}^t) - l_t(q_s^t) \right]. \end{aligned}$$

Assuming $\mu_l \leq \frac{1}{4BL} \sqrt{\frac{1}{2+1/\alpha}}$, so that the coefficient on $\|\nabla F(w^t)\|^2$ remains positive, we further simplify:

$$\begin{aligned} \mathbb{E}_t \left[F(w^{t+1}) \right] &\leq F(w^t) - \frac{\tilde{\mu}B}{8} \left\| \nabla F(w^t) \right\|_2^2 + \left(4 + \frac{2}{\alpha} \right) \tilde{\mu} \mu_l^2 B^2 L^2 \left(B \zeta_{\text{unif}}^2 + \frac{\sigma^2}{2} \right) \\ &\quad + \frac{\tilde{\mu}^2 BL}{K} \left(\frac{\sigma^2}{2\alpha} + B \mathbb{E}_t \left[V(q_s^t, w^t) \right] \right) + \frac{BL \tilde{\mu}^2}{2} \mathbb{E}_t \left[l_t(\hat{p}^t) - l_t(q_s^t) \right]. \end{aligned}$$

Rearranging, we obtain a bound on the squared gradient norm:

$$\begin{aligned} \left\| \nabla F(w^t) \right\|_2^2 &\leq \frac{8}{\tilde{\mu}B} \left(F(w^t) - \mathbb{E}_t \left[F(w^{t+1}) \right] \right) + 8 \left(4 + \frac{2}{\alpha} \right) \mu_l^2 B^2 L^2 \left(\zeta_{\text{unif}}^2 + \frac{\sigma^2}{2B} \right) \\ &\quad + \frac{8\tilde{\mu}BL}{K} \left(\frac{\sigma^2}{2B\alpha} + \mathbb{E}_t \left[V(q_s^t, w^t) \right] \right) + 4L\tilde{\mu} \mathbb{E}_t \left[l_t(\hat{p}^t) - l_t(q_s^t) \right]. \end{aligned}$$

Taking full expectation and averaging over $t = 1$ to T , we obtain:

$$\begin{aligned} &\frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[\left\| \nabla F(w^t) \right\|_2^2 \right] \\ &\leq \frac{8}{\tilde{\mu}B} \left(F(w^0) - \mathbb{E}[F(w^T)] \right) + 8 \left(4 + \frac{2}{\alpha} \right) \mu_l^2 B^2 L^2 \left(\zeta_{\text{unif}}^2 + \frac{\sigma^2}{2B} \right) \\ &\quad + \frac{8\tilde{\mu}BL}{K} \left(\frac{\sigma^2}{2B\alpha} + \frac{1}{T} \sum_{t=1}^T \mathbb{E}_t \left[V(q_s^t, w^t) \right] \right) + 4L\tilde{\mu} \cdot \frac{\text{D-Regret}_T(q^{1:T})}{T}. \end{aligned}$$

Finally, by definition of $\zeta_T^2(\alpha, \beta)$ from Equation (4.44), we have:

$$\begin{aligned} &\mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T V(q_s^t, w^t) \right] \\ &= \frac{1}{T} \mathbb{E}_{w^1, \dots, w^T} \left[\sum_{t=1}^T V(q_s^t, w^t) \right] \leq \frac{1}{T} \sup_{w^1} \cdots \sup_{w^T} \sum_{t=1}^T V(q_s^t, w^t) = \zeta_T^2(\alpha, \beta). \end{aligned}$$

Substituting this bound completes the proof:

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[\left\| \nabla F(w^t) \right\|_2^2 \right] &\leq \frac{8}{\tilde{\mu}B} \left(F(w^0) - F^\star \right) + 8 \left(4 + \frac{2}{\alpha} \right) \mu_l^2 B^2 L^2 \left(\zeta_{\text{unif}}^2 + \frac{\sigma^2}{2B} \right) \\ &\quad + \frac{8\tilde{\mu}BL}{K} \left(\frac{\sigma^2}{2B\alpha} + \zeta_T^2(\alpha, \beta) \right) + 4L\tilde{\mu} \cdot \frac{\text{D-Regret}_T(q^{1:T})}{T}. \end{aligned}$$

Conclusion. In summary, under the conditions $\tilde{\mu} \leq \frac{1}{4BL}$ and $\mu_l \leq \frac{1}{4BL} \sqrt{\frac{1}{2+1/\alpha}}$, we obtain the following bound on the expected squared gradient norm:

$$\begin{aligned} \mathbb{E} \left[\left\| \nabla F(w^R) \right\|_2^2 \right] &= \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[\left\| \nabla F(w^t) \right\|_2^2 \right] \\ &\leq \frac{8}{\tilde{\mu}BT} \left(F(w^0) - F^\star \right) + 8 \left(4 + \frac{2}{\alpha} \right) \mu_l^2 B^2 L^2 \left(\zeta_{\text{unif}}^2 + \frac{\sigma^2}{2B} \right) \\ &\quad + 4\tilde{\mu}BL \left(\frac{2\zeta_T^2(\alpha, \beta)}{K} + \frac{\sigma^2}{KB\alpha} + \frac{\text{D-Regret}_T(q_s^{1:T})}{BT} \right). \end{aligned}$$

Now suppose $\mu \geq 1$, and $\tilde{\mu} \leq \frac{1}{4BL} \sqrt{\frac{1}{1+1/(2\alpha)}}$. Letting $D^F := F(w^1) - F^\star$, we conclude that:

$$\begin{aligned} \mathbb{E} \left[\left\| \nabla F(w^R) \right\|_2^2 \right] &\leq \frac{8D^F}{\tilde{\mu}BT} + 8 \left(4 + \frac{2}{\alpha} \right) \tilde{\mu}^2 B^2 L^2 \left(\zeta_{\text{unif}}^2 + \frac{\sigma^2}{2B} \right) \\ &\quad + 4\tilde{\mu}BL \left(\frac{2\zeta_T^2(\alpha, \beta)}{K} + \frac{\sigma^2}{KB\alpha} + \frac{\text{D-Regret}_T(q_s^{1:T})}{BT} \right). \end{aligned} \tag{4.59}$$

We now turn to bounding the regret term. Recall that

$$g_m^t = \mu_l \sum_{b=0}^{B-1} \nabla \phi(w_m^{t,b}; \xi_m^{t,b}).$$

By Assumption 11, we have $\|\nabla \phi(w; \xi)\|_2 \leq G$ for all w and ξ , so it follows that

$$\left\| g_m^t \right\|_2^2 \leq B^2 \mu_l^2 G^2.$$

Therefore, using the definition $a_m^t = \frac{\|g_m^t\|_2^2}{B(M\mu_l)^2}$, we get:

$$a_m^t \leq \frac{BG^2}{M^2}, \quad a_{\max}^t \leq \frac{BG^2}{M^2}.$$

Applying Theorem 40, and noting that $q_s^t \in \mathcal{A}$ for all $t \in [T]$ and $\text{TV}(q_s^{1:T}) \leq \beta$, we obtain:

$$\text{D-Regret}_T(q_s^{1:T}) \leq \frac{\log M}{\eta} + \frac{2\beta \log(M/\alpha)}{\eta} + \frac{\eta T M^2 B^2 G^4}{2K^2 \alpha^6}.$$

Optimizing the learning rate by setting

$$\eta = \frac{K\alpha^3}{MBG^2} \sqrt{\frac{2\log M + 4\beta \log(M/\alpha)}{T}},$$

yields the bound

$$\text{D-Regret}_T(q_s^{1:T}) \leq \sqrt{T} \cdot \frac{MBG^2}{K\alpha^3} \cdot \sqrt{\frac{1}{2} \log M + \beta \log(M/\alpha)}.$$

Substituting this regret bound into the convergence result (4.59), we obtain the final result:

$$\begin{aligned} \mathbb{E} \left[\left\| \nabla F(w^R) \right\|_2^2 \right] &\leq \frac{8D^F}{\tilde{\mu}BT} + 8 \left(4 + \frac{2}{\alpha} \right) \tilde{\mu}^2 B^2 L^2 \left(\zeta_{\text{unif}}^2 + \frac{\sigma^2}{2B} \right) \\ &\quad + 4\tilde{\mu}BL \left(\frac{2\zeta_T^2(\alpha, \beta)}{K} + \frac{\sigma^2}{KB\alpha} + \frac{MG^2}{K\alpha^3} \cdot \sqrt{\frac{\frac{1}{2} \log M + \beta \log(M/\alpha)}{T}} \right). \end{aligned} \tag{4.60}$$

To optimize the final bound in (4.60), we consider two balancing strategies for the leading terms.

