

**SOME FUNDAMENTAL MACHINE LEARNING PROBLEMS IN THE  
DIFFERENTIAL PRIVACY MODEL**

by

Di Wang

August 2020

A dissertation submitted to the  
Faculty of the Graduate School of  
the University at Buffalo, State University of New York  
in partial fulfilment of the requirements for the  
degree of

Doctor of Philosophy

Department of Computer Science and Engineering

Copyright by

Di Wang

2020

All Rights Reserved

# Acknowledgments

It is time to say goodbye to my PhD studies in the State University of New York at Buffalo. It has been an exciting and memorable experience and I would like to take this opportunity to thank everyone who has helped me during my PhD studies.

First and foremost, I would like to thank my advisor, Dr. Jinhui Xu. I can confidently say that I owe to him the majority of my development as a researcher in the past five years, from problem development to note-taking, to mathematical exploration, to writing papers, to crafting presentations, and to mentoring others. I also thank him for providing me the chance to pursue a PhD degree in Computer Science when I was a Master student in Mathematics with no sufficient background in Computer Science, and for encouraging me when I had no publication in the first two and half years of my PhD studies. Moreover, I sincerely appreciate him providing me with funding for attending many conferences and visiting quite a few other universities.

I would like to express my gratitude and respect to Professor Adam Smith, who hosted me as a visiting student at Boston University while I was participating in the Privacy Tools project in Harvard University as an intern in the summer of 2018. I was quite impressed by Adam's knowledge in Differential Privacy, and the deep thinking of ERM in the Non-interactive Local DP model. I was really lucky to have the chance to work with him and had our joint work published in ALT 2019.

I own my thanks to Professor Marco Gaboardi, who helped me attend the Privacy Tools project in Harvard University and invited me as a visiting graduate student to the

Data Privacy: Foundations and Applications program at Simons Institute for the Theory of Computing, University of California Berkeley. I also thank him for writing recommendation letters for my job search. Finally, I was lucky to work with him and had our joint work published in NeurIPS 2018 and NeurIPS 2019.

I sincerely appreciate my thesis committee members, Professor Shi Li and Professor Changyou Chen. I thank Shi for writing recommendation letters for my job search and helping me overcoming the challenges of our joint work published in NeurIPS 2019. I thank Changyou for writing recommendation letters for me and providing me with some directions on our joint work published in ICML 2019.

I thank the Department of Computer Science and Engineering for providing me with the Teaching Assistantship and Instructor opportunities. My thanks also go to every one in UB who accompanied me during my PhD studies.

My gratitude also goes to all of my collaborators and members of the theory lab in the UB CSE department, including but not limited to Minwei Ye, Mengdi Huai, Adam Smith, Marco Gaboardi, Huanyu Zhang, Yang He, Chenglin Miao, Changyou Chen, Aidong Zhang, Yunus Esencayi, Xiangyu Guo, Shi Li, Chaowen Guan, Hanshen Xiao, Srinivas Devadas, Tianhang Zheng, Baochun Li, Jiahao Ding, Zeyun Xie, Miao Pan, Danyang Chen, Yangwei Liu, Ziyun Huang, Xiangyu Wang, Zihe Chen, Zheshuo Li, Jiayi Xian, Minghua Wang and Yufan Zhou. I also thank Professor Hu Ding for inviting me to visit USTC and Peng Zhao for inviting me to visit Nanjing University.

I would like to thank the National Science Foundation for the generous support during my PhD studies through grants CCF-1422324, CCF-1716400, IIS-1422591, and IIS-1910492.

Last but not least, my special thanks go to all my family, including my parents Jijun Wang and Yuhong Jia, and my grandparents, who brought me to this wonderful world, raised me up, taught me, and respected my decision to attend the Mathematics and Applied Mathematics program in Shandong University. I also need to thank my two dogs who brought me endless happiness. Finally, I owe my special thanks to my fiancé, for her

understanding, caring and support during my PhD studies. I could not complete this PhD if I were the only one pursuing it.

# Table of Contents

<b>Acknowledgments</b> . . . . .	ii
<b>List of Tables</b> . . . . .	xii
<b>List of Figures</b> . . . . .	xiv
<b>1 Introduction</b> . . . . .	1
1.1 Dissertation Contributions . . . . .	4
1.2 Dissertation Outline . . . . .	13
<b>2 Differential Privacy Background</b> . . . . .	15
2.1 Central Differential Privacy . . . . .	15
2.2 Local Differential Privacy . . . . .	19
<b>3 Empirical Risk Minimization with Convex Loss Functions in Differential Privacy Model</b> . . . . .	22
3.1 Faster Algorithms of DP-ERM under the Classical Setting . . . . .	24
3.1.1 Related Work . . . . .	26
3.1.2 Preliminaries . . . . .	27
3.1.3 Low Dimensional Case . . . . .	29
3.1.4 High Dimensional Case . . . . .	34
3.1.5 Experiments . . . . .	36
3.1.6 Omitted Proofs . . . . .	39
3.1.7 Proofs of Differential Privacy . . . . .	43
3.2 DP-ERM with Heavy-tailed Data . . . . .	58
3.2.1 Related Work . . . . .	60
3.2.2 Preliminaries . . . . .	61
3.2.3 Sample-aggregation based method . . . . .	62
3.2.4 Gradient descent based methods . . . . .	65
3.2.5 Experiments . . . . .	73
3.2.6 Omitted Proofs . . . . .	79
3.3 DP-ERM with Pairwise Loss Functions . . . . .	89
3.3.1 Private pairwise learning . . . . .	92
3.3.2 Online Private pairwise learning . . . . .	95
3.3.3 Offline Private Pairwise Learning . . . . .	102

3.3.4	Improved Upper Bounds for Offline Setting . . . . .	103
3.3.5	Experiments . . . . .	105
3.3.6	Omitted Proofs . . . . .	110
<b>4</b>	<b>Empirical Risk Minimization with Non-Convex Loss Functions in Differential Privacy Model . . . . .</b>	<b>120</b>
4.1	First Order Stationary View . . . . .	121
4.1.1	Low Dimension Case . . . . .	123
4.1.2	High Dimension Case . . . . .	128
4.1.3	Error Bounded by Norm of Gradient . . . . .	130
4.1.4	Further Reducing the Utility . . . . .	131
4.1.5	Experimental Results . . . . .	132
4.1.6	Omitted Proofs . . . . .	135
4.2	Global Minimum View . . . . .	142
4.2.1	Preliminaries . . . . .	144
4.2.2	Excess Risk of DP-ERM with Non-convex Loss Functions . . . . .	147
4.2.3	Omitted Proofs . . . . .	155
4.2.4	Proof of Theorem 4.2.6 . . . . .	172
4.3	Local Minimum/Second Order Stationary View . . . . .	175
4.3.1	Finding Approximate Local Minimum Privately Using DP-GD . . . . .	177
4.3.2	Improved Sample Complexity via DP-TR Method . . . . .	182
4.3.3	Experiments . . . . .	187
4.3.4	Omitted Proofs . . . . .	190
<b>5</b>	<b>Empirical Risk Minimization in Local Differential Privacy Model . . . . .</b>	<b>211</b>
5.1	ERM in Non-interactive LDP model . . . . .	212
5.1.1	Related Work . . . . .	216
5.1.2	Preliminaries . . . . .	218
5.1.3	LDP-ERM with Smooth Loss Functions . . . . .	222
5.1.4	LDP-ERM with Convex Generalized Linear Loss Functions . . . . .	230
5.1.5	LDP Algorithms for Learning k-way Marginals Queries and Smooth Queries . . . . .	236
5.1.6	Omitted Proofs . . . . .	241
5.1.7	Omitted Details in Section 5.1.3 . . . . .	254
5.1.8	Detailed Algorithm of SIGM in Lemma 5.1.5 . . . . .	255
5.2	ERM in a Relaxed Non-interactive LDP model . . . . .	257
5.2.1	Related Work . . . . .	259
5.2.2	Our Model . . . . .	260
5.2.3	Privately Learning Generalized Linear Models . . . . .	260
5.2.4	Privately Learning Non-linear Regressions . . . . .	265
5.2.5	Experiments . . . . .	268
5.2.6	Omitted Proofs . . . . .	273
5.3	Sparse Linear Regression in LDP model . . . . .	292
5.3.1	Related Work . . . . .	294
5.3.2	Problem Set-up . . . . .	296

5.3.3	Keeping the Whole Dataset Private . . . . .	297
5.3.4	Keeping the Responses Private . . . . .	304
5.3.5	Extension to Other Problems . . . . .	306
5.3.6	Experiments . . . . .	312
5.3.7	Omitted Proofs . . . . .	316
<b>6</b>	<b>Some Matrix Estimation Problems in Differential Privacy Model . . . . .</b>	<b>338</b>
6.1	Principal Component Analysis in Local Differential Privacy Model . . . . .	338
6.1.1	Related Work . . . . .	340
6.1.2	Preliminaries . . . . .	341
6.1.3	Low Dimensional Case . . . . .	342
6.1.4	High Dimensional Sparse Case . . . . .	346
6.1.5	Experiments . . . . .	350
6.1.6	Omitted Proofs . . . . .	353
6.2	Differentially Private Sparse Covariance Matrix Estimation . . . . .	364
6.2.1	Related Work . . . . .	365
6.2.2	Private Sparse Covariance Estimation . . . . .	367
6.2.3	Main Method in Central DP Model . . . . .	368
6.2.4	Extension to Local Differential Privacy . . . . .	371
6.2.5	Lower Bound in Local Differential Privacy Model . . . . .	372
6.2.6	Experiments . . . . .	383
6.2.7	Omitted Proofs . . . . .	386
6.3	Differentially Private Sparse Inverse Covariance Matrix Estimation . . . . .	400
6.3.1	Related Work . . . . .	402
6.3.2	Preliminaries . . . . .	403
6.3.3	Sparse Inverse Covariance Estimation . . . . .	404
6.3.4	Output Perturbation Method . . . . .	405
6.3.5	Covariance Perturbation Method . . . . .	407
6.3.6	Experiments . . . . .	411
6.3.7	Omitted Proofs . . . . .	413
<b>7</b>	<b>Some Other Machine Learning Problems . . . . .</b>	<b>416</b>
7.1	Inferring Ground Truth From Crowdsourced Data Under Local Attribute Differential Privacy . . . . .	416
7.1.1	Related Work . . . . .	419
7.1.2	Preliminaries . . . . .	419
7.1.3	Main Method . . . . .	422
7.1.4	Theoretical Guarantees . . . . .	425
7.1.5	Comparison with Private Major Voting . . . . .	428
7.1.6	Omitted Proofs . . . . .	429
7.2	Differentially Private Expectation Maximization Algorithm . . . . .	438
7.2.1	Related Work . . . . .	439
7.2.2	Preliminaries . . . . .	440
7.2.3	Main Method . . . . .	444
7.2.4	Implications for Some Specific Models . . . . .	451

7.2.5	Statistical Guarantees of DP Expectation Maximization Algorithm .	455
7.2.6	DP EM Algorithm . . . . .	457
7.2.7	Experiments . . . . .	459
7.2.8	Omitted Proofs . . . . .	463
<b>8</b>	<b>Conclusion and Future Research</b> . . . . .	<b>479</b>
8.1	Conclusion . . . . .	479
8.2	Future Research . . . . .	483
<b>Reference</b> . . . . .		<b>515</b>

# Abstract

Machine Learning has emerged as one of the most powerful tools for us to learn and extract useful information from big data. It plays a vital role in many applications, especially in those from social sciences, finance, medical sciences and genomics research. However, due to the existence of sensitive information, we cannot implement machine learning algorithms directly on such data. Traditional ad hoc approaches like anonymization have suffered from numerous high-profile failures. Thus, approaches with more privacy preserving ability are urgently needed. For this purpose, we focus our studies on differential privacy (DP), which is a strong mathematical scheme for privacy preserving rooted in cryptography. It allows for rich statistical and machine learning analysis, and is now becoming a standard for private data analysis. Despite the rapid development of differential privacy in theory, its adoption to machine learning community remains slow. This dissertation summarizes our contributions to the sub-field of differentially private machine learning and presents a number of novel algorithms, new results and limitations for a number of fundamental machine learning problems.

In part one of this dissertation, we consider the Empirical Risk Minimization (ERM) problem in the differential privacy model. Firstly, we investigate the behaviors of Convex ERM in the central DP model. For this problem, we propose several algorithms with tighter utility upper bound and less running time in different settings, such as general convex, strongly convex and high dimensional settings. We also study the problem in the case where the underlying distribution of data is heavy-tailed, and the case where the loss function of

ERM is pairwise. Secondly, we investigate the behaviors of ERM with non-convex loss functions in the central DP model. Specifically, we first generalize the expected excess empirical risk from convex to Polyak-Lojasiewicz condition. Then, we study ERM with general non-convex loss functions by considering the error measurements from the first order stationary, second order stationary and global view, respectively. Thirdly, we consider ERM in the Non-interactive Local DP (NLDP) model and show how to reduce the exponential sample complexity given by previous studies for some special loss functions. We also show that if the server is allowed to have some public but unlabeled data, the sample complexity can be further reduced to polynomial size for smooth Generalized Linear Model. Fourthly, we try to understand the limitations of high dimensional ERM in the LDP model. Particularly, we study the sparse linear regression problem and show the lower bound of its estimation error. We also show some positive results under a relaxation of the problem.

In part two of this dissertation, we consider some matrix estimation problems. Firstly, we study the problem of Principal Component Analysis (PCA) in the LDP model and show its lower bound and near optimal upper bound for both low dimension and high dimensional sparse cases. Secondly, we study the sparse covariance matrix estimation problem and show its optimal upper bound and algorithm. Finally, we provide the first study of sparse inverse covariance matrix estimation problem in the DP model.

In part three of this dissertation, we consider some other machine learning related problems. Firstly, we study the problem of Uniform Facility Location problem in the Joint Differential Privacy model, we provide its lower bound and provide a near optimal algorithm. Second, we study the the problem of inferring ground truth in the Local Attribute Differential Privacy model and provide the first theoretical result on the problem. Thirdly, we focus on the DP version of Expectation Maximization algorithm. Specifically, we propose in the first DP version of (Gradient) EM algorithm with statistical guarantees. Finally, we consider the problem of truth discovery and propose an algorithm which can generate crowdsourced data differentially privately.

# List of Tables

3.1 Comparisons with previous $(\epsilon, \delta)$ -DP algorithms. We assume that the loss function $f$ is convex, 1-smooth, differentiable (twice differentiable for objective perturbation), and 1-Lipschitz. $r(\cdot)$ is $\mu$ -strongly convex. Bound and complexity ignore multiplicative dependence on $\log(1/\delta)$ . $\kappa = \frac{L}{\mu}$ is the condition number. The lower bound is $\Omega(\min\{1, \frac{p}{n^2\epsilon^2}\})$ [29]. . . . .	27
3.2 Comparisons with previous $(\epsilon, \delta)$ -DP algorithms, where $F^r$ is not necessarily strongly convex. We assume that the loss function $f$ is convex, 1-smooth, differentiable( twice differentiable for objective perturbation), and 1-Lipschitz. Bound and complexity ignore multiplicative dependence on $\log(1/\delta)$ . The lower bound in this case is $\Omega(\min\{1, \frac{\sqrt{p}}{n\epsilon}\})$ [29]. . . . .	28
3.3 Comparisons with previous $(\epsilon, \delta)$ -DP algorithms. We assume that the loss function $f$ is convex, 1-smooth, differentiable( twice differentiable for objective perturbation), and 1-Lipschitz. The utility bound depends on $G_{\mathcal{C}}$ , which is the Gaussian width of $\mathcal{C}$ . Bound and complexity ignore multiplicative dependence on $\log(1/\delta)$ . . . . .	28
3.4 The statistics of the adopted datasets. . . . .	106
4.1 Comparisons with previous $(\epsilon, \delta)$ -DP algorithms for DP-ERM with non-convex loss function. We assume that the Lipschitz and smooth parameters are 1, and $\ \mathcal{C}\ _2 \leq 1$ . . . . .	124
4.2 Summary of Datasets used in the experiments. . . . .	188
5.1 Comparisons on the sample complexities for achieving error $\alpha$ in the empirical risk, where $c$ is a constant. We assume that $\ x_i\ _2, \ y_i\  \leq 1$ for every $i \in [n]$ and the constraint set $\ \mathcal{C}\ _2 \leq 1$ . Asymptotic statements assume $\epsilon, \delta, \alpha \in (0, 1/2)$ and ignore dependencies on $\log(1/\delta)$ . . . . .	217

6.1	Results with different sparsity $s$ for LDP-High dimensional PCA on real world datasets. For all the datasets, the target dimensions $k$ is set to be $k = 10$ and $\epsilon = 2$ .	354
6.2	Results with different privacy levels $\epsilon$ for LDP-High dimensional PCA on real world datasets. For all the datasets, the target dimensions $k$ is set to be $k = 10$ and $s = 20$ .	355
6.3	The error upper bound of methods in the paper, which is measured by $\ \Theta_{\text{priv}} - \Theta^*\ _F$ , here we assume the $\ell_2$ -norm of each $x_i$ is bounded by 1.	401
6.4	Performance comparisons of the $\epsilon$ -differentially private algorithms on both synthetic and real-world datasets.	412
6.5	Performance comparisons of the $(\epsilon, \delta)$ -differentially private algorithms on both synthetic and real-world datasets.	413

# List of Figures

3.1	Experimental results on synthetic dataset for strongly convex case. . . . .	38
3.2	Experimental results on Covertype dataset for strongly convex case. . . . .	38
3.3	Experimental results on synthetic dataset for convex case. . . . . . . . .	38
3.4	Experimental results on IJCNN dataset for convex case. . . . . . . . .	39
3.5	Experiments on synthetic datasets. Figures 3.5a and 3.5b are for ridge regressions over synthetic data with Lognormal noises. Figures 3.5c and 3.5d are for logistic regressions over synthetic data with Loglogistic noises.	75
3.6	Experiments on UCI Adult dataset. Figures 3.6a and 3.6b are for ridge regressions. Figures 3.6c and 3.6d are for logistic regressions. . . . .	76
3.7	Experiments for the impact of dimensionality. Figure 3.7a and 3.7b are for ridge regressions. Figure 3.7c and 3.7d are for logistic regressions. . . . .	77
3.8	Experiments for the impact of the size of the dataset. Figure 3.8a and 3.8b are for ridge regressions. Figure 3.8c and 3.8d are for logistic regressions. .	78
3.9	The objective value of OnPairStrC for AUC maximization. . . . . . . . .	107
3.10	The objective value of OnPairC for AUC maximization. . . . . . . . .	108
3.11	The objective value of OffPairStrC for AUC maximization. . . . . . . . .	108
3.12	The AUC measurement of OffPairC. . . . . . . . . . . . . . . . .	108
3.13	The objective value of OnPairC for metric learning task under different training sizes. . . . . . . . . . . . . . . . .	109
3.14	The classification accuracy of OffPairC for metric learning task under different training sizes. . . . . . . . . . . . . . . . .	109

4.1	Experimental results on synthetic datasets for nonconvex case. . . . .	134
4.2	Experimental results on Covertype dataset for nonconvex case. . . . .	135
4.3	Experimental results on IJCNN dataset for nonconvex case. . . . .	136
4.4	Experimental results on the norm of projected gradient w.r.t different methods. The left one is for synthetic dataset, the middle one is for Covertype dataset, and the right one is for IJCNN dataset. . . . .	137
4.5	Accuracy w.r.t privacy level on Covertype and IJCNN datasets . . . . .	190
4.6	Results of logistic regression with non-convex regularizer on Covertype dataset . . . . .	190
4.7	Results of logistic regression with non-convex regularizer on IJCNN dataset	191
4.8	Results of sigmoid regression with $\ell_2$ norm regularizer on Covertype dataset	191
4.9	Results of sigmoid regression with $\ell_2$ norm regularizer on IJCNN dataset . .	191
5.1	GLM with logistic loss under i.i.d Bernoulli design. The left plot shows the squared relative error under different levels of privacy. The right one shows relative error under different dimensionality. . . . .	268
5.2	Cubic regression with i.i.d Bernoulli design. The left plot shows the squared relative error under different level of privacy. The right one shows relative error under different dimensionality. . . . .	270
5.3	GLM with logistic loss on real dataset. The dataset we use is Covertype (left) and SUSY (right). . . . .	270
5.4	The effect of the number of public unlabeled samples. The left plot shows the relative error of GLM with logistic loss. The right one shows the relative error of cubic regression. . . . .	271
5.5	Experimental results on sparse linear regression under LDP while keeping the whole dataset private (Algorithm 5.3.41). . . . .	314
5.6	Experimental results on sparse linear regression under LDP while keeping the labels private (Algorithm 5.3.42). . . . .	314
5.7	Experimental results on Covertype dataset [90] for $\ell_0$ -constrained logistic regression under $(\epsilon, \delta)$ -DP (Algorithm 5.3.43). . . . .	314

5.8	Experimental results on rcv1 dataset [64] for $\ell_0$ -constrained logistic regression under $(\epsilon, \delta)$ -DP (Algorithm 5.3.43). . . . .	315
5.9	Experimental results for sparse regression with non-linear measurement under LDP when keeping the whole dataset private (Algorithm 5.3.44). . . . .	315
5.10	Experimental results for sparse regression with non-linear measurement under LDP when keeping the label private (Algorithm 5.3.45). . . . .	315
6.1	LDP-PCA in low dimensional case on real world datasets with different sample size. The left one is for Covertype. The middle one is for Buzz. The right one is for Year dataset. . . . .	351
6.2	LDP-PCA in low dimensional case on real world datasets at different levels of privacy. The left one is for Covertype. The middle one is for Buzz. The right one is for Year dataset. . . . .	351
6.3	LDP-PCA in low dimensional case on synthetic datasets. The left one is for different target dimensions $k$ over sample size $n$ with fixed $\epsilon = 0.5$ and $p = 40$ . The middle one is for different dimensions with fixed $n = 10^5$ and $\epsilon = 0.5$ . The right one is for different level of privacy with fixed $n = 10^5$ and $p = 40$ . . . . .	352
6.4	LDP-PCA in high dimensional case on synthetic datasets. The left one is for different target dimensions $k$ over sample size $n$ with fixed $\epsilon = 1$ and $p = 400$ . The middle one is for different dimensions with fixed $n = 2000$ and $\epsilon = 1$ . The right one is for different level of privacy with fixed $n = 2000$ and $p = 400$ . . . . .	352
6.5	Experiment results of Algorithm 6.2.48 for $\ell_2$ -norm relative error. The left one is for different sparsity levels, the middle one is for different dimensionality $p$ , and the right one is for different privacy level $\epsilon$ . . . . .	385
6.6	Experiment results of Algorithm 6.2.48 for $\ell_1$ -norm relative error. The left one is for different sparsity levels, the middle one is for different dimensionality $p$ , and the right one is for different privacy level $\epsilon$ . . . . .	385
6.7	Experiment results of Algorithm 6.2.49 for $\ell_2$ -norm relative error. The left one is for different sparsity levels, the middle one is for different dimensionality $p$ , and the right one is for different privacy level $\epsilon$ . . . . .	385
6.8	Experiment results of Algorithm 6.2.49 for $\ell_1$ -norm relative error. The left one is for different sparsity levels, the middle one is for different dimensionality $p$ , and the right one is for different privacy level $\epsilon$ . . . . .	386

7.1	Estimation error of Algorithm 7.2.55 (clipped) v.s. iteration $t$ under different clipping threshold $C$ . . . . .	460
7.2	Estimation error of GMM w.r.t privacy budget $\epsilon$ , data dimension $d$ , data size $n$ and iteration $t$ . . . . .	460
7.3	Estimation error of MRM w.r.t privacy budget $\epsilon$ , data dimension $d$ , data size $n$ and iteration $t$ . . . . .	460
7.4	Estimation error of RMC w.r.t privacy budget $\epsilon$ , data dimension $d$ , data size $n$ and iteration $t$ . . . . .	460
7.5	Estimation error of GMM w.r.t privacy budget $\epsilon$ , data dimension $d$ and data size $n$ (we set $\beta = \beta^T$ with $T = 22$ ) . . . . .	462
7.6	Estimation error of MRM w.r.t privacy budget $\epsilon$ , data dimension $d$ and data size $n$ (we set $\beta = \beta^T$ with $T = 22$ ) . . . . .	462
7.7	Estimation error of RMC w.r.t privacy budget $\epsilon$ , data dimension $d$ and data size $n$ (we set $\beta = \beta^T$ with $T = 22$ ) . . . . .	462

# Chapter 1

## Introduction

In the big data era, Machine Learning is now becoming one of the most powerful tools for us to learn and extract useful information from data. It plays an important role in many applications, especially in those from social sciences [79, 147, 136, 213] , finance [145, 80, 281, 88], medical sciences [236, 89, 190, 167] and genomics research [13, 194, 232]. However, due to the existence of sensitive information in such applications, we cannot directly apply existing machine learning algorithms to such data. For example, most of the biomedical data are held by some organizations, such as hospitals or physicians, without a proper privacy-preserving mechanism, these organizations cannot or are unwilling to share these sensitive data. Thus, it is urgently needed to develop effective machine learning algorithms which can learn these sensitive data efficiently and meanwhile protect their privacy.

To preserve the privacy of sensitive data, a commonly adopted strategy is anonymization, which simply removes any sensitive information from the original data. For example, TriNetX is a global federated research network providing statistics on Electric Medical Record (EMR) that includes various types of patient data like diagnosis, procedures, medications, laboratory results, and genomic information. TriNetX allows participating organizations and individuals to explore anonymized patient data in a browser- based, real-time

fashion [263]. Its users include a mix of hospitals, primary cares, and specialty treatment providers spanning a wide range of geographies, age groups, and income levels. At TriNetx, only statistical summaries of de-identified information are provided, but not the protected health information. Both the patients and the data-providing organizations stay anonymous.

However, such a privacy preserving approach could have a few severe flaws. One of them is that some sensitive information could still be released after removing those apparent identifiers such as name, address, and social security number. This could still happen even when only statistical summaries are released. There are quite a few anonymization failures. One of the famous examples is that Netflix made a huge database of movie recommendations available for study with the obvious identifiable information removed. But it is shown that when paired with other existing data, re-identification becomes possible [223]. The same phenomenon has been observed in other kinds of data, such as social network graphs [15], search query logs [168] and others. Moreover, releasing statistics computed on sensitive data can also be problematic; for example, Wang et al. in [333] showed that releasing R<sup>2</sup>-values computed on high-dimensional genetic data can lead to privacy breaches by an adversary who is armed with a small amount of auxiliary information.

Instead of publicizing anonymized data, even just releasing the machine learning model could still cause privacy breach. Actually, previous papers have already showed that many machine learning algorithms are exposed to several types of privacy attacks. Attacks targeting data privacy include: adversary inferring whether input examples were used to train the target model with membership inference attacks [352, 256, 259, 224], learning the global properties of training data with property inference attacks [120], and covert channel model training attacks [258]. Attacks targeting model privacy include: adversary uncovering the model details with model extraction attacks [279], and inferring hyperparameters with hyperparameter stealing attacks [294]. Thus, approaches with more privacy preserving ability are urgently needed.

An effective way to resolve these issues is to design differentially private machine

learning algorithms. Differential Privacy (DP) [107], with roots in cryptography, is a strong mathematical scheme for privacy preserving. It allows for rich statistical and machine learning analysis, and is now becoming a standard for private data analysis. Informally speaking, DP ensures that an adversary cannot infer whether or not a particular individual is participating in the database query, even with unbounded computational power and access to every entry in the database except for that particular individual’s data. DP considers a centralized setting that includes a trusted data curator, who generates the perturbed statistical information (e.g., counts and histograms) by using some randomized mechanism.

The decade and a half since the seminal differential privacy paper [107] saw an early focus on developing privacy release mechanisms and answering some basic queries, such as [35, 41, 63, 87, 97, 103, 105, 137, 138, 142, 152]. Despite the rapid development of DP in theory, its adoption to machine learning community remains slow. Possibly, the reason is that, unlike the traditional machine learning algorithms, there are three main ingredients in any differentially private learning algorithm involving sensitive data, the privacy-preserving model, the (sensitive) data, and the objective functions. Each of them could impose tremendous challenges for designing an effective privately learning algorithm. For example, the popular non-interactive local differential privacy model often requires a large (or even exponential) number of data samples (called sample complexity) in order to ensure the learning accuracy for some loss functions, making them inapplicable to real world data [257]. Many data such as biomedical data are often high dimensional and irregular (e.g., heavy-tailed due to the existence of outliers). This could cause major difficulty for designing DP algorithms. It is known that differential privacy is not achievable for some learning problems, such as linear regression, in high dimensional space [317]. Heavy-tailed data could lead to unbounded gradient, and thus fail almost all existing DP learning algorithms. Loss functions in many machine learning tasks could be non-convex (such as those used in deep neural networks) and complex (such as those pairwise loss functions used for patient similarity learning). Such types of loss functions could be difficult to optimize and thus

challenging to achieve differential privacy. Thus, two fundamental questions are

**What are the limitations of machine learning problems and how to design machine learning algorithms in the differential privacy model?**

This thesis focuses on answering the above two questions. Specifically, it investigates both theoretical and practical behaviors of several fundamental machine learning problems in different differential privacy models.

## 1.1 Dissertation Contributions

To be more precise, parts of the thesis are based on my following published joint work

- In Chapter 3, I study the Empirical Risk Minimization problem (ERM) (*i.e.*, DP-ERM) with convex loss in the differential privacy model. Specifically,
  - In Chapter 3.1, I focus on DP-ERM in the central  $(\epsilon, \delta)$ -DP model. For smooth (strongly) convex loss function with or without (non)-smooth regularization, I give algorithms that achieve either optimal or near optimal utility bounds with less gradient complexity compared with previous work. For ERM with smooth convex loss function in high-dimensional ( $p \gg n$ ) setting, I give an algorithm which achieves the upper bound with less gradient complexity than previous ones. At last, I generalize the expected excess empirical risk from convex loss functions to non-convex ones satisfying the Polyak-Lojasiewicz condition and give a tighter upper bound on the utility than the one in [356]. Part of this work appeared in our published work in Conference on Neural Information Processing Systems (NIPS/NeurIPS) 2017 [327].
  - In Chapter 3.2, I study Differentially Private Stochastic Convex Optimization, which is a generalization of DP-ERM, with heavy-tailed data. For this problem, I provide a comprehensive study of DP-SCO under various settings. First, I

consider the case where the loss function is strongly convex and smooth. For this case, I propose a method based on the sample-and-aggregate framework, which has an excess population risk of  $\tilde{O}(\frac{d^3}{n\epsilon^4})$  (after omitting other factors), where  $n$  is the sample size and  $d$  is the dimensionality of the data. Then, I show that with some additional assumptions on the loss functions, it is possible to reduce the *expected* excess population risk to  $\tilde{O}(\frac{d^2}{n\epsilon^2})$ . To lift these additional conditions, I also provide a gradient smoothing and trimming based scheme to achieve excess population risks of  $\tilde{O}(\frac{d^2}{n\epsilon^2})$  and  $\tilde{O}(\frac{d^{\frac{3}{2}}}{(n\epsilon^2)^{\frac{1}{3}}})$  for strongly convex and general convex loss functions, respectively, *with high probability*. Experiments suggest that these algorithms can effectively deal with the challenges caused by data irregularity. Part of this work appeared in our published work in The 37th International Conference on Machine Learning (ICML 2020) [331].

- In Chapter 3.3, I generalize the classical DP-ERM setting to the case where the loss functions are pairwise loss instead of pointwise loss. I propose several differentially private pairwise learning algorithms for both online and offline settings. Specifically, for the online setting, I first introduce a differentially private algorithm (called OnPairStrC) for strongly convex loss functions. Then, I extend this algorithm to general convex loss functions and give another differentially private algorithm (called OnPairC). For the offline setting, I also present two differentially private algorithms (called OffPairStrC and OffPairC) for strongly and general convex loss functions, respectively. These proposed algorithms can not only learn the model effectively from the data but also provide strong privacy protection guarantee for sensitive information in the training set. Extensive experiments on real-world datasets are conducted to evaluate the proposed algorithms and the experimental results support my theoretical analysis. Part of this work appeared in our published work in The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2020) [154].

- Beyond the convex loss functions, in Chapter 4, I investigate the theoretical behaviors of DP-ERM with non-convex loss functions. I will study the problem from three perspectives.
  - In Chapter 4.1, I study the behavior of the problem from the first order stationary view (that is, I use some first order stationary measurement to measure the private estimator). For DP-ERM with non-smooth regularizer, we generalize an existing work by measuring the utility using  $\ell_2$  norm of the projected gradient. Also, I extend the error bound measurement, for the first time, from empirical risk to population risk by using the expected  $\ell_2$  norm of the gradient. I then investigate the problem in high dimensional space, and show that by measuring the utility with Frank-Wolfe gap, it is possible to bound the utility by the Gaussian Width of the constraint set, instead of the dimensionality  $p$  of the underlying space. I further demonstrate that the advantages of this result can be achieved by the measure of  $\ell_2$  norm of the projected gradient. A somewhat surprising discovery is that although the two kinds of measurements are quite different, their induced utility upper bounds are asymptotically the same under some assumptions. I also show that the utility of some special non-convex loss functions can be reduced to a level (*i.e.*, depending only on  $\log p$ ) similar to that of convex loss functions. Finally, I test the proposed algorithms on both synthetic and real world datasets and the experimental results confirm those theoretical analysis. Part of this work appeared in our published work in The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI 2019) [308].
  - Next, in Chapter 4.2, I study the problem with the measurement of the excess empirical risk or population risk, which was primarily used as the utility to measure the quality for convex loss functions. Specifically, I show that the excess empirical (or population) risk can be upper bounded by  $\tilde{O}(\frac{d \log(1/\delta)}{\log n \epsilon^2})$  in the  $(\epsilon, \delta)$ -DP settings, where  $n$  is the data size and  $d$  is the dimensionality of

the space. The  $\frac{1}{\log n}$  term in the empirical risk bound can be further improved to  $\frac{1}{n^{\Omega(1)}}$  (when  $d$  is a constant) by a highly non-trivial analysis on the time-average error. Next, I show how to improve the bounds for some specific problems. Particularly, we focus on the generalized linear model with non-convex loss functions and the robust regressions problem with additional assumptions, and present an  $(\epsilon, \delta)$ -DP algorithm for them with population risk  $O(\frac{\sqrt{d}}{\sqrt{n}\epsilon})$ . Part of this work appeared in our published work in The 36th International Conference on Machine Learning (ICML 2019) [296] and Conference on Neural Information Processing Systems (NIPS/NeurIPS) 2017 [327].