First, set

$$\frac{8D^F}{\tilde{\mu}BT} = 8 \left(4 + \frac{2}{\alpha} \right) \tilde{\mu}^2 B^2 L^2 \left(\zeta_{\text{unif}}^2 + \frac{\sigma^2}{2B} \right),$$

which yields the optimal step size

$$\tilde{\mu} = \frac{(D^F)^{1/3}}{\left(4 + \frac{2}{\alpha}\right)^{1/3} BL^{2/3} \left(\zeta_{\text{unif}}^2 + \frac{\sigma^2}{2B}\right)^{1/3} T^{1/3}}.$$

This leads to the bound

$$\frac{8D^F}{\tilde{\mu}BT} + 8 \left(4 + \frac{2}{\alpha}\right) \tilde{\mu}^2 B^2 L^2 \left(\zeta_{\text{unif}}^2 + \frac{\sigma^2}{2B}\right) \leq \frac{16 \left(4 + \frac{2}{\alpha}\right)^{1/3} (D^F)^{2/3} L^{2/3} \left(\zeta_{\text{unif}}^{2/3} + \frac{\sigma^{2/3}}{(2B)^{1/3}}\right)}{T^{2/3}}.$$

Second, balance the first term against the final error term:

$$\frac{8D^F}{\tilde{\mu}BT} = 4\tilde{\mu}BL \left(\frac{2\zeta_T^2(\alpha, \beta)}{K} + \frac{\sigma^2}{KB\alpha} + \frac{MG^2}{K\alpha^3} \sqrt{\frac{\frac{1}{2} \log M + \beta \log(M/\alpha)}{T}} \right),$$

which gives the optimal step size

$$\tilde{\mu} = \frac{\sqrt{2D^F}}{B\sqrt{L} \sqrt{\frac{2\zeta_T^2(\alpha, \beta)}{K} + \frac{\sigma^2}{KB\alpha} + \sqrt{\frac{\frac{1}{2} \log M + \beta \log(M/\alpha)}{T}} \sqrt{T}}}.$$

Substituting this yields:

$$\frac{8D^F}{\tilde{\mu}BT} + 4\tilde{\mu}BL(\dots) \leq \frac{8\sqrt{2D^F}L}{\sqrt{T}} \left(\frac{\sqrt{2}\zeta_T(\alpha, \beta)}{\sqrt{K}} + \frac{\sigma}{\sqrt{KB\alpha}} + \left(\frac{\frac{1}{2} \log M + \beta \log(M/\alpha)}{T} \right)^{1/4} \right).$$

Putting the pieces together: under $\mu \geq 1$ and

$$\tilde{\mu} = \min \left\{ \frac{1}{4BL} \sqrt{\frac{1}{2 + 1/\alpha}}, \frac{(D^F)^{1/3}}{\left(4 + \frac{2}{\alpha}\right)^{1/3} BL^{2/3} \left(\zeta_{\text{unif}}^2 + \frac{\sigma^2}{2B}\right)^{1/3} T^{1/3}}, \frac{\sqrt{2D^F}}{B\sqrt{L}\sqrt{\dots}\sqrt{T}} \right\},$$

we obtain the convergence guarantee

$$\begin{aligned} & \mathbb{E} \left[\left\| \nabla F(w^R) \right\|_2^2 \right] \\ & \lesssim \frac{D^F L \sqrt{2 + \frac{1}{\alpha}}}{T} + \frac{\left(4 + \frac{2}{\alpha}\right)^{1/3} (D^F L)^{2/3} \zeta_{\text{unif}}^{2/3}}{T^{2/3}} + \frac{\left(4 + \frac{2}{\alpha}\right)^{1/3} (D^F L)^{2/3} \sigma^{2/3}}{B^{1/3} T^{2/3}} \\ & \quad + \frac{\sqrt{D^F L} \zeta_T(\alpha, \beta)}{\sqrt{TK}} + \frac{\sqrt{D^F L} \sigma}{\sqrt{TKB\alpha}} + \frac{\sqrt{D^F L}}{\sqrt{T}} \left(\frac{\frac{1}{2} \log M + \beta \log(M/\alpha)}{T} \right)^{1/4}. \end{aligned}$$

4.10.5 Proof of Theorem 40

We begin by stating a key proposition that serves as a building block for proving Theorem 40. The main distinction between Theorem 40 and Proposition 52 lies in the admissibility of the comparator sequence: in Proposition 52, the comparator sequence is assumed to lie entirely in the set $\mathcal{A}$, and hence no projection error arises.

Proposition 52. *Suppose the conditions of Theorem 40 hold. Then for any comparator sequence $q^{1:T}$ with $q^t \in \mathcal{A}$ for all $t \in [T]$, we have*

$$D\text{-Regret}_T(q^{1:T}) \leq \frac{\log M}{\eta} + \frac{2 \log(M/\alpha)}{\eta} \mathbb{E} \left[TV(q^{1:T}) \right] + \frac{\eta M^6}{2K^2 \alpha^6} \sum_{t=1}^T \mathbb{E} \left[\left(a_{\max}^t \right)^2 \right].$$

Proof. By Lemma 46 and the definition of $\hat{p}^{t+1}$ in Step 7 of Algorithm 5, we have

$$\langle \hat{p}^{t+1} - q^t, \nabla \hat{l}_t(\hat{p}^t; \hat{p}^t) \rangle \leq \frac{1}{\eta} \langle \nabla \Phi(\hat{p}^t) - \nabla \Phi(\hat{p}^{t+1}), \hat{p}^{t+1} - q^t \rangle. \quad (4.61)$$

Using the convexity of $\hat{l}_t(\cdot; \hat{p}^t)$, we obtain

$$\begin{aligned} \hat{l}_t(\hat{p}^t; \hat{p}^t) - \hat{l}_t(q^t; \hat{p}^t) & \leq \langle \nabla \hat{l}_t(\hat{p}^t; \hat{p}^t), \hat{p}^t - q^t \rangle \\ & = \langle \nabla \hat{l}_t(\hat{p}^t; \hat{p}^t), \hat{p}^{t+1} - q^t \rangle + \langle \nabla \hat{l}_t(\hat{p}^t; \hat{p}^t), \hat{p}^t - \hat{p}^{t+1} \rangle. \end{aligned}$$

Substituting inequality (4.61), we get

$$\hat{l}_t(\hat{p}^t; \hat{p}^t) - \hat{l}_t(q^t; \hat{p}^t) \leq \frac{1}{\eta} \langle \nabla \Phi(\hat{p}^t) - \nabla \Phi(\hat{p}^{t+1}), \hat{p}^{t+1} - q^t \rangle + \langle \nabla \hat{l}_t(\hat{p}^t; \hat{p}^t), \hat{p}^t - \hat{p}^{t+1} \rangle.$$

By the three-point identity for Bregman divergence,

$$D_\Phi(x_1 \| x_2) = D_\Phi(x_3 \| x_2) + D_\Phi(x_1 \| x_3) + \langle \nabla \Phi(x_2) - \nabla \Phi(x_3), x_3 - x_1 \rangle, \quad \forall x_1, x_2, x_3 \in \mathcal{D},$$

we derive:

$$\begin{aligned} & \hat{l}_t(\hat{p}^t; \hat{p}^t) - \hat{l}_t(q^t; \hat{p}^t) \\ & \leq \frac{1}{\eta} \left[D_\Phi(q^t \| \hat{p}^t) - D_\Phi(q^t \| \hat{p}^{t+1}) - D_\Phi(\hat{p}^{t+1} \| \hat{p}^t) \right] \\ & \quad + \langle \nabla \hat{l}_t(\hat{p}^t; \hat{p}^t), \hat{p}^t - \hat{p}^{t+1} \rangle \\ & = \frac{1}{\eta} \left[D_\Phi(q^t \| \hat{p}^t) - D_\Phi(q^{t+1} \| \hat{p}^{t+1}) \right] + \frac{1}{\eta} \left[D_\Phi(q^{t+1} \| \hat{p}^{t+1}) - D_\Phi(q^t \| \hat{p}^{t+1}) \right] \\ & \quad - \frac{1}{\eta} D_\Phi(\hat{p}^{t+1} \| \hat{p}^t) + \langle \nabla \hat{l}_t(\hat{p}^t; \hat{p}^t), \hat{p}^t - \hat{p}^{t+1} \rangle. \end{aligned} \tag{4.62}$$

We now bound the second term in (4.62). Using the convexity of Φ , Hölder's inequality, and the uniform lower bound α on elements of $\mathcal{A}$, we have:

$$\begin{aligned} D_\Phi(q^{t+1} \| \hat{p}^{t+1}) - D_\Phi(q^t \| \hat{p}^{t+1}) & = \Phi(q^{t+1}) - \Phi(q^t) - \langle \nabla \Phi(\hat{p}^{t+1}), q^{t+1} - q^t \rangle \\ & \leq \langle \nabla \Phi(q^{t+1}) - \nabla \Phi(\hat{p}^{t+1}), q^{t+1} - q^t \rangle \\ & \leq \|\nabla \Phi(q^{t+1}) - \nabla \Phi(\hat{p}^{t+1})\|_\infty \cdot \|q^{t+1} - q^t\|_1 \\ & \leq 2 \log(M/\alpha) \cdot \|q^{t+1} - q^t\|_1. \end{aligned} \tag{4.63}$$

We now bound the final two terms in (4.62). Since Φ is 1-strongly convex with respect

to $\|\cdot\|_1$ (by Pinsker's inequality), we have:

$$-\frac{1}{\eta}D_{\Phi}(\hat{p}^{t+1}||\hat{p}^t) + \langle \nabla \hat{l}_t(\hat{p}^t; \hat{p}^t), \hat{p}^t - \hat{p}^{t+1} \rangle \leq -\frac{1}{2\eta} \|\hat{p}^{t+1} - \hat{p}^t\|_1^2 + \|\nabla \hat{l}_t(\hat{p}^t; \hat{p}^t)\|_{\infty} \cdot \|\hat{p}^t - \hat{p}^{t+1}\|_1.$$

Applying Young's inequality $ab \leq \frac{a^2}{2\epsilon} + \frac{\epsilon b^2}{2}$, we get:

$$-\frac{1}{\eta}D_{\Phi}(\hat{p}^{t+1}||\hat{p}^t) + \langle \nabla \hat{l}_t(\hat{p}^t; \hat{p}^t), \hat{p}^t - \hat{p}^{t+1} \rangle \leq \frac{\eta}{2} \|\nabla \hat{l}_t(\hat{p}^t; \hat{p}^t)\|_{\infty}^2.$$