- Finally, in Chapter 4.3, I study the behavior of the problem from the second order stationary view (that is, I use some second order stationary measurement to measure the private estimator). Specifically, I consider the connection between achieving differential privacy and finding approximate local minimum. Particularly, I show that when the size  $n$  is large enough, there are  $(\epsilon, \delta)$ -DP algorithms which can find an approximate local minimum of the empirical risk with high probability in both the constrained and non-constrained settings. These results indicate that one can escape saddle points privately. To deal with the issues of high sample complexity and non-scalable, I propose a new method called Differentially Private Trust Region, and show that it outputs a second-order stationary point with high probability and less sample complexity, compared to the existing one. Moreover, I also provide a stochastic version of the method (along with some theoretical guarantees) to make it faster and more scalable. Experiments on benchmark datasets suggest that these methods are indeed more efficient and practical. Part of this work appeared in our published work in The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD 2020) and The 36th International Conference on Machine Learning (ICML 2019) [296, 313].

- Instead of the central DP model, in Chapter 5 I will study ERM in the Local Differential Privacy model. Particularly, I focus on the theoretical behaviors of ERM in the non-interactive local model and high dimensional sparse linear regression problem. Specifically,

- In Chapter 5.1 I study ERM in the non-interactive local model. Previous research on this problem [257] indicates that the sample complexity, to achieve error  $\alpha$ , needs to be exponentially depending on the dimensionality  $p$  for general loss functions. In this chapter, I make two attempts to resolve this issue by investigating conditions on the loss functions that allow us to remove such a limit. In the first attempt, I show that if the loss function is  $(\infty, T)$ -smooth, by using the Bernstein polynomial approximation we can avoid the exponential dependency in the term of  $\alpha$ . I then propose player-efficient algorithms with 1-bit communication complexity and  $O(1)$  computation cost for each player. The error bound of these algorithms is asymptotically the same as the original one. With some additional assumptions, we also give an algorithm which is more efficient for the server. In the second attempt, I show that for any 1-Lipschitz generalized linear convex loss function, there is an  $(\epsilon, \delta)$ -LDP algorithm whose sample complexity for achieving error  $\alpha$  is only linear in the dimensionality  $p$ . Finally, motivated by the idea of using polynomial approximation and based on different types of polynomial approximations, I propose (efficient) non-interactive locally differentially private algorithms for learning the set of k-way marginal queries and the set of smooth queries. Part of this work appeared in our published work in Conference on Neural Information Processing Systems (NIPS/NeurIPS) 2018 [300] and The 30th International Conference on Algorithmic Learning Theory (ALT 2019) [306].
- To alleviate the issues of practice and exponential sample complexity. In Chapter 5.2 I relax the non-interactive LDP model. Different from its classical setting,

my new model allows the server to access some additional public but unlabeled data. I first show that there is an  $(\epsilon, \delta)$ -NLDP algorithm for GLM (under some mild assumptions), if each data record is i.i.d sampled from some sub-Gaussian distribution with bounded  $\ell_1$ -norm. Then with high probability, the sample complexity of the public and private data, for the algorithm to achieve an  $\alpha$  estimation error (in  $\ell_\infty$ -norm), is  $O(p^2\alpha^{-2})$  and  $O(p^2\alpha^{-2}\epsilon^{-2})$ , respectively, if  $\alpha$  is not too small (*i.e.*,  $\alpha \geq \Omega(\frac{1}{\sqrt{p}})$ ), where  $p$  is the dimensionality of the data. We then extend our idea to the non-linear regression problem and show a similar phenomenon for it. To my best knowledge, this is the first paper showing the existence of efficient and effective algorithms for GLM and non-linear regression in the NLDP model with public unlabeled data. Note that this is my unpublished work.

- Instead of ERM in the non-interactive local model. In Chapter 5.3 I will study the high dimensional ERM in the general local model, and I will concentrate on the most simplest problem, *i.e.*, Sparse Linear Regression. I first show that polynomial dependency on the dimensionality  $p$  of the space is unavoidable for the estimation error in both non-interactive and sequential interactive local models, if the privacy of the whole dataset needs to be preserved. Similar limitations also exist for other types of error measurements and in the relaxed local models. This indicates that differential privacy in high dimensional space is unlikely achievable for the problem. With the understanding of this limitation, then I present two algorithmic results. The first one is a sequential interactive LDP algorithm for the low dimensional sparse case, called Locally Differentially Private Iterative Hard Thresholding (LDP-IHT), which achieves a near optimal upper bound. This algorithm is actually rather general and can be used to solve quite a few other problems, such as (Local) DP-ERM with sparsity constraints and sparse regression with non-linear measurements. The second one is for

the restricted (high dimensional) case where only the privacy of the responses (labels) needs to be preserved. For this case, we show that the optimal rate of the error estimation can be made logarithmically dependent on  $p$  (i.e.,  $\log p$ ) in the local model, where an upper bound is obtained by a label-privacy version of LDP-IHT. Part of this work appeared in our published work in The 36th International Conference on Machine Learning (ICML 2019) [317].

- In the second part of this dissertation, beyond the vector estimation in ERM problem, I will focus on some estimation or inference statistical problems that are related to matrix. Specifically, in Chapter 6, I will study three canonical matrix estimation problems,
  - In Chapter 6.1, I study the Principal Component Analysis (PCA) problem under the non-interactive local differential privacy model. For the low dimensional case (*i.e.*,  $p \ll n$ ), I will show that the optimal rate of  $\Theta(\frac{kp}{n\epsilon^2})$  (omitting the eigenvalue terms) for the private minimax risk of the  $k$ -dimensional PCA using the squared subspace distance as the measurement, where  $n$  is the sample size and  $\epsilon$  is the privacy parameter. For the high dimensional (*i.e.*,  $p \gg n$ ) row sparse case, I first give a lower bound of  $\Omega(\frac{ks \log p}{n\epsilon^2})$  on the private minimax risk, where  $s$  is the underlying sparsity parameter. Then we provide an efficient algorithm to achieve the upper bound of  $O(\frac{s^2 \log p}{n\epsilon^2})$ . Experiments on both synthetic and real world datasets confirm my theoretical guarantees. Part of this work appeared in our published work in The 28th International Joint Conference on Artificial Intelligence (IJCAI 2019) [320] and Theoretical Computer Science [322].
  - Next, in Chapter 6.2, I will study the problem of estimating the covariance matrix under differential privacy, where the underlying covariance matrix is assumed to be sparse and of high dimensions. Firstly, I propose a new method, called DP-Thresholding, to achieve a non-trivial  $\ell_2$ -norm based error bound

*i.e.,*  $O\left(\frac{s^2 \log p \log \frac{1}{\delta}}{n\epsilon^2}\right)$  where  $s$  is the row sparsity of the underlying covariance matrix,  $n$  is the sample size, and  $p$  is the dimensionality of the data, and it is significantly better than the existing ones from adding noise directly to the empirical covariance matrix. I also extend the  $\ell_2$ -norm based error bound to a general  $\ell_w$ -norm based one for any  $1 \leq w \leq \infty$ , and show that they share the same upper bound asymptotically. My approach can be easily extended to local differential privacy. Secondly, I show that the upper bound of the problem in LDP model is actually tight. My main technique for achieving this lower bound is a general framework, called General Private Assouad Lemma, which is a considerable generalization of the previous private Assouad lemma and can be used as a general method for bounding the private minimax risk of matrix-related estimation problems. Experiments on the synthetic datasets show consistent results with our theoretical claims. Part of this work appeared in our published work in The 53rd Annual Conference on Information Sciences and Systems (CISS 2019) [312], The 28th International Joint Conference on Artificial Intelligence (IJCAI 2019) [316] and Theoretical Computer Science [323].

- Finally, in Chapter 6.3 I give the first study of sparse inverse covariance estimation problem under differential privacy. Firstly, we propose an  $\epsilon$ -differentially private algorithm via output perturbation, which is based on the sensitivity of the optimization problem and Wishart mechanism. Based on the idea of that, I propose a general covariance perturbation method, and then for  $\epsilon$ -differential privacy, I analyze Laplacian and Wishart mechanisms, for  $(\epsilon, \delta)$ -differential privacy I analyze Gaussian and Wishart mechanisms. Moreover, I extend the covariance perturbation algorithm to distributed setting and local differential privacy. Experiments on synthetic and benchmark datasets are also support these theoretical analysis. Part of this work appeared in our published work in

2018 6th IEEE Global Conference on Signal and Information Processing (2018 GlobalSip) [303].

- In Chapter 7, I will study some other problems of Machine Learning in DP model, which ranges from clustering, truth discovery, latent variable models and generating synthetic dataset. Specifically,
  - Chapter 7.1 is my unpublished work. In this chapter I focus on studying the ground truth inference problem under local attribute differential privacy (LADP) model, which is a relaxation of LDP model, and propose a new algorithm called private Dawid-Skene method, which is motivated by the classical Dawid-Skene method. Specifically, I first provide the estimation errors for both ability of users and the ground truth under some assumptions of the problem if the algorithm start with some appropriate initial vector. Moreover, I also propose an explicit instance and show that the estimation error of the ground truth achieved by the private major voting algorithm is always greater than the error achieved by previous method.
  - Chapter 7.2 is also an unpublished work. I propose in this chapter the first DP version of (Gradient) EM algorithm with statistical guarantees. Moreover, I apply the general framework to three canonical models: Gaussian Mixture Model (GMM), Mixture of Regressions Model (MRM) and Linear Regression with Missing Covariates (RMC). Specifically, for GMM in the DP model, my estimation error is near optimal in some cases. For the other two models, I provide the first finite sample statistical guarantees. My theory is also supported by thorough numerical experiments.

Furthermore, some other work during my PhD not included in the thesis include four of our published papers [315, 332, 330, 314] and one submitted manuscript [364].

## 1.2 Dissertation Outline

The rest of the chapters go as follows:

- In Chapter 2 I will review some definitions, mechanisms, properties and lemmas of Differential Privacy (DP) and its local version, Local Differential Privacy (LDP) that will be used throughout the whole dissertation.
- In Chapter 3 I will study the Empirical Risk Minimization with convex loss functions in the DP model (DP-ERM). Chapter 3.1 studies how to design faster algorithms for DP-ERM in the  $(\epsilon, \delta)$ -DP model. Chapter 3.2 focuses on the stochastic version of DP-ERM, *i.e.*, DP-SCO, in  $(\epsilon, \delta)$ -DP model where the data distribution is heavy-tailed. Chapter 3.3 studies DP-ERM with pairwise loss functions.
- Instead of convex loss functions, Chapter 4 studies DP-ERM with non-convex loss functions. Chapter 4.1 is about theoretical behaviors of private estimator under the first order stationary measurement. In Chapter 4.2 I provide some upper bounds of errors using the excess empirical or population risk. In Chapter 4.3 I show how to escape saddle points of the Empirical Risk function in DP model.
- Instead of the central model, in Chapter 5 I study ERM in LDP model. I first study ERM in the non-interactive LDP model in Chapter 5.1. Then I relax the non-interactive LDP model and study Generalized Linear Models in the non-interactive LDP model with some public but unlabeled data. Finally, I study the high dimensionality issue of ERM in LDP model via studying sparse linear regression.
- Chapter 6 focuses on some matrix related estimation problems in (Local) DP model. In Chapter 6.1 I study Principal Component Analysis in LDP model. In Chapter 6.2 I study Sparse Covariance Matrix estimation in DP and LDP model. Finally in Chapter 6.3 I study Sparse Inverse Covariance Matrix estimation in DP model.

- Chapter 7 I study other machine learning problems. In Chapter 7.1 I study ground truth inference in the Local Attribute Differential Privacy model. Chapter 7.2 focuses on the statistical guarantees of DP version of Expectation Maximization algorithm.
- In Chapter 8, I will conclude the dissertation and discusses some potential directions for future research.

# Chapter 2

## Differential Privacy Background

In this Chapter, I will introduce some definitions, properties and mechanisms of Differential Privacy (DP) that will be used throughout the whole dissertation. More details could be found in [104].

### 2.1 Central Differential Privacy

Informally speaking, Differential Privacy (DP) ensures that an adversary cannot infer whether or not a particular individual is participating in the database query, even with unbounded computational power and access to every entry in the database except for that particular individual's data. DP considers a centralized setting that includes a trusted data curator, who generates the perturbed statistical information (e.g., counts and histograms) by using some randomized mechanism. It works by injecting random noise into the statistical results obtained from sensitive data so that the distribution of the perturbed results is insensitive to any single element (i.e., a data point/item) change in the original dataset. Formally, it can be defined as follows.

We say that two datasets  $D$  and  $D'$  are neighbors to each other if they differ by only one entry, denoted as  $D \sim D'$ .

**Definition 2.1.1** (Differential Privacy [107]). Given a data universe  $\mathcal{X}$ , we say that two datasets  $D, D' \subseteq \mathcal{X}$  are neighbors if they differ by only one entry, which is denoted as  $D \sim D'$ . A randomized algorithm  $\mathcal{A}$  is  $(\epsilon, \delta)$ -differentially private (DP) if for all neighboring datasets  $D, D'$  and for all events  $S$  in the output space of  $\mathcal{A}$ , we have

$$\mathbb{P}(\mathcal{A}(D) \in S) \leq e^\epsilon \mathbb{P}(\mathcal{A}(D') \in S) + \delta.$$

In practice,  $\epsilon \approx 0.1$  and  $\delta \approx 1/n^{\omega(1)}$  are often good enough choices for  $(\epsilon, \delta)$ -differential privacy, where  $n$  is the number of samples in the dataset.

It is notable that DP enjoys the post-processing and sub-sampling properties, which are commonly used in machine learning related problems.

**Lemma 2.1.1** (Post-processing Property of DP). Let  $\mathcal{M}$  be an  $(\epsilon, \delta)$ -DP mechanism, and  $f : \text{Range}(M) \mapsto \mathcal{R}$  be an arbitrary randomized mapping. Then  $f \circ \mathcal{M}$  is also  $(\epsilon, \delta)$ -DP.

**Lemma 2.1.2** (Sub-sampling Property of DP [29]). Over a domain of datasets  $\mathcal{X}^n$ , if an algorithm  $\mathcal{A}$  is  $(\epsilon, \delta)$ -DP, then for any  $n$ -size dataset  $D$ , executing  $\mathcal{A}$  on uniformly random  $\gamma n$  entries of  $D$  ensures  $(2\gamma\epsilon, \delta)$ -DP.

**Definition 2.1.2** (Laplacian Mechanism). Given a function  $q : \mathcal{X}^n \rightarrow \mathbb{R}^p$ , the Laplacian Mechanism is defined as:  $\mathcal{M}_L(D, q, \epsilon) = q(D) + (Y_1, Y_2, \dots, Y_p)$ , where  $Y_i$  is i.i.d. drawn from a Laplacian Distribution  $\text{Lap}(\frac{\Delta_1(q)}{\epsilon})$ , where  $\Delta_1(q)$  is the  $\ell_1$ -sensitivity of the function  $q$ , i.e.,  $\Delta_1(q) = \sup_{D \sim D'} \|q(D) - q(D')\|_1$ . For a parameter  $\lambda$ , the Laplacian distribution has the density function:

$$\text{Lap}(x|\lambda) = \frac{1}{2\lambda} \exp(-\frac{x}{\lambda}).$$

Laplacian Mechanism preserves  $\epsilon$ -differentially private.

**Definition 2.1.3** (Gaussian Mechanism). Given a function  $q : \mathcal{X}^n \rightarrow \mathbb{R}^p$ , the Gaussian Mechanism is defined as:  $\mathcal{M}_G(D, q, \epsilon) = q(D) + Y$ , where  $Y$  is drawn from a Gaussian Distribution  $\mathcal{N}(0, \sigma^2 I_p)$  with  $\sigma \geq \frac{\sqrt{2 \ln(1.25/\delta)} \Delta_2(q)}{\epsilon}$ .  $\Delta_2(q)$  is the  $\ell_2$ -sensitivity of the function  $q$ ,

i.e.,  $\Delta_2(q) = \sup_{D \sim D'} \|q(D) - q(D')\|_2$ . Gaussian Mechanism preserves  $(\epsilon, \delta)$ -differentially private.

**Definition 2.1.4** (Exponential Mechanism). The Exponential Mechanism allows differentially private computation over arbitrary domains and range  $\mathcal{R}$ , parametrized by a score function  $u(D, r)$  which maps a pair of input data set  $D$  and candidate result  $r \in \mathcal{R}$  to a real valued score. With the score function  $u$  and privacy budget  $\epsilon$ , the mechanism yields an output with exponential bias in favor of high scoring outputs. Let  $\mathcal{M}(D, x, R)$  denote the exponential mechanism, and  $\Delta$  be the sensitivity of  $u$  in the range  $R$ ,  $\Delta = \max_{r \in \mathcal{R}} \max_{D \sim D'} |u(D, r) - u(D', r)|$ . Then if  $\mathcal{M}(D, x, R)$  selects and outputs an element  $r \in \mathcal{R}$  with probability proportional to  $\exp(\frac{\epsilon u(D, r)}{2\Delta u})$ , it preserves  $\epsilon$ -differential privacy.

**Lemma 2.1.3** ([104]). For the exponential mechanism  $\mathcal{M}(D, u, \mathcal{R})$ , we have

$$\Pr\{u(\mathcal{M}(D, u, \mathcal{R})) \leq OPT_u(x) - \frac{2\Delta u}{\epsilon}(\ln |\mathcal{R}| + t)\} \leq e^{-t},$$

where  $OPT_u(x)$  is the highest score in the range  $\mathcal{R}$ , i.e.  $\max_{r \in \mathcal{R}} u(D, r)$ .

**Lemma 2.1.4** (Basic Composition Theorem). Let  $\mathcal{M}_i$  be an  $(\epsilon_i, \delta_i)$  DP mechanism, then the composition mechanism  $\mathcal{M}^T = (\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_T)$  will be  $(\sum_{i=1}^T \epsilon_i, \sum_{i=1}^T \delta_i)$  DP.

Thus, given target privacy parameters  $0 < \epsilon < 1$  and  $0 < \delta < 1$ , to ensure  $(\epsilon, \delta)$ -DP over  $T$  mechanisms, it suffices that each mechanism is  $(\epsilon', \delta')$ -DP, where  $\epsilon' = \frac{\epsilon}{T}$  and  $\delta' = \frac{\delta}{T}$ .

In addition to allowing the parameters to degrade more slowly, we would like our theorem to be able to handle more complicated forms of composition.

**Lemma 2.1.5** (Advanced Composition Theorem). Given target privacy parameters  $0 < \epsilon < 1$  and  $0 < \delta < 1$ , to ensure  $(\epsilon, T\delta' + \delta)$ -DP over  $T$  mechanisms, it suffices that each mechanism is  $(\epsilon', \delta')$ -DP, where  $\epsilon' = \frac{\epsilon}{2\sqrt{2T \ln(2/\delta)}}$  and  $\delta' = \frac{\delta}{T}$ .

The Moments Accountant method proposed in [1] is a technique to accumulate the privacy cost which has tighter bound for  $\epsilon$  and  $\delta$ .

**Lemma 2.1.6** (Moments Accountant). Given target privacy parameters  $0 < \epsilon < 1$  and  $0 < \delta < 1$ , to ensure  $(\epsilon, T\delta' + \delta)$ -DP over  $T$  mechanisms, it suffices that each mechanism is  $(\epsilon', \delta')$ -DP, where  $\epsilon' = \frac{\epsilon}{2\sqrt{2T\ln(2/\delta)}}$  and  $\delta' = \delta$ .

Roughly speaking, when we use the Gaussian Mechanism on the (stochastic) gradient descent, we can save a factor of  $\sqrt{\ln(T/\delta)}$  in the asymptotic bound of standard deviation of noise compared with the advanced composition theorem in [105].

**Lemma 2.1.7.** [1] For any  $G$ -Lipschitz loss function, there exist constants  $c_1$  and  $c_2$  so that given the sampling probability  $q = l/n$  and the number of steps  $T$ , for any  $\epsilon < c_1 q^2 T$ , a DP stochastic gradient algorithm with batch size  $l$  that injects Gaussian Noise with standard deviation  $\frac{G}{n}\sigma$  to the gradients (Algorithm 1 in [1]) is  $(\epsilon, \delta)$ -differentially private for any  $\delta > 0$  if  $\sigma \geq c_2 \frac{q\sqrt{T\ln(1/\delta)}}{\epsilon}$ .

More details of how to use this lemma could be found in Chapter 3 and 4.

Besides the classical Differential Privacy, additionally, we also use zero Concentrated Differential Privacy (zCDP) [52] and its composition property to guarantee  $(\epsilon, \delta)$ -DP. Compared to directly using the composition property of DP, it has many advantages (see [192, 309] for more details).

**Definition 2.1.5.** A randomized mechanism  $\mathcal{A}$  is  $\rho$ -zCDP if, for all neighboring dataset  $D, D'$  and all  $\alpha \in (1, \infty)$ ,

$$D_\alpha(\mathcal{A}(D) || \mathcal{A}(D')) \leq \rho\alpha,$$

where  $D_\alpha(\cdot || \cdot)$  is the  $\alpha$ -Rényi Divergence <sup>1</sup>.

The following three lemmas are some properties of zCDP, which will be used in the proofs of our theorems.

**Lemma 2.1.8** ([52]). Suppose that two mechanisms satisfy  $\rho_1$ -zCDP and  $\rho_2$ -zCDP, respectively. Then, their composition is  $(\rho_1 + \rho_2)$ -zCDP.

---

<sup>1</sup>For two distributions  $P$  and  $Q$  on  $\Omega$  and  $\alpha \in (1, \infty)$ , the  $\alpha$ -Rényi Divergence between  $P, Q$  is defined as  $D_\alpha(P || Q) = \frac{1}{\alpha-1} \log \int_{\Omega} P(x)^\alpha Q(x)^{1-\alpha} dx$ .

**Lemma 2.1.9** ([52]). For a Gaussian mechanism  $q(D) + Y$  with  $Y \sim \mathcal{N}(0, \sigma^2 I_d)$ , it satisfies  $(\frac{\Delta_2(q)}{2\sigma^2})$ -zCDP.

**Lemma 2.1.10** ([52]). If a mechanism is  $\rho$ -zCDP, then it is  $(\rho + 2\sqrt{\rho \log \frac{1}{\delta}}, \delta)$ -DP for any  $\delta > 0$ .

## 2.2 Local Differential Privacy

Instead of the trusted curator, in Local Differential Privacy model, each data provider perturb his/her private data record locally via some differentially private mechanisms before sending it to the curator.

Since we will consider the sequential interactive and non-interactive local models in this dissertation, we follow the definitions in [97].

We assume that  $\{Z_i\}_{i=1}^n$  are the private observations transformed from  $\{X_i\}_{i=1}^n$  through some privacy mechanisms. We say that the mechanism is **sequentially interactive**, when it has the following conditional independence structure:

$$\{X_i, Z_1, \dots, Z_{i-1}\} \mapsto Z_i, \text{ and } Z_i \text{ is independent with } X_j \mid \{X_i, Z_1, \dots, Z_{i-1}\}$$

for all  $j \neq i$  and  $i \in [n]$ . The full conditional distribution can be specified in terms of conditionals  $Q_i(Z_i \mid X_i = x_i, Z_{1:i-1} = z_{1:i-1})$ . The full privacy mechanism can be specified by a collection  $Q = \{Q_i\}_{i=1}^n$ .

When  $Z_i$  is depending only on  $X_i$ , the mechanism is called **non-interactive** and in this case we have a simpler form for the conditional distributions  $Q_i(Z_i \mid X_i = x_i)$ . We now define local differential privacy by restricting the conditional distribution  $Q_i$ .

**Definition 2.2.1** ([97]). For given privacy parameters  $\epsilon > 0, \delta \geq 0$ , the random variable  $Z_i$  is an  $(\epsilon, \delta)$  sequentially locally differentially private view of  $X_i$  if for all  $z_1, z_2, \dots, z_{i-1}$

and  $x, x' \in \mathcal{X}$  we have the following for all the events  $S$ :

$$Q_i(Z_i \in S \mid X_i = x_i, Z_{1:i-1} = z_{1:i-1}) \leq e^\epsilon Q_i(Z_i \in S \mid X_i = x'_i, Z_{1:i-1} = z_{1:i-1}) + \delta.$$

If  $\delta = 0$ , we will omit the term of  $\delta$  (the same for other definitions).

We say that the random variable  $Z_i$  is an  $(\epsilon, \delta)$  non-interactively locally differentially private view of  $X_i$  if

$$Q_i(Z_i \in S \mid X_i = x_i) \leq e^\epsilon Q_i(Z_i \in S \mid X_i = x'_i) + \delta.$$

We say that the privacy mechanism  $Q = \{Q_i\}_{i=1}^n$  is  $(\epsilon, \delta)$ -sequentially (non-interactively) locally differentially private (LDP) if each  $Z_i$  is a sequentially (non-interactively) locally differentially private view.

Note that the LDP can be regarded as a special case of traditional DP where each dataset only contains one tuple. Thus, for the same privacy parameter  $\epsilon$ , LDP provides a stronger guarantee than DP.

Since all of our lower bounds are in the form of private minimax risk, we first introduce the classical statistical minimax risk before discussing the locally private version.

Let  $\mathcal{P}$  be a class of distributions over a data universe  $\mathcal{X}$ . For each distribution  $p \in \mathcal{P}$ , there is a deterministic function  $\theta(p) \in \Theta$ , where  $\Theta$  is the parameter space. Let  $\rho : \Theta \times \Theta \rightarrow \mathbb{R}_+$  be a semi-metric function on the space  $\Theta$  and  $\Phi : \mathbb{R}_+ \mapsto \mathbb{R}_+$  be a non-decreasing function with  $\Phi(0) = 0$  (in this paper, we assume that  $\rho(x, y) = |x - y|$  and  $\Phi(x) = x^2$  unless specified otherwise). We further assume that  $\{X_i\}_{i=1}^n$  are  $n$  i.i.d observations drawn according to some distribution  $p \in \mathcal{P}$ , and  $\hat{\theta} : \mathcal{X}^n \mapsto \Theta$  be some estimator. Then the minimax risk in metric  $\Phi \circ \rho$  is defined by the following saddle point problem:

$$\mathcal{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) := \inf_{\hat{\theta}} \sup_{p \in \mathcal{P}} \mathbb{E}_p[\Phi(\rho(\hat{\theta}(X_1, \dots, X_n), \theta(p))],$$

where the supremum is taken over distributions  $p \in \mathcal{P}$  and the infimum over all estimators  $\hat{\theta}$ .

### Private Minimax Risk

For a given privacy parameter  $\epsilon > 0$ , let  $\mathcal{Q}_\epsilon$  be the set of conditional distributions that have the  $\epsilon$ -LDP property. For a given set of samples  $\{X_i\}_{i=1}^n$ , let  $\{Z_i\}_{i=1}^n$  be the set of observations produced by any distribution  $Q \in \mathcal{Q}_\epsilon$ . Then, our estimator will be based on  $\{Z_i\}_{i=1}^n$ , that is,  $\hat{\theta}(Z_1, \dots, Z_n)$ . This yields a modified version of the minimax risk:

$$\mathcal{M}_n(\theta(\mathcal{P}), \Phi \circ \rho, Q) := \inf_{\hat{\theta}} \sup_{p \in \mathcal{P}} \mathbb{E}_p[\Phi(\rho(\hat{\theta}(Z_1, \dots, Z_n)), \theta(p))].$$

From the above definition, it is natural for us to seek the mechanism  $Q \in \mathcal{Q}_\epsilon$  that has the smallest value for the minimax risk. This allows us to define functions that characterize the optimal rate of estimation in terms of privacy parameter  $\epsilon$ .

**Definition 2.2.2.** Given a family of distributions  $\theta(\mathcal{P})$  and a privacy parameter  $\epsilon > 0$ , the  $\epsilon$  sequential private minimax risk in the metric  $\Phi \circ \rho$  is:

$$\mathcal{M}_n^{\text{Int}}(\theta(\mathcal{P}), \Phi \circ \rho, \epsilon) := \inf_{Q \in \mathcal{Q}_\epsilon} \mathcal{M}_n(\theta(\mathcal{P}), \Phi \circ \rho, Q),$$

where  $\mathcal{Q}_\epsilon$  is the set of all  $\epsilon$  sequentially locally differentially private mechanisms. Moreover, the  $\epsilon$  non-interactive private minimax risk in the metric  $\Phi \circ \rho$  is:

$$\mathcal{M}_n^{\text{Nint}}(\theta(\mathcal{P}), \Phi \circ \rho, \epsilon) := \inf_{Q \in \mathcal{Q}_\epsilon} \mathcal{M}_n(\theta(\mathcal{P}), \Phi \circ \rho, Q),$$

where  $\mathcal{Q}_\epsilon$  is the set of all  $\epsilon$  non-interactively locally differentially private mechanisms.

# Chapter 3

## Empirical Risk Minimization with Convex Loss Functions in Differential Privacy Model

Empirical Risk Minimization (ERM) is one of the most fundamental problem in supervised learning which encompasses a large family of classical models such as linear regression, LASSO, ridge regression, SVM, logistic regression, sigmoid regression, and neural networks. Due to its importance, its differentially private version ( called DP-ERM) has become one of the core problems in both machine learning and differential privacy communities [66]. In this chapter, we will revisit the classical setting DP-ERM with convex loss functions and its stochastic version, *i.e.*, DP Stochastic Convex Optimization (DP-SCO) in the central DP model. Specifically, in Chapter 3.1, we will study DP-ERM from optimization perspective. Particularly, we will focus on designing faster algorithm to achieve (near) optimal error under different settings. In Chapter 3.2, we will study DP-SCO in the setting where the dataset may follows some heavy-tailed distribution. Finally, in Chapter 3.3 we will generalize DP-ERM with pointwise loss functions to pairwise loss functions. To make each chapter independent and self-contained, we will review the definition of DP-ERM in each Chapter.

We also note that the notations of loss function, constraint set and parameter space may be different across different chapters. We first review some definitions in convex optimization.

## Convex Optimization and Convex Geometry

**Definition 3.0.1** (Lipschitz Function). A loss function  $f : \mathcal{C} \times \mathcal{X} \rightarrow \mathbb{R}$  is G-Lipschitz (under  $\ell_2$ -norm) over  $\theta$ , if for any  $z \in \mathcal{X}$  and  $\theta_1, \theta_2 \in \mathcal{C}$ , we have  $|f(\theta_1, z) - f(\theta_2, z)| \leq G\|\theta_1 - \theta_2\|_2$ .

**Definition 3.0.2** (L-smooth Function). A loss function  $f : \mathcal{C} \times \mathcal{X} \rightarrow \mathbb{R}$  is L-smooth over  $\theta$  with respect to the norm  $\|\cdot\|$  if for any  $z \in \mathcal{X}$  and  $\theta_1, \theta_2 \in \mathcal{C}$ , we have

$$\|\nabla f(\theta_1, z) - \nabla f(\theta_2, z)\|_* \leq L\|\theta_1 - \theta_2\|,$$

where  $\|\cdot\|_*$  is the dual norm of  $\|\cdot\|$ . If  $f$  is differentiable, this yields

$$f(\theta_1, z) \leq f(\theta_2, z) + \langle \nabla f(\theta_2, z), \theta_1 - \theta_2 \rangle + \frac{L}{2}\|\theta_1 - \theta_2\|^2.$$

**Definition 3.0.3** (Strongly Convex). The loss function  $f(x)$  is  $\mu$ -strongly convex with respect to norm  $\|\cdot\|$  if for any  $x, y \in \text{dom}(f)$  and  $z \in \mathcal{X}$ , there exists  $\mu > 0$  such that

$$f(\theta_1, z) \geq f(\theta_2, z) + \langle \partial f(\theta_2, z), \theta_1 - \theta_2 \rangle + \frac{\mu}{2}\|\theta_1 - \theta_2\|^2,$$

where  $\partial f(\theta_2, z)$  is any subgradient on  $\theta_2$  of  $f(\cdot, z)$ .

Next, we define the gradient complexity of finite sum function (3.1). Before that, we first let IFO denote incremental first-order oracle, which is widely used in optimization theory [4].

## 3.1 Faster Algorithms of DP-ERM under the Classical Setting

We will start from the most classical setting of DP-ERM. As we mentioned previously, ERM is the most important model in Supervised Machine Learning, thus, DP-ERM is also one core problem in the Differentially Private Machine Learning coomunity. It can be formally defined as follows.

**Definition 3.1.1** (DP-ERM). Given a dataset  $D = \{z_1, \dots, z_n\}$  from a data universe  $\mathcal{X}$  and a closed convex set  $\mathcal{C} \subseteq \mathbb{R}^p$ , DP-ERM is to find  $x^{\text{priv}} \in \mathcal{C}$  so as to minimize the empirical risk, *i.e.*

$$F^r(x, D) = \frac{1}{n} \sum_{i=1}^n f(x, z_i) + r(x), \quad (3.1)$$

with the guarantee of being differentially private, where  $f$  is the loss function and  $r$  is some simple (non-)smooth convex function called **regularizer**<sup>1</sup>. When the inputs are drawn i.i.d from an unknown underlying distribution  $\mathcal{P}$  on  $\mathcal{X}$ , we also consider the population risk  $\mathbb{E}_{z \sim \mathcal{P}}[f(x, z)]$ . If the loss function is convex, the utility of the algorithm is measured by the expected excess empirical risk, that is

$$\mathbb{E}_{\mathcal{A}}[F^r(x^{\text{priv}}, D)] - \min_{x \in \mathcal{C}} F^r(x, D),$$

or the expected excess population risk (generalization error), that is

$$\mathbb{E}_{z \sim \mathcal{P}, \mathcal{A}}[f(x^{\text{priv}}, z)] - \min_{x \in \mathcal{C}} \mathbb{E}_{z \sim \mathcal{P}}[f(x, z)],$$

where the expectation of  $\mathcal{A}$  is taking over all the randomness of the algorithm.

Due to its importance, DP-ERM has received a great deal of attentions in recent years. Most of them have been focused on convex loss functions. A number of approaches have

---

<sup>1</sup>If there is no regularizer, we will simply denote the ERM as  $F(x, D)$ .

been proposed for DP-ERM with convex loss functions, which can be roughly classified into three categories. The first type of approaches is to perturb the output of a non-DP algorithm. [66] first proposed the output perturbation approach which is extended by [356]. The second type of approaches is to perturb the objective function [66]. We referred to it as objective perturbation approach. The third type of approaches is to perturb gradients in first order optimization algorithms. [29] proposed the gradient perturbation approach and gave a lower bound on the utility of both general convex and strongly convex loss functions. Later, [269] showed that this bound can actually be broken by adding more restrictions on the convex domain  $\mathcal{C}$  of the problem. As shown in Tables 3.1, 3.2 and 3.3<sup>2</sup>, the output perturbation approach [356] can achieve the optimal bound of utility for strongly convex case, but cannot be generalized to the case with non-smooth regularizer; also, the gradient complexity of this approach is often too high, making it impractical. [159] extends the output perturbation approach in [66]. However, their method is only applicable to the unconstrained case and is not robust to the case with non-smooth regularizer. The objective perturbation approach needs to obtain the optimal solution to ensure both differential privacy and utility, which is often intractable in practice, and cannot achieve the optimal bound [159]. The gradient perturbation approach can overcome all the issues and thus is preferred in practice. However, its current results are all based on Gradient Descent (GD) or Stochastic Gradient Descent (SGD), which could be slow for large datasets. In this section, we will focus on the gradient perturbation based approach. Specifically, we will focus on how to design faster algorithms in both theory and practice, while also could achieve (near) optimal (expected) excess empirical or population risk.

Below is a summary of our results on DP-ERM with convex loss functions.

1. For strongly convex loss functions, we first propose a differentially private version of SVRG [166], *i.e.*, DP-SVRG, and show that it could achieve a near optimal error bound with less gradient complexity, meaning that it runs much faster than the previous

---

<sup>2</sup>Bound and complexity ignore multiplicative dependence on  $\log(1/\delta)$ .

ones to achieve the near optimal bound. Moreover, combining with the Katyusha momentum [8] and Variance Reduction methods, we introduce an accelerated version of DP-SVRG, *i.e.*, DP-Katyusha, and show that it can further reduce the gradient complexity while achieving a near optimal upper bound (see Table 3.1 for details).

2. For general convex loss functions, we also propose a DP version of variance reduction method, *i.e.*, *DP-SVRG++*, which is a DP version of SVRG++ [9]. We show that our method can achieve the optimal error bound with significantly less gradient complexity compared to previous ones. See Table 3.2 for details.
3. In high dimensions, for smooth and convex loss functions, we propose an algorithm called DP-AccMD, which is motivated by the Nesterov’s accelerated version of Mirror Descent. Our algorithm has significantly less gradient complexity than the previous one to achieve an upper bound of error which depends only on the Gaussian width of the underlying constraint set. More details are in Table 3.3.

### 3.1.1 Related Work

There is a long list of works on differentially private ERM in the last decade which attack the problem from different perspectives, such as [160, 276, 336, 305, 299]. We compare to those that are most related to ours from the utility and gradient complexity (*i.e.*, the number (complexity) of times that the first order oracle  $(f(x, z_i), \nabla f(x, z_i))$  is called) points of view. **Table 3.1** is the comparisons for the case that the loss function is strongly convex and 1-smooth. Our algorithm achieves near optimal bound with less gradient complexity compared to previous ones. It is also robust against non-smooth regularizers.

**Tables 3.2 and 3.3** show that for non-strongly convex loss functions and in high dimensional space, our algorithms outperform other existing methods. Particularly, we improve the gradient complexity from  $O(n^2)$  to  $O(n \log n)$  while preserving the optimal bound for non-strongly convex case. For the high dimensional case, the gradient complexity of

	Method	Utility Upper Bd	Gradient Complexity	Non smooth Regularizer?
[67][66]	Objective Perturbation	$O(\frac{p}{n^2\epsilon^2})$	N/A	No
[181]	Objective Perturbation	$O(\frac{p}{n^2\epsilon^2} + \frac{\lambda\ x_*\ ^2}{n\epsilon})$	N/A	Yes
[29]	Gradient Perturbation	$O(\frac{p\log^2(n)}{n^2\epsilon^2})$	$O(n^2)$	Yes
[356]	Output Perturbation	$O(\frac{p}{n^2\epsilon^2})$	$O(n\kappa\log(\frac{n\epsilon}{\kappa}))$	No
[159]	Output Perturbation	$O(\frac{p}{n^2\epsilon^2})$	N/A	No
<b>Algorithm 3.1.1</b>	Gradient Perturbation	$O(\frac{p\log(n)}{n^2\epsilon^2})$	$O((n + \kappa)\log(\frac{n\epsilon}{\sqrt{p}}))$	Yes
<b>Algorithm 3.1.2</b>	Gradient Perturbation	$O(\frac{p\log^2(n)}{n^2\epsilon^2})$	$O((n + \sqrt{\kappa n})\log(\frac{n\epsilon}{\sqrt{p}}))$	Yes

Table 3.1: Comparisons with previous  $(\epsilon, \delta)$ -DP algorithms. We assume that the loss function  $f$  is convex, 1-smooth, differentiable (twice differentiable for objective perturbation), and 1-Lipschitz.  $r(\cdot)$  is  $\mu$ -strongly convex. Bound and complexity ignore multiplicative dependence on  $\log(1/\delta)$ .  $\kappa = \frac{L}{\mu}$  is the condition number. The lower bound is  $\Omega(\min\{1, \frac{p}{n^2\epsilon^2}\})$  [29].

our method is reduced from  $O(n^3)$  to  $O(n^{1.5})$ . Note that [177] also considered the high dimensional case via a dimension reduction method. But their method requires the optimal value in the dimension-reduced space; in addition, they considered the loss functions under a more stricter condition than the  $\ell_2$ -norm Lipschitz requirement.

### 3.1.2 Preliminaries

**Definition 3.1.2.** Given some  $x \in \mathbb{R}^p$  and  $i \in [n]$ , the IFO returns a pair  $(f(x, z_i), \nabla f(x, z_i))$ . The gradient complexity of an algorithm is the complexity of IFO in the algorithm.

For convenience, we let  $F(x) = \frac{1}{n} \sum_{i=1}^n f(x, z_i)$  and  $F^r(x) = F^r(x, D)$ , and denote by  $x_* = \arg \min_{x \in \mathcal{C}} F^r(x)$ .

**Assumption 3.1.1.** The loss function  $f(\cdot, z)$  is assumed to be differentiable,  $L$ -smooth over  $x$  with respect to  $\ell_2$  norm and is  $G$ -Lipschitz over  $x$  with respect to  $\ell_2$ -norm for all  $z \in \mathcal{X}$ .

The following definitions and lemmas will be used in the high dimensional case,

	Method	Utility Upper Bd	Gradient Complexity	Non smooth Regularizer?
[181]	Objective Perturbation	$O\left(\frac{\sqrt{p}}{n\epsilon}\right)$	N/A	Yes
[29]	Gradient Perturbation	$O\left(\frac{\sqrt{p} \log^{3/2}(n)}{n\epsilon}\right)$	$O(n^2)$	Yes
[356]	Output Perturbation	$O\left(\frac{\sqrt{p}}{n\epsilon}^{\frac{2}{3}}\right)$	$O(n[\frac{n\epsilon}{d}]^{\frac{2}{3}})$	No
<b>Algorithm 3.1.3</b>	Gradient Perturbation	$O\left(\frac{\sqrt{p}}{n\epsilon}\right)$	$O\left(\frac{n\epsilon}{\sqrt{p}} + n \log\left(\frac{n\epsilon}{p}\right)\right)$	Yes

Table 3.2: Comparisons with previous  $(\epsilon, \delta)$ -DP algorithms, where  $F^r$  is not necessarily strongly convex. We assume that the loss function  $f$  is convex, 1-smooth, differentiable (twice differentiable for objective perturbation), and 1-Lipschitz. Bound and complexity ignore multiplicative dependence on  $\log(1/\delta)$ . The lower bound in this case is  $\Omega(\min\{1, \frac{\sqrt{p}}{n\epsilon}\})$  [29].

	Method	Utility Upper Bd	Gradient Complexity	Non-smooth Regularizer
[269]	Gradient Perturbation	$O\left(\frac{\sqrt{G_{\mathcal{C}}^2 + \ \mathcal{C}\ ^2} \log(n)}{n\epsilon}\right)$	$O\left(\frac{n^3 \epsilon^2}{(G_{\mathcal{C}}^2 + \ \mathcal{C}\ ^2) \log^2(n)}\right)$	Yes
[269]	Objective Perturbation	$O\left(\frac{G_{\mathcal{C}} + \lambda \ \mathcal{C}\ ^2}{n\epsilon}\right)$	N/A	No
[270]	Gradient Perturbation	$O\left(\frac{(G_{\mathcal{C}}^{\frac{2}{3}} \log^2(n))}{(n\epsilon)^{\frac{2}{3}}}\right)$	$O\left(\frac{(n\epsilon)^{\frac{2}{3}}}{G_{\mathcal{C}}^{\frac{2}{3}}}\right)$	Yes
<b>Algorithm 3.1.4</b>	Gradient Perturbation	$O\left(\frac{\sqrt{G_{\mathcal{C}}^2 + \ \mathcal{C}\ ^2}}{n\epsilon}\right)$	$O\left(\frac{n^{1.5} \sqrt{\epsilon}}{(G_{\mathcal{C}}^2 + \ \mathcal{C}\ ^2)^{\frac{1}{4}}}\right)$	No

Table 3.3: Comparisons with previous  $(\epsilon, \delta)$ -DP algorithms. We assume that the loss function  $f$  is convex, 1-smooth, differentiable (twice differentiable for objective perturbation), and 1-Lipschitz. The utility bound depends on  $G_{\mathcal{C}}$ , which is the Gaussian width of  $\mathcal{C}$ . Bound and complexity ignore multiplicative dependence on  $\log(1/\delta)$ .

**Definition 3.1.3** (Minkowski Norm). The Minkowski norm (denoted by  $\|\cdot\|_{\mathcal{C}}$ ) with respect to a centrally symmetric convex set  $\mathcal{C} \subseteq \mathbb{R}^p$  is defined as follows. For any vector  $v \in \mathbb{R}^p$ ,  $\|\cdot\|_{\mathcal{C}} = \min\{r \in \mathbb{R}^+ : v \in r\mathcal{C}\}$ . The dual norm of  $\|\cdot\|_{\mathcal{C}}$  is denoted as  $\|\cdot\|_{\mathcal{C}^*}$ ; for any vector  $v \in \mathbb{R}^p$ ,  $\|v\|_{\mathcal{C}^*} = \max_{w \in \mathcal{C}} |\langle w, v \rangle|$ .

**Definition 3.1.4** (Gaussian Width). Let  $b \sim \mathcal{N}(0, I_p)$  be a Gaussian random vector in  $\mathbb{R}^p$ . The Gaussian width for a set  $\mathcal{C}$  is defined as  $G_{\mathcal{C}} = \mathbb{E}_b[\sup_{w \in \mathcal{C}} \langle b, w \rangle]$ .

Compared with the dimensionality  $p$ , Gaussian Width of a convex set  $\mathcal{C} \subseteq \mathbb{R}^p$  could be much smaller. For example, when  $\mathcal{C}$  is  $\ell_1$ -norm unit ball,  $G_{\mathcal{C}} = O(\sqrt{\log p})$ ; when  $\mathcal{C}$  is the set of all unit  $s$ -sparse vectors on  $\mathbb{R}^p$ ,  $G_{\mathcal{C}} = O(\sqrt{s \log(p/s)})$ .

### 3.1.3 Low Dimensional Case

Since the constraint set can be represented as a indication function, in this section we will consider ERM with (non)-smooth regularizer<sup>3</sup>, i.e.

$$\min_{x \in \mathbb{R}^p} F^r(x, D) = F(x, D) + r(x) = \frac{1}{n} \sum_{i=1}^n f(x, z_i) + r(x). \quad (3.2)$$

The loss function  $f$  is convex for every  $z$ . We define the proximal operator as

$$\text{prox}_r(y) = \arg \min_{x \in \mathbb{R}^p} \left\{ \frac{1}{2} \|x - y\|_2^2 + r(x) \right\}.$$

Note that for many specified non-smooth regularizer  $r(\cdot)$ , such as  $\ell_1$ -norm or elastic net, there are efficient or closed forms of solution for the operator.

Before showing our algorithm, we first introduce SVGR. SVRG is a general technique called variance reduction method, which has been studied considerably in recent years [8, 9, 166, 347]. All these results have showed that SGD converges faster if one makes a better choice of the gradient estimator  $v_t$  so that its variance reduces as  $k$  increases. In

---

<sup>3</sup>All the algorithms and theorems in this section are applicable to closed convex set  $\mathcal{C}$  rather than  $\mathbb{R}^p$ .

SVRG [166], the estimator behaves as follows. It first keeps a snapshot vector  $\tilde{x}$  that is updated every  $m$  iterations, and computes the full gradient  $\nabla F(\tilde{x})$ . Then it sets  $v_t = \nabla f(x_t, z_{i_t}) - \nabla f(\tilde{x}, z_{i_t}) + \nabla F(\tilde{x}, D)$  as an unbiased estimator of the gradient  $\nabla F(x_t, D)$ . After that, it updates  $x_t$ , using the gradient descent on  $v_t$ , *i.e.*,  $x_{t+1} = x_t - \eta_t v_t$ , where  $\eta_t$  is the step size.

The basic idea of our algorithms is to inject Gaussian noise to this unbiased estimator  $v_t$ ; it can be shown that the estimator is still unbiased after injecting noise, which means that it attains all the advantages (*i.e.*, faster convergence) of the original one except for some slightly increased variance.

---

**Algorithm 3.1.1 DP-SVRG**


---

**Input:**  $f(x, z)$  is G-Lipschitz and L-smooth.  $r(\cdot)$  is  $\mu$ -strongly convex w.r.t  $\ell_2$ -norm.  $\tilde{x}_0$  is the initial point,  $\eta$  is the step size, and  $T, m$  are the iteration numbers.

```

1: for  $s = 1, 2, \dots, T$  do
2:    $\tilde{x} = \tilde{x}_{s-1}$ 
3:    $\tilde{v} = \nabla F(\tilde{x})$ 
4:    $x_0^s = \tilde{x}$ 
5:   for  $t = 1, 2, \dots, m$  do
6:     Pick  $i_t^s \in [n]$ 
7:      $v_t^s = \nabla f(x_{t-1}^s, z_{i_t^s}) - \nabla f(\tilde{x}, z_{i_t^s}) + \tilde{v} + u_t^s$ , where  $u_t^s \sim \mathcal{N}(0, \sigma^2 I_p)$ 
8:      $x_t^s = \text{prox}_{\eta r}(x_{t-1}^s - \eta v_t^s)$ 
9:   end for
10:   $\tilde{x}_s = \frac{1}{m} \sum_{k=1}^m x_k^s$ 
11: end for
12: return  $\tilde{x}_T$ 

```

---

### Strongly convex case

We first consider the case that  $F^r(x, D)$  is  $\mu$ -strongly convex. As mentioned earlier, **Algorithm 3.1.1** is based on the Prox-SVRG [347], which is much faster than SGD or GD. We will show that DP-SVRG is also faster than DP-SGD or DP-GD in terms of the gradient complexity needed to achieve a near optimal excess empirical risk bound.

**Theorem 3.1.1. DP-SVRG** (Algorithm 3.1.1) is  $(\epsilon, \delta)$ -differentially private, where  $0 < \epsilon \leq c_1 \frac{Tm}{n^2}$  for some constant  $c_1$  and  $\delta > 0$  is a constant, if the following condition holds for some

constant  $c$

$$\sigma^2 = c \frac{G^2 T m \ln(\frac{1}{\delta})}{n^2 \epsilon^2}. \quad (3.3)$$

**Remark 3.1.1.** The constraint on  $\epsilon$  in Theorems 3.1.1 and 3.1.4 comes from Lemma 2.1.7. This constraint can be removed if the noise  $\sigma$  is amplified by a factor of  $O(\ln(T/\delta))$  in (3.3) and (3.7). But accordingly, there will be a factor of  $\tilde{O}(\log(\frac{Tm}{\delta}))$  in the utility bound in (3.4), (3.6) and (3.8). In this case, the differential privacy guarantee is achieved by advanced composition theorem and privacy amplification via sampling [29].

The following theorem shows that the output in Algorithm 3.1.1 achieves a near optimal error bound.

**Theorem 3.1.2.** Under Assumption 3.1.1 and further assuming that  $r(x)$  is  $\mu$ -strongly convex w.r.t  $\ell_2$ -norm, the output of **DP-SVRG** (Algorithm 3.1.1) has the following error bound after  $T = O\left(\log\left(\frac{n^2\epsilon^2\mu}{pG^2\ln(1/\delta)}\right)\right)$  iterations

$$\mathbb{E}[F^r(\tilde{x}_T)] - F^r(x_*) \leq \tilde{O}\left(\frac{p \log(n) G^2 \log(1/\delta)}{n^2 \epsilon^2 \mu}\right), \quad (3.4)$$

if  $\sigma$  is chosen as in (3.3),  $\eta$  is set as  $\eta = \Theta(\frac{1}{L}) \leq \frac{1}{12L}$ , and  $m = \Theta(\frac{L}{\mu})$  is sufficiently large so that they satisfy inequality

$$\frac{1}{\eta(1 - 8\eta L)\mu m} + \frac{8L\eta(m+1)}{m(1 - 8L\eta)} < \frac{1}{2}, \quad (3.5)$$

where some insignificant logarithmic terms are hiding in the  $\tilde{O}$ -notation. The total gradient complexity is  $O\left((n + \frac{L}{\mu}) \log \frac{ne}{\sqrt{p}}\right)$ .

From Table 3.1, we can see that the gradient complexity of DP-SGD is  $O(n^2)$ , which means that our method is much faster when  $\frac{L}{\mu} \ll n$ . We will verify this in the experimental section.

With the above theorem, a natural question is whether we can further reduce the gradient complexity. Recently, [8] proposed the Katyusha technique to accelerate the stochastic vari-

ance reduced methods and achieved the best-known gradient complexity  $O((n + \sqrt{\frac{L}{\mu} n}) \log \frac{1}{\epsilon})$  for strongly convex loss functions. Combining this technique with Algorithm 3.1.1, we can obtain a differentially private version of Katyusha, DP-Katyusha, and show that it can indeed improve the gradient complexity in Theorem 3.1.2. See Algorithm 3.1.2 for details.

---

**Algorithm 3.1.2 DP-Katyusha**


---

**Input:**  $f(x, z)$  is G-Lipschitz and L-smooth.  $r(x)$  is  $\mu$ -strongly convex w.r.t  $\ell_2$ -norm.  $x_0$  is the initial point,  $\eta$  is the step size, and  $T, m$  are the iteration numbers. Parameter  $\theta$

```

1: Let  $\tilde{x}_0 = x_0^1 = x_0$ ,  $w = 1 + \eta\mu$ .
2: for  $s = 1, 2, \dots, T$  do
3:    $\tilde{v} = \nabla F(\tilde{x}_{s-1})$ .
4:   for  $t = 1, 2, \dots, m$  do
5:     Pick  $i_t^s \in [n]$  uniformly.
6:     Let  $y_{t-1} = \theta x_{t-1}^s + (1 - \theta)\tilde{x}_{s-1}$ 
7:      $v_t^s = \nabla f(y_{t-1}, z_{i_t^s}) - \nabla f(\tilde{x}_{s-1}, z_{i_t^s}) + \tilde{v}$ .
8:      $x_t^s = \text{prox}_{\eta r}(x_{t-1}^s - \eta v_t^s)$ .
9:   end for
10:  Let  $\tilde{x}_s = \theta(\sum_{j=0}^{m-1} w^j)^{-1} \sum_{j=0}^{m-1} w^j x_{j+1}^s + (1 - \theta)\tilde{x}_{s-1}$ .
11:   $x_0^{s+1} = x_m^s$ 
12: end for return  $\tilde{x}_T$ .
```

---

**Theorem 3.1.3.** Under Assumption 3.1.1 and taking  $\sigma$  as in Theorem 3.1.1, Algorithm 3.1.2 is  $(\epsilon, \delta)$ -DP. Furthermore if  $r(x)$  is  $\mu$ -strongly convex w.r.t  $\ell_2$ -norm and the parameters are chosen in the following way: 1) if  $n \leq \frac{L}{\mu}$ , set  $m = \frac{3}{4}n$ ,  $\eta = \sqrt{\frac{1}{3\mu m L}}$ ,  $\theta = \sqrt{\frac{m\mu}{3L}}$  and  $T = O(\sqrt{\frac{L}{\mu n}} \log \frac{n\epsilon}{\sqrt{p}})$ ; 2) if  $n \geq \frac{L}{\mu}$ , set  $m = \frac{3}{4}\frac{L}{\mu}$ ,  $\eta = \frac{2}{3L}$ ,  $\theta = \frac{1}{2}$ ,  $T = O(\log(\frac{n\epsilon}{\sqrt{p}}))$ , then the output of Algorithm 3.1.2 has the following error bound

$$\mathbb{E}[F^r(\tilde{x}_T)] - F^r(x_*) \leq \tilde{O}\left(\frac{p \log^2 n G^2 \log \frac{1}{\delta}}{n^2 \epsilon^2 \mu}\right). \quad (3.6)$$

In other words, the overall gradient complexity of DP-Katyusha is  $O\left((n + \sqrt{n \frac{L}{\mu}}) \log \frac{n\epsilon}{\sqrt{p}}\right)$ .

Compared with the gradient complexity in Theorem 3.1.1, we can see that in the ill-conditioned problem where the condition number  $\frac{L}{\mu} \gg n$ , the gradient complexity in Theorem 3.1.3 is less.

**Remark 3.1.2.** Note that compared with DP-SVRG (Algorithm 3.1.1), DP-Katyusha (Algorithm 3.1.2) has additional variables  $y_t$ , which are linear combinations of  $x_t$  and  $\tilde{x}$ . Actually, this corresponds to a special case of Katyusha momentum [8], *i.e.*, the case of  $1 - \tau_1 - \tau_2 = 0$ . We note that this special case has also been studied in [368]. We can easily see that the updating of  $\tilde{x}_s$  can be written as  $\tilde{x}_s = (\sum_{j=0}^{m-1} w^j)^{-1} \sum_{j=o}^{m-1} w^j y_{j+1}$ .

### Non-strongly convex case

In some cases,  $F^r(x)$  may not be strongly convex. For such cases, [10] has recently showed that SVRG++ has less gradient complexity than Accelerated Gradient Descent. Following the idea of DP-SVRG, we present algorithm DP-SVRG++ for the non-strongly convex case. Unlike the previous one, this algorithm can achieve the optimal utility bound.

Compared with DP-SVRG (Algorithm 3.1.1) and DP-Katyusha (Algorithm 3.1.2), there are some differences in DP-SVRG++. The first one is that the inner iteration number  $m_s$  is doubled when the outer loop iteration number  $s$  increases, while it is a fixed number in both DP-SVRG and DP-Katyusha. The second one is that the starting vector  $x_0^{s+1}$  in each epoch is the ending vector of the last epoch  $x_{m_s}^s$ , which is similar to the one in DP-Katyusha, but not the average as in DP-SVRG.

**Theorem 3.1.4. DP-SVRG++** (Algorithm 3.1.3) is  $(\epsilon, \delta)$ -differentially private, where  $0 < \epsilon \leq c_1 \frac{2^T m}{n^2}$  for some constant  $c_1$  and  $\delta > 0$  is a constant, if the following condition holds for some constant  $c$

$$\sigma^2 = c \frac{G^2 2^T m \ln(\frac{2}{\delta})}{n^2 \epsilon^2}. \quad (3.7)$$

**Theorem 3.1.5.** Under Assumption 3.1.1 and further assuming that  $F^r(x)$  is convex, the output of **DP-SVRG++** (Algorithm 3.1.3) has the following error bound after  $T = O\left(\log\left(\frac{n\epsilon}{G\sqrt{p}\sqrt{\log(1/\delta)}}\right)\right)$  iterations

$$\mathbb{E}[F^r(\tilde{x}_T)] - F^r(x_*) \leq O\left(\frac{G\sqrt{p\ln(1/\delta)}}{n\epsilon}\right), \quad (3.8)$$

---

**Algorithm 3.1.3 DP-SVRG++**


---

**Input:**  $f(x, z)$  is G-Lipschitz, and L-smooth over  $x \in \mathcal{C}$ .  $\tilde{x}_0$  is the initial point,  $\eta$  is the step size, and  $T, m$  are the iteration numbers.

```

1:  $x_0^1 = \tilde{x}_0$ 
2: for  $s = 1, 2, \dots, T$  do
3:    $\tilde{v} = \nabla F(\tilde{x}_{s-1})$ 
4:    $m_s = 2^s m$ 
5:   for  $t = 1, 2, \dots, m_s$  do
6:     Pick  $i_t^s \in [n]$ 
7:      $v_t^s = \nabla f(x_{t-1}^s, z_{i_t^s}) - \nabla f(\tilde{x}_{s-1}, z_{i_t^s}) + \tilde{v} + u_t^s$ , where  $u_t^s \sim \mathcal{N}(0, \sigma^2 I_p)$ 
8:      $x_t^s = \text{prox}_{\eta r}(x_{t-1}^s - \eta v_t^s)$ 
9:   end for
10:   $\tilde{x}_s = \frac{1}{m_s} \sum_{k=1}^{m_s} x_k^s$ 
11:   $x_0^{s+1} = x_{m_s}^s$ 
12: end for
13: return  $\tilde{x}_T$ 

```

---

if  $\sigma$  is chosen as in (3.7),  $\eta = \frac{1}{13L}$ , and  $m = \Theta(L)$  is sufficiently large. The gradient complexity is  $O\left(\frac{nL\epsilon}{\sqrt{p}} + n \log\left(\frac{n\epsilon}{p}\right)\right)$ .

Note that only near optimal error bound has been achieved for strongly convex loss functions as shown in (3.4), while optimal bound has been obtained for general convex loss functions. It is not clear whether our method can achieve optimal error bound for strongly convex loss functions. Another problem is to determine whether the gradient complexity in Theorem 3.1.5 can be further improved by using DP-Katyusha. We leave both problems as future research.

### 3.1.4 High Dimensional Case

The utility bounds in Section 3.1.3 depend polynomially on the dimensionality  $p$ . In high-dimensional (i.e.,  $p \gg n$ ) space, such a dependence could be too large and thus not very desirable. To alleviate this issue, we can usually get rid of the dependence on dimensionality by reformulating the problem so that the goal is to find the parameter in some closed centrally symmetric convex set  $\mathcal{C} \subseteq \mathbb{R}^p$  (such as  $l_1$ -norm ball), i.e.,  $\min_{x \in \mathcal{C}} F(x, D) = \frac{1}{n} \sum_{i=1}^n f(x, z_i)$ , where the loss function is convex.

Recently, [269] and [270] showed that the  $\sqrt{p}$  term in (3.4), (3.6) and (3.8) can be replaced by the Gaussian Width of  $\mathcal{C}$ , which is no larger than  $O(\sqrt{p})$  and can be significantly smaller for many special cases, such as unit  $\ell_1$ -norm ball. However, one issue of their methods is that the gradient complexity to achieve this upper bound is  $\tilde{O}(n^3)$ , which is quite large. In this section, we propose a faster algorithm to achieve the same upper utility bound.

---

**Algorithm 3.1.4 DP-AccMD**

**Input:**  $f(x, z)$  is G-Lipschitz, and L-smooth over  $x \in \mathcal{C}$ .  $\|\mathcal{C}\|_2$  is the  $\ell_2$  norm diameter of the convex set  $\mathcal{C}$ .  $w$  is a function that is 1-strongly convex w.r.t  $\|\cdot\|_{\mathcal{C}}$ .  $x_0$  is the initial point, and  $T$  is the iteration number.

- 1: Define  $V(y, x) = w(y) - \langle \nabla w(x), y - x \rangle - w(x)$  as the Bregman divergence associated with  $w$ .
  - 2:  $y_0, z_0 = x_0$
  - 3: **for**  $k = 0, \dots, T - 1$  **do**
  - 4:      $\alpha_{k+1} = \frac{k+2}{4L}$  and  $r_k = \frac{1}{2\alpha_{k+1}L}$
  - 5:      $x_{k+1} = r_k z_k + (1 - r_k) y_k$
  - 6:      $y_{k+1} = \arg \min_{y \in \mathcal{C}} \left\{ \frac{L\|\mathcal{C}\|_2^2}{2} \|y - x_{k+1}\|_{\mathcal{C}}^2 + \langle \nabla F(x_{k+1}), y - x_{k+1} \rangle \right\}$
  - 7:
  - 8:      $z_{k+1} = \arg \min_{z \in \mathcal{C}} \{V(z, z_k) + \alpha_{k+1} \langle \nabla F(x_{k+1}) + b_{k+1}, z - z_k \rangle\}$ , where  $b_{k+1} \sim \mathcal{N}(0, \sigma^2 I_p)$
  - 9: **end for**
  - 10: **return**  $y_T$
- 

Our algorithm **DP-AccMD** is based on the Accelerated Mirror Descent method, which was studied in [9] and [229]. Since there is an additional noise injected to the gradient in this method, the parameters  $\alpha_k$  and  $r_k$  are quite different from the original one, which makes the proof much more challenging. Before showing our result, we first introduce the Bregman divergence.

**Definition 3.1.5.** A function  $w : \mathcal{C} \rightarrow \mathbb{R}$  is said to be a distance generating function with modulus  $\alpha > 0$  (w.r.t.  $\|\cdot\|$  norm), if  $w$  is continuously differentiable and strongly convex satisfying the following inequality for any  $x, z \in \mathcal{C}$ ,  $\langle x - z, \nabla w(x) - \nabla w(z) \rangle \geq \alpha \|x - z\|^2$ . The Bregman Divergence associated with  $w$  is defined as  $V(x, z) = w(x) - w(z) - \langle \nabla w(z), x - z \rangle$ .

**Theorem 3.1.6. DP-AccMD** (Algorithm 3.1.4) is  $(\epsilon, \delta)$ -differentially private for constants  $\epsilon, \delta > 0$ , if the following holds

$$\sigma^2 = c \frac{G^2 T \ln(1/\delta)}{n^2 \epsilon^2} \quad (3.9)$$

for some constant  $c$ .

**Theorem 3.1.7.** Under Assumption 3.1.1 and further assuming that the loss function is convex, the output of DP-AccMD (Algorithm 3.1.4) has the following error bound

$$\mathbb{E}[F(y_T)] - F(x_*) \leq O\left(\frac{\sqrt{G_{\mathcal{C}}^2 + \|\mathcal{C}\|_2^2} G \sqrt{\ln(1/\delta)}}{n\epsilon}\right)$$

after  $T$  iterations, where

$$T^2 = O\left(\frac{L \|\mathcal{C}\|_2^2 \sqrt{V(x_*, x_0)} n \epsilon}{G \sqrt{\ln(1/\delta)} \sqrt{G_{\mathcal{C}}^2 + \|\mathcal{C}\|_2^2}}\right),$$

if  $\sigma$  is chosen as in (3.9) and  $w$  is function that is 1-strongly convex with respect to  $\|\cdot\|_{\mathcal{C}}$ . The total gradient complexity is  $O\left(\frac{n^{1.5} \sqrt{\epsilon L}}{(G_{\mathcal{C}}^2 + \|\mathcal{C}\|_2^2)^{\frac{1}{4}}}\right)$ , where  $\|\mathcal{C}\|_2$  is the  $\ell_2$ -norm of the diameter of  $\mathcal{C}$ , i.e.,  $\|\mathcal{C}\|_2 = \max_{x,y \in \mathcal{C}} \|x - y\|_2$ .

We note that compared with DP-Mirror Descent in [269], our method can improve a factor of  $O(n^{1.5})$  in the gradient complexity. However, we need to assume that the loss function is  $L$ -smooth while in [269] it is required only to be convex.

A remaining issue in our algorithm is that in Steps 6 and 7 it needs to solve a sub-problem in each iteration. This could be costly for some general convex set  $\mathcal{C}$ . We leave it as an open problem for future research.

### 3.1.5 Experiments

In this section, we study the practical performance of some of our proposed algorithms on both synthetic and real-world datasets. As we will see later, all experimental results support

our theoretical analysis.

We will use logistic regression as an example to study the practical performance of our algorithms. Particularly, we test DP-SVRG (Algorithm 3.1.1) and DP-Katyusha (Algorithm 3.1.2) for logistic regression with  $\ell_2$ -norm regularizer and DP-SVRG++ (Algorithm 3.1.3) for logistic regression with  $\ell_1$ -norm regularizer:

$$\min_{\theta \in \mathbb{R}^p} F^r(\theta, D) = \frac{1}{n} \sum_{i=1}^n \log(1 + \exp(-y_i \langle x_i, \theta \rangle)) + \frac{\lambda}{2} \|\theta\|_2^2,$$

$$\min_{\theta \in \mathbb{R}^p} F^r(\theta, D) = \frac{1}{n} \sum_{i=1}^n \log(1 + \exp(-y_i \langle x_i, \theta \rangle)) + \frac{\lambda}{2} \|\theta\|_1,$$

where  $\{x_i\}_{i=1}^n$  are the feature vectors and  $\{y_i\}_{i=1}^n$  are the corresponding labels.

## Experimental Settings

For all the experiments, we set  $\lambda = 10^{-4}$ . We compare the optimality gap with the gradient complexity in different settings, where the optimal value is obtained through gradient descent. For both strongly and general convex cases, we compare our methods with DP-SGD [29] and DP-GD [356]. The synthetic dataset is generated by  $\Pr(y_i|x_i) = \frac{1}{1+\exp(-y_i \langle \theta^*, x_i \rangle)}$  for some  $\theta^*$ . That is, we first randomly choose  $\theta^*$ , and then for each random vector  $x_i$ , we set  $y_i = 1$  if  $\frac{1}{1+\exp(-y_i \langle \theta^*, x_i \rangle)} > \frac{1}{2}$ . The size of the synthetic dataset is  $(10^5, 50)$ . For the real-world datasets, we use Covertype and IJCNN, which are commonly used in binary classification. The sizes of the training sets are  $(5 \times 10^5, 54)$  and  $(5 \times 10^4, 22)$ , respectively. We normalize all the above datasets as pre-processing so that the loss functions are 1-Lipschitz.