Let

$$Q_t := \frac{M^3 a_{\max}^t}{K \alpha^3}.$$

Then for all $m \in [M]$, we have

$$\left| \left[\nabla \hat{l}_t(q; p^t) \right]_m \right| = \frac{1}{K^2} \cdot \frac{a_m^t}{q_m^2 p_m^t} \cdot \mathbb{I}\{m \in S^t\} \leq \frac{M^3 a_{\max}^t}{K \alpha^3} = Q_t.$$

Thus, $\|\nabla \hat{l}_t(q; p^t)\|_{\infty} \leq Q_t$, which gives

$$-\frac{1}{\eta}D_{\Phi}(\hat{p}^{t+1}||\hat{p}^t) + \langle \nabla \hat{l}_t(\hat{p}^t; \hat{p}^t), \hat{p}^t - \hat{p}^{t+1} \rangle \leq \frac{\eta}{2} Q_t^2. \quad (4.64)$$

Combining (4.62)–(4.64), we obtain:

$$\hat{l}_t(\hat{p}^t; \hat{p}^t) - \hat{l}_t(q^t; \hat{p}^t) \leq \frac{D_{\Phi}(q^t||\hat{p}^t) - D_{\Phi}(q^{t+1}||\hat{p}^{t+1})}{\eta} + \frac{2 \log(M/\alpha)}{\eta} \|q^{t+1} - q^t\|_1 + \frac{\eta}{2} Q_t^2.$$

Summing over $t = 1$ to T , we get:

$$\begin{aligned} \sum_{t=1}^T \hat{l}_t(\hat{p}^t; \hat{p}^t) - \sum_{t=1}^T \hat{l}_t(q^t; \hat{p}^t) &\leq \frac{D_{\Phi}(q^1||\hat{p}^1)}{\eta} + \frac{2 \log(M/\alpha)}{\eta} \sum_{t=1}^T \|q^{t+1} - q^t\|_1 + \frac{\eta}{2} \sum_{t=1}^T Q_t^2 \\ &\leq \frac{\log M}{\eta} + \frac{2 \log(M/\alpha)}{\eta} \text{TV}(q^{1:T}) + \frac{\eta}{2} \sum_{t=1}^T Q_t^2, \end{aligned} \quad (4.65)$$

where we use $D_{\Phi}(q^1 \|\hat{p}^1) \leq \log M$ due to the initialization $\hat{p}^1 = p^{\text{unif}}$.

Finally, by taking expectation and applying the unbiasedness of $\hat{l}_t$:

$$\mathbb{E} \left[\sum_{t=1}^T \hat{l}_t(\hat{p}^t; \hat{p}^t) - \hat{l}_t(q^t; \hat{p}^t) \right] = \sum_{t=1}^T \mathbb{E} \left[l_t(\hat{p}^t) - l_t(q^t) \right] = \text{D-Regret}_T(q^{1:T}).$$

Taking expectations on both sides of (4.65) yields the desired result. □

We are now ready to prove Theorem 40.

Proof of Theorem 40. Let $q^{1:T}$ be any comparator sequence with $q^t \in \mathcal{P}_{M-1}$ for all $t \in [T]$.

We begin by constructing an auxiliary sequence $\tilde{q}^{1:T}$ defined by

$$\tilde{q}_m^t = \begin{cases} \alpha/M & \text{if } q_m^t < \alpha/M, \\ q_m^t - \omega(q^t, \alpha) \left(q_m^t - \frac{\alpha}{M} \right) & \text{if } q_m^t \geq \alpha/M, \end{cases} \quad (4.66)$$

where $\omega(q^t, \alpha)$ is defined in (4.19).

Step 1: Verify that $\tilde{q}^t \in \mathcal{A}$. First, we show $\tilde{q}_m^t \geq \alpha/M$ for all $m \in [M]$. This holds by construction. Second, to show normalization, we compute:

$$\begin{aligned} \sum_{m=1}^M \tilde{q}_m^t &= \sum_{m=1}^M \frac{\alpha}{M} \mathbb{1} \left\{ q_m^t < \frac{\alpha}{M} \right\} + \sum_{m=1}^M \left(q_m^t - \omega(q^t, \alpha) \left(q_m^t - \frac{\alpha}{M} \right) \right) \mathbb{1} \left\{ q_m^t \geq \frac{\alpha}{M} \right\} \\ &= \sum_{m=1}^M q_m^t \mathbb{1} \left\{ q_m^t \geq \frac{\alpha}{M} \right\} + \sum_{m=1}^M \frac{\alpha}{M} \mathbb{1} \left\{ q_m^t < \frac{\alpha}{M} \right\} \\ &\quad - \omega(q^t, \alpha) \sum_{m=1}^M \left(q_m^t - \frac{\alpha}{M} \right) \mathbb{1} \left\{ q_m^t \geq \frac{\alpha}{M} \right\} \\ &= 1 - \sum_{m=1}^M \left(\frac{\alpha}{M} - q_m^t \right) \mathbb{1} \left\{ q_m^t < \frac{\alpha}{M} \right\} + \sum_{m=1}^M \frac{\alpha}{M} \mathbb{1} \left\{ q_m^t < \frac{\alpha}{M} \right\} \\ &= 1. \end{aligned}$$

To ensure non-negativity, we note that

$$\sum_{m=1}^M \left(q_m^t - \frac{\alpha}{M} \right) \mathbb{1} \left\{ q_m^t \geq \frac{\alpha}{M} \right\} \geq \sum_{m=1}^M \left(\frac{\alpha}{M} - q_m^t \right) \mathbb{1} \left\{ q_m^t < \frac{\alpha}{M} \right\},$$

which guarantees $0 \leq \omega(q^t, \alpha) \leq 1$. Hence, $\tilde{q}^t \in \mathcal{A}$.

Step 2: Decompose the regret. We decompose the regret as

$$\text{D-Regret}_T(q^{1:T}) = \mathbb{E} \left[\sum_{t=1}^T l_t(\hat{p}^t) - \sum_{t=1}^T l_t(\tilde{q}^t) + \sum_{t=1}^T l_t(\tilde{q}^t) - \sum_{t=1}^T l_t(q^t) \right]. \quad (4.67)$$

Applying Proposition 52 to $\tilde{q}^{1:T} \in \mathcal{A}$ yields:

$$\begin{aligned} \mathbb{E} \left[\sum_{t=1}^T l_t(\hat{p}^t) - \sum_{t=1}^T l_t(\tilde{q}^t) \right] &\leq \frac{\log M}{\eta} + \frac{2 \log(M/\alpha)}{\eta} \mathbb{E} [\text{TV}(q^{1:T})] \\ &\quad + \frac{\eta M^6}{2K^2 \alpha^6} \sum_{t=1}^T \mathbb{E} [(a_{\max}^t)^2] + \frac{2 \log(M/\alpha)}{\eta} \mathbb{E} [\text{TV}(\tilde{q}^{1:T}) - \text{TV}(q^{1:T})]. \end{aligned} \quad (4.68)$$

Step 3: Bound loss difference $\sum_t l_t(\tilde{q}^t) - l_t(q^t)$. For each $m \in [M]$, if $q_m^t < \alpha/M$, then $\tilde{q}_m^t > q_m^t$, implying $1/\tilde{q}_m^t < 1/q_m^t$. Otherwise, when $q_m^t \geq \alpha/M$, we compute:

$$\frac{1}{\tilde{q}_m^t} - \frac{1}{q_m^t} = \frac{1}{q_m^t} \cdot \frac{\omega(q^t, \alpha) \left(1 - \frac{\alpha}{M q_m^t} \right)}{1 - \omega(q^t, \alpha) \left(1 - \frac{\alpha}{M q_m^t} \right)} \leq \frac{\phi(q^t, \alpha)}{q_m^t}.$$

Thus,

$$\begin{aligned} \sum_{t=1}^T [l_t(\tilde{q}^t) - l_t(q^t)] &= \frac{1}{K} \sum_{t=1}^T \sum_{m=1}^M a_m^t \left(\frac{1}{\tilde{q}_m^t} - \frac{1}{q_m^t} \right) \\ &\leq \frac{1}{K} \sum_{t=1}^T \phi(q^t, \alpha) \sum_{m=1}^M \frac{a_m^t}{q_m^t} \mathbb{1} \left\{ q_m^t \geq \alpha/M \right\} \leq \frac{1}{K} \sum_{t=1}^T \phi(q^t, \alpha) l_t(q^t). \end{aligned} \quad (4.69)$$

Step 4: Bound TV difference. Using the triangle inequality:

$$\begin{aligned}
\text{TV}(\tilde{q}^{1:T}) &= \sum_{t=2}^T \|\tilde{q}^t - \tilde{q}^{t-1}\|_1 \\
&\leq \sum_{t=2}^T \|\tilde{q}^t - q^t\|_1 + \sum_{t=2}^T \|q^t - q^{t-1}\|_1 + \sum_{t=2}^T \|q^{t-1} - \tilde{q}^{t-1}\|_1 \\
&\leq \text{TV}(q^{1:T}) + 2 \sum_{t=1}^T \|\tilde{q}^t - q^t\|_1.
\end{aligned}$$

We now bound $\|\tilde{q}^t - q^t\|_1$:

$$\|\tilde{q}^t - q^t\|_1 = 2\psi(q^t, \alpha) \quad \Rightarrow \quad \text{TV}(\tilde{q}^{1:T}) - \text{TV}(q^{1:T}) \leq 4 \sum_{t=1}^T \psi(q^t, \alpha).$$

Step 5: Combine results. Substituting (4.67), (4.68), (4.69), and the above bound on the TV difference, and taking expectations, we conclude the proof. □