## Parameter Settings

For the strongly convex case, the outer and inner iteration numbers are chosen to be  $2\lceil \log n \rceil$  and 100, respectively, in DP-SVRG, while they are set to be  $\lceil \log n \rceil$  and  $\frac{3}{4}n$ , respectively, in DP-Katyusha. Since the stepsize does not affect the privacy, we use the Barzilai-Borwein stepsize strategy to determine the stepsize in each iteration [271] for DP-SVRG and DP-

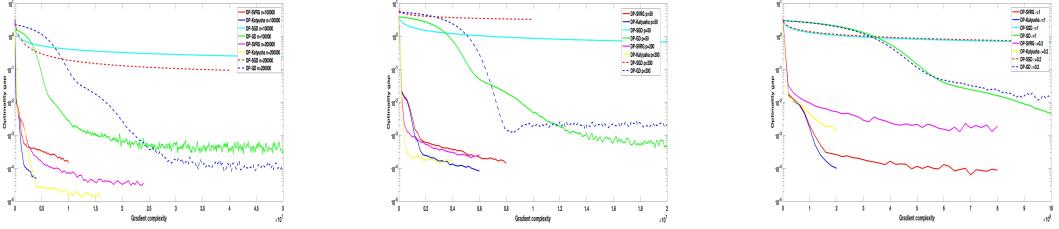


Figure 3.1: Experimental results on synthetic dataset for strongly convex case.

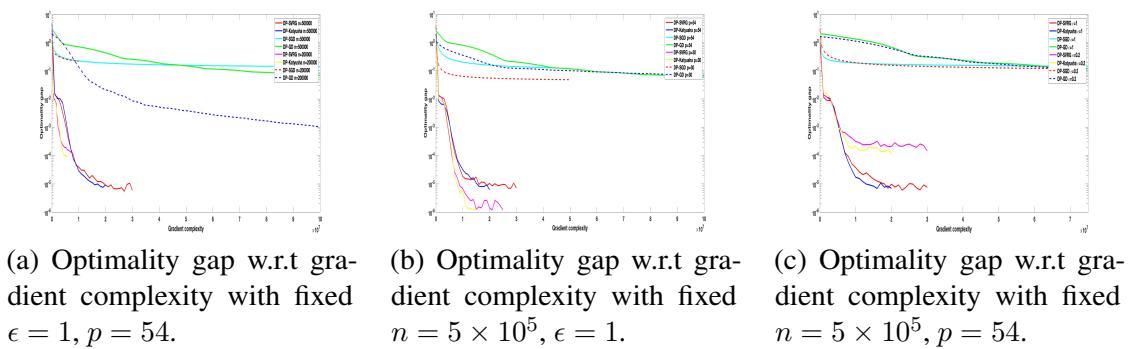


Figure 3.2: Experimental results on Covertype dataset for strongly convex case.

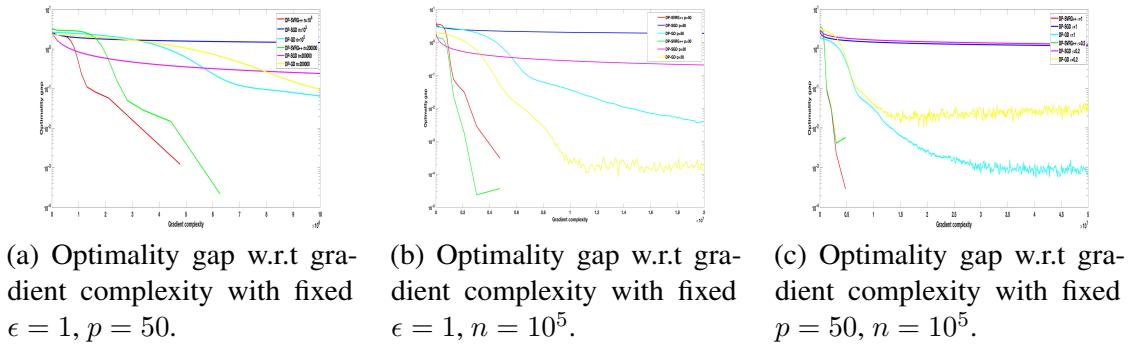
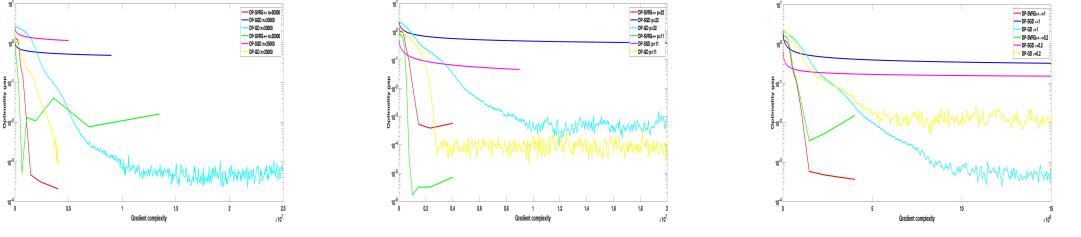


Figure 3.3: Experimental results on synthetic dataset for convex case.

Katyusha. The initial stepsize is  $\eta = 0.1$ . For the privacy parameters, we choose  $\epsilon = 0.2, 1$  and a fixed  $\delta = 10^{-4}$ . All experiments are performed on MATLAB.

## Results

Figure 3.1 and 3.2 show the results of  $\ell_2$ -norm regularized logistic regression on synthetic and Covertype dataset, respectively. Firstly, from the figures we can see that DP-Katyusha



(a) Optimality gap w.r.t gradient complexity with fixed  $\epsilon = 1, p = 22$ . (b) Optimality gap w.r.t gradient complexity with fixed  $\epsilon = 1, n = 5 \times 10^4$ . (c) Optimality gap w.r.t gradient complexity with fixed  $p = 22, n = 5 \times 10^4$ .

Figure 3.4: Experimental results on IJCNN dataset for convex case.

and DP-SVRG not only have lower gradient complexities in all cases than other existing methods, but also achieve the lowest optimality gap. This suggests that our methods are more practical and effective, which is consistent with our theoretical analysis. Comparing DP-SVRG and DP-Katyusha, we can see that both of them can achieve almost the same optimality gap, but DP-Katyusha has considerably lower gradient complexity. Secondly, we can see that when the sample size becomes smaller, the optimality gap increases, which is mainly due to the added noise in each iteration. Thirdly, when the dimensionality increases, the optimality gap also increases. This is due to the fact that the error bound is linearly depending on the dimensionality. Finally, we know that when the privacy parameter  $\epsilon$  increases, which means less privacy, the optimality gap decreases.

Figure 3.3 and 3.4 depict the results of logistic regression on synthetic and IJCNN dataset, respectively. Firstly, we can see that in all the cases, DP-SVRG++ has significantly lower gradient complexity than other methods, and also achieves a comparable optimality gap with DP-GD. Secondly, when the sample size decreases, the optimality gap increases. Thirdly, when the dimensionality increases, the optimality gap also increases. Finally, we know that with a larger privacy parameter  $\epsilon$ , the optimality gap decreases.

### 3.1.6 Omitted Proofs

For simplicity, we omit the superscripts of iterations in the same epoch  $s$ , *i.e.* use  $x_t$  to denote  $x_t^s$ , unless otherwise specified.

## Useful Lemmas

**Lemma 3.1.1.** Suppose that each component function  $f(x, z_i)$  is  $L$ -smooth. Let  $v = \nabla f(y_{t-1}, z_{i_t}) - \nabla f(\tilde{x}, z_{i_t}) + \nabla F(\tilde{x}) + u_t$ , where  $u_t \sim \mathcal{N}(0, \sigma^2 I_p)$  is independent of  $i_t$ . Then, the following inequality holds

$$\mathbb{E}_{i_t, u_t} \|\nabla F(y_{t-1}) - v\|_2^2 \leq 2L(F(\tilde{x}) - F(y_{t-1}) - \langle \nabla F(y_{t-1}), \tilde{x} - y_{j-1} \rangle) + p\sigma^2,$$

where the expectation is taking over  $i_t$  and  $u$ .

*Proof.* Let  $\tilde{v} = \nabla f(y_{t-1}, z_{i_t}) - \nabla f(\tilde{x}, z_{i_t}) + \nabla F(\tilde{x})$ . Then,  $\mathbb{E}\langle \nabla F(y_{t-1}) - \tilde{v}, u_t \rangle = 0$ . For the term  $\mathbb{E}\|\nabla F(y_{t-1}) - v\|_2^2$ , by a tighter upper bound on the gradient estimator variance in [8], we have

$$\mathbb{E}_{i_t} \|\nabla F(y_{t-1}) - \tilde{v}\|_2^2 \leq 2L(F(\tilde{x}) - F(y_{t-1}) - \langle \nabla F(y_{t-1}), \tilde{x} - y_{j-1} \rangle).$$

Thus, we get the proof.  $\square$

**Lemma 3.1.2.** Assume that  $z^*$  is an optimal solution to the following problem

$$\min_x \frac{\gamma}{2} \|x - z_0\|^2 + \phi(x),$$

where  $\gamma > 0$ , and  $\phi(x)$  is a convex function (possibly non-differentiable). Then for all  $z \in \mathbb{R}^p$ , there exists a vector  $\mathcal{G} \in \partial\phi(z^*)$  with

$$\langle \mathcal{G}, z - z^* \rangle = \frac{\gamma}{2} \|z_0 - z^*\|^2 - \frac{\gamma}{2} \|z - z_0\|^2 + \frac{\gamma}{2} \|z - z^*\|^2.$$

*Proof.* By the optimality of  $z^*$ , there exists a vector  $\mathcal{G} \in \partial\phi(z^*)$  which satisfies

$$\gamma(z^* - z_0) + \mathcal{G} = \mathbf{0}.$$

Thus, for all  $\in \mathbb{R}^p$ , we have

$$\begin{aligned}
0 &= \langle \gamma(z^* - z_0) + \mathcal{G}, z^* - z \rangle \\
&= \gamma \langle z^* - z_0, z^* - z \rangle + \langle \mathcal{G}, z^* - z \rangle \\
&= \frac{\gamma}{2} \|z^* - z_0\|^2 - \frac{\gamma}{2} \|z - z_0\|^2 + \frac{\gamma}{2} \|z - z^*\|^2 + \langle \mathcal{G}, z^* - z \rangle.
\end{aligned}$$

□

**Lemma 3.1.3.** If two vectors  $x_j, x_{j-1} \in \mathbb{R}^p$  satisfies  $x_j = \text{Prox}_{\eta r}(x_{j-1} - \eta v)$  with a constant vector  $v$  and a general convex function  $r(x)$ , then for all  $u \in \mathbb{R}^p$ , we have

$$\langle v, x_j - u \rangle \leq -\frac{1}{2\eta} \|x_{j-1} - x_j\|^2 + \frac{1}{2\eta} \|x_{j-1} - u\|^2 - \frac{1}{2\eta} \|x_j - u\|^2 + r(u) - r(x_j).$$

Moreover, if  $r(x)$  is  $\mu$ -strongly convex, the above inequality becomes

$$\langle v, x_j - u \rangle \leq -\frac{1}{2\eta} \|x_{j-1} - x_j\|^2 + \frac{1}{2\eta} \|x_{j-1} - u\|^2 - \frac{1 + \eta\mu}{2\eta} \|x_j - u\|^2 + r(u) - r(x_j).$$

*Proof.* By the definition of the proximal operator  $\text{Prox}(\cdot)$ , we can see that  $x_j = \text{Prox}_{\eta r}(x_{j-1} - \eta v)$  is equivalent to

$$x_j = \arg \min_x \left\{ \frac{1}{2\eta} \|x - x_{j-1}\|^2 + \langle v, x \rangle + r(x) \right\}.$$

Applying Lemma 3.1.2 with  $z = u, z_0 = x_{j-1}, z^* = x_j, \gamma = \frac{1}{\eta}$  and  $\phi(x) = \langle v, x \rangle + r(x)$ , then there exists a vector  $\mathcal{G} \in \partial r(x)$  satisfying

$$\langle v, u - x_j \rangle + \langle \mathcal{G}, u - x_j \rangle = \frac{1}{2\eta} \|x_{j-1} - x_j\|^2 - \frac{1}{2\eta} \|x_{j-1} - u\|^2 + \frac{1}{2\eta} \|x_j - u\|^2.$$

Using the convexity of  $r(\cdot)$ , we get  $r(u) - r(x_j) \geq \langle \mathcal{G}, u - x_j \rangle$ . After rearranging, we have the first inequality.

If  $r(x)$  is  $\mu$ -strongly convex, we have  $r(u) - r(x_j) \geq \langle \mathcal{G}, u - x_j \rangle + \frac{\mu}{2} \|x_j - u\|^2$ .

□

**Lemma 3.1.4.** Let  $r$  be a closed convex function on  $\mathbb{R}^p$ . Then for any  $x, y \in \text{dom}(R)$

$$\|\text{prox}_r(x) - \text{prox}_r(y)\| \leq \|x - y\|.$$

**Lemma 3.1.5.** Let  $w$  be a distance generating function with modulus  $\alpha$  w.r.t.  $\|\cdot\|$  norm, and  $x^+ = \arg \min_{u \in \mathcal{C}} \{\langle \nabla F(x) + \epsilon, u \rangle + \frac{1}{\gamma} V(u, x) + h(u)\}$ . Then the following is true

$$\langle \nabla F(x), x - x^+ \rangle \geq \frac{\alpha}{\gamma} \|x^+ - x\|^2 + r(x^+) - r(x) + \langle \epsilon, x^+ - x \rangle.$$

*Proof.* By the optimality of  $x^+$ , we know that there exists a  $p \in \partial r(x^+)$  such that

$$\langle \nabla F(x) + \epsilon + \frac{1}{\gamma} [\nabla w(x^+) - \nabla w(x)] + p, u - x^+ \rangle \geq 0, \forall x \in \mathcal{C}. \quad (3.10)$$

Letting  $u = x$  in above inequality, we have

$$\langle \nabla F(x), x - x^+ \rangle \geq \frac{1}{\gamma} \langle \nabla w(x^+) - \nabla w(x), x^+ - x \rangle + \langle p + \epsilon, x^+ - x \rangle.$$

By the strongly convexity of  $w$  and  $\langle p, x^+ - x \rangle \geq r(x^+) - r(x)$ , we get the proof. □

**Lemma 3.1.6.** For any vector  $v$ , we have  $\|v\|_2 \leq \|\mathcal{C}\|_2 \|v\|_{\mathcal{C}}$ , where  $\|\mathcal{C}\|_2$  is the  $\ell_2$ -diameter and  $\|\mathcal{C}\|_2 = \sup_{x,y \in \mathcal{C}} \|x - y\|_2$ .

Lemma 3.1.6 implies that any smooth convex function  $F(\theta)$ , which is L-smooth with respect to  $\ell_2$  norm, is  $L\|\mathcal{C}\|_2^2$ -smooth with respect to  $\|\cdot\|_{\mathcal{C}}$  norm, which is the motivation of our algorithm.

*Proof.* If  $v = 0$ , this is trivially true. Otherwise, we will show that  $\frac{\|v\|_2}{\|\mathcal{C}\|_2} \leq \|v\|_{\mathcal{C}}$ . This is equivalent to show that  $v \notin \frac{\|v\|_2}{\|\mathcal{C}\|_2} \mathcal{C}$ . Taking any  $y \in \mathcal{C}$ , since  $\|\frac{\|v\|_2}{\|\mathcal{C}\|_2} y\|_2 = \frac{\|v\|_2}{\|\mathcal{C}\|_2} \|y\|_2$ , we know that  $\|y\|_2 < \|\mathcal{C}\|_2$ . Thus,  $\|\frac{\|v\|_2}{\|\mathcal{C}\|_2} y\|_2 < \|v\|_2$ . We get  $v \notin \frac{\|v\|_2}{\|\mathcal{C}\|_2} \mathcal{C}$ . □

**Lemma 3.1.7.** [269] For  $W = (\max_{w \in \mathcal{C}} \langle w, v \rangle)^2$ , where  $v \sim \mathcal{N}(0, I_p)$ , we have  $\mathbb{E}_v[W] = O(G_{\mathcal{C}}^2 + \|\mathcal{C}\|_2^2)$ .

### 3.1.7 Proofs of Differential Privacy

We will show that Algorithm 3.1.1, 3.1.2 and 3.1.3 are  $(\epsilon, \delta)$ -DP. The proof of Algorithm 3.1.4 is just based on the Moment in Lemma 2.1.7.

W.l.o.g, we assume  $G = 1$ , i.e.,  $\|\nabla f\| \leq 1$  (otherwise we can rescale  $f$ ). We will mainly focus on the proof of Theorem 3.1.1, the Proof of Theorem 3.1.3 and Theorem 3.1.4 are the same, instead of the iteration number (or number of queries). Let the difference data of  $D, D'$  be the n-th data. Now, consider the i-th query:

$$M_i = \nabla f(x_{t-1}^s, z_{i_t^s}) - \nabla f(\tilde{x}, z_{i_t^s}) + \frac{1}{n} \sum_{i=1}^n \nabla f(\tilde{x}, z_i) + u_t^s, u_t^s \sim \mathcal{N}(0, \sigma^2 I_p),$$

where  $i_t^s \in [n]$  is a uniform sample. This query can be thought as the composition of two queries:

$$M_{i,1} = \nabla f(x_{t-1}^s, z_{i_t^s}) - \nabla f(\tilde{x}, z_{i_t^s}) + \mathcal{N}(0, \sigma_1^2 I_p) \quad (3.11)$$

and

$$M_{i,2} = \nabla f(\tilde{x}, D) + \mathcal{N}(0, \sigma_2^2 I_p) = \frac{1}{n} \sum_{i=1}^n \nabla f(\tilde{x}, z_i) + \mathcal{N}(0, \sigma_2^2 I_p) \quad (3.12)$$

for some  $\sigma_1, \sigma_2$ . By **Theorem 2.1** in [1] we have  $\alpha_{M_i}(\lambda) \leq \alpha_{M_{i,1}}(\lambda) + \alpha_{M_{i,2}}(\lambda)$ . Now we bound  $\alpha_{M_{i,1}}(\lambda)$  and  $\alpha_{M_{i,2}}(\lambda)$ .

For  $\alpha_{M_{i,1}}$ , we can use **Lemma 3** in [1] directly, where  $q = \frac{1}{n}$ ,  $f(\cdot) = \nabla f(x_{t-1}^s, \cdot) - \nabla f(\tilde{x}, \cdot)$ . For some constant  $c_1$  and any integer  $\lambda \leq \sigma_1^2 \ln(n/\sigma_1)$ , we have

$$\alpha_{M_{i,1}}(\lambda) \leq c_1 \frac{\lambda^2}{n^2 \sigma_1^2} + O\left(\frac{\lambda^3}{n^3 \sigma_1^3}\right). \quad (3.13)$$

For  $\alpha_{M_{i,2}}(\lambda)$ , we use the relationship between moment account and Rényi divergence. By

Definition 2.1 in [52] we have:

$$\alpha_{M_{i,2}}(\lambda) = \lambda D_{\lambda+1}(P||Q), \quad (3.14)$$

where  $P = \nabla F(\tilde{x}, D) + \mathcal{N}(0, \sigma_2^2 I_p) = \mathcal{N}(\nabla F(\tilde{x}, D), \sigma_2^2)$  and  $Q = \nabla F(\tilde{x}, D') + \mathcal{N}(0, \sigma_2^2 I_p) = \mathcal{N}(\nabla F(\tilde{x}, D'), \sigma_2^2)$ . By Lemma 2.5 in [52], we have for some  $c_2$ :

$$\lambda D_{\lambda+1}(P||Q) = \frac{\lambda(\lambda+1)\|\nabla F(\tilde{x}, D) - \nabla F(\tilde{x}, D')\|^2}{2\sigma^2} \leq \frac{2\lambda(\lambda+1)}{n^2\sigma_2^2} \leq \frac{c_1\lambda^2}{n^2\sigma_2^2}. \quad (3.15)$$

Combining (3.13), (3.14) and (3.15), we have

$$\alpha_{M_i}(\lambda) \leq c_1 \frac{\lambda^2}{n^2\sigma_2^2} + c_2 \frac{\lambda^2}{n^2\sigma_1^2} + O\left(\frac{\lambda^3}{n^3\sigma_1^3}\right). \quad (3.16)$$

After  $T$  iterations, we have for some  $c_1, c_2$ ,

$$\alpha_M \leq \sum_{i=1}^T \alpha_{M_i} \leq c_1 \frac{\lambda^2}{n^2\sigma_2^2} + c_2 \frac{\lambda^2}{n^2\sigma_1^2}. \quad (3.17)$$

To be  $(\epsilon, \delta)$ -differentially private, by Theorem 2.2 in [1], it suffices to show that

$$c_1 \frac{T\lambda^2}{n^2\sigma_2^2} + c_2 \frac{T\lambda^2}{n^2\sigma_1^2} \leq \frac{\lambda\epsilon}{2}$$

and

$$\exp\left(\frac{-\lambda\epsilon}{2}\right) \leq \delta.$$

In addition, we need

$$\lambda \leq \sigma_1^2 \ln(n/\sigma_1). \quad (3.18)$$

It can be verified that when  $\epsilon \leq c_3 \frac{T}{n^2}$  for some constant  $c_3$ , we have

$$\sigma_1 = c_4 \frac{\sqrt{T \log(1/\delta)}}{n\epsilon} \quad (3.19)$$

and

$$\sigma_2 = c_5 \frac{\sqrt{T \log(1/\delta)}}{n\epsilon}. \quad (3.20)$$

For some constants  $c_4, c_5$ , all the conditions can be satisfied. Since the sum of two Gaussian distributions is still a Gaussian distribution, and  $M_i = M_{i,1} + M_{i,2}$ , we have  $\sigma = c \frac{\sqrt{T \log(1/\delta)}}{n\epsilon}$  for some  $c$ . Thus, T-fold of the queries

$$M_i = \nabla f(x_{t-1}^s, z_{i_t^s}) - \nabla f(\tilde{x}, z_{i_t^s}) + \frac{1}{n} \sum_{i=1}^n \nabla f(\tilde{x}, z_i) + \mathcal{N}(0, \sigma^2 I_p)$$

will guarantee  $(\epsilon, \delta)$ -differential private when  $\epsilon \leq c_3 \frac{T}{n^2}$ .

For Algorithm 3.1.1 and 3.1.3,  $T = Tm$ , while for Algorithm 3.1.2,  $T = 2^{T+1}m$ .

### Proof of Theorem 3.1.2

Let  $g_t = \frac{1}{\eta}(x_{t-1} - \text{prox}_{\eta r}(x_{t-1} - \eta v_t))$ . Then we have  $x_t = x_{t-1} - \eta g_t$ . Thus

$$\|x_t - x_*\|_2^2 = \|x_{t-1} - \eta g_t - x_*\|_2^2 = \|x_{t-1} - x_*\|_2^2 - 2\eta \langle g_t^s, x_{t-1} - x_* \rangle + \eta^2 \|g_t\|_2^2. \quad (3.21)$$

By Lemma 3 in [347], we have the following inequality

$$\begin{aligned} -\langle g_t, x_{t-1} - x_* \rangle + \frac{\eta}{2} \|g_t\|_2^2 &\leq F^r(x_*) - F^r(x_t^s) - \frac{\mu_F}{2} \|x_{t-1} - x_*\|_2^2 \\ &\quad - \frac{\mu_r}{2} \|x_t - x_*\|_2^2 - \langle v_t - \nabla F(x_{t-1}), x_t - x_* \rangle. \end{aligned} \quad (3.22)$$

Plugging (3.21) into (3.22), we have

$$\|x_t - x_*\|_2^2 \leq \|x_{t-1} - x_*\|_2^2 - 2\eta[F^r(x_t) - F^r(x_*)] - 2\eta \langle v_t - \nabla F(x_{t-1}), x_t - x_* \rangle. \quad (3.23)$$

Next we bound  $-2\eta \langle v_t - \nabla F(x_{t-1}), x_t - x_* \rangle$ . Denote  $\hat{x}_t = \text{prox}_{\eta r}(x_{t-1} - \eta \nabla F(x_{t-1}^s))$ .

Then we have

$$\begin{aligned}
& -2\eta \langle v_t - \nabla F(x_{t-1}), x_t - x_* \rangle \\
&= -2\eta \langle v_t - \nabla F(x_{t-1}), x_t - \hat{x}_t \rangle - 2\eta \langle v_t - \nabla F(x_{t-1}), \hat{x}_t - x_* \rangle \\
&\leq 2\eta \|v_t - \nabla F(x_{t-1})\|_2 \|x_t - \hat{x}_t\|_2 - 2\eta \langle v_t - \nabla F(x_{t-1}), \hat{x}_t - x_* \rangle \\
&\leq 2\eta \|v_t - \nabla F(x_{t-1})\|_2 \|x_{t-1} - \eta v_t - (x_{t-1} - \nabla F(x_{t-1}))\|_2 - 2\eta \langle v_t - \nabla F(x_{t-1}), \hat{x}_t - x_* \rangle \\
\end{aligned} \tag{3.24}$$

$$\leq 2\eta^2 \|v_t - \nabla F(x_{t-1})\|_2^2 - 2\eta \langle v_t - \nabla F(x_{t-1}), \hat{x}_t - x_* \rangle \tag{3.25}$$

We can easily get  $\mathbb{E}_{u_t, i_t}(v_t - \nabla F(x_{t-1})) = 0$ , since  $u_t^s$  is independent with  $v_{t-1}^s$ . Also by Lemma 3.1.1, we have

$$\mathbb{E}\|v_t - \nabla F(x_{t-1})\|_2^2 \leq 2L[F^r(x_{t-1}) - F^r(x_*) + F^r(\tilde{x}) - F^r(x_*)] + \sigma^2 p. \tag{3.26}$$

Plugging (3.26) into (3.25) and taking the expectation over  $i_t, u_t$ , we have

$$\begin{aligned}
\mathbb{E}\|x_t - x_*\|_2^2 &\leq \|x_{t-1} - x_*\|_2^2 - 2\eta [\mathbb{E}(F^r(x_t) - F^r(x_*))] \\
&\quad + 16\eta^2 L [F^r(x_{t-1}) - F^r(x_*) + F^r(\tilde{x}) - F^r(x_*)] + 4\eta^2 \sigma^2 p.
\end{aligned} \tag{3.27}$$

Summing over  $t = 1, 2, \dots, m$  and taking the expectation, we have

$$\begin{aligned}
&\mathbb{E}[\|x_m - x_*\|_2^2] + 2\eta(1 - 8\eta L) \sum_{t=1}^m [\mathbb{E}(F^r(x_t)) - F^r(x_*)] \\
&\leq \|\tilde{x} - x_*\|^2 + 16L\eta^2(m+1)[F^r(\tilde{x}) - F^r(x_*)] + 4m\eta^2 \sigma^2 p.
\end{aligned}$$

Since  $F^r$  is  $\mu$  strongly convex, we have  $\|\tilde{x} - x_*\|^2 \leq \frac{2}{\mu}(F^r(\tilde{x}) - F^r(x_*))$ . Dividing

$2m\eta(1 - 8L\eta)$  from both sides, we get

$$\mathbb{E}[F^r(\tilde{x}^s)] - F^r(x_*) \leq \left( \frac{1}{\eta(1 - 8\eta L)\mu m} + \frac{8L\eta(m+1)}{m(1 - 8L\eta)} \right) (\mathbb{E}[F^r(\tilde{x}_{s-1})] - F^r(x_*)) + \frac{2\eta}{1 - 8L\eta} \sigma^2 p. \quad (3.28)$$

Thus, we can choose  $\eta = \Theta(\frac{1}{L}) < \frac{1}{12L}$  and  $m = \Theta(\frac{L}{\mu})$  to make

$$A = \frac{1}{\eta(1 - 8\eta L)\mu m} + \frac{8L\eta(m+1)}{m(1 - 8L\eta)} < \frac{1}{2}$$

and  $\frac{2\eta}{1 - 8L\eta} < \frac{1}{2L}$ . By (3.28) and summing over  $s = 1, 2, \dots, T$ , we get

$$\begin{aligned} & \mathbb{E}[F^r(\tilde{x}^T)] - F^r(x_*) \\ & \leq A^T [F^r(x_0) - F^r(x_*)] + \frac{\sigma^2 p}{L} \\ & = A^T [F^r(x_0) - F^r(x_*)] + O\left(\frac{pG^2 T m \ln(1/\delta)}{n^2 \epsilon^2 L}\right) \\ & = A^T [F^r(x_0) - F^r(x_*)] + O\left(\frac{pG^2 T \ln(1/\delta)}{n^2 \epsilon^2 \mu}\right). \end{aligned}$$

Thus, if we take  $T$  such that  $A^T [F^r(x_0) - F^r(x_*)] = O\left(\frac{pG^2 \ln(1/\delta)}{n^2 \epsilon^2 \mu}\right)$ , i.e.,

$$T = O\left(\log\left(\frac{n^2 \epsilon^2 \mu}{pG^2 \ln(1/\delta)}\right)\right),$$

we have

$$\mathbb{E}[F^r(\tilde{x}^T)] - F^r(x_*) \leq O\left(\frac{pG^2 \ln(n\epsilon\mu/pG) \ln(1/\delta)}{n^2 \epsilon^2 \mu}\right),$$

where the big-O notation omits the other  $\ln$  term.

### Proof of Theorem 3.1.3

We first impose the following constraint on parameters  $\eta, \theta$

$$L\theta + \frac{L\theta}{1 - \theta} \leq \frac{1}{\eta}. \quad (3.29)$$

By the convexity of  $F(\cdot)$ , we have

$$\begin{aligned}
F(y_{t-1}) - F(u) &\leq \langle \nabla F(y_{t-1}), y_{t-1} - u \rangle \\
&= \langle \nabla F(y_{t-1}), y_{t-1} - x_{t-1} \rangle + \langle \nabla F(y_{t-1}), x_{t-1} - u \rangle \\
&= \frac{1-\theta}{\theta} \langle \nabla F(y_{t-1}), \tilde{x}_{s-1} - x_{t-1} \rangle + \langle \nabla F(y_{t-1}), x_{t-1} - u \rangle,
\end{aligned} \tag{3.30}$$

where the last equality is by the definition of  $y_{t-1}$ .

For the term  $\langle \nabla F(y_{t-1}), x_{t-1} - u \rangle$ , we expand it as

$$\langle \nabla F(y_{t-1}), x_{t-1} - u \rangle = \langle \nabla F(y_{t-1}) - v_t, x_{t-1} - u \rangle + \langle v_t, x_{t-1} - x_t \rangle + \langle v_t, x_t - u \rangle. \tag{3.31}$$

Since  $F(\cdot)$  is  $L$ -smooth, we have

$$\begin{aligned}
F(y_t) - F(y_{t-1}) &\leq \langle \nabla F(y_{t-1}), y_t - y_{t-1} \rangle + \frac{L}{2} \|y_t - y_{t-1}\|_2^2 \\
&= \theta \langle \nabla F(y_{t-1}), x_t - x_{t-1} \rangle + \frac{L\theta^2}{2} \|x_t - x_{t-1}\|_2^2 \\
&= \theta (\langle \nabla F(y_{t-1}) - v_t, x_t - x_{t-1} \rangle + \langle v_t, x_t - x_{t-1} \rangle) + \frac{L\theta^2}{2} \|x_t - x_{t-1}\|_2^2.
\end{aligned} \tag{3.32}$$

Thus, we have

$$\langle v_t, x_t - x_{t-1} \rangle \leq \frac{1}{\theta} (F(y_{t-1}) - F(y_t)) + \langle \nabla F(y_{t-1}) - v_t, x_t - x_{t-1} \rangle + \frac{L\theta}{2} \|x_t - x_{t-1}\|_2^2. \tag{3.33}$$

By (3.29), we have

$$\begin{aligned}
\langle v_t, x_t - x_{t-1} \rangle &\leq \frac{1}{\theta} (F(y_{t-1}) - F(y_t)) + \langle \nabla F(y_{t-1}) - v_t, x_t - x_{t-1} \rangle \\
&\quad + \frac{1}{2\eta} \|x_t - x_{t-1}\|_2^2 - \frac{L\theta}{2(1-\theta)} \|x_t - x_{t-1}\|_2^2.
\end{aligned} \tag{3.34}$$

Combining this with (3.30), (3.31), (3.32), (3.33) and (3.34), as well as by Lemma 3.1.3

(here  $r(x)$  is  $\mu$ -strongly convex), we have

$$\begin{aligned}
F(y_{t-1}) - F(u) &\leq \frac{1-\theta}{\theta} \langle \nabla F(y_{t-1}), \tilde{x}_{s-1} - y_{t-1} \rangle \\
&+ \langle \nabla F(y_{t-1}) - v_t, x_t - u \rangle + \frac{1}{\theta} (F(y_{t-1}) - F(y_t)) - \frac{L\theta}{2(1-\theta)} \|x_t - x_{t-1}\|_2^2 \\
&+ \frac{1}{2\eta} \|x_{t-1} - u\|_2^2 - \frac{1+\eta\mu}{2\eta} \|x_t - u\|_2^2 + r(u) - r(x_t).
\end{aligned}$$

Taking the expectation w.r.t  $i_t, u_t$ , we have

$$\begin{aligned}
F(y_{t-1}) - F(u) &\leq \frac{1-\theta}{\theta} \langle \nabla F(y_{t-1}), \tilde{x}_{s-1} - y_{t-1} \rangle + \mathbb{E} \langle \nabla F(y_{t-1}) - v_t, x_t - u \rangle + \frac{1}{\theta} (F(y_{t-1}) - \mathbb{E} F(y_t)) \\
&- \frac{L\theta}{2(1-\theta)} \|x_t - x_{t-1}\|_2^2 + \frac{1}{2\eta} \|x_{t-1} - u\|_2^2 - \frac{1+\eta\mu}{2\eta} \mathbb{E} \|x_t - u\|_2^2 + r(u) - \mathbb{E} r(x_t) \\
&\leq \frac{1-\theta}{\theta} \langle \nabla F(y_{t-1}), \tilde{x}_{s-1} - y_{t-1} \rangle + \frac{1}{2\beta} \mathbb{E} \|\nabla F(y_{t-1}) - v_t\|_2^2 + \frac{\beta}{2} \mathbb{E} \|x_t - x_{t-1}\|_2^2 \\
&+ \frac{1}{\theta} (F(y_{t-1}) - \mathbb{E} F(y_t)) - \frac{L\theta}{2(1-\theta)} \|x_t - x_{t-1}\|_2^2 + \frac{1}{2\eta} \|x_{t-1} - u\|_2^2 - \frac{1+\eta\mu}{2\eta} \mathbb{E} \|x_t - u\|_2^2 \\
&+ r(u) - \mathbb{E} r(x_t)
\end{aligned}$$

Applying Lemma 3.1.1, we have

$$\begin{aligned}
F(y_{t-1}) - F(u) &\leq \frac{1-\theta}{\theta} \langle \nabla F(y_{t-1}), \tilde{x}_{s-1} - y_{t-1} \rangle + \frac{L}{\beta} (F(\tilde{x}_{s-1}) - F(y_{t-1})) \\
&- \langle \nabla F(y_{t-1}), \tilde{x}_{s-1} - y_{t-1} \rangle + \frac{p\sigma^2}{2\beta} + \frac{\beta}{2} \mathbb{E} \|x_t - x_{t-1}\|_2^2 + \frac{1}{\theta} (F(y_{t-1}) - \mathbb{E} F(y_t)) \\
&- \frac{L\theta}{2(1-\theta)} \|x_t - x_{t-1}\|_2^2 + \frac{1}{2\eta} \|x_{t-1} - u\|_2^2 - \frac{1+\eta\mu}{2\eta} \mathbb{E} \|x_t - u\|_2^2 + r(u) - \mathbb{E} r(x_t).
\end{aligned}$$

Let  $\beta = \frac{L\theta}{1-\theta} > 0$ . By rearranging the above inequality, we obtain the following

$$0 \leq \frac{1-\theta}{\theta} F(\tilde{x}_{s-1}) - \frac{1}{\theta} \mathbb{E}F(y_j) + F^r(u) - \mathbb{E}r(x_t) + \frac{1}{2\eta} \|x_{t-1} - u\|_2^2 - \frac{1+\eta\mu}{2\eta} \mathbb{E}\|x_t - u\|_2^2 \\ (3.35)$$

$$\leq \frac{1-\theta}{\theta} F^r(\tilde{x}_{s-1}) - \frac{1}{\theta} \mathbb{E}F^r(y_t) + F(u) + \frac{1}{2\eta} \|x_{t-1} - u\|_2^2 - \frac{1+\eta\mu}{2\eta} \mathbb{E}\|x_t - u\|_2^2 + \frac{1-\theta}{2L\theta} p\sigma^2, \\ (3.36)$$

where inequality (3.36) is by the definition of  $y_{t-1}$  and the convexity of  $r(\cdot)$ , which leads to

$$-r(x_t) \leq \frac{1-\theta}{\theta} r(\tilde{x}_{s-1}) - \frac{1}{\theta} r(y_t).$$

This is equivalent to

$$\frac{1}{\theta} (\mathbb{E}F^r(y_t) - F^r(u)) \leq \frac{1-\theta}{\theta} (F^r(\tilde{x}_{s-1}) - F^r(u)) \\ + \frac{1}{2\eta} \|x_{t-1} - u\|_2^2 - \frac{1+\eta\mu}{2\eta} \mathbb{E}\|x_t - u\|_2^2 + \frac{1-\theta}{2L\theta} p\sigma^2. \quad (3.37)$$

Let  $u = x_*$ . Using  $w = 1 + \eta\mu$  to sum (3.37) over  $t = 1, \dots, m$  with increasing  $w^{t-1}$ , and taking the expectation, we have

$$\frac{1}{\theta} \sum_{t=0}^{m-1} w^t (\mathbb{E}F^r(y_{t+1}) - F^r(x_*)) + \frac{w^m}{2\eta} \mathbb{E}\|x_m - x_*\|_2^2 \leq \\ \frac{1-\theta}{\theta} \sum_{t=0}^{m-1} w^t (F^r(\tilde{x}_{s-1}) - F^r(x_*)) + \frac{1}{2\eta} \|x_0 - x_*\|_2^2 + \sum_{t=0}^{m-1} w^t \frac{1-\theta}{2L\theta} p\sigma^2. \quad (3.38)$$

Using Jensen's inequality and

$$\tilde{x}_s = \theta \left( \sum_{t=0}^{m-1} w^t \right)^{-1} \sum_{t=0}^{m-1} w^t x_{t+1} + (1-\theta) \tilde{x}_{s-1} \\ = \left( \sum_{t=0}^{m-1} w^t \right)^{-1} \sum_{t=0}^{m-1} w^t y_{t+1}$$

we have

$$\begin{aligned} & \left( \frac{1}{\theta} \sum_{t=0}^{m-1} w^t \right) (\mathbb{E} F^r(\tilde{x}_s) - F^r(x_*)) + \frac{w^m}{2\eta} \mathbb{E} \|x_m - x_*\|_2^2 \leq \\ & \frac{1-\theta}{\theta} \sum_{t=0}^{m-1} w^t (F^r(\tilde{x}_{s-1}) - F^r(x_*)) + \frac{1}{2\eta} \|x_0 - x_*\|_2^2 + \sum_{t=0}^{m-1} w^t \frac{1-\theta}{2L\theta} p\sigma^2. \end{aligned} \quad (3.39)$$

Consider the first case with  $m \leq \frac{3L}{4\mu}$ . Denote by  $\kappa = \frac{L}{\mu}$ . We set  $\eta = \sqrt{\frac{1}{3\mu m L}}$ ,  $\theta = \sqrt{\frac{m}{3\kappa}} \leq \frac{1}{2}$  and  $m = \Theta(n)$ . By (3.29), we know that  $\sqrt{\frac{m}{\kappa}} \leq \frac{\sqrt{3}}{2}$ .

For the term of  $(1-\theta)w^m$ , we have

$$(1-\theta)w^m = (1 - \sqrt{\frac{m}{3\kappa}})(1 + \sqrt{\frac{1}{3m\kappa}})^m.$$

Let  $\zeta = \sqrt{\frac{m}{\kappa}} \in (0, \frac{\sqrt{3}}{2}]$ . We denote

$$\phi(\zeta) = (1 - \frac{\sqrt{3}}{3}\zeta)(1 + \frac{\sqrt{3}}{3}\frac{\zeta}{m})^m$$

as a function of  $\zeta$ . We can easily get that  $\phi(\zeta)$  is monotonically decreasing on  $[0, \frac{\sqrt{3}}{2}]$  for any  $m \geq 0$ , which means that  $(1-\theta)w^m \leq \phi(0) = 1$ . Thus, we have  $\frac{1}{\theta} \geq \frac{1-\theta}{\theta}w^m$ .

$$\begin{aligned} & \left( \frac{1-\theta}{\theta} \sum_{t=0}^{m-1} w^t \right) (\mathbb{E} F^r(\tilde{x}_s) - F^r(x_*)) + \frac{w^m}{2\eta} \mathbb{E} \|x_m - x_*\|_2^2 \leq \\ & w^{-m} \left( \frac{1-\theta}{\theta} \sum_{t=0}^{m-1} w^t (F^r(\tilde{x}_{s-1}) - F^r(x_*)) + \frac{1}{2\eta} \|x_0 - x_*\|_2^2 + \sum_{t=0}^{m-1} w^t \frac{1-\theta}{2L\theta} p\sigma^2 \right). \end{aligned} \quad (3.40)$$

Dividing  $\frac{1-\theta}{\theta} \sum_{t=0}^{m-1} w^t$ , we get

$$\begin{aligned} & \mathbb{E} F^r(\tilde{x}_s) - F^r(x_*) + \frac{\theta}{2\eta(1-\theta) \sum_{t=0}^{m-1} w^t} \mathbb{E} \|x_m - x_*\|_2^2 \leq \\ & w^{-m} (F^r(\tilde{x}_{s-1}) - F^r(x_*) + \frac{\theta}{2\eta(1-\theta) \sum_{t=0}^{m-1} w^t} \mathbb{E} \|x_0 - x_*\|_2^2 + \frac{1}{2L} p\sigma^2). \end{aligned} \quad (3.41)$$

Summing the above inequality over  $s = 1, \dots, T$ , we have

$$\begin{aligned} \mathbb{E}F^r(\tilde{x}_T) - F^r(x_*) &\leq w^{-Tm}(F^r(\tilde{x}_0) - F^r(x_*)) + \frac{\theta}{2\eta(1-\theta)\sum_{t=0}^{m-1}w^t}\mathbb{E}\|x_0 - x_*\|_2^2 \\ &+ \frac{T}{2L}p\sigma^2. \end{aligned} \quad (3.42)$$

Since  $F^r(\cdot)$  is  $\mu$ -strongly convex, we have  $\|x_0 - x_*\|_2^2 \leq \frac{2}{\mu}(F^r(x_0) - F^r(x_*)).$  By substituting with our parameters and  $\sigma^2 = O\left(\frac{G^2 \log \frac{1}{\delta} T^m}{n^2 \epsilon^2}\right)$ , we have

$$\mathbb{E}F^r(\tilde{x}_T) - F^r(x_*) \leq (O(1 + \sqrt{\frac{1}{3n\kappa}}))^{-Tm}O(F^r(x_0) - F^r(x_*)) + \frac{T^2 mp G^2 \log \frac{1}{\delta}}{2Ln^2 \epsilon^2}. \quad (3.43)$$

Let  $Tm = \log_{O(1+\sqrt{\frac{1}{3n\kappa}})}\left(\frac{n^2\epsilon^2}{p}\right) = O(\sqrt{n\kappa} \log \frac{n\epsilon}{\sqrt{p}}).$  Since  $m = \Theta(n)$ , we have  $T = O(\sqrt{\frac{\kappa}{n}} \log \frac{n\epsilon}{\sqrt{p}}).$  Thus, we get

$$\mathbb{E}F^r(\tilde{x}_T) - F^r(x_*) \leq O\left(\frac{\log^2 n G^2 \log \frac{1}{\delta} p}{n^2 \epsilon^2 \mu}\right).$$

For the other case with  $\frac{n}{k} \geq \frac{3}{4}$ , we set  $\eta = \frac{2}{3L}$ ,  $\theta = \frac{1}{2}$  and  $m = \frac{3}{4}\kappa.$  Since  $w^m = (1 + \frac{2}{3\kappa})^m \geq 1 + \frac{2m}{3\kappa} \geq \frac{3}{2}$ , substituting the parameters into (3.39), we have

$$\begin{aligned} &\sum_{t=0}^{m-1} w^t (\mathbb{E}F^r(\tilde{x}_s) - F^r(x_*)) + \frac{3L}{4} \mathbb{E}\|x_m - x_*\|^2 \\ &\leq \frac{2}{3} \left( \left( \sum_{t=0}^{m-1} w^t \right) (\mathbb{E}F^r(\tilde{x}_{s-1}) - F^r(x_*)) + \frac{3L}{4} \|x_0 - x_*\|_2^2 + \sum_{t=0}^{m-1} w^t \frac{p\sigma^2}{2L} \right) \end{aligned} \quad (3.44)$$

$$\leq \frac{2}{3} \left( \left( \sum_{t=0}^{m-1} w^t \right) (\mathbb{E}F^r(\tilde{x}_{s-1}) - F^r(x_*)) + \frac{3L}{4} \|x_0 - x_*\|_2^2 \right) + O\left(\sum_{t=0}^{m-1} w^t \frac{p\sigma^2}{L}\right). \quad (3.45)$$

Dividing  $\sum_{t=0}^{m-1} w^t$  on both sides, we get

$$\mathbb{E}F^r(\tilde{x}_T) - F^r(x_*) \leq \left(\frac{2}{3}\right)^T O(F^r(x_0) - F^r(x_*)) + O\left(\frac{pTG^2 \log \frac{1}{\delta}}{n^2 \epsilon^2 \mu}\right). \quad (3.46)$$

Taking  $T = O(\log(\frac{n\epsilon}{\sqrt{p}}))$ , we have  $\mathbb{E}F^r(\tilde{x}_s) - F^r(x_*) \leq O(\frac{G^2 p \log n \log \frac{1}{\delta}}{n^2 \epsilon^2 \mu})$ . The total gradient complexity is thus  $O(T(m+n)) = O(nT) = O(n \log(\frac{n\epsilon}{\sqrt{p}}))$

### Proof of Theorem 3.1.5

$$\begin{aligned}
& \mathbb{E}_{i_t^s, u_t^s}[F^r(x_t^s) - F^r(x_*)] = \mathbb{E}_{i_t^s, u_t^s}[F(x_t^s) - F(x_*) + r(x_t^s) - r(x_*)] \\
& \leq \mathbb{E}_{i_t^s, u_t^s}[F(x_{t-1}^s) + \langle \nabla F(x_{t-1}^s), x_t^s - x_{t-1}^s \rangle + \frac{L}{2} \|x_t^s - x_{t-1}^s\|^2 - F(x_*) + r(x_t^s) - r(x_*)] \\
& \leq \mathbb{E}_{i_t^s, u_t^s}[\langle \nabla F(x_{t-1}^s), x_{t-1}^s - x_* \rangle] + \langle \nabla F(x_{t-1}^s), x_t^s - x_{t-1}^s \rangle + \frac{L}{2} \|x_t^s - x_{t-1}^s\|^2 + r(x_t^s) - r(x_*)] \\
& = \mathbb{E}_{i_t^s, u_t^s}[\langle v_t^s, x_{t-1}^s - x_* \rangle] + \langle \nabla F(x_{t-1}^s), x_t^s - x_{t-1}^s \rangle + \frac{L}{2} \|x_t^s - x_{t-1}^s\|^2 + r(x_t^s) - r(x_*)].
\end{aligned} \tag{3.47}$$

The last equality is due to the fact that  $\mathbb{E}_{i_t^s, u_t^s}[v_t^s] = \nabla F(x_{t-1}^s)$ . By [10], we have

$$\begin{aligned}
\langle v_t^s, x_{t-1}^s - x_* \rangle + r(x_t^s) - r(x_*) & \leq \langle v_t^s, x_{t-1}^s - x_t^s \rangle + \frac{\|x_{t-1}^s - x_*\|^2}{2\eta} \\
& \quad - \frac{\|x_t^s - x_*\|^2}{2\eta} - \frac{\|x_t^s - x_{t-1}^s\|^2}{2\eta}.
\end{aligned} \tag{3.48}$$

Plugging (3.48) into (3.47), we have

$$\begin{aligned}
& LHS \\
& \leq \mathbb{E}_{i_t^s, u_t^s}[\langle v_t^s - \nabla F(x_{t-1}^s), x_{t-1}^s - x_t^s \rangle - \frac{1-\eta L}{2\eta} \|x_t^s - x_{t-1}^s\|^2 + \frac{\|x_{t-1}^s - x_*\|^2 - \|x_t^s - x_*\|^2}{2\eta}] \\
& \leq \mathbb{E}_{i_t^s, u_t^s} \frac{\eta}{2(1-\eta L)} \|v_t^s - \nabla F(x_{t-1}^s)\|^2 + \frac{\|x_{t-1}^s - x_*\|^2 - \mathbb{E}_{i_t^s, u_t^s}[\|x_t^s - x_*\|^2]}{2\eta} \\
& \leq \frac{4\eta L}{1-\eta L} [F^r(x_{t-1}^s) - F^r(x_*) + F^r(\tilde{x}_{s-1}) - F^r(x_*)] + \frac{\eta}{1-\eta L} p \sigma^2 \\
& \quad + \frac{\|x_{t-1}^s - x_*\|^2 - \mathbb{E}_{i_t^s, u_t^s}[\|x_t^s - x_*\|^2]}{2\eta}.
\end{aligned}$$

Choosing  $\eta = \frac{1}{13L}$ , summing over  $t = 1, \dots, m_s$ , dividing  $m_s$ , and taking the expectation,

we have

$$\begin{aligned}\mathbb{E}\left[\frac{1}{m_s} \sum_{t=1}^{m_s} F^r(x_t^s) - F^r(x_*)\right] &\leq \frac{1}{3} \mathbb{E}\left[\frac{1}{m_s} \sum_{t=0}^{m_s-1} [F^r(x_t^s) - F^r(x_*) + F^r(\tilde{x}_{s-1}) - F^r(x_*)]\right. \\ &\quad \left.+ \frac{\|x_0^s - x_*\|^2 - \mathbb{E}[\|x_{m_s}^s - x_*\|^2]}{2\eta m_s} + \frac{1}{12L}\sigma^2 p.\right]\end{aligned}$$

By the definitions of  $x_0^{s+1}$  and  $\tilde{x}_s$ , we have

$$\begin{aligned}2\mathbb{E}[F^r(\tilde{x}_s) - F^r(x_*)] &\leq \mathbb{E}\left[\frac{F^r(x_0^s) - F^r(x_*) - (F^r(x_0^{s+1}) - F^r(x_*))}{m_s}\right. \\ &\quad \left.+ F^r(\tilde{x}_{s-1}) - F^r(x_*) + \frac{\|x_0^s - x_*\|^2 - \|x_0^{s+1} - x_*\|^2}{2\eta/3m_s}\right] + \frac{1}{4L}\sigma^2 p, \quad (3.49)\end{aligned}$$

which implies that

$$\begin{aligned}2(\mathbb{E}[F^r(\tilde{x}_s) - F^r(x_*) + \frac{\|x_0^{s+1} - x_*\|^2}{4\eta/3m_s} + \frac{F^r(x_0^{s+1}) - F^r(x_*)}{2m_s}]) &\leq \\ \mathbb{E}[F^r(\tilde{x}_{s-1}) - F^r(x_*) + \frac{\|x_0^s - x_*\|^2}{4\eta/3m_{s-1}} + \frac{F^r(x_0^s) - F^r(x_*)}{2m_{s-1}}] + \frac{\sigma^2 p}{4L}. \quad (3.50)\end{aligned}$$

Summing over  $s = 1, \dots, T$ , we get

$$\mathbb{E}[F^r(\tilde{x}_T) - F^r(x_*)] \leq \frac{F^r(\tilde{x}_0) - F^r(x_*)}{2^{T-1}} + \frac{\|\tilde{x}_0 - x_*\|^2}{2^T 4\eta/3m} + \frac{1}{4L}\sigma^2 p.$$

Thus, if we take  $m = \Theta(L)$  to make  $A = 2F^r(\tilde{x}_0) - F^r(x_*) + \frac{\|\tilde{x}_0 - x_*\|^2}{4\eta/3m}$  independent of  $T, n, p, \sigma, L$ , and plug  $\sigma$  into (3.50), we have

$$\begin{aligned}\mathbb{E}[F^r(\tilde{x}_T)] - F^r(x_*) &\leq \frac{A}{2^T} + O\left(\frac{G^2 p 2^T m \ln 2/\delta}{n^2 \epsilon^2 L}\right) \\ &= \frac{A}{2^T} + O\left(\frac{G^2 p 2^T \ln(1/\delta)}{n^2 \epsilon^2}\right).\end{aligned}$$

Let  $T = O(\log(\frac{n\epsilon}{G\sqrt{p}\sqrt{1/\delta}}))$ . We have

$$\mathbb{E}[F^r(\tilde{x}_s)] - F^r(x_*) \leq O\left(\frac{G\sqrt{p\ln(1/\delta)}}{n\epsilon}\right).$$

The gradient complexity is  $O(2^s m + Tn) = O(\frac{nL\epsilon}{G\sqrt{p}} + n\log(\frac{n\epsilon}{G\sqrt{p}}))$ .

### Proof of Theorem 3.1.7

We use  $\|\cdot\|$  and  $\|\cdot\|_*$  instead of  $\|\cdot\|_{\mathcal{C}}$  and  $\|\cdot\|_{\mathcal{C}^*}$ . Also, w.l.o.g we assume that  $\|\mathcal{C}\|_2 = 1$  (for the general case, just replace  $L$  by  $L\|\mathcal{C}\|_2^2$ ). Since  $b_{k+1}$  is independent of  $x_{k+1}$ , we have for any  $u$

$$\begin{aligned} & \mathbb{E}_{b_{k+1}}[\langle \alpha_{k+1} \nabla F(x_{k+1}), z_k - u \rangle] \\ &= \mathbb{E}_{b_{k+1}}[\langle \alpha_{k+1} (\nabla F(x_{k+1}) + b_{k+1}), z_k - u \rangle] \\ &= \mathbb{E}_{b_{k+1}}[\langle \alpha_{k+1} (\nabla F(x_{k+1}) + b_{k+1}), z_k - z_{k+1} \rangle] + \mathbb{E}_{b_{k+1}}[\langle \alpha_{k+1} (\nabla F(x_{k+1}) + b_{k+1}), z_{k+1} - u \rangle]. \end{aligned} \tag{3.51}$$

Since

$$z_{k+1} = \arg \min_{z \in \mathcal{C}} \{V(z, z_k) + \alpha_{k+1} \langle \nabla F(x_{k+1}) + b_{k+1}, z - z_k \rangle\},$$

which implies that

$$\langle \nabla V(z_{k+1}, z_k) + \alpha_{k+1} (\nabla F(x_{k+1}) + b_{k+1}), u - z_{k+1} \rangle \geq 0$$

for every  $u \in \mathcal{C}$ . So we can get

$$\begin{aligned} & \mathbb{E}_{b_{k+1}}[\langle \alpha_{k+1} (\nabla F(x_{k+1}) + b_{k+1}), z_{k+1} - u \rangle] \\ & \leq \mathbb{E}_{b_{k+1}}[\langle -\nabla V(z_{k+1}, z_k), z_{k+1} - u \rangle] \\ &= \mathbb{E}_{b_{k+1}}[V(u, z_k) - V(u, z_{k+1}) - V(z_{k+1}, z_k)], \end{aligned} \tag{3.52}$$

where the equality is due to the triangle equality of Bregman divergence. Since  $w$  is 1-strong convex with respect to  $\|\cdot\|$ , we have  $-V(z_{k+1}, z_k) \leq -\frac{1}{2}\|z_{k+1} - z_k\|^2$ . Plugging this into (3.51), we have

$$\begin{aligned}
& \mathbb{E}_{b_{k+1}}[\langle \alpha_{k+1} \nabla F(x_{k+1}), z_k - u \rangle] \\
& \leq \mathbb{E}_{b_{k+1}}[\langle \alpha_{k+1}(\nabla F(x_{k+1}) + b_{k+1}), z_k - z_{k+1} \rangle - \frac{1}{2}\|z_{k+1} - z_k\|^2] \\
& \quad + V(u, z_k) - \mathbb{E}_{b_{k+1}}[V(u, z_{k+1})] \\
& \leq \mathbb{E}_{b_{k+1}}[\langle \alpha_{k+1} \nabla F(x_{k+1}), z_k - z_{k+1} \rangle - \frac{1}{4}\|z_{k+1} - z_k\|^2] + \alpha_{k+1}^2 \mathbb{E}_{b_{k+1}}[\|b_{k+1}\|_*^2] \\
& \quad + V(u, z_k) - \mathbb{E}_{b_{k+1}}[V(u, z_{k+1})]. \tag{3.53}
\end{aligned}$$

The last inequality is due to Cauchy-Schwartz Inequality. Thus, we have  $\langle \alpha_{k+1} b_{k+1}, z_k - z_{k+1} \rangle \leq \alpha_{k+1}^2 \|b_{k+1}\|_*^2 + \frac{1}{4}\|z_k - z_{k+1}\|^2$ . Now, we want to bound  $\mathbb{E}_{b_{k+1}}[\langle \alpha_{k+1} \nabla F(x_{k+1}), z_k - z_{k+1} \rangle - \frac{1}{4}\|z_{k+1} - z_k\|^2]$ . Define  $v = r_k z_{k+1} + (1 - r_k) y_k \in \mathcal{C}$  so that  $x_{k+1} - v = r_k(z_k - z_{k+1})$ .

We have

$$\begin{aligned}
& \langle \alpha_{k+1} \nabla F(x_{k+1}), z_k - z_{k+1} \rangle - \frac{1}{4}\|z_{k+1} - z_k\|^2 \\
& = \langle \frac{\alpha_{k+1}}{r_k} \nabla F(x_{k+1}), x_{k+1} - v \rangle - \frac{1}{4r_k^2} \|x_{k+1} - v\|^2 \\
& = 2\alpha_{k+1}^2 L(\langle F(x_{k+1}), x_{k+1} - v \rangle - \frac{L}{2} \|x_{k+1} - v\|^2) \\
& \leq 2\alpha_{k+1}^2 L(-\min_{y \in \mathcal{C}} \{\frac{L}{2} \|y - x_{k+1}\|^2 + \langle F(x_{k+1}), y - x_{k+1} \rangle\}) \\
& = 2\alpha_{k+1}^2 L(-\{\frac{L}{2} \|y_{k+1} - x_{k+1}\|^2 + \langle F(x_{k+1}), y_{k+1} - x_{k+1} \rangle\}) \\
& \leq 2\alpha_{k+1}^2 L(F(x_{k+1}) - F(y_{k+1})). \tag{3.54}
\end{aligned}$$

The last inequality is due to the fact that  $F$  is  $L\|\mathcal{C}\|_2^2$ -smooth (note that  $\|\mathcal{C}\|_2 = 1$ ) in  $\|\cdot\|$  norm and the definition of  $y_{k+1}$ . Thus, we get the following

$$\begin{aligned}
& \mathbb{E}_{b_{k+1}}[\langle \alpha_{k+1} \nabla F(x_{k+1}), z_k - u \rangle] \\
&= \mathbb{E}_{b_{k+1}}[\langle \alpha_{k+1} (\nabla F(x_{k+1}) + b_{k+1}), z_k - u \rangle] \\
&\leq 2\alpha_{k+1}^2 L(F(x_{k+1}) - F(y_{k+1})) + V(u, z_k) - \mathbb{E}_{b_{k+1}}[V(u, z_{k+1})] + \alpha_{k+1}^2 \mathbb{E}_{b_{k+1}}\|b_{k+1}\|_*^2.
\end{aligned} \tag{3.55}$$

By using the Concentration of Gaussian Width, Lemma 3.1.7 shows that  $\mathbb{E}_{b_{k+1}}\|b_{k+1}\|_*^2 = \sigma^2 O(G_{\mathcal{C}}^2 + \|\mathcal{C}\|_2^2)$ , where  $G_{\mathcal{C}}$  is the Gaussian Width of  $\mathcal{C}$ . From this, we have

$$\begin{aligned}
& \mathbb{E}_{b_{k+1}}[\alpha_{k+1}(F(x_{k+1}) - F(u))] \\
&\leq \mathbb{E}_{b_{k+1}}[\langle \alpha_{k+1} \nabla F(x_{k+1}), x_{k+1} - u \rangle] \\
&= \mathbb{E}_{b_{k+1}}([\langle \alpha_{k+1} \nabla F(x_{k+1}), x_{k+1} - z_k \rangle] + [\langle \alpha_{k+1} \nabla F(x_{k+1}), z_k - u \rangle]) \\
&\leq \frac{\alpha_{k+1}(1 - r_k)}{r_k} \langle \nabla F(x_{k+1}), y_k - x_{k+1} \rangle + \mathbb{E}_{b_{k+1}}[\langle \alpha_{k+1} \nabla F(x_{k+1}), z_k - u \rangle] \\
&\leq \frac{\alpha_{k+1}(1 - r_k)}{r_k} (F(y_k) - F(x_{k+1}) + \mathbb{E}_{b_{k+1}}[\langle \alpha_{k+1} \nabla F(x_{k+1}), z_k - u \rangle]) \\
&\leq (2\alpha_{k+1}^2 L - \alpha_{k+1})(F(y_k) - F(x_{k+1}) + 2\alpha_{k+1}^2 L(F(x_{k+1}) - F(y_{k+1}))) \\
&\quad + V(u, z_k) - \mathbb{E}_{b_{k+1}}[V(u, z_{k+1})] + \alpha_{k+1}^2 \mathbb{E}_{b_{k+1}}\|b_{k+1}\|_*^2.
\end{aligned}$$

Thus, we obtain

$$\begin{aligned}
& 2\alpha_{k+1}^2 L F(y_{k+1}) - (2\alpha_{k+1}^2 L - \alpha_{k+1}) F(y_k) + \mathbb{E}(V(u, z_{k+1}) - V(u, z_k)) \\
&\leq \alpha_{k+1} F(u) + \alpha_{k+1}^2 \sigma^2 O(G_{\mathcal{C}}^2 + \|\mathcal{C}\|_2^2).
\end{aligned} \tag{3.56}$$

By the definition of  $\alpha_{k+1}$ , we have  $2\alpha_k^2 L = 2\alpha_{k+1}^2 L - \alpha_{k+1} + \frac{1}{8L}$ . Summing over  $k = 0 \dots, T-1$  and setting  $u = x_*$ , by the definition of  $\alpha_k$  we have  $\sum_{k=1}^T \alpha_k^2 = O(T^3)$ . After

taking the expectation we get

$$\begin{aligned} & 2\alpha_T^2 L \mathbb{E}[F(y_T)] + \frac{1}{8L} \mathbb{E}\left[\sum_{k=1}^{T-1} F(y_k)\right] + \mathbb{E}[V(x_*, z_{T-1})] - V(x_*, z_0) \\ & \leq \sum_{k=1}^T \alpha_k F(x_*) + O(T^3 \sigma^2 (G_C^2 + \|\mathcal{C}\|_2^2)/L^2). \end{aligned} \quad (3.57)$$

Plugging  $\alpha_k = \frac{k+1}{4L}$  into the above, dividing both sides by a factor of  $2\alpha_T^2 L$ , and by the fact that  $V \geq 0$ , we finally get

$$\mathbb{E}[F(y_T)] - F(x_*) \leq \frac{8LV(x_*, x_0)}{(T+1)^2} + O(T\sigma^2(G_C^2 + \|\mathcal{C}\|_2^2)/L). \quad (3.58)$$

Since  $\sigma^2 = O(\frac{G^2 T \ln(1/\delta)}{n^2 \epsilon^2})$ , if choose

$$T^2 = O\left(\frac{L\sqrt{V(x_*, x_0)}n\epsilon}{G\sqrt{\ln(1/\delta)}\sqrt{G_C^2 + \|\mathcal{C}\|_2^2}}\right), \quad (3.59)$$

we have the bound

$$\mathbb{E}[F(y_T)] - F(x_*) \leq O\left(\frac{\sqrt{V(x_*, x_0)}\sqrt{G_C^2 + \|\mathcal{C}\|_2^2}G\sqrt{\ln(1/\delta)}}{n\epsilon}\right).$$

## 3.2 DP-ERM with Heavy-tailed Data

It is worth noting that all previous results of DP-ERM or DP-SCO need to assume that either the loss function is  $O(1)$ -Lipschitz or each data sample has bounded  $\ell_2$  or  $\ell_\infty$  norm. This is particularly true for those output perturbation based [67] and objective or gradient perturbation based [29] DP methods. However, such assumptions may not always hold when dealing with real-world datasets, especially those from biomedicine and finance, implying that existing algorithms may fail. The main reason is that in such applications, the datasets are often unbounded or even heavy-tailed [345, 39, 157]. As pointed out by Mandelbrot and Fama in their influential finance papers [211, 111], asset prices in the early

1960s exhibit some power-law behavior. The heavy-tailed data could lead to unbounded gradient and thus violate the Lipschitz condition. For example, consider the linear squared loss  $\ell(w, x, y) = (w^T x - y)^2$ . When  $x$  is heavy-tailed, the gradient of  $\ell(w, x, y)$  becomes unbounded.

With the above understanding, our questions now are: **What is the behavior of DP-SCO on heavy-tailed data and is there any effective method for the problem?**

To answer these questions, we will conduct, in this section, a comprehensive study of the DP-SCO problem. Our contributions can be summarized as follows.

1. We first consider the case where the loss function is strongly convex and smooth.

For this case, we propose an  $(\epsilon, \delta)$ -DP method based on the sample-and-aggregate framework by [233] and show that under some assumptions, with high probability, the excess population risk of the output is  $\tilde{O}(\frac{d^3}{n\epsilon^4} L_{\mathcal{D}}(w^*))$ , where  $n$  is the sample size,  $d$  is the dimensionality and  $L_{\mathcal{D}}(w^*)$  is the minimal value of the population risk.

2. Then, we study the case with the additional assumptions: each coordinate of the

gradient of the loss function is sub-exponential and Lipschitz. For this case, we introduce an  $(\epsilon, \delta)$ -DP algorithm based on the gradient descent method and a recent algorithm on private 1-dimensional mean estimation [51] (*i.e.*, Algorithm 3.2.7).

We show that the expected excess population risk for this case can be improved to  $\tilde{O}(\frac{d^2 \log \frac{1}{\delta}}{n\epsilon^2})$ .

3. We also consider the general case, where the loss function does not need the above

additional assumptions and can be general convex, instead of strongly convex. For this case, we present a gradient descent method based on the strategy of trimming the unbounded gradient (Algorithm 3.2.8). We show that if each coordinate of the gradient of the loss function has bounded second-order moment, then with high probability, the output of our algorithm achieves excess population risks of  $\tilde{O}(\frac{d^2 \log \frac{1}{\delta}}{n\epsilon^2})$  and  $\tilde{O}(\frac{\log \frac{1}{\delta} d^2}{(n\epsilon^2)^{\frac{1}{3}}})$  for strongly convex and general convex loss functions, respectively. It is notable that

compared with Algorithm 3.2.8, Algorithm 3.2.7 uses stronger assumptions and yields weaker results.

4. Finally, we test our proposed algorithms on both synthetic and real-world datasets. Experimental results are consistent with our theoretical claims and reveal the effectiveness of our algorithms in handling heavy-tailed datasets.

### 3.2.1 Related Work

As mentioned earlier, there is a long list of works on DP-SCO or DP-ERM. However, none of them considers the case with heavy-tailed data. Recently, a number of works have studied the SCO and ERM problems with heavy-tailed data [46, 220, 151, 191]. However, all of them focus on the non-private version of the problem. It is not clear whether they can be adapted to private versions. To our best knowledge, the work presented in this paper is the first one on general DP-SCO with heavy-tailed data.

The works that are most related to ours are perhaps those dealing with unbounded sensitivity. [103] proposed a general framework called propose-test-release and applied it to mean estimation. They obtained asymptotic results which are incomparable with ours. Also, it is not clear whether such a framework can be applied to our problem. In our second result, we adopt the private mean estimation procedure in [51]. However, their results are in expectation form, which is not preferred in robust estimation [46]. For this reason, we propose a new algorithm which yields theoretically guaranteed bounds with high probability. [176] considered the confidence interval estimation problem for Gaussian distributions which was later extended to general distributions [114]. However, it was unknown how to extend them to the DP-SCO problem. [1] proposed a DP-SGD method based on truncating the gradient, which could deal with the infinity sensitivity issue. However, there is no theoretical guarantees on the excess population risk.