4.10.6 Proof of Theorem 42

Given a comparator sequence $q^{1:T}$, where q^t is allowed to be random, such that $q^t \in \mathcal{A}$ for all $t \in [T]$ and $\mathbb{E}[\text{TV}(q^{1:T})] \leq \beta$, let

$$\eta^\star = \frac{K\alpha^3}{M^3 A_{\max}} \sqrt{\frac{2 \log M + 4\beta \log(M/\alpha)}{T}}. \quad (4.70)$$

The proof proceeds in two steps. First, we show that there exists an expert learning rate $\eta_e \in \mathcal{E}$ such that the regret bound for $\hat{p}_e^{1:T}$ is close to (4.21). That is, we show that there exists $\eta_e \in \mathcal{E}$ such that

$$\mathbb{E} \left[\sum_{t=1}^T \hat{l}_t(\hat{p}_e^t; \hat{p}^t) - \sum_{t=1}^T l_t(q^t) \right] \leq \frac{3M^3 A_{\max}}{K\alpha^3} \sqrt{T \left[\frac{1}{2} \log M + \beta \log(M/\alpha) \right]}. \quad (4.71)$$

Note that $S^t \sim \hat{p}^t$. Second, we show that the output of meta-algorithm can track the best expert with small regret. That is, we show that

$$\mathbb{E} \left[\sum_{t=1}^T l_t(\hat{p}^t) \right] - \mathbb{E} \left[\sum_{t=1}^T \hat{l}_t(\hat{p}_e^t; \hat{p}^t) \right] \leq \frac{M}{\alpha} \sqrt{\frac{TA_{\max}}{8K}} (1 + 2 \log E), \quad e \in [E]. \quad (4.72)$$

The theorem then follows by combining (4.71) and (4.72).

We first prove (4.71). Since $0 \leq \beta \leq 2(T-1)$, we have

$$\min \mathcal{E} = \frac{K\alpha^3}{M^3 A_{\max}} \sqrt{\frac{2 \log M}{T}} \leq \eta^* \leq \frac{K\alpha^3}{M^3 A_{\max}} \sqrt{\frac{2 \log M + 8 \log(M/\alpha)(T-1)}{T}} \leq \max \mathcal{E},$$

where η^* is defined as in (4.70). Thus, there exists $\eta_e \in \mathcal{E}$, such that $\eta_e \leq \eta^* \leq 2\eta_e$.

Repeating the proof of (4.65), we can show that

$$\sum_{t=1}^T \hat{l}_t(\hat{p}_e^t; \hat{p}^t) - \sum_{t=1}^T \hat{l}_t(q^t; \hat{p}^t) \leq \frac{\log M}{\eta_e} + \frac{2 \log(M/\alpha)}{\eta_e} \text{TV}(q^{1:T}) + \frac{\eta_e M^6}{2K^2 \alpha^6} \sum_{t=1}^T \left(a_{\max}^t \right)^2,$$

which then implies that

$$\begin{aligned} \mathbb{E} \left[\sum_{t=1}^T \hat{l}_t(\hat{p}_e^t; \hat{p}^t) - \sum_{t=1}^T \hat{l}_t(q^t; \hat{p}^t) \right] &\leq \frac{\log M}{\eta_e} + \frac{2 \log(M/\alpha)}{\eta_e} \mathbb{E} [\text{TV}(q^{1:T})] \\ &\quad + \frac{\eta_e M^6}{2K^2 \alpha^6} \sum_{t=1}^T \mathbb{E} \left[\left(a_{\max}^t \right)^2 \right] \\ &\leq \frac{\log M}{\eta_e} + \frac{2\beta \log(M/\alpha)}{\eta_e} + \frac{\eta_e M^6 T A_{\max}^2}{2K^2 \alpha^6}. \end{aligned}$$

Since $\eta^*/2 \leq \eta_e \leq \eta^*$, we further have

$$\begin{aligned}
& \mathbb{E} \left[\sum_{t=1}^T \hat{l}_t(\hat{p}_e^t; \hat{p}^t) - \sum_{t=1}^T \hat{l}_t(q^t; \hat{p}^t) \right] \\
& \leq \frac{2 \log M}{\eta^*} + \frac{4\beta \log(M/\alpha)}{\eta^*} + \frac{\eta^* M^6 T A_{\max}^2}{2K^2 \alpha^6} \\
& = \frac{3M^3 A_{\max}}{K \alpha^3} \sqrt{T \left[\frac{1}{2} \log M + \beta \log(M/\alpha) \right]}.
\end{aligned}$$

Now, (4.71) follows, since

$$\mathbb{E} \left[\sum_{t=1}^T \hat{l}_t(q^t; \hat{p}^t) \right] = \sum_{t=1}^T \mathbb{E} \left[\mathbb{E}_{S^t} \left[\hat{l}_t(q^t; \hat{p}^t) \right] \right] = \sum_{t=1}^T \mathbb{E} \left[l_t(q^t) \right].$$

We prove (4.72) next. Let

$$\hat{L}_t^e = \sum_{s=1}^t \hat{l}_s(\hat{p}_e^s; \hat{p}^s) \quad e \in [E], t \in [T].$$

Recall the update for θ_e^t in Step 11 of Alg 9. We have

$$\theta_e^t = \frac{\theta_e^1 \exp(-\gamma \hat{L}_{t-1}^e)}{\sum_{b=1}^E \theta_b^1 \exp(-\gamma \hat{L}_{t-1}^b)}, \quad t = 2, \dots, T.$$

Let $\Theta_t = \sum_{b=1}^E \theta_b^1 \exp\{-\gamma \hat{L}_t^b\}$. Then

$$\log \Theta_1 = \log \left(\sum_{b=1}^E \theta_b^1 \exp\{-\gamma \hat{L}_1^b\} \right)$$

and, for $t \geq 2$,

$$\begin{aligned} \log \left(\frac{\Theta_t}{\Theta_{t-1}} \right) &= \log \left(\frac{\sum_{b=1}^E \theta_b^1 \exp \left\{ -\gamma \hat{L}_{t-1}^b \right\} \exp \left\{ -\gamma \hat{l}_t(\hat{p}_b^t; \hat{p}^t) \right\}}{\sum_{b=1}^E \theta_b^1 \exp \left\{ -\gamma \hat{L}_{t-1}^b \right\}} \right) \\ &= \log \left(\sum_{b=1}^E \theta_b^t \exp \left\{ -\gamma \hat{l}_t(\hat{p}_b^t; \hat{p}^t) \right\} \right). \end{aligned}$$

We have

$$\begin{aligned} \log \Theta_T &= \log \Theta_1 + \sum_{t=1}^T \log \left(\frac{\Theta_t}{\Theta_{t-1}} \right) \\ &= \sum_{t=1}^T \log \left(\sum_{b=1}^E \theta_b^t \exp \left\{ -\gamma \hat{l}_t(\hat{p}_b^t; \hat{p}^t) \right\} \right) \\ &\leq \sum_{t=1}^T \left(-\gamma \sum_{b=1}^E \theta_b^t \hat{l}_t(\hat{p}_b^t; \hat{p}^t) + \frac{\gamma^2 M^2 a_{\max}^t}{8K\alpha^2} \right) && \text{(Lemma 50)} \\ &\leq -\gamma \sum_{t=1}^T \hat{l}_t(\hat{p}^t; \hat{p}^t) + \frac{\gamma^2 M^2 \left(\sum_{t=1}^T a_{\max}^t \right)}{8K\alpha^2} && \text{(Jensen's inequality)} \end{aligned}$$

and

$$\begin{aligned} \log (\Theta_T) &= \log \left(\sum_{b=1}^E \theta_b^1 \exp \left\{ -\gamma \hat{L}_T^b \right\} \right) \\ &\geq \log \left(\max_{1 \leq b \leq E} \theta_b^1 \exp \left\{ -\gamma \hat{L}_T^b \right\} \right) = -\gamma \min_{1 \leq b \leq E} \left\{ \hat{L}_T^b + \frac{1}{\gamma} \log \frac{1}{\theta_b^1} \right\}. \end{aligned}$$

Combining the last two displays, we have

$$-\gamma \min_{1 \leq b \leq E} \left\{ \hat{L}_T^b + \frac{1}{\gamma} \log \frac{1}{\theta_b^1} \right\} \leq -\gamma \sum_{t=1}^T \hat{l}_t(\hat{p}^t; \hat{p}^t) + \frac{\gamma^2 M^2 \left(\sum_{t=1}^T a_{\max}^t \right)}{8K\alpha^2},$$

which implies that

$$\sum_{t=1}^T \hat{l}_t(\hat{p}^t; \hat{p}^t) - \hat{L}_T^e \leq \frac{\gamma M^2 \left(\sum_{t=1}^T a_{\max}^t \right)}{8K\alpha^2} + \frac{1}{\gamma} \log \frac{1}{\theta_e^1} \leq \frac{\gamma M^2 T A_{\max}}{8K\alpha^2} + \frac{1}{\gamma} \log \frac{1}{\theta_e^1}, \quad e \in [E].$$

Taking expectation on both sides, we then have

$$\mathbb{E} \left[\sum_{t=1}^T \hat{l}_t(\hat{p}^t; \hat{p}^t) - \hat{L}_T^e \right] \leq \frac{\gamma M^2 T A_{\max}}{8K\alpha^2} + \frac{1}{\gamma} \log \frac{1}{\theta_e^1}.$$

Since $\theta_e^1 \geq \frac{1}{E^2}$, $\log 1/\theta_e^1 \leq 2 \log E$. Let $\gamma = \sqrt{8K\alpha^2/(TM^2 A_{\max})}$ to minimize the right hand side of the above inequality with $\log 1/\theta_e^1$ substituted by 1. Then

$$\begin{aligned} \mathbb{E} \left[\sum_{t=1}^T \hat{l}_t(\hat{p}^t; \hat{p}^t) - \hat{L}_T^e \right] &= \mathbb{E} \left[\sum_{t=1}^T \hat{l}_t(\hat{p}^t; \hat{p}^t) - \sum_{t=1}^T \hat{l}_t(\hat{p}_e^t; \hat{p}^t) \right] \\ &\leq \frac{M}{\alpha} \sqrt{\frac{TA_{\max}}{8K}} (1 + 2 \log E), \quad e \in [E]. \end{aligned}$$

4.10.7 Proof of Theorem 43

For Mini-batch SGD, following (4.48) and Theorem 42, we have

$$\begin{aligned} &\mathbb{E} \left[\left\| \nabla F(w^R) \right\|_2^2 \right] \\ &\leq \frac{2D^F}{T\mu} + L\mu \left(\frac{2\sigma^2}{KB\alpha} + \frac{2\zeta_T^2(\alpha, \beta)}{K} + \frac{MG^2}{K\alpha^3} \sqrt{\frac{\frac{1}{2} \log M + \beta \log(M/\alpha)}{T}} \right. \\ &\quad \left. + \frac{G}{\alpha} \sqrt{\frac{1}{8KT}} (1 + 2 \log E) \right). \end{aligned}$$

Let

$$\frac{2D^F}{T\mu} = \frac{\mu LG}{\alpha} \sqrt{\frac{1}{8KT}} (1 + 2 \log E),$$

we have

$$\mu = \sqrt{\frac{2\alpha D^F}{LG}} \left(\frac{8K}{T}\right)^{\frac{1}{2}} \sqrt{\frac{1}{1+2\log E}},$$

and

$$\frac{2D^F}{T\mu} + \frac{\mu LG}{\alpha} \sqrt{\frac{1}{8KT}} (1+2\log E) \leq \sqrt{\frac{2D^F LG}{\alpha}} \left(\frac{1}{8K}\right)^{\frac{1}{4}} \left(\frac{1}{T}\right)^{\frac{3}{4}} \sqrt{1+2\log E}.$$

Then follow the same argument as in the proof of Theorem 38, when

$$\mu = \min \left\{ \frac{1}{L}, \frac{1}{\sigma} \sqrt{\frac{D^F KB\alpha}{LT}}, \frac{1}{\zeta_T(\alpha, \beta)} \sqrt{\frac{D^F K}{LT}}, \right. \\ \left. \frac{\sqrt{D^F K} \alpha^{\frac{3}{2}}}{\sqrt{LMT}^{\frac{1}{4}} G \left(\frac{1}{2} \log M + \beta \log(M/\alpha)\right)^{\frac{1}{4}}}, \sqrt{\frac{\alpha D^F}{LG}} \left(\frac{K}{T}\right)^{\frac{1}{2}} \sqrt{\frac{1}{1+2\log E}} \right\},$$

we have

$$\mathbb{E} \left[\left\| \nabla F(w^R) \right\|^2 \right] \\ \lesssim \frac{D^F L}{T} + \frac{\sigma \sqrt{D^F L}}{\sqrt{TKB\alpha}} + \frac{\zeta_T(\alpha, \beta) \sqrt{D^F L}}{\sqrt{TK}} + \frac{\sqrt{D^F L} M^{\frac{1}{2}} G}{T^{\frac{3}{4}} K^{\frac{1}{2}} \alpha^{\frac{3}{2}}} \left(\frac{1}{2} \log M + \beta \log(M/\alpha)\right)^{\frac{1}{4}} \\ + \sqrt{\frac{D^F LG}{\alpha}} \left(\frac{1}{K}\right)^{\frac{1}{4}} \left(\frac{1}{T}\right)^{\frac{3}{4}} \sqrt{1+2\log E}.$$

For FedAvg, following (4.59) and Theorem 42, we have

$$\mathbb{E} \left[\left\| \nabla F(w^R) \right\|_2^2 \right] \leq \frac{8D^F}{\tilde{\mu}BT} + 8 \left(2 + \frac{1}{\alpha} \right) \tilde{\mu}^2 B^2 L^2 \left(\zeta_{\text{unif}}^2 + \frac{\sigma^2}{2B} \right) \\ + 4\tilde{\mu}BL \left(\frac{2\zeta_T^2(\alpha, \beta)}{K} + \frac{\sigma^2}{KB\alpha} + \frac{3MG^2}{K\alpha^3} \sqrt{\frac{\frac{1}{2} \log M + \beta \log(M/\alpha)}{T}} \right. \\ \left. + \frac{G}{\alpha} \sqrt{\frac{1}{8KBT}} (1+2\log E) \right).$$

Let

$$\frac{8D^F}{\tilde{\mu}BT} = 4\tilde{\mu}BL \times \frac{G}{\alpha} \sqrt{\frac{1}{8KBT}} (1 + 2 \log E),$$

we have

$$\tilde{\mu} = \sqrt{\frac{2\alpha D^F}{LG(1+2\log E)}} \left(\frac{1}{B}\right)^{\frac{3}{4}} \left(\frac{8K}{T}\right)^{\frac{1}{4}},$$

and

$$\frac{8D^F}{\tilde{\mu}BT} + 4\tilde{\mu}BL \times \frac{G}{\alpha} \sqrt{\frac{1}{8KBT}} (1 + 2 \log E) = 8^{\frac{3}{4}} \sqrt{2} \sqrt{\frac{D^F LG}{\alpha}} \left(\frac{1}{KB}\right)^{\frac{1}{4}} \left(\frac{1}{T}\right)^{\frac{3}{4}} \sqrt{1 + 2 \log E}.$$

Then follow the same argument as in the proof of Theorem 39, when $\mu \geq 1$ and

$$\tilde{\mu} = \min \left\{ \frac{1}{4BL} \sqrt{\frac{1}{1+1/(2\alpha)}}, \frac{(D^F)^{\frac{1}{3}}}{\left(2 + \frac{1}{\alpha}\right)^{\frac{1}{3}} BL^{\frac{2}{3}} \left(\zeta_{\text{unif}}^2 + \frac{\sigma^2}{2B}\right)^{\frac{1}{3}} T^{\frac{1}{3}}}, \frac{\sqrt{2D^F}}{B\sqrt{L} \sqrt{\frac{2\zeta_T^2(\alpha, \beta)}{K} + \frac{\sigma^2}{KB\alpha} + \sqrt{\frac{\frac{1}{2} \log M + \beta \log(M/\alpha)}{T}} \sqrt{T}}}, \sqrt{\frac{2\alpha D^F}{LG(1+2\log E)}} \left(\frac{1}{B}\right)^{\frac{3}{4}} \left(\frac{8K}{T}\right)^{\frac{1}{4}} \right\},$$

we have

$$\begin{aligned} \mathbb{E} \left[\left\| \nabla F(w^R) \right\|_2^2 \right] &\lesssim \frac{D^F L \sqrt{1 + \frac{1}{2\alpha}}}{T} + \frac{\left(2 + \frac{1}{\alpha}\right)^{\frac{1}{3}} (D^F L)^{\frac{2}{3}} \zeta_{\text{unif}}^{\frac{2}{3}}}{T^{\frac{2}{3}}} + \frac{\left(2 + \frac{1}{\alpha}\right)^{\frac{1}{3}} (D^F L)^{\frac{2}{3}} \sigma^{\frac{2}{3}}}{B^{\frac{1}{3}} T^{\frac{2}{3}}} \\ &+ \frac{\sqrt{D^F L} \zeta_T(\alpha, \beta)}{\sqrt{TK}} + \frac{\sqrt{D^F L} \sigma}{\sqrt{TKB\alpha}} + \frac{\sqrt{D^F L}}{\sqrt{T}} \left(\frac{\frac{1}{2} \log M + \beta \log(M/\alpha)}{T} \right)^{\frac{1}{4}} \\ &+ \sqrt{\frac{D^F LG}{\alpha}} \left(\frac{1}{KB}\right)^{\frac{1}{4}} \left(\frac{1}{T}\right)^{\frac{3}{4}} \sqrt{1 + 2 \log E}. \end{aligned}$$

4.10.8 Proof of Theorem 44

Recall that $T_b = 2^{b-1}$. Let $B = \lceil \log_2(T+1) \rceil$, we then have $T_B \leq T \leq T_{B+1} - 1$, which implies that $1 \leq T - T_B + 1 \leq T_{B+1} - T_B = 2^B$.