### 3.2.2 Preliminaries

**Definition 3.2.1** (DP-SCO). Given a dataset  $D = \{x_1, \dots, x_n\}$  from a data universe  $\mathcal{X}$  where  $x_i$  are i.i.d. samples from some unknown distribution  $\mathcal{D}$ , a convex loss function  $\ell(\cdot, \cdot)$ , and a convex constraint set  $\mathcal{W} \subseteq \mathbb{R}^d$ , Differentially Private Stochastic Convex Optimization (DP-SCO) is to find  $w^{\text{priv}}$  so as to minimize the population risk, *i.e.*,  $L_{\mathcal{D}}(w) = \mathbb{E}_{x \sim \mathcal{D}}[\ell(w, x)]$  with the guarantee of being differentially private. The utility of the algorithm is measured by the (*expected*) excess population risk, that is  $\mathbb{E}_{\mathcal{A}}[L_{\mathcal{D}}(w^{\text{priv}})] - \min_{w \in \mathcal{W}} L_{\mathcal{D}}(w)$ , where the expectation of  $\mathcal{A}$  is taken over all the randomness of the algorithm. Besides the population risk, we can also measure the *empirical risk* of dataset  $D$ :  $\hat{L}(w, D) = \frac{1}{n} \sum_{i=1}^n \ell(w, x_i)$ .

**Definition 3.2.2.** A random variable  $X$  with mean  $\mu$  is called  $\tau$ -sub-exponential if

$$\mathbb{E}[\exp(\lambda(X - \mu))] \leq \exp\left(\frac{1}{2}\tau^2\lambda^2\right), \forall |\lambda| \leq \frac{1}{\tau}.$$

**Assumption 3.2.1.** For the loss function and the population risk, we assume the following.

1. The loss function  $\ell(w, x)$  is non-negative, differentiable and convex for all  $w \in \mathcal{W}$  and  $x \in \mathcal{X}$ .
2. The population risk  $L_{\mathcal{D}}(w)$  is  $\beta$ -smooth.
3. The convex constraint set  $\mathcal{W}$  is bounded with diameter  $\Delta = \max_{w, w' \in \mathcal{W}} \|w - w'\|_2 < \infty$ .
4. The optimal solution  $w^* = \arg \min_{w \in \mathcal{W}} L_{\mathcal{D}}(w)$  satisfies  $\nabla L_{\mathcal{D}}(w^*) = 0$ .

**Assumption 3.2.2.** There exists a number  $n_\alpha$  such that when the sample size  $|D| \geq n_\alpha$ , the empirical risk  $\hat{L}(\cdot, D)$  is  $\alpha$ -strongly convex with probability at least  $\frac{5}{6}$  over the choice of i.i.d. samples in  $D$ .

We note that Assumptions 3.2.1 and 3.2.2 are commonly used in the studies on the

problem of Stochastic Strongly Convex Optimization with heavy-tailed data, such as [151, 148]. Also the probability of  $\frac{5}{6}$  in Assumption 3.2.2 is only for convenience.

**Assumption 3.2.3.** We assume the following for the loss functions.

1. For any  $w \in \mathcal{W}$  and each coordinate  $j \in [d]$ , we assume that the random variable  $\nabla_j \ell(w, x)$  is  $\tau$ -sub-exponential and  $\beta_j$ -Lipschitz (that is  $\ell_j(w, x)$  is  $\beta_j$ -smooth), where  $\nabla_j$  represents the  $j$ -th coordinate of the gradient.
2. There are known constants  $a, b = O(1)$  such that  $a \leq \mathbb{E}[\nabla_j \ell(w, x)] \leq b$  for all  $w \in \mathcal{W}$ .

**Assumption 3.2.4.** For any  $w \in \mathcal{W}$  and each coordinate  $j \in [d]$ , we have  $\mathbb{E}[(\nabla_j \ell(w, x))^2] \leq v = O(1)$ , where  $v$  is some known constant.

We can see that, compared with Assumption 3.2.3, Assumption 3.2.4 needs fewer assumptions on the loss functions, because we only need to assume the gradient of the loss function has bounded second-order moment. We also note that Assumption 3.2.4 is more suitable to the problem of Stochastic Convex Optimization with heavy-tailed data and has been used in some previous works such as [149, 46].

### 3.2.3 Sample-aggregation based method

In this section we first summarize the sample-aggregate framework introduced in [233].

Most of the existing privacy-preserving frameworks are based on the notion of *global sensitivity*, which is defined as the maximum output perturbation  $\|f(D) - f(D')\|_\xi$ , where the maximum is over all neighboring datasets  $D, D'$  and  $\xi = 1, 2$ . However, in some problems such as clustering [233, 337] the sensitivity could be very high and thus ruin the utility of the algorithm.

To circumvent this issue, [233] introduced the sample-aggregate framework based on a smooth version of *local sensitivity*. Unlike the global sensitivity, local sensitivity measures

the maximum perturbation  $\|f(D) - f(D')\|_\xi$  over all databases  $D'$  neighboring the input database  $D$ . The proposed sample-aggregate framework (Algorithm 3.2.5) enjoys local sensitivity and comes with the following guarantee:

**Theorem 3.2.1** (Theorem 4.2 in [233]). Let  $f : \mathcal{D} \mapsto \mathbb{R}^d$  be a function where  $\mathcal{D}$  is the collection of all databases and  $d$  is the dimensionality of the output space. Let  $d_{\mathcal{M}}(\cdot, \cdot)$  be a semi-metric on the output space of  $f$ . Set  $\epsilon > \frac{2d}{\sqrt{m}}$  and  $m = \omega(\log^2 n)$ . The sample-aggregate algorithm  $\mathcal{A}$  in Algorithm 3.2.5 is an efficient  $(\epsilon, \delta)$ -DP algorithm.<sup>4</sup> Furthermore, if  $f$  and  $m$  are chosen such that the  $\ell_1$  norm of the output of  $f$  is bounded by  $\Lambda$  and

$$\Pr_{D_S \subseteq D}[d_{\mathcal{M}}(f(D_S), c) \leq r] \geq \frac{3}{4} \quad (3.60)$$

for some  $c \in \mathbb{R}^d$  and  $r > 0$ , then the standard deviation of Gaussian noise added is upper bounded by  $O\left(\frac{r}{\epsilon} + \frac{\Lambda}{\epsilon} e^{-\Omega(\frac{\epsilon\sqrt{m}}{d})}\right)$ . In addition, when  $m = \omega\left(\frac{d^2 \log^2(r/\Lambda)}{\epsilon^2}\right)$ , with high probability each coordinate of  $\mathcal{A}(D) - \bar{c}$  is upper bounded by  $O\left(\frac{r}{\epsilon}\right)$ , where  $\bar{c}$  depending on  $\mathcal{A}(D)$  satisfies  $d_{\mathcal{M}}(c, \bar{c}) = O(r)$ .

---

**Algorithm 3.2.5** Sample-aggregate Framework [233]

---

**Input:**  $D = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$ , number of subsets  $m$ , privacy parameters  $\epsilon, \delta$ ;  $f, d_{\mathcal{M}}$ .

- 1: **Initialize:**  $s = \sqrt{m}$ ,  $\gamma = \frac{\epsilon}{5\sqrt{2\log(2/\delta)}}$  and  $\beta = \frac{\epsilon}{4(d+\log(2/\delta))}$ .
  - 2: **Subsampling:** Select  $m$  random subsets of size  $\frac{n}{m}$  of  $D$  independently and uniformly at random without replacement. Repeat this step until no single data point appears in more than  $\sqrt{m}$  of the sets. Mark the subsampled subsets  $D_{S_1}, D_{S_2}, \dots, D_{S_m}$ .
  - 3: Compute  $\mathcal{S} = \{s_i\}_{i=1}^m$ , where  $s_i = f(D_{S_i})$ .
  - 4: Compute  $g(\mathcal{S}) = s_{i^*}$ , where  $i^* = \arg \min_{i=1}^m r_i(t_0)$  with  $t_0 = \frac{m+s}{2} + 1$ . Here  $r_i(t_0)$  denotes the distance  $d_{\mathcal{M}}(\cdot, \cdot)$  between  $s_i$  and the  $t_0$ -th nearest neighbor to  $s_i$  in  $\mathcal{S}$ .
  - 5: **Noise Calibration:** Compute  $S(\mathcal{S}) = 2 \max_k (\rho(t_0 + (k+1)s) \cdot e^{-\beta k})$ , where  $\rho(t)$  is the mean of the top  $\lceil \frac{s}{\beta} \rceil$  values in  $\{r_1(t), \dots, r_m(t)\}$ .
  - 6: Return  $\mathcal{A}(D) = g(\mathcal{S}) + \frac{S(\mathcal{S})}{\gamma} u$ , where  $u$  is a standard Gaussian random vector.
- 

We have the following Lemma 3.2.1, which shows that the minimum of the empirical risk satisfies (3.60).

---

<sup>4</sup>Here the efficiency means that the time complexity is polynomial in all terms.

**Lemma 3.2.1.** Let  $w_D = f(D) = \arg \min_{w \in \mathcal{W}} \hat{L}(w, D)$  where  $|D| = n$ . Then, under Assumptions 3.2.1 and 3.2.2, if  $n \geq n_\alpha$ , the following holds

$$\Pr[\|w_D - w^*\|_2 \leq \eta] \geq \frac{3}{4}, \quad (3.61)$$

where  $\eta = O(\sqrt{\frac{\mathbb{E}\|\nabla \ell(w^*, x)\|_2^2}{n\alpha^2}})$ .

Combining Lemma 3.2.1 and Theorem 3.2.1, we get the following upper bound for DP-SCO with heavy-tailed data and strongly convex loss functions.

**Theorem 3.2.2.** Under Assumptions 3.2.1 and 3.2.2, for any  $\epsilon, \delta > 0$ , if  $n \geq \tilde{\Omega}(\frac{n_\alpha d^2}{\epsilon^2})$ ,  $m \geq \tilde{\omega}(\frac{d^2}{\epsilon^2})$ ,  $f(D) = \arg \min_{w \in \mathcal{W}} \hat{L}(w, D)$  and  $d_{\mathcal{M}}(x, y) = \|x - y\|_2$ , then Algorithm 3.2.5 is  $(\epsilon, \delta)$ -DP. Moreover, with high probability the output of  $\mathcal{A}(D)$  ensures that

$$L_{\mathcal{D}}(\mathcal{A}(D)) - L_{\mathcal{D}}(w^*) \leq \tilde{O}\left((\frac{\beta}{\alpha})^2 \frac{d^3}{n\epsilon^4} L_{\mathcal{D}}(w^*)\right), \quad (3.62)$$

where the Big- $\tilde{O}$ ,  $\Omega$  and small- $\omega$  notations omit the logarithmic terms.

**Remark 3.2.1.** For DP-SCO with Lipschitz and strongly-convex loss function and bounded data, [29, 328, 31] showed that the upper bound of the excess population risk is  $O(\frac{\sqrt{d}}{n\epsilon})$ , and the lower bound is  $\Omega(\frac{d}{n^2\epsilon^2})$ <sup>5</sup>. This suggests that the bound in Theorem 3.2.2 has some additional factors related to  $d$  and  $\frac{1}{\epsilon}$ . We note that the upper bound in Theorem 3.2.2 has a multiplicative term of  $L_{\mathcal{D}}(w^*)$ . This means that when  $L_{\mathcal{D}}(w^*)$  is small, our bound is better. For example, when  $L_{\mathcal{D}}(w^*) = 0$ , our algorithm can recover  $w^*$  exactly and results in an excess risk of 0. Notice that there is no previous work on DP-ERM or DP-SCO that has a multiplicative error with respect to  $L_{\mathcal{D}}(w^*)$ .

---

<sup>5</sup>[29] only shows the lower bound of the excess empirical risk. We can obtain the lower bound of the excess population risk by using the reduction from private ERM to private SCO [31].

### 3.2.4 Gradient descent based methods

There are several issues in the sample-aggregation based method presented in last section. Firstly, function  $f(D)$  in Theorem 3.2.2 needs to solve the optimization problem exactly, which could be quite inefficient in practice. Second, previous empirical evidence suggests that sample-aggregation based methods often suffer from poor utility in practice [265, 337]. Thirdly, Theorem 3.2.2 needs to assume strong convexity for the empirical risk and it is unclear whether it can be extended to the general convex case. Finally, from Eq.(3.62) we can see that when  $L_D(w^*) = \Theta(1)$ , the excess population risk is quite large as compared to the ones in [29]. Thus, an immediate question is whether we can further lower the upper bound. To answer this question and resolve the above issues, we propose in this section two DP algorithms based on the Gradient Descent method under different assumptions.

Recently, [51] studied the problem of estimating the mean of a 1-dimensional heavy-tailed distribution and proposed algorithms based on the idea of truncating the empirical mean and the local sensitivity. Motivated by this DP algorithm that has the capability of handling heavy-tailed data, we plan to develop a new method by borrowing some ideas from the work [51] and robust gradient descent. Our method is inspired by their theorem that follows and uses the Arsinh-Normal mechanism (see Algorithm 3.2.6 and Prop. 5 in [51]).

**Theorem 3.2.3** (Theorem 7 in [51]). Let  $0 < \epsilon, \delta \leq 1$  be two constants and  $n$  be some integer  $\geq O(\log(\frac{n(b-a)/\sigma}{\epsilon}))$ . Then, there exists a  $\frac{1}{2}\epsilon^2$ -zero concentrated Differentially Private (zCDP) (see Appendix for the definition of zCDP) algorithm (Algorithm 3.2.6)  $M : \mathbb{R}^n \mapsto \mathbb{R}$  such that the following holds: Let  $\mathcal{D}$  be a distribution with mean  $\mu \in [a, b]$ , where  $a, b$  are given constants and unknown variance  $\sigma^2$ . Then,

$$\mathbb{E}_{X \sim \mathcal{D}^n, Z}[(M(X) - \mu)^2] \leq O\left(\frac{\sigma^2 \log n}{n\epsilon^2}\right).$$

The key idea of our algorithm is that, in each iteration, after getting  $w^{t-1}$ , we use the mechanism in Theorem 3.2.3 on each coordinate of  $\nabla \ell(w, x_i)$ . See Algorithm 3.2.7 for

details. By the composition theorem and the relationship between  $zCDP$  and  $(\epsilon, \delta)$ -DP

---

**Algorithm 3.2.6** Mechanism  $\mathcal{M}$  in [51]

---

**Input:**  $D = \{x_i\}_{i=1}^n \subset \mathbb{R}, \epsilon, a, b$ .

- 1: Let  $t = \frac{\epsilon^2}{16}$  and  $s = \frac{\epsilon}{4}$ . Sort  $\{x_i\}_{i=1}^n$  in the ascending order as  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ . Calculate the upper bound of the smooth sensitivity for the trimming and truncating step:

$$S_{[\text{trim}_m(\cdot)]_{[a,b]}}^t(D) = \max\left\{\frac{x_{(n)} - x_{(1)}}{n - 2m}, e^{-mt}(b - a)\right\},$$

where  $m = O(1) \leq \frac{n}{2}$  is a constant.

- 2: Do the average trimming and truncating step:

$$[\text{Trim}_m(D)]_{[a,b]} = \left[\frac{x_{(m+1)} + \dots + x_{(n-m)}}{n - 2m}\right]_{[a,b]},$$

where  $[x]_{[a,b]} = x$  if  $a \leq x \leq b$ , equals to  $a$  if  $x < a$  and otherwise equals to  $b$ .

- 3: Output  $[\text{Trim}_m(D)]_{[a,b]} + \frac{1}{s} S_{[\text{trim}_m(\cdot)]_{[a,b]}}^t(D) \cdot Z$ , where  $Z = \sinh(Y) = \frac{e^Y - e^{-Y}}{2}$  and  $Y$  is the Standard Gaussian.
- 

**Algorithm 3.2.7** Heavy-tailed DP-SCO with known mean

---

**Input:**  $D = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$ , privacy parameters  $\epsilon, \delta$ ; loss function  $\ell(\cdot, \cdot)$ , initial parameter  $w^0, a, b$  which satisfy Assumption 3.2.3, and the number of iterations  $T$  (to be specified later).

- 1: Let  $\tilde{\epsilon} = \sqrt{2 \log \frac{1}{\delta}} + 2\epsilon - \sqrt{2 \log \frac{1}{\delta}}$ .
  - 2: **for**  $t = 1, 2, \dots, T$  **do**
  - 3:     For each  $j \in [d]$ , calculate  

$$D_{t-1,j}(w^{t-1}) = \{\nabla_j \ell(w^{t-1}, x_i)\}_{i=1}^n$$
  - 4:     Run Algorithm 3.2.6 for each  $D_{t-1,j}$  and denote the output  

$$\tilde{\nabla}_{t-1,j}(w^{t-1}) = (\mathcal{M}(D_{t-1,j}(w^{t-1})), \frac{\tilde{\epsilon}}{\sqrt{dT}}, a, b)$$
. Denote  

$$\nabla \tilde{L}(w^{t-1}, D) = (\tilde{\nabla}_{t-1,1}(w^{t-1}), \dots, \tilde{\nabla}_{t-1,d}(w^{t-1})).$$
  - 5:     Updating  $w^t = \mathcal{P}_{\mathcal{W}}(w^{t-1} - \eta_{t-1} \nabla \tilde{L}(w^{t-1}, D))$ , where  $\eta_{t-1}$  is some step size and  $\mathcal{P}_{\mathcal{W}}$  is the projection operator.
  - 6: **end for**
- 

[52], we have the DP guarantee.

**Theorem 3.2.4.** For any  $0 < \epsilon, \delta \leq 1$ , Algorithm 3.2.7 is  $(\epsilon, \delta)$ -differentially private.

To show the *expected* excess population risk of Algorithm 3.2.7, we cannot use the upper bound in Theorem 3.2.3 directly for the following reasons. First, since the upper bound is

for the expectation w.r.t.  $X$  and  $Z$  while the *expected* excess population risk depends only on the randomness of the algorithm instead of the data. Thus, we need to obtain an upper bound for  $\mathbb{E}_Z[(M(X) - \mu)^2]$  (with high probability w.r.t.  $X$ ). Secondly, to get an upper bound, it is sufficient to analyze the term  $\|\nabla \tilde{L}(w^{t-1}, D) - \nabla L_{\mathcal{D}}(w^{t-1})\|_2$  in each iteration. However, since the parameter  $w^{t-1}$  at any step depends on the random draw of the dataset  $\{x_i\}_{i=1}^n$ , upper bounds on the estimation error need to be uniform in  $w \in \mathcal{W}$  in order to capture all contingencies. To resolve these two issues, we use the same technique as in [74, 289] (under Assumption 3.2.3) to obtain the following lemma.

**Lemma 3.2.2.** Under Assumption 3.2.3, with probability at least  $1 - \frac{2dn}{(1+n\hat{\beta}\Delta)^d}$  the following holds for all  $w \in \mathcal{W}$ ,

$$\mathbb{E}_Z \|\nabla \tilde{L}(w, D) - \nabla L_{\mathcal{D}}(w)\|_2 \leq O\left(\frac{\tau d \sqrt{T \log n}}{\sqrt{n\tilde{\epsilon}}}\right), \quad (3.63)$$

where  $\hat{\beta} = \sqrt{\beta_1^2 + \dots + \beta_d^2}$ , the expectation is w.r.t. the random variables  $\{Z_i\}_{i=1}^d$  and the Big- $O$  notation omits other factors.

Next, we show the expected excess population risk for strongly convex loss functions.

**Theorem 3.2.5** (Strongly-convex case). Under Assumptions 3.2.1 and 3.2.3, if the population risk is  $\alpha$ -strongly convex and  $T$  and  $\eta$  are set to be  $T = O(\frac{\beta}{\alpha} \log n)$  and  $\eta = \frac{1}{\beta}$ , respectively, in Algorithm 3.2.7, then with probability at least  $1 - \Omega(\frac{\beta}{\alpha} \frac{2dn \log n}{(1+n\hat{\beta}\Delta)^d})$  the output satisfies the following for all  $D \sim \mathcal{D}^n$ ,

$$\mathbb{E}[L_{\mathcal{D}}(w^T)] - L_{\mathcal{D}}(w^*) \leq O\left(\frac{\Delta^2 \beta^2 \tau^2 d^2 \log^2 n \log \frac{1}{\delta}}{\alpha^3 n \epsilon^2}\right).$$

Compared with the bound in Theorem 3.2.2, we can see that the bound in Theorem 3.2.5 improves a factor of  $\tilde{O}(\frac{d}{\epsilon^2})$  (if we omit other terms). However, there are more assumptions on the distribution and the loss functions. Specifically, in Assumption 3.2.3 we need to assume the sub-exponential property, *i.e.*, the moment of  $\nabla_j \ell(w, x)$  exists for every order.

Also, we need to assume that  $\nabla_j \ell(w, x)$  is Lipschitz and the range of its mean is known. These assumptions are quite strong, compared to those used in the literature of learning with heavy-tailed data, such as [149, 46, 151, 220].

To improve the above result, we consider the following. First, we would like to relax those assumptions in the theorem. Second, in the problem of ERM with heavy-tailed data, it is expected to have an excess population risk bound that is in the form of *with high probability* instead of its *expectation* [46]. However, it is unclear whether Algorithm 3.2.7 can achieve a high probability bound. This is due to the fact that the noise added in each iteration is a combination of log-normal distributions, which is non-sub-exponential and thus is hard to get tail bounds. Third, Algorithm 3.2.7 depends on the local sensitivity and thus cannot be extended to the distributed settings or local differential privacy model. Finally, the practical performance of Algorithm 3.2.7 has poor utility and is unstable due to the noise added in each iteration (see Section 6 for details), which means that Algorithm 3.2.7 is still impractical. To resolve all these issues and still keeping (approximately) the same upper bound, we propose a new algorithm that is simply based on the Gaussian mechanism.

In the following we will study the problem under Assumptions 1 and 3.2.4. Note that compared with Assumption 3.2.3, we only need to assume that the second-order moment of  $\nabla_j \ell(w, x)$  exists for all  $w \in \mathcal{W}$  and  $j \in [d]$  and its upper bound is known.

Our method is motivated by the robust mean estimator given in [148]. To be self-contained, we first review their estimator. Now, we consider 1-dimensional random variable  $x$  and assume that  $x_1, x_2, \dots, x_n$  are i.i.d. sampled from  $x$ . The estimator consists of the following steps:

**Scaling and Truncation** For each sample  $x_i$ , we first re-scale it by dividing  $s$  (which will be specified later). Then, we apply the re-scaled one to some soft truncation function  $\phi$ .

Finally, we put the truncated mean back to the original scale. That is,

$$\frac{s}{n} \sum_{i=1}^n \phi\left(\frac{x_i}{s}\right) \approx \mathbb{E}X. \quad (3.64)$$

Here, we use the function given in [62],

$$\phi(x) = \begin{cases} x - \frac{x^3}{6}, & -\sqrt{2} \leq x \leq \sqrt{2} \\ \frac{2\sqrt{2}}{3}, & x > \sqrt{2} \\ -\frac{2\sqrt{2}}{3}, & x < -\sqrt{2}. \end{cases} \quad (3.65)$$

Note that a key property for  $\phi$  is that  $\phi$  is bounded, that is,  $|\phi(x)| \leq \frac{2\sqrt{2}}{3}$ .

**Noise Multiplication** Let  $\eta_1, \eta_2, \dots, \eta_n$  be random noise generated from a common distribution  $\eta \sim \chi$  with  $\mathbb{E}\eta = 0$ . We multiply each data  $x_i$  by a factor of  $1 + \eta_i$ , and then perform the scaling and truncation step on the term  $x_i(1 + \eta_i)$ . That is,

$$\tilde{x}(\eta) = \frac{s}{n} \sum_{i=1}^n \phi\left(\frac{x_i + \eta_i x_i}{s}\right). \quad (3.66)$$

**Noise Smoothing** In this final step, we smooth the multiplicative noise by taking the expectation w.r.t. the distributions. That is,

$$\hat{x} = \mathbb{E}\tilde{x}(\eta) = \frac{s}{n} \sum_{i=1}^n \int \phi\left(\frac{x_i + \eta_i x_i}{s}\right) d\chi(\eta_i). \quad (3.67)$$

Computing the explicit form of each integral in (3.67) depends on the function  $\phi(\cdot)$  and the distribution  $\chi$ . Fortunately, [62] showed that when  $\phi$  is in (3.65) and  $\chi \sim \mathcal{N}(0, \frac{1}{\beta})$  (where  $\beta$  will be specified later), we have for any  $a, b$

$$\mathbb{E}_\eta \phi(a + b\sqrt{\beta}\eta) = a\left(1 - \frac{b^2}{2}\right) - \frac{a^3}{6} + C(a, b), \quad (3.68)$$

where  $C(a, b)$  is a correction form which is easy to implement and its explicit form can be calculated as the followings: We first define the following notations:

$$V_- := \frac{\sqrt{2} - a}{b}, V_+ = \frac{\sqrt{2} + a}{b} \quad (3.69)$$

$$F_- := \Phi(-V_-), F_+ := \Phi(-V_+) \quad (3.70)$$

$$E_- := \exp\left(-\frac{V_-^2}{2}\right), E_+ := \exp\left(-\frac{V_+^2}{2}\right), \quad (3.71)$$

where  $\Phi$  denotes the CDF of the standard Gaussian distribution. Then we have

$$C(a, b) = T_1 + T_2 + \cdots + T_5, \quad (3.72)$$

where

$$T_1 := \frac{2\sqrt{2}}{3}(F_- - F_+) \quad (3.73)$$

$$T_2 := -(a - \frac{a^3}{6})(F_- + F_+) \quad (3.74)$$

$$T_3 := \frac{b}{\sqrt{2\pi}}(1 - \frac{a^2}{2})(E_+ - E_-) \quad (3.75)$$

$$T_4 := \frac{ab^2}{2} \left( F_+ + F_- + \frac{1}{\sqrt{2\pi}}(V_+ E_+ + V_- E_-) \right) \quad (3.76)$$

$$T_5 := \frac{b^3}{6\sqrt{2\pi}} ((2 + V_-^2)E_- - (2 + V_+^2)E_+). \quad (3.77)$$

[148] showed the following estimation error for the mean estimator  $\hat{x}$  after these three steps.

**Lemma 3.2.3** (Lemma 5 in [148]). Let  $x_1, x_2, \dots, x_n$  be i.i.d. samples from distribution  $x \sim \mu$ . Assume that there is some known upper bound on the second-order moment, *i.e.*,  $\mathbb{E}_\mu x^2 \leq v$ . For a given failure probability  $\delta'$ , if set  $\beta = 2 \log \frac{1}{\delta'}$  and  $s = \sqrt{\frac{nv}{2 \log \frac{1}{\delta'}}}$ , then with

---

**Algorithm 3.2.8** Heavy-tailed DP-SCO with known variance

---

**Input:**  $D = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$ , privacy parameters  $\epsilon, \delta$ , loss function  $\ell(\cdot, \cdot)$ , initial parameter  $w^0, v$  which satisfies Assumption 3.2.4, the number of iterations  $T$  (to be specified later), and failure probability  $\delta'$ .

- 1: Let  $\tilde{\epsilon} = (\sqrt{\log \frac{1}{\delta}} + \epsilon) - \sqrt{\log \frac{1}{\delta}}$ ,  $s = \sqrt{\frac{nv}{2 \log \frac{1}{\delta'}}}$ ,  $\beta = \log \frac{1}{\delta'}$ .
- 2: **for**  $t = 1, 2, \dots, T$  **do**
- 3:     For each  $j \in [d]$ , calculate the robust gradient by (3.66)-(3.68), that is

$$\begin{aligned} g_j^{t-1}(w^{t-1}) &= \frac{1}{n} \sum_{i=1}^n \left( \nabla_j \ell(w^{t-1}, x_i) \left( 1 - \frac{\nabla_j^2 \ell(w^{t-1}, x_i)}{2s^2 \beta} \right) - \frac{\nabla_j^3 \ell(w^{t-1}, x_i)}{6s^2} \right) \\ &\quad + \frac{s}{n} \sum_{i=1}^n C \left( \frac{\nabla_j \ell(w^{t-1}, x_i)}{s}, \frac{|\nabla_j \ell(w^{t-1}, x_i)|}{s\sqrt{\beta}} \right) + Z_j^{t-1}, \end{aligned} \quad (3.78)$$

where  $Z_j^{t-1} \sim \mathcal{N}(0, \sigma^2)$  with  $\sigma^2 = \frac{8vdT}{9 \log \frac{1}{\delta'} n \tilde{\epsilon}}$ .

- 4:     Let vector  $g^{t-1}(w^{t-1}) \in \mathbb{R}^d$  to denote  $g^{t-1}(w^{t-1}) = (g_1^{t-1}(w^{t-1}), g_2^{t-1}(w^{t-1}), \dots, g_d^{t-1}(w^{t-1}))$ .
  - 5:     Update  $w^t = \mathcal{P}_W(w^{t-1} - \eta_{t-1} g^{t-1})$ .
  - 6: **end for**
- 

probability at least  $1 - \delta'$  the following holds

$$|\hat{x} - \mathbb{E}x| \leq O\left(\sqrt{\frac{v \log \frac{1}{\delta'}}{n}}\right). \quad (3.79)$$

To obtain an  $(\epsilon, \delta)$ -DP estimator, the key observation is that the bounded function  $\phi$  in (3.65) also makes the integral form of (3.68) bounded by  $\frac{2\sqrt{2}}{3}$ . Thus, we know that the  $\ell_2$ -norm sensitivity is  $\frac{s}{n} \frac{4\sqrt{2}}{3}$ . Hence, the query

$$\mathcal{A}(D) = \hat{x} + Z, Z \sim \mathcal{N}(0, \sigma^2), \sigma^2 = O\left(\frac{s^2 \log \frac{1}{\delta}}{\epsilon^2 n^2}\right) \quad (3.80)$$

will be  $(\epsilon, \delta)$ -DP, which leads to the following theorem.

**Theorem 3.2.6.** Under the assumptions in Lemma 3.2.3, with probability at least  $1 - \delta'$  the following holds

$$|\mathcal{A}(D) - \mathbb{E}(x)| \leq O\left(\sqrt{\frac{v \log \frac{1}{\delta} \log \frac{1}{\delta'}}{n \epsilon^2}}\right). \quad (3.81)$$

Comparing with Theorem 3.2.3, we can see that the upper bound in Theorem 3.2.6 is in the form of ‘with high probability’ (after transferring zCDP to  $(\epsilon, \delta)$ -DP [52]). Moreover, we improve by a factor of  $O(\log n)$  in the error bound.

Inspired by Theorem 3.2.6 and Algorithm 3.2.7, we propose a new method (Algorithm 3.2.8), which uses our private mean estimator (3.80) on each coordinate of the gradient in each iteration. The following theorem shows the error bound when the loss function is strongly convex.

**Theorem 3.2.7.** For any  $0 < \epsilon, \delta < 1$ , Algorithm 3.2.8 is  $(\epsilon, \delta)$ -DP. Under Assumptions 3.2.1 and 3.2.4, if the population risk is  $\alpha$ -strongly convex and  $\eta_t$  and  $T$  in Algorithm 3.2.8 are set to be  $\eta_t = \frac{1}{\beta}$  and  $T = O(\frac{\beta}{\alpha} \log n)$ , respectively, then for any  $\delta' > 0$ , with probability at least  $1 - 2\delta'T$  the output  $w^T$  satisfies

$$L_{\mathcal{D}}(w^T) - L_{\mathcal{D}}(w^*) \leq O\left(\frac{v\Delta^2\beta^4d^2\log^2 n \log \frac{1}{\delta} \log \frac{1}{\delta'}}{\alpha^3 n \epsilon^2}\right).$$

Comparing with Theorem 3.2.7 and 3.2.5, we can see that if we omit other terms, the bounds are asymptotically the same and Theorem 3.2.7 needs fewer assumptions.

With the high probability guarantee on the error in Theorem 3.2.6, we can actually get an upper bound for general convex loss functions. For this general convex case, we need the following mild technical assumption on the constraint set  $\mathcal{W}$ .

**Assumption 3.2.5.** The constraint set  $\mathcal{W}$  contains the following  $\ell_2$ -ball centered at  $w^*$ :

$$\{w : \|w - w^*\|_2 \leq 2\|w^0 - w^*\|_2\}.$$

**Theorem 3.2.8 (Convex case).** Under Assumptions 3.2.1, 3.2.4 and 3.2.5, if we take  $\eta = \frac{1}{\beta}$  and  $T = \tilde{O}\left(\frac{\|w^0 - w^*\|_2 \sqrt{n} \sqrt{\epsilon}}{d}\right)^{\frac{2}{3}}$  in Algorithm 3.2.8, then for any given failure probability  $\delta'$ , with probability at least  $1 - T\delta'$  the following holds

$$L_{\mathcal{D}}(w^T) - L_{\mathcal{D}}(w^*) \leq \tilde{O}\left(\frac{\log^{\frac{1}{3}} \frac{1}{\delta} \sqrt{\log \frac{1}{\delta'} d^{\frac{2}{3}}}}{(n\epsilon^2)^{\frac{1}{3}}}\right) \quad (3.82)$$

when  $n \geq \tilde{\Omega}(\frac{d^2}{\epsilon^2})$ , where the Big- $\tilde{O}$  notation omits other logarithmic factors and the term of  $v, \beta$ .

The problem is far from being closed. First, it is unclear whether the upper bounds of the excess population risk for strongly convex and general convex loss functions can be further improved. The second open problem is that we do not know what the lower bound for the excess population risk for these two cases is. Finally, it is an open problem to determine whether we can further relax the assumptions in our previous theorems. We leave these open problems for future research.

### 3.2.5 Experiments

**Baseline Methods** As mentioned earlier, sample-aggregation based methods often have poor practical performance. Thus, we will not conduct experiments on Algorithm 3.2.5. Moreover, as this is the first paper studying DP-SCO with heavy-tailed data and almost all previous methods on DP-SCO that have theoretical guarantees fail to provide DP guarantees, we do not compare our methods with them, and instead focus on comparing the performance of Algorithm 3.2.7 and Algorithm 3.2.8. To show the effectiveness of our methods, we use the non-private heavy-tailed SCO method in [148], denoted by (stochastic) RGD in the following, as our baseline method.

**Experimental Settings** For synthetic data, we consider the linear and binary logistic models. Specifically, we generate the synthetic datasets in the following way. Each dataset has a size of  $1 \times 10^5$  and each data point  $(x_i, y_i)$  is generated by the model of  $y_i = \langle \omega^*, x_i \rangle + e_i$  and  $y_i = \text{sign}[\frac{1}{1+e^{\langle \omega^*, x_i \rangle + e_i}} - \frac{1}{2}]$ , respectively, where  $x_i \in \mathbb{R}^{10}$  and  $y_i \in \mathbb{R}$ . In the first model, the zero mean noise  $e_i$  is generated as follows. We first generate a noise  $\Delta_i$  from the  $(\mu, \sigma)$  log-normal distribution, *i.e.*,  $\mathbb{P}(\Delta_i = x) = \frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$ , and then let  $e_i = \Delta_i - \mathbb{E}[\Delta_i]$ . For the second model, we first generate a noise  $\Delta_i$  from the  $(\mu, \sigma)$  log-logistic distribution, *i.e.*,  $\mathbb{P}(\Delta_i = x) = \frac{e^z}{\sigma x(1+e^z)^2}$ , where  $x > 0$  and  $z = \frac{\log(x) - \mu}{\sigma}$ . Then, we let  $e_i = \Delta_i - \mathbb{E}[\Delta_i]$ .