Note that $\hat{p}^{T_b}$ is reinitialized as the uniform distribution. Let

$$\text{D-Regret}_b = \mathbb{E} \left[\sum_{t=T_b}^{T_{b+1}-1} l_t(\hat{p}^t) - \sum_{t=T_b}^{T_{b+1}-1} l_t(q^t) \right].$$

Similar to the proof of (4.71) and (4.72), we have

$$\begin{aligned} & \text{D-Regret}_b \\ & \leq \frac{3M^3 A_{\max}}{K\alpha^3} \sqrt{\left[\frac{1}{2} \log M + \log(M/\alpha) \mathbb{E} \left[\text{TV} \left(q^{T_b:(T_{b+1}-1)} \right) \right] \right]} (T_{b+1} - T_b) \\ & \quad + \frac{M}{\alpha} \sqrt{\frac{(T_{b+1} - T_b) A_{\max}}{8K}} (1 + 2 \log E_b) \\ & = \frac{3M^3 A_{\max}}{K\alpha^3} \sqrt{\left[\frac{1}{2} \log M + \log(M/\alpha) \mathbb{E} \left[\text{TV} \left(q^{T_b:(T_{b+1}-1)} \right) \right] \right]} (\sqrt{2})^{b-1} \\ & \quad + \frac{M}{\alpha} \sqrt{\frac{A_{\max}}{8K}} (1 + 2 \log E_b) (\sqrt{2})^{b-1} \\ & \leq \frac{3M^3 A_{\max}}{K\alpha^3} \sqrt{\left[\frac{1}{2} \log M + \log(M/\alpha) \beta \right]} (\sqrt{2})^{b-1} + \frac{M}{\alpha} \sqrt{\frac{A_{\max}}{8K}} (1 + 2 \log E) (\sqrt{2})^{b-1}, \end{aligned}$$

where E_b is defined in (4.27) and E is defined in (4.23). We can similarly obtain

$$\begin{aligned} \mathbb{E} \left[\sum_{t=T_B}^T l_t(\hat{p}^t) - \sum_{t=T_B}^T l_t(q^t) \right] & \leq \frac{3M^3 A_{\max}}{K\alpha^3} \sqrt{\left[\frac{1}{2} \log M + \log(M/\alpha) \beta \right]} (\sqrt{2})^B \\ & \quad + \frac{M}{\alpha} \sqrt{\frac{A_{\max}}{8K}} (1 + 2 \log E) (\sqrt{2})^B. \end{aligned}$$

Combine the last two displays, we have

$$\begin{aligned}
\text{D-Regret}_T(q^{1:T}) &= \mathbb{E} \left[\sum_{t=1}^T l_t(\hat{p}^t) - \sum_{t=1}^T l_t(q^t) \right] \\
&= \sum_{b=1}^{B-1} \mathbb{E} \left[\sum_{t=T_b}^{T_{b+1}-1} l_t(\hat{p}^t) - \sum_{t=T_b}^{T_{b+1}-1} l_t(q^t) \right] + \mathbb{E} \left[\sum_{t=T_B}^T l_t(\hat{p}^t) - \sum_{t=T_B}^T l_t(q^t) \right] \\
&\leq \frac{3M^3 A_{\max}}{K\alpha^3} \sqrt{\left[\frac{1}{2} \log M + \log(M/\alpha) \beta \right]} \left(\sum_{b=1}^{B-1} (\sqrt{2})^{b-1} + (\sqrt{2})^B \right) \\
&\quad + \frac{M}{\alpha} \sqrt{\frac{A_{\max}}{8K}} (1 + 2 \log E) \left(\sum_{b=1}^{B-1} (\sqrt{2})^{b-1} + (\sqrt{2})^B \right) \\
&= \frac{3M^3 A_{\max}}{K\alpha^3} \sqrt{\left[\frac{1}{2} \log M + \log(M/\alpha) \beta \right]} \frac{(\sqrt{2})^{B+1} - \sqrt{2}}{\sqrt{2} - 1} \\
&\quad + \frac{M}{\alpha} \sqrt{\frac{A_{\max}}{8K}} (1 + 2 \log E) \frac{(\sqrt{2})^{B+1} - \sqrt{2}}{\sqrt{2} - 1} \\
&\leq \frac{3M^3 A_{\max}}{K\alpha^3} \sqrt{\left[\frac{1}{2} \log M + \log(M/\alpha) \beta \right]} \frac{2}{\sqrt{2} - 1} \sqrt{T} \\
&\quad + \frac{M}{\alpha} \sqrt{\frac{A_{\max}}{8K}} (1 + 2 \log E) \frac{2}{\sqrt{2} - 1} \sqrt{T}.
\end{aligned}$$

CHAPTER 5

CONCLUSION

This dissertation advances a unified perspective on statistical learning and optimization under distribution shift, centered on the principle of *adaptation*. At the heart of this framework is the distinction between *preserved components*—aspects of the problem that remain stable across domains—and *adapted components*—those that vary and must be re-estimated. By systematically disentangling what can be transferred from what must be modified, each chapter develops domain-specific methodologies that leverage this structure to improve performance under distributional heterogeneity.

Chapter 2 addresses high-dimensional precision matrix estimation in heterogeneous biological settings. The proposed method, **Trans-Glasso**, exploits a shared sparsity pattern across source and target domains as the preserved component, while adapting to domain-specific deviations through a two-stage procedure. This approach enhances estimation accuracy in data-scarce target domains and illustrates how structural similarity can be systematically incorporated via multi-task learning and refinement.

Chapter 3 focuses on multi-task learning under distribution shift. The **SMART** framework assumes that the spectral subspaces of the regression coefficient matrix are preserved across domains, while the projection weights are task-specific and adapted to the target. This separation guides a nonconvex, regularized optimization procedure that enables reliable recovery of the regression matrix, even in low-sample regimes. The method demonstrates how spectral priors can facilitate targeted adaptation in complex multi-task learning scenarios.

Chapter 4 turns to federated optimization with heterogeneous and evolving client contributions. In this setting, the preserved component is the underlying informativeness of clients, which changes slowly over time, while the adapted component is the client sampling strategy, which must dynamically adjust to maximize training efficiency. By framing this as an online learning problem with bandit feedback, the proposed OSMD-based algorithm

achieves strong theoretical guarantees and empirical performance, exemplifying the role of adaptation in decentralized optimization.

Collectively, these contributions demonstrate how the principle of adaptation—leveraging domain-invariant structure while accommodating domain-specific variation—can guide the design of effective algorithms across a broad spectrum of statistical learning and optimization problems. This perspective enhances robustness and generalization, offering a principled foundation for data-efficient adaptation under distributional shift.

Future Directions

Building on the ideas developed in this dissertation, several promising directions for future research emerge:

- **Inference under adaptation.** While the current work focuses on estimation and optimization, extending adaptive methods to support statistical inference—such as constructing confidence intervals or conducting hypothesis tests under distribution shift—remains an important and largely open challenge.
- **Adaptation in dynamic environments.** The present frameworks primarily address a single target and multiple static sources. Future research may consider more dynamic scenarios, such as continual learning and temporal adaptation. In particular, handling temporal distribution shift—where data distributions evolve gradually or change abruptly over time—poses methodological and theoretical challenges that warrant deeper investigation.
- **Adaptation for large-scale neural models.** With the increasing prominence of foundation models, there is growing interest in incorporating structured adaptation principles into large-scale neural architectures. This includes developing methods for efficient fine-tuning, domain-specific calibration, and robust transfer, while leveraging the expressive representations encoded by pre-trained networks.

- **Adaptation in non-Euclidean domains.** Many real-world datasets reside in non-Euclidean spaces, such as graphs and manifolds. Extending the adaptation framework to these settings—where preserved structures may be geometric or topological—opens the door to novel methodological advances and broader applications.
- **Adaptation in unstructured domains.** A key challenge lies in adapting models to unstructured data modalities such as text, images, and audio, where the underlying structure is implicit. Effectively transferring knowledge in these settings requires new strategies for representation learning, alignment, and regularization. This direction is especially timely given the rise of powerful foundation models—such as large language models and vision transformers—which offer rich pre-trained embeddings but also raise new questions about how to adapt them reliably under shift.

Taken together, these directions underscore the broader potential of adaptation as a unifying framework—not only for enhancing robustness and efficiency in statistical learning and optimization, but also for enabling principled generalization across increasingly diverse, complex, and heterogeneous data environments.

BIBLIOGRAPHY

- Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *ArXiv Preprint ArXiv:1603.04467*, 2016.
- Theodore Wilbur Anderson. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics*, pages 327–351, 1951.
- A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2:183–202, 2009.
- Zalan Borsos, Andreas Krause, and Kfir Y. Levy. Online variance reduction for stochastic optimization. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Conference On Learning Theory (COLT)*, 2018.
- Zalán Borsos, Sebastian Curi, Kfir Yehuda Levy, and Andreas Krause. Online variance reduction with mixtures. In *International Conference on Machine Learning (ICML)*, 2019.
- Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.
- Stephen Boyd, Stephen P Boyd, and Lieven Vandenbergh. *Convex Optimization*. Cambridge university press, 2004.
- Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.
- Florentina Bunea, Yiyuan She, and Marten H Wegkamp. Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics*, pages 1282–1309, 2011.
- Florentina Bunea, Yiyuan She, and Marten H. Wegkamp. Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. *The Annals of Statistics*, 40(5): 2359–2388, 2012.
- Changxiao Cai, T Tony Cai, and Hongzhe Li. Transfer learning for contextual multi-armed bandits. *The Annals of Statistics*, 52(1):207–232, 2024a.
- T. Tony Cai and Hongji Wei. Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *The Annals of Statistics*, 49(1):100 – 128, 2021.
- T. Tony Cai, Weidong Liu, and Xi Luo. A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494): 594–607, 2011.

- T. Tony Cai, Weidong Liu, and Harrison H. Zhou. Estimating sparse precision matrix: optimal rates of convergence and adaptive estimation. *The Annals of Statistics*, 44(2): 455–488, 2016.
- T Tony Cai, Dongwoo Kim, and Hongming Pu. Transfer learning for functional mean estimation: Phase transition and adaptive algorithms. *The Annals of Statistics*, 52(2): 654–678, 2024b.
- Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečný, H Brendan McMahan, Virginia Smith, and Ameet Talwalkar. Leaf: A benchmark for federated settings. *ArXiv Preprint ArXiv:1812.01097*, 2018.
- Rich Caruana. Multitask learning. *Machine learning*, 28:41–75, 1997.
- Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.
- Zachary Charles and Jakub Konečný. On the outsized importance of learning rates in local update methods. *ArXiv preprint ArXiv:2007.00878*, 2020.
- Kun Chen and Wenjie Wang. *rrpack: Reduced-Rank Regression*, 2022. R package version 0.1-13.
- Kun Chen, Hongbo Dong, and Kung-Sik Chan. Reduced rank regression via adaptive nuclear norm penalization. *Biometrika*, 100(4):901–920, 2013.
- Lisha Chen and Jianhua Z Huang. Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association*, 107(500): 1533–1545, 2012.
- Mouxian Chen, Lefei Shen, Han Fu, Zhuo Li, Jianling Sun, and Chenghao Liu. Calibration of time-series forecasting: Detecting and adapting context-driven distribution shift. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024.
- Wenlin Chen, Samuel Horváth, and Peter Richtárik. Optimal client sampling for federated learning. *Transactions on Machine Learning Research*, 2022.
- Yuxin Chen, Jianqing Fan, Cong Ma, and Kaizheng Wang. Spectral method and regularized mle are both optimal for top-k ranking. *Annals of statistics*, 47(4):2204, 2019.
- Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, et al. Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14(5):566–806, 2021.
- Yae Jee Cho, Samarth Gupta, Gauri Joshi, and Osman Yagan. Bandit-based communication-efficient client selection strategies for federated learning. In *Asilomar Conference on Signals, Systems, and Computers (ACSCC)*, 2020a.