Accordingly, we implement Algorithm 3.2.7 and Algorithm 3.2.8, together with RGD, on the ridge and logistic regressions.

For the synthetic data generation, we select the parameters  $(\mu = 1, \sigma = 1)$  and  $(\mu = 0.2, \sigma = 0.2)$  for the Lognormal and Loglogistic noises underlying, respectively. The step size of Algorithm 3 is set to 0.01 where  $m = 0.05n$ . As for algorithm 4,  $v = 5$ , failure probability  $\delta' = 0.01$  and the step size is set to 0.1. For the stochastic Algorithm 4, the step size is selected as  $\frac{1}{\sqrt{t}}$ , where  $t$  is the iteration number. Accordingly,  $\bar{w}^T = \frac{\sum_{t=1}^T w^t}{T}$ . Corresponding to Fig. 1 and 2, we present the results which also mark the difference between the best and the worst performances as follows.

To measure the impact from dimension on performances, we fix  $n = 10^5$  and test  $d$  varying from 10 to 50 through stochastic Algorithm 4 and RGD under the same setup as above. To test the impact from the size of the dataset, we fix  $d = 20$  and test  $n$  varying from  $2 \times 10^4$  to  $10^5$ .

For real-world data, we use the Adult dataset from the UCI Repository [94]. We aim to predict whether the annual income of an individual is above 50,000. We select 30,000 samples, 28,000 amongst which are used as the training set and the rest are used for test.

For the privacy parameters, we will choose  $\epsilon = \{0.1, 0.5, 1\}$  and  $\delta = O(\frac{1}{n})$ . See Appendix for the selections of other parameters. For Algorithm 3.2.7, the strength of prior knowledge is modeled by  $\kappa = b - a$ .

**Experimental Results** Figure 3.5 and 3.6 show the results of ridge and logistic regressions on synthetic and real datasets w.r.t iteration, respectively. Since there is no ground truth in the real dataset, we use the empirical risk on test data as the measurement. To test scalability of Algorithm 3.2.8 dealing with large-scaling data, experiments on stochastic versions of Algorithm 3.2.8 and RGD with minibatch size 1000 are also conducted. We can see that the performance of Algorithm 3.2.7 bears a larger variation compared to Algorithm 4, since we have to apply a heavy-tailed noise to fit the smooth sensitivity. Moreover, the performance of

Algorithm 3.2.7 is sensitive to the parameter  $\kappa$ . Thus, these results show that Algorithm 3.2.7 has poor performance and the results of Algorithm 3.2.8 are comparable to the non-private ones. In Figure 3.7 and 3.8 we test the estimation error w.r.t different dimensionality  $d$  and sample size  $n$ , respectively. From these results we can see that when  $n$  increases or  $d$  decreases, the estimation error will decrease. Also, with fixed  $n$  and  $d$ , we can see that the estimation error will decrease as  $\epsilon$  becomes larger. Thus, all these results confirm our previous theoretical analysis.

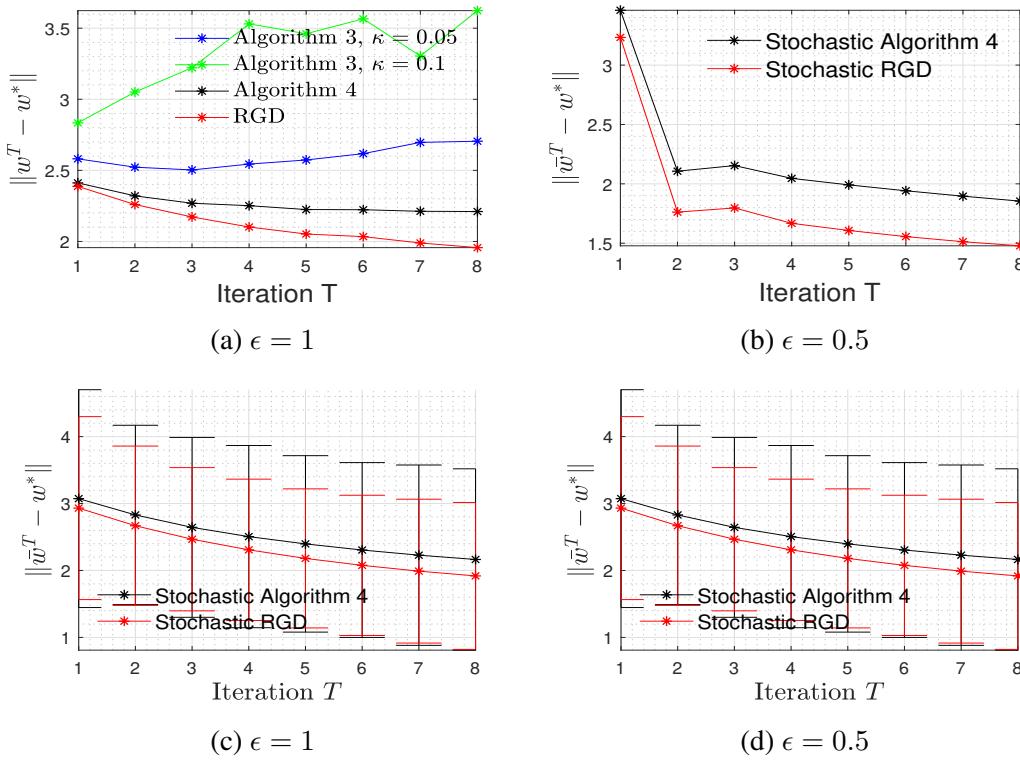


Figure 3.5: Experiments on synthetic datasets. Figures 3.5a and 3.5b are for ridge regressions over synthetic data with Lognormal noises. Figures 3.5c and 3.5d are for logistic regressions over synthetic data with Loglogistic noises.

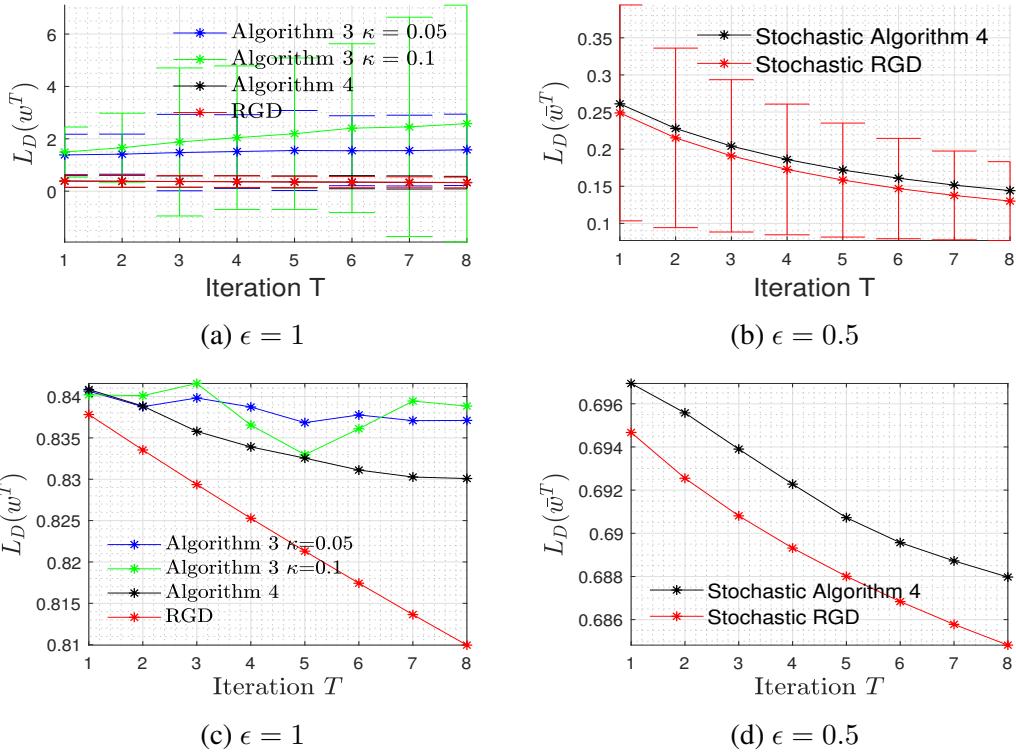


Figure 3.6: Experiments on UCI Adult dataset. Figures 3.6a and 3.6b are for ridge regressions. Figures 3.6c and 3.6d are for logistic regressions.

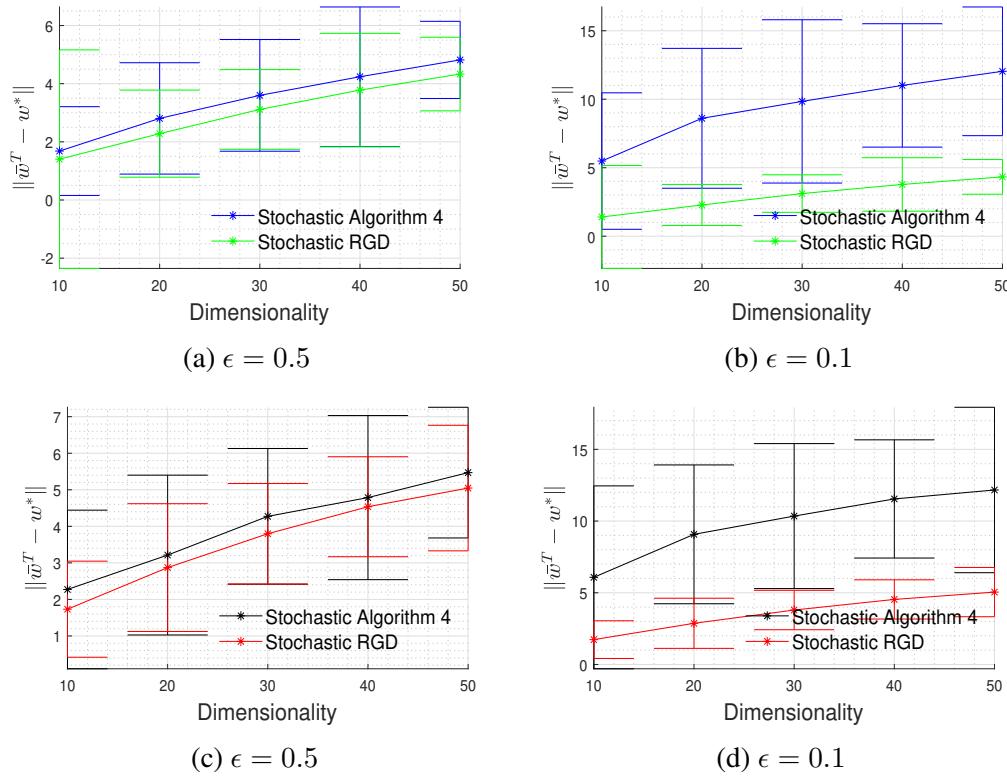


Figure 3.7: Experiments for the impact of dimensionality. Figure 3.7a and 3.7b are for ridge regressions. Figure 3.7c and 3.7d are for logistic regressions.

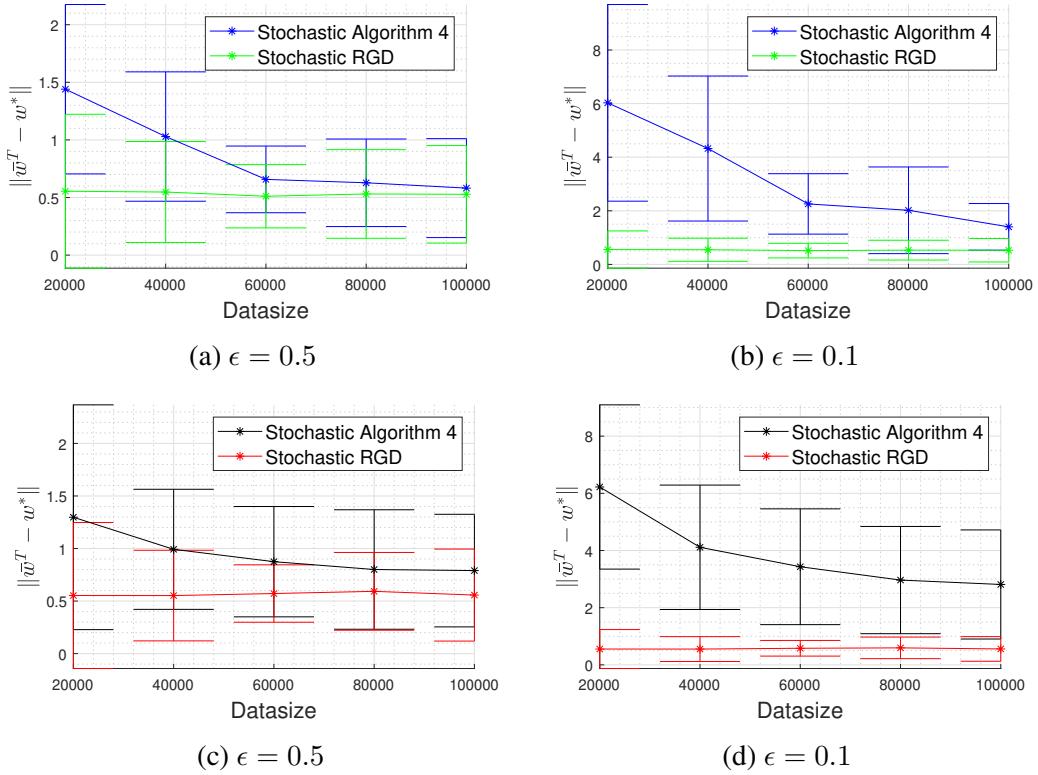


Figure 3.8: Experiments for the impact of the size of the dataset. Figure 3.8a and 3.8b are for ridge regressions. Figure 3.8c and 3.8d are for logistic regressions.

### 3.2.6 Omitted Proofs

#### Proof of 3.2.1

Before the proof, we recall the following two lemmas

**Lemma 3.2.4** ([262]). If a non-negative function  $f : \mathcal{W} \mapsto \mathbb{R}_+$  is  $\beta$ -smooth, then  $\|\nabla f(w)\|_2^2 \leq 4\beta f(w)$  for all  $w \in \mathcal{W}$ .

**Lemma 3.2.5** ([171]). Let  $X_1, X_2, \dots, X_n$  be independent copies of a zero-mean random vector  $X$ , then  $\mathbb{E}\left\|\frac{1}{n} \sum_{i=1}^n X_i\right\|_2^2 \leq \frac{1}{n} \mathbb{E}\|X\|_2^2$ .

Consider  $w = w^*$ . Then by Assumption 3.2.1, we have  $\nabla L(w^*) = \mathbb{E}[\nabla \ell(w^*, x)] = 0$ . Thus, by Lemma 3.2.5 we have

$$\mathbb{E}\|\nabla \hat{L}(w^*, D)\|_2^2 \leq \frac{1}{n} \mathbb{E}[\|\nabla \ell(w^*, x)\|_2^2].$$

By Markov's inequality, we get

$$\Pr[\|\nabla \hat{L}(w^*, D)\|_2^2 \leq \frac{10}{n} \mathbb{E}[\|\nabla \ell(w^*, x)\|_2^2]] \geq \frac{9}{10}.$$

Since  $n \geq n_\alpha$ , by the assumption we have with probability at least  $\frac{5}{6}$  that  $\hat{L}(w, D)$  is  $\alpha$  strongly convex. Thus, we get

$$\begin{aligned} \frac{\alpha}{2} \|w_D - w^*\|_2^2 &\leq -\langle \nabla \hat{L}(w^*, D), w_D - w^* \rangle + \hat{L}(w_D, D) - \hat{L}(w^*, D) \\ &\leq \|\nabla \hat{L}(w^*, D)\|_2 \|w_D - w^*\|_2. \end{aligned}$$

In total, with probability at least  $\frac{3}{4}$ , we have

$$\|w_D - w^*\|_2 \leq \sqrt{\frac{40 \mathbb{E}[\|\nabla \ell(w^*, x)\|_2^2]}{n \alpha^2}}.$$

### Proof of Theorem 3.2.2

For each subsample set  $D_{S_i}$ , by the assumption we have its size  $\frac{n}{m} \geq n_\alpha$ . Thus, Lemma 3.2.1 holds with  $n = \frac{n}{m}$ . That is, (3.60) holds with  $r = \sqrt{\frac{40m\mathbb{E}\|\nabla\ell(w^*, x)\|_2^2}{n\alpha^2}}$ . Hence, by Theorem 3.2.1 we have

$$\|\mathcal{A}(D) - w^*\|_2 \leq O\left(\frac{\sqrt{dr}}{\epsilon}\right) = O\left(\sqrt{\frac{dm\mathbb{E}\|\nabla\ell(w^*, x)\|_2^2}{n\epsilon^2\alpha^2}}\right).$$

Since  $L_{\mathcal{D}}(w)$  is  $\beta$ -smooth and  $\nabla L_{\mathcal{D}}(w^*) = 0$ , we have  $L_{\mathcal{D}}(\mathcal{A}(D)) - L_{\mathcal{D}}(w^*) \leq \frac{\beta}{2}\|\mathcal{A}(D) - w^*\|_2^2$ . Also, by Lemma 3.2.1 and the non-negative property we get

$$L_{\mathcal{D}}(\mathcal{A}(D)) - L_{\mathcal{D}}(w^*) \leq O\left((\frac{\beta}{\alpha})^2 \frac{dm}{n\epsilon^2} L_{\mathcal{D}}(w^*)\right).$$

Taking  $m = \tilde{\Theta}(\frac{d^2}{\epsilon^2})$ , we get the proof.

### Proof of Theorem 3.2.4

We first give the definition of zCDP in [52].

**Definition 3.2.3.** A randomized algorithm  $\mathcal{A} : \mathcal{X}^n \mapsto \mathcal{Y}$  is  $\rho$ -zero Concentrated Differentially Private (zCDP) if for all neighboring datasets  $D \sim D'$  and all  $\alpha \in (1, \infty)$ ,

$$D_\alpha(\mathcal{A}(D)\|\mathcal{A}(D')) \leq \rho\alpha,$$

where  $D_\alpha(P\|Q) = \frac{1}{\alpha-1} \log \mathbb{E}_{X \sim P}[(\frac{P(X)}{Q(X)})^{\alpha-1}]$  denotes the Rényi divergence of order  $\alpha$ .

We first convert  $(\epsilon, \delta)$ -DP to  $\frac{1}{2}\tilde{\epsilon}^2$ -zCDP by using the following lemma

**Lemma 3.2.6** ([52]). Let  $M : \mathcal{X}^n \mapsto \mathcal{Y}$  be a randomized algorithm. If  $M$  is  $\frac{1}{2}\epsilon^2$ -zCDP, it is  $(\frac{1}{2}\epsilon^2 + \epsilon \cdot \sqrt{2 \log \frac{1}{\delta}}, \delta)$ -DP for all  $\delta > 0$ .

Thus, it suffices to show that Algorithm 3.2.7 is  $\frac{1}{2}\tilde{\epsilon}^2$ -zCDP. We note that in each iteration

and each coordinate, outputting  $\nabla_{t-1,j}$  will be  $\frac{1}{2} \frac{\tilde{\epsilon}^2}{dT}$ -zCDP by Theorem 3.2.3. Thus by the composition property of CDP, we know that it is  $\frac{1}{2} \tilde{\epsilon}^2$ -zCDP.

### Proof of Lemma 3.2.2

By assumption, we know that  $\mathcal{W}$  is closed and bounded, and hence it is compact. By [204] we know that its covering number with radius  $\delta$  (will be specified later) is bounded from above as  $N_\delta \leq (\frac{3\Delta}{2\delta})^d$ . Denote the center of this  $\delta$ -net as  $\tilde{\mathcal{W}} = \{\tilde{w}_1, \tilde{w}_2, \dots, \tilde{w}_{N_\delta}\}$ .

We first fix  $j \in [d]$  and consider  $|\tilde{\nabla}_j(w) - \nabla_j L_D(w)|$  (we omit the subscript  $t-1$ ). Then, we have

$$\begin{aligned} & \mathbb{E}_{Z_j}(\tilde{\nabla}_j(w) - \nabla_j L_D(w))^2 \\ &= \mathbb{E}([\text{Trim}_m(D_j(w))]_{[a,b]} + \frac{1}{S} S_{[\text{trim}(\cdot)]_{[a,b]}}^t(D_j(w)) \cdot Z_j - \nabla_j L_D(w))^2 \\ &\leq O(([\text{Trim}_m(D_j(w))]_{[a,b]} - \nabla_j L_D(w))^2 + \mathbb{E}(\frac{1}{S} S_{[\text{trim}(\cdot)]_{[a,b]}}^t(D_j(w)) \cdot Z_j)^2) \\ &\leq O((\text{Trim}_m(D_j(w)) - \nabla_j L_D(w))^2 + \mathbb{E}(\frac{1}{S} S_{[\text{trim}_m(\cdot)]_{[a,b]}}^t(D(w)) \cdot Z_j)^2), \end{aligned} \quad (3.83)$$

where  $D_j(w) = \{\nabla_j \ell(w, x_i)\}_{i=1}^n$  and the last inequality is due to the property that the truncation operation reduces error.

**Lemma 3.2.7.** Let  $a \leq \mu \leq b$  and  $X$  be a random variable. Then

$$([X]_{[a,b]} - \mu)^2 \leq (x - \mu)^2.$$

By the proof of Theorem 51 in [51] and the fact that  $\epsilon = \frac{\tilde{\epsilon}}{\sqrt{dT}}$ , we have ( $m, a, b = O(1)$ )

$$\mathbb{E}_Z(\frac{1}{S} S_{[\text{trim}_m(\cdot)]_{[a,b]}}^t(D_j(w)) \cdot Z)^2 \leq O(\frac{\tau^2 d T \log n}{n \tilde{\epsilon}^2}), \quad (3.84)$$

where the  $O$ -notation omits the  $\log \sigma^2$  and  $\log(b-a)$  factors.

Next, we bound the first term of (3.83). Before showing that, we first give the following

estimation error on the trimming operation for sub-exponential random variables.

**Lemma 3.2.8.** Suppose that  $x_i$  are i.i.d  $v$ -sub-exponential with mean  $\mu$ . Then, the following holds for any  $t \geq 0$ ,

$$\mathbb{P}\left\{\frac{1}{n} \sum_{i=1}^n x_i - \mu \geq t\right\} \leq 2 \exp\left(-n \min\left\{\frac{t}{2v}, \frac{t^2}{2v^2}\right\}\right),$$

and for any  $s \geq 0$ ,

$$\mathbb{P}\left[\max_{i \in [n]} \{|x_i - \mu|\} \geq s\right] \leq 2n \exp\left(-\min\left\{\frac{s}{2v}, \frac{s^2}{2v^2}\right\}\right),$$

and for any  $m \geq 0$ , under the above two events,

$$|\text{Trim}_m(\{x_i\}_{i=1}^n) - \mu| \leq \frac{nt + ms}{n - 2m}.$$

*Proof of Lemma 3.2.8.* Note that the first two inequalities are just the Bernstein's Inequality. We only prove the last inequality.

Let  $\mathcal{T} \subset [n]$  denote the set of all trimmed variables and  $\mathcal{U} = [n] \setminus \mathcal{T}$ . Then, we know that  $\text{Trim}_m(\{x_i\}_{i=1}^n) = \frac{\sum_{i \in \mathcal{U}} x_i}{n - 2m}$ . Thus, we have

$$\begin{aligned} \left|\frac{\sum_{i \in \mathcal{U}} x_i}{n - 2m} - \mu\right| &= \frac{1}{n - 2m} \left| \sum_{i \in [n]} (x_i - \mu) - \sum_{i \in \mathcal{T}} (x_i - \mu) \right| \\ &\leq \frac{1}{n - 2m} \left( \left| \sum_{i \in [n]} (x_i - \mu) \right| + \left| \sum_{i \in \mathcal{T}} (x_i - \mu) \right| \right). \end{aligned} \tag{3.85}$$

For the second term of (3.85), we have  $\left| \sum_{i \in \mathcal{T}} (x_i - \mu) \right| \leq m \max\{|x_i - \mu|\}$ . Plugging the inequalities into (3.85) we get the proof.  $\square$

Now, fix any  $w \in \mathcal{W}$ , we know that there exists a  $\tilde{w}$  which is in the  $\delta$ -net, *i.e.*,  $\|\tilde{w} - w\|_2 \leq \delta$ . Then by using the Bernstein inequality and the sub-exponential assumption and taking the union bound, we can see that with probability at least  $1 - 2dN_\delta \exp(-n \min\{\frac{t}{2\tau}, \frac{t^2}{2\tau^2}\})$ ,

we have the following for all  $j \in [d]$  and  $\tilde{w} \in \tilde{\mathcal{W}}$

$$\left| \sum_{i=1}^n \frac{\nabla_j \ell(\tilde{w}, x_i)}{n} - \nabla_j L_{\mathcal{D}}(\tilde{w}) \right| \leq t, \quad (3.86)$$

and with probability at least  $1 - 2dnN_{\delta} \exp(-\min\{\frac{s}{2\tau}, \frac{s^2}{2\tau^2}\})$ , we get the following for all  $j \in [d]$  and  $\tilde{w} \in \tilde{W}$ ,

$$\max_{i \in [n]} |\nabla_j \ell(\tilde{w}, x_i) - \nabla_j L_{\mathcal{D}}(\tilde{w})| \leq s. \quad (3.87)$$

By the  $\beta_j$ -smoothness of  $\ell_j(\cdot, x)$  we have

$$\left| \sum_{i=1}^n \frac{\nabla_j \ell(\tilde{w}, x_i)}{n} - \sum_{i=1}^n \frac{\nabla_j \ell(w, x_i)}{n} \right| \leq \beta_j \|w - \tilde{w}\|_2 \leq \beta_j \delta, \quad (3.88)$$

$$|\nabla_j L_{\mathcal{D}}(\tilde{w}) - \nabla_j L_{\mathcal{D}}(w)| \leq \beta_j \delta. \quad (3.89)$$

Thus, we get

$$\left| \sum_{i=1}^n \frac{\nabla_j \ell(w, x_i)}{n} - \nabla_j L_{\mathcal{D}}(w) \right| \leq t + 2\beta_j \delta \quad (3.90)$$

$$\max_{i \in [n]} |\nabla_j \ell(w, x_i) - \nabla_j L_{\mathcal{D}}(w)| \leq s + 2\beta_j \delta. \quad (3.91)$$

By Lemma 3.2.8 we have for all  $j \in [d]$  and  $w \in \mathcal{W}$

$$|\text{Trim}_m(D_j(w)) - \nabla_j L_{\mathcal{D}}(w)| \leq \frac{nt + ms}{n - 2m} + \frac{m + n}{n - 2m} 2\beta_j \delta.$$

Combining this with (3.84) we have the following for all  $j \in [d]$  with probability at least

$$1 - 2dnN_{\delta} \exp(-\min\{\frac{s}{2\tau}, \frac{s^2}{2\tau^2}\}) - 2dN_{\delta} \exp(-n \min\{\frac{t}{2\tau}, \frac{t^2}{2\tau^2}\})$$

and  $\tilde{w} \in \tilde{\mathcal{W}}$ ,

$$\mathbb{E}\|\nabla \tilde{L}(w, D) - \nabla L_{\mathcal{D}}(w)\|_2 \leq O\left(\sqrt{d}\frac{nt + ms}{n - 2m} + \hat{\beta}\delta\frac{m + n}{n - 2m} + \frac{\tau d\sqrt{T \log n}}{\sqrt{n\tilde{\epsilon}}}\right), \quad (3.92)$$

where  $\hat{\beta} = \sqrt{\beta_1^2 + \dots + \beta_d^2}$ . Thus, let  $\delta = \frac{1}{n\hat{\beta}}$ ,  $m = O(1)$ ,

$$t = O\left(\tau \max\left\{\frac{d}{n} \log(n\hat{\beta}\Delta), \sqrt{\frac{d}{n} \log(n\hat{\beta}\Delta)}\right\}\right),$$

$$s = O(\tau d \log(\hat{\beta}n\Delta)).$$

Then, we get the proof.

### Proof of Theorem 3.2.5

In the  $t$ -th iteration, let

$$\hat{w}^t = w^{t-1} - \eta \nabla \tilde{L}(w^{t-1}, D).$$

Then, by the property of Euclidean project we have

$$\|w^t - w^{t-1}\|_2 \leq \|\hat{w}^t - w^{t-1}\|_2.$$

Hence, we have

$$\begin{aligned} \|\hat{w}^t - w^*\|_2 &\leq \|w^{t-1} - \eta \nabla \tilde{L}(w^{t-1}, D) - w^*\|_2 \\ &\leq \|w^{t-1} - \eta \nabla L_{\mathcal{D}}(w^{t-1}) - w^*\|_2 \\ &\quad + \eta \|\nabla \tilde{L}(w^{t-1}, D) - \nabla L_{\mathcal{D}}(w^{t-1})\|_2. \end{aligned}$$

For the first term, by the co-coercivity of strongly convex functions [47], we have

$$\langle w^{t-1} - w^*, \nabla L_{\mathcal{D}}(w^{t-1}) \rangle \geq \frac{\alpha\beta}{\alpha + \beta} \|w^{t-1} - w^*\|_2^2 + \frac{1}{\alpha + \beta} \|\nabla L_{\mathcal{D}}(w^{t-1})\|_2^2.$$

Thus we obtain the following by taking  $\eta = \frac{1}{\beta}$

$$\begin{aligned}
& \|w^{t-1} - \eta \nabla L_{\mathcal{D}}(w^{t-1}) - w^*\|_2^2 \leq \\
& (1 - \frac{2\alpha}{\alpha + \beta}) \|w^{t-1} - w^*\|_2^2 - \frac{2}{\beta(\beta + \alpha)} \|\nabla L_{\mathcal{D}}(w^{t-1})\|_2^2 + \frac{1}{\beta^2} \|\nabla L_{\mathcal{D}}(w^{t-1})\|_2^2 \\
& \leq (1 - \frac{2\alpha}{\alpha + \beta}) \|w^{t-1} - w^*\|_2^2.
\end{aligned} \tag{3.93}$$

Taking the expectation w.r.t  $Z_{t-1}$  and using the inequality of  $\sqrt{1-x} \leq 1 - \frac{x}{2}$  and Lemma 4, we have

$$\mathbb{E} \|\hat{w}^t - w^*\|_2 \leq (1 - \frac{\alpha}{\alpha + \beta}) \mathbb{E} \|w^{t-1} - w^*\|_2 + O(\frac{\tau d \sqrt{T \log n}}{\beta \sqrt{n} \tilde{\epsilon}}). \tag{3.94}$$

That is,

$$\mathbb{E} \|\hat{w}^T - w^*\|_2 \leq (1 - \frac{\alpha}{\beta + \alpha})^T \Delta + O(\frac{\beta}{\alpha} \frac{\tau d \sqrt{T \log n}}{\beta \sqrt{n} \tilde{\epsilon}}).$$

Thus, taking  $T = O(\frac{\beta}{\alpha} \log n)$ , we have the following with probability at least  $1 - \Omega(\frac{2dn \log n}{(1+n\hat{L}\Delta)^d})$

$$\mathbb{E} \|\hat{w}^t - w^*\|_2 \leq O(\sqrt{\frac{\beta}{\alpha}} \frac{\Delta \tau d \log n}{\alpha \sqrt{n} \tilde{\epsilon}}).$$

Since  $\tilde{\epsilon} = \sqrt{2 \log \frac{1}{\delta} + 2\epsilon} - \sqrt{2 \log \frac{1}{\delta}}$ , by using the Taylor series of the function  $\sqrt{x+1} - \sqrt{x}$ , we have  $\tilde{\epsilon} = O(\frac{\epsilon}{\sqrt{\log \frac{1}{\delta}}})$ . Since  $L_{\mathcal{D}}(w)$  is  $\beta$ -smooth we have  $\mathbb{E} L_{\mathcal{D}}(w^T) - L_{\mathcal{D}}(w^*) \leq \frac{\beta}{2} \mathbb{E} \|w^T - w^*\|_2^2$ . Thus we get the proof.

### Proof of Theorem 3.2.7

The proof of  $(\epsilon, \delta)$ -DP is the same as in the proof of Theorem 3. The  $\ell_2$  sensitivity is  $\frac{s}{n} \frac{4\sqrt{2}}{3}$ .

Next, we show the upper bound. The key lemma on the uniform converge rate is the

following. For convenience, we denote by

$$\begin{aligned}\hat{g}_j(w) = & \frac{1}{n} \sum_{i=1}^n (\nabla_j \ell(w, x_i) \left(1 - \frac{\nabla_j^2 \ell(w, x_i)}{2s^2 \beta}\right) \\ & - \frac{\nabla_j^3 \ell(w, x_i)}{6s^2}) + \frac{1}{n} \sum_{i=1}^n C \left( \frac{\nabla_j \ell(w, x_i)}{s}, \frac{|\nabla_j \ell(w, x_i)|}{s\sqrt{\beta}} \right)\end{aligned}$$

and  $\hat{g}_j(w) = (\hat{g}_1(w), \hat{g}_2(w), \dots, \hat{g}_d(w))$ .

**Lemma 3.2.9** (Lemma 8 in [148]). Under Assumptions 1 and 4, with probability at least  $1 - \delta'$ , the following holds for any  $w \in \mathcal{W}$ ,

$$\|\hat{g}_j(w) - \mathbb{E}[\nabla \ell(w, x)]\|_2 \leq O\left(\frac{\beta d \sqrt{v \log(\frac{1}{\delta'} \Delta n)}}{\sqrt{n}}\right). \quad (3.95)$$

Thus, we have the following lemma.

**Lemma 3.2.10.** Under the assumptions in the previous lemma, the following holds with probability at least  $1 - 2\delta'$  for any  $w \in \mathcal{W}$

$$\|g_j(w) - \mathbb{E}[\nabla \ell(w, x)]\|_2 \leq O\left(\frac{\beta d \sqrt{vT \log(\frac{1}{\delta'} \Delta n)}}{\sqrt{n} \sqrt{\epsilon}}\right). \quad (3.96)$$

The remaining proof is almost the same as the proof of Theorem 3.2.5 by using Lemma 3.2.10. We omit it here for convenience.