- Yae Jee Cho, Jianyu Wang, and Gauri Joshi. Client selection in federated learning: Convergence analysis and power-of-choice selection strategies. *ArXiv Preprint ArXiv:2010.01243*, 2020b.
- Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. *ArXiv Preprint ArXiv:1812.01718*, 2018.
- Dominik Csiba and Peter Richtárik. Importance sampling for minibatches. *Journal of Machine Learning Research*, 19(27):1–21, 2018.
- Patrick Danaher, Pei Wang, and Daniela M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 76(2):373–397, 2014.
- Amit Daniely, Alon Gonen, and Shai Shalev-Shwartz. Strongly adaptive online learning. In *International Conference on Machine Learning (ICML)*, 2015.
- Chang Deng, Kevin Bello, Bryon Aragam, and Pradeep Kumar Ravikumar. Optimizing notears objectives via topological swaps. In *International Conference on Machine Learning*, 2023a.
- Chang Deng, Kevin Bello, Pradeep Ravikumar, and Bryon Aragam. Global optimality in bivariate gradient-based dag learning. *Advances in Neural Information Processing Systems*, 2023b.
- Chang Deng, Kevin Bello, Pradeep Ravikumar, and Bryon Aragam. Markov equivalence and consistency in differentiable structure learning. *Advances in Neural Information Processing Systems*, 2024.
- Cynthia Dwork. Differential privacy: A survey of results. In *Theory and Applications of Models of Computation (TAMC)*, 2008.
- Vladmir Estivill-Castro and Derick Wood. A survey of adaptive sorting algorithms. *ACM Computing Surveys (CSUR)*, 24(4):441–476, 1992.
- Jianqing Fan, Kunpeng Li, and Yuan Liao. Recent developments in factor models and applications in econometric learning. *Annual Review of Financial Economics*, 13(1):401–430, 2021.
- Farideh Fazayeli and Arindam Banerjee. Generalized direct change estimation in ising model structure. In *International Conference on Machine Learning*, 2016.
- Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

- Jiawei Ge, Shange Tang, Jianqing Fan, Cong Ma, and Chi Jin. Maximum likelihood estimation is all you need for well-specified covariate shift. In *International Conference on Learning Representations*, 2024.
- Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.
- Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.
- Siddharth Gopal. Adaptive sampling for SGD by exploiting side information. In *International Conference on Machine Learning (ICML)*, 2016.
- Tian Gu, Yi Han, and Rui Duan. Robust angle-based transfer learning in high dimensions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkae111, 2024.
- Jian Guo, Elizaveta Levina, George Michailidis, and Ji Zhu. Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15, 2011.
- Eric C Hall and Rebecca M Willett. Online convex optimization in dynamic environments. *IEEE Journal of Selected Topics in Signal Processing*, 9(4):647–662, 2015.
- Ayoub El Hanchi and David A. Stephens. Adaptive importance sampling for finite-sum optimization and sampling with decreasing step-sizes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- Filip Hanzely, Boxin Zhao, and Mladen Kolar. Personalized federated learning: A unified framework and universal optimization techniques. *Transactions on Machine Learning Research*, 2023.
- Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.
- Yong He, Zeyu Li, Dong Liu, Kangxiang Qin, and Jiahui Xie. Representational transfer learning for matrix completion. *arXiv preprint arXiv:2412.06233*, 2024a.
- Zelin He, Ying Sun, and Runze Li. Transfusion: Covariate-shift robust transfer learning for high-dimensional regression. In *International Conference on Artificial Intelligence and Statistics*, pages 703–711, 2024b.
- Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis*, 5(2):248–264, 1975.
- Ali Jalali, Sujay Sanghavi, Chao Ruan, and Pradeep Ravikumar. A dirty model for multi-task learning. *Advances in neural information processing systems*, 23, 2010.

- Tyler B. Johnson and Carlos Guestrin. Training deep models faster with robust, approximate importance sampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.
- Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021.
- Minoru Kanehisa, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. Kegg for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, 40(D1):D109–D114, 2012.
- Sai Praneeth Karimireddy, Martin Jaggi, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Mime: Mimicking centralized stochastic algorithms in federated learning. *ArXiv Preprint ArXiv:2008.03606*, 2020a.
- Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. SCAFFOLD: stochastic controlled averaging for federated learning. In *International Conference on Machine Learning (ICML)*, 2020b.
- Taehyeon Kim, Sangmin Bae, Jin-woo Lee, and Seyoung Yun. Accurate and fast federated learning via combinatorial multi-armed bandits. *ArXiv Preprint ArXiv:2012.03270*, 2020.
- Steven M Kornblau, Raoul Tibes, Yi Hua Qiu, Wenjing Chen, Hagop M Kantarjian, Michael Andreeff, Kevin R Coombes, and Gordon B Mills. Functional proteomic profiling of aml predicts response and survival. *Blood, The Journal of the American Society of Hematology*, 113(1):154–164, 2009.
- Zhaobin Kuang, Sinong Geng, and David Page. A screening rule for l1-regularized ising model estimation. In *Advances in Neural Information Processing Systems*, 2017.
- Viraj Kulkarni, Milind Kulkarni, and Aniruddha Pant. Survey of personalization techniques for federated learning. In *World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*, 2020.
- Clifford Lam and Jianqing Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of statistics*, 37(6B):4254, 2009.
- Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.
- Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.
- Yann LeCun and Corinna Cortes. MNIST handwritten digit database. *CoRR*, 2010.
- Sai Li, T Tony Cai, and Hongzhe Li. Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):149–173, 2022.

- Sai Li, T Tony Cai, and Hongzhe Li. Transfer learning in large-scale gaussian graphical models with false discovery rate control. *Journal of the American Statistical Association*, 118(543):2171–2183, 2023a.
- Sai Li, Linjun Zhang, T Tony Cai, and Hongzhe Li. Estimation and inference for high-dimensional generalized linear models with knowledge transfer. *Journal of the American Statistical Association*, 0(0):1–12, 2023b.
- Haotian Lin and Matthew Reimherr. Transfer learning for functional linear regression with structural interpretability. *arXiv preprint arXiv:2206.04277*, 2022.
- Haotian Lin and Matthew Reimherr. On hypothesis transfer learning of functional linear models. In *International Conference on Machine Learning*, 2024.
- Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10(80):2295–2328, 2009.
- Han Liu, Fang Han, Ming Yuan, John Lafferty, and Larry Wasserman. High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293 – 2326, 2012a.
- Han Liu, Fang Han, and Cun-hui Zhang. Transelliptical graphical models. *Advances in neural information processing systems*, 25, 2012b.
- Song Liu, John A Quinn, Michael U Gutmann, Taiji Suzuki, and Masashi Sugiyama. Direct learning of sparse changes in markov networks by density ratio estimation. *Neural computation*, 26(6):1169–1197, 2014.
- Malte Luecken, Daniel Burkhardt, Robrecht Cannoodt, Christopher Lance, Aditi Agrawal, Hananeh Aliee, Ann Chen, Louise Deconinck, Angela Detweiler, Alejandro Granados, Shelly Huynh, Laura Isacco, Yang Kim, Dominik Klein, BONY DE KUMAR, Sunil Kuppasani, Heiko Lickert, Aaron McGeever, Joaquin Melgarejo, Honey Mekonen, Maurizio Morri, Michaela Müller, Norma Neff, Sheryl Paul, Bastian Rieck, Kaylie Schneider, Scott Steelman, Michael Sterr, Daniel Treacy, Alexander Tong, Alexandra-Chloe Villani, Guilin Wang, Jia Yan, Ce Zhang, Angela Pisco, Smita Krishnaswamy, Fabian Theis, and Jonathan M Bloom. A sandbox for prediction and integration of dna, rna, and proteins in single cells. In *Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.
- Cong Ma, Junwei Lu, and Han Liu. Inter-subject analysis: A partial gaussian graphical model approach. *Journal of the American Statistical Association*, 116(534):746–755, 2021.
- Cong Ma, Reese Pathak, and Martin J Wainwright. Optimally tackling covariate shift in rkhs-based nonparametric regression. *The Annals of Statistics*, 51(2):738–761, 2023.
- Zhuang Ma, Zongming Ma, and Tingni Sun. Adaptive estimation in two-way sparse reduced-rank regression. *Statistica Sinica*, 30(4):2179–2201, 2020.

- Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.
- Sidhant Misra, Marc Vuffray, and Andrey Y. Lokhov. Information theoretic optimal learning of gaussian graphical models. In *Conference on Learning Theory*, 2020.
- Ashin Mukherjee and Ji Zhu. Reduced rank ridge regression and its kernel extensions. *Statistical analysis and data mining: the ASA data science journal*, 4(6):612–622, 2011.
- Ashin Mukherjee, Kun Chen, Naisyin Wang, and Ji Zhu. On the degrees of freedom of reduced-rank estimators in multivariate regression. *Biometrika*, 102(2):457–477, 2015.
- Hongseok Namkoong, Aman Sinha, Steve Yadlowsky, and John C. Duchi. Adaptive sampling probabilities for non-smooth optimization. In *International Conference on Machine Learning (ICML)*, 2017.
- Deanna Needell, Nathan Srebro, and Rachel Ward. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. *Mathematical Programming*, 1(155):549–573, 2016.
- Yurii Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.
- Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.
- Mark EJ Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 74(3):036104, 2006.
- Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14, 2001.
- Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer New York, 2006.
- Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.
- Neal Parikh and Stephen P. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, 2014.
- Kumar Kshitij Patel, Lingxiao Wang, Blake E Woodworth, Brian Bullins, and Nati Srebro. Towards optimal communication complexity in distributed non-convex optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- Reese Pathak, Cong Ma, and Martin Wainwright. A new similarity measure for covariate shift with applications to nonparametric regression. In *International Conference on Machine Learning*, 2022.

- Dmytro Perekrestenko, Volkan Cevher, and Martin Jaggi. Faster coordinate descent via adaptive importance sampling. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.
- Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. Technical report, Technical University of Denmark, 2008.
- Christine Peterson, Francesco C Stingo, and Marina Vannucci. Bayesian inference of multiple gaussian graphical models. *Journal of the American Statistical Association*, 110(509):159–174, 2015.
- Nan Qiao, Haowei Jiang, and Cunjie Lin. Rank-based transfer learning for high-dimensional survival data with application to sepsis data. *arXiv preprint arXiv:2504.11270*, 2025.
- Xinghao Qiao, Shaojun Guo, and Gareth M James. Functional graphical models. *Journal of the American Statistical Association*, 114(525):211–222, 2019.
- Zhe Qu, Rui Duan, Lixing Chen, Jie Xu, Zhuo Lu, and Yao Liu. Context-aware online client selection for hierarchical federated learning. *IEEE Transactions on Parallel and Distributed Systems*, 33(12):4353–4367, 2022.
- Aravind Rajeswaran, Chelsea Finn, Sham M. Kakade, and Sergey Levine. Meta-learning with implicit gradients. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- Pradeep Ravikumar, Martin J. Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935 – 980, 2011.
- Zhao Ren, Tingni Sun, Cun-Hui Zhang, and Harrison H. Zhou. Asymptotic normality and optimality in estimation of large Gaussian graphical models. *The Annals of Statistics*, 43(3):991–1026, 2015.
- Monica Ribero and Haris Vikalo. Communication-efficient federated learning via optimal client sampling. *ArXiv Preprint ArXiv:2007.15197*, 2020.
- Adam J. Rothman, Peter J. Bickel, Elizaveta Levina, and Ji Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494 – 515, 2008.
- Farnood Salehi, L Elisa Celis, and Patrick Thiran. Stochastic optimization with bandit sampling. *ArXiv Preprint ArXiv:1708.02544*, 2017.
- Farnood Salehi, Patrick Thiran, and L. Elisa Celis. Coordinate descent with bandit sampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.
- Sebastian U. Stich, Anant Raj, and Martin Jaggi. Safe adaptive importance sampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

- Xin Tang, Jiawei Zhang, Yichun He, Xinhe Zhang, Zuwan Lin, Sebastian Partarrieu, Emma Bou Hanna, Zhaolin Ren, Hao Shen, Yuhong Yang, et al. Explainable multi-task learning for multi-modality biological data analysis. *Nature communications*, 14(1): 2546, 2023.
- Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023.
- Ye Tian and Yang Feng. Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association*, 118(544):2684–2697, 2023.
- James Townsend, Niklas Koep, and Sebastian Weichwald. Pymanopt: A python toolbox for optimization on manifolds using automatic differentiation. *Journal of Machine Learning Research*, 17(137):1–5, 2016.
- Katherine Tsai, Oluwasanmi Koyejo, and Mladen Kolar. Joint gaussian graphical model estimation: A survey. *Wiley Interdisciplinary Reviews: Computational Statistics*, 14(6): e1582, 2022.
- Katherine Tsai, Boxin Zhao, Sanmi Koyejo, and Mladen Kolar. Latent multimodal functional graphical model estimation. *Journal of the American Statistical Association*, 119(547): 2217–2229, 2024.
- Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer New York, NY, 1st edition, 2008.
- Jitendra K Tugnait. Estimation of high-dimensional differential graphs from multi-attribute data. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.
- Flora Tzifi, Christina Economopoulou, Dimitrios Gourgiotis, Alexandros Ardavanis, Sotirios Papageorgiou, and Andreas Scorilas. The role of bcl2 family of apoptosis regulator proteins in acute and chronic leukemias. *Advances in Hematology*, 2012, 2012.
- Yoshimasa Uematsu, Yingying Fan, Kun Chen, Jinchi Lv, and Wei Lin. Sofar: large-scale association network learning. *IEEE transactions on information theory*, 65(8):4924–4939, 2019.
- Tim van Erven and Wouter M. Koolen. Metagrad: Multiple learning rates in online learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.
- Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.
- Eric Vivier, David Artis, Marco Colonna, Andreas Diefenbach, James P Di Santo, Gérard Eberl, Shigeo Koyasu, Richard M Locksley, Andrew NJ McKenzie, Reina E Mebius, et al. Innate lymphoid cells: 10 years on. *Cell*, 174(5):1054–1066, 2018.

- Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.
- Hao Wang, Zakhary Kaplan, Di Niu, and Baochun Li. Optimizing federated learning on non-iid data with reinforcement learning. In *IEEE Conference on Computer Communications (INFOCOM)*, 2020.
- Kaizheng Wang. Pseudo-labeling for kernel ridge regression under covariate shift. *arXiv preprint arXiv:2302.10160*, 2023.
- Lingxiao Wang, Boxin Zhao, and Mladen Kolar. Differentially private matrix completion through low-rank matrix factorization. In *International Conference on Artificial Intelligence and Statistics*, pages 5731–5748, 2023.
- Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *ArXiv Preprint ArXiv:1708.07747*, 2017.
- Miao Yang, Ximin Wang, Hongbin Zhu, Haifeng Wang, and Hua Qian. Federated learning with class imbalance reduction. In *European Signal Processing Conference (EUSIPCO)*, 2021.
- Tianbao Yang, Lijun Zhang, Rong Jin, and Jinfeng Yi. Tracking slowly moving clairvoyant: Optimal dynamic regret of online learning with true and noisy gradient. In *International Conference on Machine Learning (ICML)*, 2016.
- Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.
- Huili Yuan, Ruibin Xi, Chong Chen, and Minghua Deng. Differential network analysis via lasso penalized d-trace loss. *Biometrika*, 104(4):755–770, 2017.
- Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.
- Ming Yuan, Ali Ekici, Zhaosong Lu, and Renato Monteiro. Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(3):329–346, 2007.
- Lijun Zhang, Shiyin Lu, and Zhi-Hua Zhou. Adaptive online learning in dynamic environments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.
- Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE transactions on knowledge and data engineering*, 34(12):5586–5609, 2021.
- Boxin Zhao, Y. Samuel Wang, and Mladen Kolar. Direct estimation of differential functional graphical models. In *Advances in Neural Information Processing Systems*, 2019.

- Boxin Zhao, Y. Samuel Wang, and Mladen Kolar. Fudge: A method to estimate a functional differential graph in a high-dimensional setting. *Journal of Machine Learning Research*, 23:82:1–82:82, 2022.
- Boxin Zhao, Boxiang Lyu, Raul Castro Fernandez, and Mladen Kolar. Addressing budget allocation and revenue allocation in data market environments using an adaptive sampling algorithm. In *International Conference on Machine Learning*, pages 42081–42097, 2023a.
- Boxin Zhao, Boxiang Lyu, and Mladen Kolar. L-svrg and l-katyusha with adaptive sampling. *Transactions on Machine Learning Research*, 2023b.
- Boxin Zhao, Cong Ma, and Mladen Kolar. Trans-glasso: A transfer learning approach to precision matrix estimation. *arXiv preprint arXiv:2411.15624*, 2024a.
- Boxin Zhao, Weishi Wang, Dingyuan Zhu, Ziqi Liu, Dong Wang, Zhiqiang Zhang, Jun Zhou, and Mladen Kolar. Personalized binomial dags learning with network structured covariates. *arXiv preprint arXiv:2406.06829*, 2024b.
- Boxin Zhao, Percy S Zhai, Y Samuel Wang, and Mladen Kolar. High-dimensional functional graphical model structure learning via neighborhood selection approach. *Electronic Journal of Statistics*, 18(1):1042–1129, 2024c.
- Boxin Zhao, Lingxiao Wang, Ziqi Liu, Zhiqiang Zhang, Jun Zhou, Chaochao Chen, and Mladen Kolar. Adaptive client sampling in federated learning via online learning with bandit feedback. *Journal of Machine Learning Research*, 26(8):1–67, 2025.
- Peilin Zhao and Tong Zhang. Stochastic optimization with importance sampling for regularized loss minimization. In *International Conference on Machine Learning (ICML)*, 2015.
- Sihai Dave Zhao, T Tony Cai, and Hongzhe Li. Direct estimation of differential networks. *Biometrika*, 101(2):253–268, 2014.
- Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31, 2018.
- Yunzhang Zhu, Xiaotong Shen, and Wei Pan. Structural pursuit over multiple undirected graphs. *Journal of the American Statistical Association*, 109(508):1683–1696, 2014.
- Zeyuan Allen Zhu, Zheng Qu, Peter Richtárik, and Yang Yuan. Even faster accelerated coordinate descent using non-uniform sampling. In *International Conference on Machine Learning (ICML)*, 2016.