### Proof of Theorem 3.2.8

Let  $\hat{w}^t$  denote the same notation as in the proof of Theorem 3.2.5. Then, we have

$$\begin{aligned}\|\hat{w}^t - w^*\|_2 &\leq \|w^{t-1} - \eta g^{t-1}(w^{t-1}) - w^*\|_2 \\ &\leq \|w^{t-1} - \eta \nabla L_{\mathcal{D}}(w^{t-1}) - w^*\|_2 + \eta \|g^{t-1}(w^{t-1}) - L_{\mathcal{D}}(w^{t-1})\|_2,\end{aligned}$$

and

$$\begin{aligned}
& \|w^{t-1} - \eta \nabla L_{\mathcal{D}}(w^{t-1}) - w^*\|_2^2 \leq \|w^{t-1} - w^*\|_2^2 \\
& - 2\eta \langle \nabla L_{\mathcal{D}}(w^{t-1}), w^{t-1} - w^* \rangle + \eta^2 \|\nabla L_{\mathcal{D}}(w^{t-1})\|_2^2 \\
& \leq \|w^{t-1} - w^*\|_2^2 - 2\eta \frac{1}{\beta} \|\nabla L_{\mathcal{D}}(w^{t-1})\|_2^2 + \eta^2 \|\nabla L_{\mathcal{D}}(w^{t-1})\|_2^2 \\
& \leq \|w^{t-1} - w^*\|_2^2.
\end{aligned}$$

Thus by Lemma 3.2.10 we have with probability at least  $1 - 2\delta'$

$$\|\hat{w}^t - w^*\|_2 \leq \|w^{t-1} - w^*\|_2 + O\left(\frac{d\sqrt{vT \log(\frac{1}{\delta'} \Delta n)}}{\sqrt{n}\sqrt{\tilde{\epsilon}}}\right). \quad (3.97)$$

Hence, when  $O\left(\frac{dT\sqrt{vT \log(\frac{1}{\delta'} \Delta n)}}{\sqrt{n}\sqrt{\tilde{\epsilon}}}\right) \leq \|w^0 - w^*\|_2$ , we have  $\hat{w}^t \in \mathcal{W}$  for all  $t = \{1, \dots, T\}$  with probability at least  $1 - 2\delta' T$ . This means that  $\hat{w}^t = w^t$  for all  $t \in [T]$ . Hence, we proceed to study the algorithm without projection. Let  $D_t = \|w^0 - w^*\|_2 + O\left(\frac{dt\sqrt{vT \log(\frac{1}{\delta'} \Delta n)}}{\sqrt{n}\sqrt{\tilde{\epsilon}}}\right)$  for  $t = \{0, 1, \dots, T\}$ . By the smoothness of  $L_{\mathcal{D}}(\cdot)$  we have

$$\begin{aligned}
L_{\mathcal{D}}(w^t) & \leq L_{\mathcal{D}}(w^{t-1}) + \langle \nabla L_{\mathcal{D}}(w^{t-1}), w^t - w^{t-1} \rangle + \frac{\beta}{2} \|w^t - w^{t-1}\|_2^2 \\
& = L_{\mathcal{D}}(w^{t-1}) + \eta \langle \nabla L_{\mathcal{D}}(w^{t-1}), -g^{t-1}(w^{t-1}) + \nabla L_{\mathcal{D}}(w^{t-1}) - \nabla L_{\mathcal{D}}(w^{t-1}) \rangle \\
& \quad + \eta^2 \frac{\beta}{2} \|g^{t-1}(w^{t-1}) - \nabla L_{\mathcal{D}}(w^{t-1}) + \nabla L_{\mathcal{D}}(w^{t-1})\|_2^2.
\end{aligned}$$

Since  $\eta = \frac{1}{\beta}$ , by simple calculation we have

$$L_{\mathcal{D}}(w^t) \leq L_{\mathcal{D}}(w^{t-1}) - \frac{1}{2\beta} \|\nabla L_{\mathcal{D}}(w^{t-1})\|^2 + O\left(\frac{\beta d^2 v T \log(\frac{1}{\delta'} \Delta n)}{n \tilde{\epsilon}}\right). \quad (3.98)$$

Next we show the following lemma

**Lemma 3.2.11.** Assume that events (3.96) hold for all  $t = \{1, \dots, T\}$ . Then there exists at

least one  $t \in \{1, \dots, T\}$  such that

$$L_{\mathcal{D}}(w^t) - L_{\mathcal{D}}(w^*) \leq 16D_0\chi,$$

where  $\chi = O\left(\frac{\beta d \sqrt{vT \log(\frac{1}{\delta'} \Delta n)}}{\sqrt{n} \sqrt{\tilde{\epsilon}}}\right)$ .

*Proof.* We note that  $D_t \leq 2D_0$  for all  $t = 0, \dots, T$ . Thus we have

$$L_{\mathcal{D}}(w) - L_{\mathcal{D}}(w^*) \leq \|\nabla L_{\mathcal{D}}(w)\|_2 \|w - w^*\|_2,$$

which implies that

$$\|\nabla L_{\mathcal{D}}(w)\|_2 \geq \frac{L_{\mathcal{D}}(w) - L_{\mathcal{D}}(w^*)}{\|w - w^*\|_2}.$$

Suppose that there exists  $t \in \{1, 2, \dots, T\}$  such that  $\|\nabla L_{\mathcal{D}}(w^t)\|_2 < \sqrt{2}\chi$ . Then, we have  $L_{\mathcal{D}}(w^t) - L_{\mathcal{D}}(w^*) \leq \|\nabla L_{\mathcal{D}}(w^t)\|_2 \|w^t - w^*\|_2 \leq 2\sqrt{2}D_0\chi$ .

Otherwise suppose that for all  $\{1, 2, \dots, T\}$ ,  $\|\nabla L_{\mathcal{D}}(w^t)\|_2 \geq \sqrt{2}\chi$ . Then, we have the following for all  $t \leq T$ ,

$$\begin{aligned} L_{\mathcal{D}}(w^t) - L_{\mathcal{D}}(w^*) &\leq L_{\mathcal{D}}(w^{t-1}) - L_{\mathcal{D}}(w^*) - \frac{1}{4\beta} \|\nabla L_{\mathcal{D}}(w^{t-1})\|_2^2 \\ &\leq L_{\mathcal{D}}(w^{t-1}) - L_{\mathcal{D}}(w^*) - \frac{1}{4\beta D_{t-1}^2} (L_{\mathcal{D}}(w^{t-1}) - L_{\mathcal{D}}(w^*)). \end{aligned}$$

Multiplying both side by  $[(L_{\mathcal{D}}(w^t) - L_{\mathcal{D}}(w^*)) (L_{\mathcal{D}}(w^{t-1}) - L_{\mathcal{D}}(w^*))]^{-1}$  we get

$$\begin{aligned} \frac{1}{L_{\mathcal{D}}(w^t) - L_{\mathcal{D}}(w^*)} &\geq \frac{1}{L_{\mathcal{D}}(w^{t-1}) - L_{\mathcal{D}}(w^*)} + \frac{1}{4\beta D_{t-1}^2} \frac{L_{\mathcal{D}}(w^{t-1}) - L_{\mathcal{D}}(w^*)}{L_{\mathcal{D}}(w^t) - L_{\mathcal{D}}(w^*)} \\ &\geq \frac{1}{L_{\mathcal{D}}(w^{t-1}) - L_{\mathcal{D}}(w^*)} + \frac{1}{16\beta D_0^2}, \end{aligned}$$

where the last inequality is due to the facts that  $D_t \leq 2D_0$  and  $L_{\mathcal{D}}(w^{t-1}) \geq L_{\mathcal{D}}(w^t)$ .

Hence, we have

$$\frac{1}{L_{\mathcal{D}}(w^T) - L_{\mathcal{D}}(w^*)} \geq \frac{T}{16\beta D_0^2} \geq \frac{1}{16D_0\chi} \quad (3.99)$$

using the fact that  $T = \frac{\beta D_0}{\chi}$ , that is,  $T = \tilde{O}\left(\frac{\|w^0 - w^*\|_2 \sqrt{n} \sqrt{\tilde{\epsilon}}}{d}\right)^{\frac{2}{3}}$ . Thus  $\chi = \tilde{O}(\Delta \frac{d^{\frac{2}{3}}}{(n\tilde{\epsilon})^{\frac{1}{3}}})$ .  $\square$

Next we show that

$$L_{\mathcal{D}}(w^T) - L_{\mathcal{D}}(w^*) \leq 16D_0\chi + \frac{1}{2\beta}\chi^2. \quad (3.100)$$

Let  $t = t_0$  be the first time that  $L_{\mathcal{D}}(w^T) - L_{\mathcal{D}}(w^*) \leq 16D_0\chi$ . We show that for any  $t \geq t_0$ ,  $L_{\mathcal{D}}(w^t) - L_{\mathcal{D}}(w^*) \leq 16D_0\chi + \frac{1}{2\beta}\chi^2$ . If not, let  $t_1$  be the first time that  $L_{\mathcal{D}}(w^t) - L_{\mathcal{D}}(w^*) > 16D_0\chi + \frac{1}{2\beta}\chi^2$ . Then, we must have  $L_{\mathcal{D}}(w^{t_1}) > L_{\mathcal{D}}(w^{t_1-1})$ . By (3.98) we have

$$L_{\mathcal{D}}(w^{t_1-1}) - L_{\mathcal{D}}(w^*) \geq L_{\mathcal{D}}(w^{t_1}) - L_{\mathcal{D}}(w^*) - \frac{1}{2\beta}\chi^2 > 16D_0\chi.$$

Thus, we have

$$\|\nabla L_{\mathcal{D}}(w^{t_1-1})\|_2 \geq \frac{L_{\mathcal{D}}(w^{t_1-1}) - L_{\mathcal{D}}(w^*)}{\|w^{t_1-1} - w^*\|_2} \geq 8\chi.$$

By (3.98) we have  $L_{\mathcal{D}}(w^{t_1}) \leq L_{\mathcal{D}}(w^{t_1-1})$  which is a contradiction.

### 3.3 DP-ERM with Pairwise Loss Functions

In the previous two chapters we studied DP-ERM with pointwise loss functions. However, recently, much more attention has been paid to an important family of learning problems called *pairwise learning*. The main difference between pairwise learning and traditional pointwise learning (e.g., classification and regression) is that pairwise learning takes pairs of samples as the input of its loss function while pointwise learning involves only individual samples as the input. Thus, pairwise learning has more advantage in modeling the relative

relationship between pairs of samples. Its importance has been demonstrated in many real-world applications. For example, in patient similarity learning, the learner (e.g., a doctor/hospital) can learn a clinically meaningful similarity metric to measure the proximity between a pair of patients through formulating the learning task as a pairwise learning problem [153]. Additionally, many other machine learning problems can also be categorized as pairwise learning, such as metric learning [61, 165], AUC maximization [360, 225], ranking [272] and multiple kernel learning [186].

Existing pairwise learning algorithms can be roughly divided into two categories: *online* and *offline*. The online pairwise learning algorithms process the input data records in a sequential manner and iteratively update the model upon the arrival of each sample [360, 175]. In contrast, the offline pairwise learning algorithms require the entire training dataset ready before the learning process starts and take it as whole to update the model [61, 165].

Despite their tremendous success in many real-world applications, existing pairwise learning algorithms fail to take into consideration an important issue in their designs, that is, the protection of sensitive information in the training set. The training datasets for pairwise learning are often collected from individual users and thus may contain private personal information. The models learned by such algorithms can implicitly memorize some details of the sensitive information, which undesirably offers opportunity for malicious parties to compromise the users' privacy. Taking the above patient similarity learning task as example, a hospital may want to train a universal patient similarity learning model from patients (crossing many hospitals) so as to obtain a better understanding of the diseases and diagnoses. Due to trust to the hospital, patients may be willing to provide necessary information for such a learning process. However, without a proper mechanism, the patients' privacy may be bleached when the trained model by the hospital is provided to other parties (such as medical research institutes or drug makers). This is because these parties can infer patients' private information using various attack techniques, such as model inversion attack [116] and membership attack [256]. Thus, without a convincing privacy-preserving mechanism,

the patients may not be willing to participate in such a learning task. Hence, a big challenge facing pairwise learning is how to learn a model privately such that sensitive information in the training set cannot be inferred from the learned model.

To the best of our knowledge, no existing work has addressed the above challenge. This motivates us to design, in this section, methods of DP-ERM with pairwise loss functions which can not only keep the sensitive information private but also guarantee good generalization performance. Although various DP-ERM methods exist for (online) pointwise learning, such as objective perturbation [66, 301] or DP-SGD [29, 160, 329, 318, 296, 307], they cannot be applied to pairwise learning algorithms directly. This is mainly because the training sample pairs in pairwise learning algorithms are not i.i.d. and the loss function depends on more than one data records. In the light of the above challenges, in this section, we propose efficient differentially private algorithms for the aforementioned two types of pairwise learning problems. The contributions of this section can be summarized as follows:

- Firstly, we consider the pairwise learning problem in online settings, and propose an  $(\epsilon, \delta)$ -DP algorithm called online pairwise private GIGA-Strongly convex method (**OnPairStrC**). This algorithm can achieve a regret upper bound of  $\tilde{O}(\frac{\sqrt{d}\sqrt{n}}{\epsilon})$  when the loss functions are strongly convex, where  $d$  is the dimensionality of the data and  $n$  is the size of the data sequence. We then extend this algorithm to general convex loss functions by proposing an algorithm called online pairwise private GIGA-convex method (**OnPairC**), which has a regret upper bound of  $\tilde{O}(\frac{\sqrt{dn}^{\frac{3}{4}}}{\epsilon})$ .
- Secondly, to deal with the computational/storage issue in online learning case, we then extend our algorithms to the finite-buffer online setting, where the buffer updates in stream oblivious. Specifically, we show that, with RS-x algorithm as the buffer updating, our algorithms can achieve a regret bound of  $\tilde{O}(\frac{\sqrt{d}\sqrt{n}}{\epsilon} + \frac{n\sqrt{d}}{\sqrt{s}})$  and  $\tilde{O}(\frac{\sqrt{dn}^{\frac{3}{4}}}{\epsilon} + \frac{n\sqrt{d}}{\sqrt{s}})$  for strongly convex and convex loss function case, respectively, where  $s$  is the capacity of the buffer.

- Thirdly, we study the pairwise learning problem in offline settings. We show that it is possible to achieve generalization errors of  $\tilde{O}(\frac{\sqrt{d}}{\sqrt{n}\epsilon})$  and  $\tilde{O}(\frac{\sqrt{d}}{\sqrt[4]{n}\epsilon})$  for strongly and general convex loss functions respectively by adopting the results in the online settings. We then improve these bounds by proposing an offline pairwise private GIGA-Strongly convex algorithm (**OffPairStrC**) and an offline pairwise private GIGA-convex algorithm (**OffPairC**) for the two types of loss functions. Particularly, in the case of general convex loss functions, our improved algorithm can achieve a generalization error of  $\tilde{O}(\frac{\sqrt{d}}{\sqrt{n}\epsilon})$ .
- Finally, we take two pairwise learning tasks (i.e., AUC maximization and metric learning) as examples and conduct extensive experiments on real-world datasets to evaluate the performance of the proposed algorithms. The experimental results not only verify our theoretical analysis but also show the effectiveness of our proposed algorithms in real-world applications.

### 3.3.1 Private pairwise learning

Different from the pointwise loss function  $\ell : \mathcal{C} \times \mathcal{D} \mapsto \mathbb{R}$ , a pairwise loss function is a function on pairs of data records, *i.e.*,  $\ell : \mathcal{C} \times \mathcal{D} \times \mathcal{D} \mapsto \mathbb{R}$ , where  $\mathcal{D}$  is the data universe. Given a dataset  $D = \{z_1, z_2, \dots, z_n\} \subseteq \mathcal{D}^n$  and a loss function  $\ell(\cdot; \cdot, \cdot)$ , its empirical risk can be defined as:

$$L(w; D) = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i} \ell(w; z_i, z_j). \quad (3.101)$$

When the inputs are drawn i.i.d from an unknown underlying distribution  $\mathcal{P}$  on  $\mathcal{D}$ , the population risk is

$$L_{\mathcal{P}}(w) = \mathbb{E}_{z_i, z_j \sim \mathcal{P}} [\ell(w; z_i, z_j)]. \quad (3.102)$$

We define private pairwise learning as follows.

**Definition 3.3.1** (DP-ERM with pairwise loss functions). Let  $\mathcal{C} \subseteq \mathbb{R}^d$  be a convex, closed and bounded constraint set,  $\mathcal{D}$  be a data universe, and  $\ell : \mathcal{C} \times \mathcal{D} \times \mathcal{D} \mapsto \mathbb{R}$  be a pairwise loss function. Also, let  $D = \{z_1 = (x_1, y_1), z_2 = (x_2, y_2), \dots, z_n = (x_n, y_n)\} \subseteq \mathcal{D}^n$  be a dataset with data records  $\{x_i\}_{i=1}^n \subset \mathbb{R}^d$  and labels (responses)  $\{y_i\}_{i=1}^n \subset [-1, 1]^n$ . Private pairwise learning is to find a private estimator  $w_{\text{priv}} \in \mathcal{C}$  so that the algorithm is  $(\epsilon, \delta)$  or  $\epsilon$  differential privacy and the error is minimized, where the error for an estimator  $w$  can be measured by either the optimality gap  $\text{Err}_D(w) = L(w; D) - \min_{w \in \mathcal{C}} L(w; D)$  or the generalization error  $\text{Err}_{\mathcal{P}}(w) = L_{\mathcal{P}}(w_{\text{priv}}) - \min_{w \in \mathcal{C}} L_{\mathcal{P}}(w)$ .

In this paper, we will focus on a special class of pairwise loss functions<sup>6</sup> which contains the loss functions of metric learning, AUC maximization and bipartite ranking.

**Assumption 3.3.1.** For the loss function, we assume that it has the form of  $\ell(w; z, z') = \phi(Y(y, y')h(w; x, x'))$ , and  $\ell$  is a  $G$ -Lipschitz and  $L$ -smooth convex function over  $w$ , where  $Y(y, y') = y - y'$  or  $Y(y, y') = yy'$ . In the experimental part, we will let  $\phi$  be the logistic function, i.e.,  $\phi(x) = \log(1 + e^{-x})$ .

**Example 1: Metric Learning [61]** The goal here is to learn a Mahalanobis metric  $M_W^2(x, x') = (x - x')^T W (w - x')$  using loss function  $\ell(W; z, z') = \phi(yy'(1 - M_W^2(x, x')))$ , where  $y, y' \in \{-1, +1\}$ . The constraint set  $\mathcal{C}$  is  $\mathcal{C} = \{W : W \in \mathbb{S}^d, \|W\|_F \leq 1\}$ , where  $\mathbb{S}^d$  is the set of  $d \times d$  positive symmetric matrices.

**Example 2: AUC Maximization [360], Bipartite Ranking [76]** The goal here is to maximize the area under the ROC curve for a linear classification problem with the constraint of  $\|w\|_2 \leq 1$ . Here  $h(w; x, x') = w^T(x - x')$  and  $\ell(w; z, z') = \phi((y - y')h(w; x, x'))$ , where  $y, y' \in \{-1, +1\}$ .

Like in the pointwise loss function case, in the following we will introduce the Rademacher average for a class of pairwise loss function functions. Specifically, we denote the Rademacher

---

<sup>6</sup>We note that all the  $(\epsilon, \delta)$ -DP algorithms in this paper can be extended to general pairwise loss functions, although the upper bounds of the generalization errors may differ.

averages of the pairwise loss functions class  $\ell \circ \mathcal{C} := \{(z, z') \mapsto \ell(w; z, z'), w \in \mathcal{C}\}$  by the following [175]:

$$\mathcal{R}_n(\ell \circ \mathcal{C}) = \mathbb{E}[\sup_{w \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^n \epsilon_i \ell(w; z, z_i)], \quad (3.103)$$

where  $\{\epsilon_i\}_{i=1}^n$  are the Rademacher variables, *i.e.*,  $\epsilon = \pm 1$  with probability  $\frac{1}{2}$ , and the expectation is over  $\{\epsilon_i\}_{i=1}^n, z, \{z_i\}_{i=1}^n$ .

Note that there are many classes of pairwise loss functions whose Rademacher average is  $\mathcal{R}_n(\ell \circ \mathcal{C}) = O(\frac{\sqrt{d}}{\sqrt{n}})$ , such as Example 1 and Example 2 [175], where  $d$  is the dimensionality of the parameter space.

### Online private pairwise learning

Here we follow online pairwise learning [175]. An online learning algorithm  $\mathcal{A}$  is given sequential access to a stream of elements  $z_1, z_2, z_3, \dots, z_n$ . At each time step  $t = 2, 3, \dots, n$ , the algorithm selects a parameter  $w_{t-1} \in \mathcal{C}$  upon which the data record  $z_t$  is revealed, and the algorithm incurs the following penalty

$$\hat{L}_t(w_{t-1}, D_t) = \frac{1}{t-1} \sum_{i=1}^{t-1} \ell(w_{t-1}; z_t, z_i), \quad (3.104)$$

where  $D_t = \{z_1, z_2, \dots, z_t\}$ . Thus, the online algorithm  $\mathcal{A}$  maps a data sequence  $\{z_1, z_2, \dots, z_n\}$  to a sequence of parameters  $\{w_1, w_2, \dots, w_{n-1}\}$ . In the non-private case, the goal is to select  $\{w_1, w_2, \dots, w_{n-1}\}$  so as to minimize the **regret**, *i.e.*,

$$\mathcal{R}_{\mathcal{A}}(n, D) = \sum_{t=2}^n \hat{L}_t(w_{t-1}, D_t) - \min_{w \in \mathcal{C}} \sum_{t=2}^n \hat{L}_t(w, D_t). \quad (3.105)$$

Moreover, if all the data records are chosen i.i.d from the distribution  $\mathcal{P}$ , we also want to minimize the **generalized regret**, *i.e.*,

$$\mathcal{R}_{\mathcal{P}, \mathcal{A}}(n) = \sum_{t=2}^n L_{\mathcal{P}}(w_{t-1}) - (n-1) \min_{w \in \mathcal{C}} L_{\mathcal{P}}(w). \quad (3.106)$$

Note that if the loss function  $\ell$  is convex, then from (3.106) we have the parameter  $\bar{w} = \frac{w_1 + \dots + w_{n-1}}{n-1}$  satisfies the following generalization error:

$$L_{\mathcal{P}}(\bar{w}) - \min_{w \in \mathcal{C}} L_{\mathcal{P}}(w) \leq \frac{\mathcal{R}_{\mathcal{P}, \mathcal{A}}(n)}{n-1}. \quad (3.107)$$

However, under the differential privacy model, we need to guarantee that the output sequence  $\{w_1, \dots, w_{n-1}\}$  is DP. Thus, private pairwise learning in the online setting can be defined as follows:

**Definition 3.3.2** (Online private pairwise learning). Let  $Z = \{z_1, z_2, \dots, z_n\}$  be any sequence of data records in the data universe  $\mathcal{D}$ . Let the sequence of outputs by algorithm  $\mathcal{A}$  be  $\mathcal{A}(Z) = \{w_1, w_2, \dots, w_{n-1}\}$ . Then,  $\mathcal{A}$  is  $(\epsilon, \delta)$  differentially private if given any other data sequence  $Z'$  which differs in at most one entry with  $Z$ , for all events  $S$ , we have  $\Pr[\mathcal{A}(Z) \in S] \leq e^\epsilon \Pr[\mathcal{A}(Z') \in S] + \delta$ . The goal of online private pairwise learning is to select private outputs  $\{w_1, w_2, \dots, w_{n-1}\}$  that minimizes the (generalized) regret.

From above discussions on (3.106) and (3.107), we know that if the generalized regret is low, the algorithm will have a good performance on generalization theoretically. From this view, the online setting is more general. Thus, in the paper, we will first consider the online private pairwise learning and provide (generalized) regrets for both strongly and general convex loss functions. After that, in the following sections, we will study the problem in the finite-buffer online and offline settings.

### 3.3.2 Online Private pairwise learning

We first consider the case that the loss function is strongly convex. After that, we will use the regularization perturbation strategy [276, 160] to extend the resulting algorithm to general convex loss functions.

## Strongly convex case

Our algorithm is inspired by the ideas in the stability of Generalized Infinitesimal Gradient Ascent (GIGA) [371, 160], which is a well-known online convex algorithm (see Remark 3.3.1 for discussions on the difference of our algorithm with the previous ones). The main steps of our algorithm are given in Algorithm 3.3.9. We call the above algorithm excluding

---

### Algorithm 3.3.9 Online Pairwise Private GIGA-Strongly Convex (OnPairStrC)

---

- 1: **Input:** Privacy parameters  $\epsilon$  and  $\delta$ , sequence of data record  $\{z_1, z_2, \dots, z_n\}$ , constrained convex set  $\mathcal{C} \subset \mathbb{R}^d$ , and pairwise loss function  $\ell(\cdot; \cdot, \cdot)$ .
  - 2: **Parameters:**  $\ell$  is  $G$ -Lipschitz,  $L$ -smooth and  $\alpha$ -strongly convex over  $w$ . Step time  $T_1 = \max\{\lceil \frac{16L^2}{\alpha^2} \rceil, 7\}$ .
  - 3: Compute  $\rho$  which satisfies  $\rho + 2\sqrt{\rho \log(\frac{1}{\delta})} = \epsilon$ .
  - 4: **for**  $t = 1, \dots, T_1$  **do**
  - 5:     Receive the data record  $z_t$  (incurs penalty  $\hat{L}_t(w_{t-1}, D_t)$  when  $t \geq 2$ ).
  - 6:     Randomly choose a parameter  $w_t \in \mathcal{C}$ .
  - 7: **end for**
  - 8: **for**  $t = T_1 + 1, \dots, n$  **do**
  - 9:     Receive the data record  $z_t$  (incurs penalty  $\hat{L}_t(w_{t-1}, D_t)$ ).
  - 10:     Set step size  $\eta_t = \frac{t-1}{t-2} \frac{2}{\alpha t}$
  - 11:      $w_t = \Pi_{\mathcal{C}}[w_{t-1} - \eta_t \nabla \hat{L}_t(w_{t-1}, D_t)]$ , where  $\Pi_{\mathcal{C}}$  is the projection onto the set  $\mathcal{C}$ .
  - 12:     Set  $\sigma_t^2 = \frac{32G^2(n-T_1)}{\alpha^2 t^2 \rho}$ . Let  $\tilde{w}_t = w_t + n_t$ , where  $n_t \sim \mathcal{N}(0, \sigma_t^2 I_d)$ .
  - 13:     Output  $w_t = \arg \min_{w \in \mathcal{C}} \|w - \tilde{w}_t\|_2^2$ .
  - 14: **end for**
- 

the portion of random perturbation (*i.e.*, steps 12 and 13) **Pairwise GIGA**. The following lemma gives an upper bound on the  $\ell_2$ -norm sensitivity of the output in the  $t$ -th iteration of Pairwise GIGA, which is the key to ensure  $(\epsilon, \delta)$ -differential privacy.

**Lemma 3.3.1.** Let  $\mathcal{A}_t(D_t)$  denote the output of Pairwise GIGA in the  $t$ -th iteration. Then, under the assumption of Algorithm 3.3.9, for any  $t \geq 1$  and  $D_t \sim D'_t$ ,

$$\|\mathcal{A}_t(D_t) - \mathcal{A}_t(D'_t)\|_2 \leq \frac{8G}{\alpha t}.$$

Theorem 3.3.1 shows that Algorithm 3.3.9 is differentially private.

**Theorem 3.3.1.** Under Assumption 3.3.1 and the assumption that the loss function  $\ell$  is  $\alpha$ -strongly convex, for any  $0 < \epsilon, \delta \leq 1$ , Algorithm 3.3.9 is  $(\epsilon, \delta)$ -differentially private.

Note that to guarantee DP, we first transfer  $(\epsilon, \delta)$ -DP to  $\rho$ -zCDP by Lemma 2.1.10, and then use composition theorem to make Algorithm 3.3.9 be  $\rho$ -zCDP (*i.e.*, we make each iteration  $T_1 + 1 \leq t \leq n$  be  $\frac{\rho}{n-T_1}$ -zCDP). It is easy to see that in this case the variance of the noise satisfies  $\sigma_t^2 = \frac{32G^2(n-T_1)}{\alpha^2 t^2 (\sqrt{\log(1/\delta)} + \epsilon - \sqrt{\log(1/\delta)})^2}$ . When  $\frac{\epsilon}{\log(1/\delta)} \ll 1$  (this case will always holds since in practice we select  $\epsilon = 0.1 \sim 5$  and  $\delta = \frac{1}{n}$ ), by Taylor expansion of  $\sqrt{1+x}$ , we have  $(\sqrt{\log(1/\delta)} + \epsilon - \sqrt{\log(1/\delta)})^2 \simeq \frac{\epsilon^2}{4\log(1/\delta)}$ . Thus in total, we have  $\sigma_t^2 \simeq \frac{128G^2(n-T_1)\log(1/\delta)}{\alpha^2 t^2 \epsilon^2}$ . We note that this idea has also been used in [192]. The main difference is that in our online setting, the iteration number equals to the size of the sequence, which is fixed and thus needs us to equally allocate the privacy budget, while in [192], the iteration number is not fixed, which allows them to allocate the budget in a more efficient way. This makes the two algorithms significantly different in their analysis and thus incomparable.

The following theorem shows an upper bound on the (expected) regret of Algorithm 3.3.9, which can be transformed to (expected) generalized error (we will show it in the following section).

**Theorem 3.3.2.** Under the assumptions in Theorem 3.3.1 and the additional condition of  $\frac{\epsilon}{\log \frac{1}{\delta}} \ll 1$ , Algorithm 3.3.9 has the following (expected) upper bound on the regret of its outputs

$$\mathcal{R}_{\mathcal{A}}(n, D) \leq O\left(\frac{G^2 \sqrt{d} \log^{1.5} \frac{n}{\zeta} \sqrt{n} \sqrt{\log \frac{1}{\delta}}}{\alpha \epsilon} + \frac{GL^2}{\alpha^2} \|\mathcal{C}\|_2 + \frac{G^2 \log n}{\alpha}\right) \quad (3.108)$$

with probability at least  $1 - \zeta$ , and

$$\mathbb{E} \mathcal{R}_{\mathcal{A}}(n) \leq O\left(\frac{G^2 \sqrt{d} \log n \sqrt{n} \sqrt{\log \frac{1}{\delta}}}{\alpha \epsilon} + \frac{GL^2}{\alpha^2} \|\mathcal{C}\|_2 + \frac{G^2 \log n}{\alpha}\right), \quad (3.109)$$

where  $\|\mathcal{C}\|_2 = \max_{w,w' \in \mathcal{C}} \|w - w'\|_2$  is the diameter of the set  $\mathcal{C}$ <sup>7</sup>.

**Remark 3.3.1.** We note that [160] also used the differentially private version of GIGA and IGD [184] in their DP pointwise learning. But their Private GIGA or IGD [160] is quite different from our method of OnPairStrC (Algorithm 3.3.9). Firstly, [160] needs to assume that each loss function  $\hat{L}_t$  is independent (see the proofs of Lemma 4 and Lemma 5 in [160]), which means that it is only applicable to pointwise loss functions. However, in our problem, the penalty function (3.104) depends on previous data records, which means that it is much more complicated than the case in [160]. Thus, we need a much finer and more different analysis on the stability of Pairwise GIGA. Also, the parameters of the step size  $\eta_t$  and time step  $T_1$  are quite different from those in [160] (see Appendix for details). Additionally, in order to show the power of our method, we also consider the case with additional finite buffer constraint, which has not been studied in [160]. Thus, our method is more general.

Secondly, the upper bound (3.109) on the expected regret of our algorithm is less than that in [160] with a factor of  $\log \frac{n}{\delta}$ . This is due to the fact that we use the composition property of zCDP instead of advanced composition theorem of DP [105].

Thirdly, since the definition of regret in our paper is different from that in pointwise learning [160], the same upper bound (*i.e.*,  $\tilde{O}(\frac{\sqrt{dn}}{\epsilon})$ ) on the regret for strongly convex loss functions are actually incomparable.

We now use the perturbation strategy [276] to obtain results for general convex pairwise loss functions.

---

**Algorithm 3.3.10** Online Pairwise Private GIGA-Convex (OnPairC)

---

- 1: **Input:** Privacy parameters  $\epsilon$  and  $\delta$ , sequence of data record  $\{z_1, z_2, \dots, z_n\}$ , constrained convex set:  $\mathcal{C}$ , pairwise loss function  $\ell(\cdot; \cdot, \cdot)$ , and a parameter  $\alpha$  to be defined later.
  - 2: **Parameters:**  $\ell$  is  $G$ -Lipschitz,  $L$ -smooth and convex over  $w$ .
  - 3: Randomly select a point  $w_0 \in \mathcal{C}$ . Let  $\tilde{\ell}(w; z, z') = \ell(w; z, z') + \frac{\alpha}{2} \|w - w_0\|_2^2$ .
  - 4: Run Algorithm 3.3.9 with loss function  $\tilde{\ell}$ , which is  $\tilde{G} = G + \alpha \|\mathcal{C}\|_2$ -Lipschitz,  $\tilde{L} = L + \alpha$ -smooth and  $\alpha$ -strongly convex.
- 

<sup>7</sup>If  $\mathcal{C} = \mathbb{R}^d$ , then we can take  $\mathcal{C} = \{w : \|w\|_2 \leq \|w^*\|_2\}$ .

**Theorem 3.3.3.** Let  $\ell$  be a pairwise loss function satisfying Assumption 3.3.1. Then, for any  $0 < \epsilon, \delta \leq 1$ , Algorithm 3.3.10 is  $(\epsilon, \delta)$ -DP. Moreover, if  $\frac{\epsilon}{\log \frac{1}{\delta}} \ll 1$  and take  $\alpha = O(\frac{1}{\sqrt[4]{n}})$ , then with probability at least  $1 - \zeta$ , the following upper bound on regret for the outputs holds:

$$\mathcal{R}_{\mathcal{A}}(n, D) \leq O\left(\frac{L^2 G^2 \|\mathcal{C}\|_2^2 \sqrt{d} \log^{1.5} \frac{n}{\zeta} n^{\frac{3}{4}} \sqrt{\log \frac{1}{\delta}}}{\epsilon}\right). \quad (3.110)$$

Similar result also holds for the expected regret.

Comparing (3.110) with (3.108), we can see that for strongly convex pairwise loss functions, the average regret, *i.e.*,  $\frac{\mathcal{R}_{\mathcal{A}}(n)}{n-1}$ , is upper bounded by  $\tilde{O}(\frac{\sqrt{d}}{\sqrt{n}\epsilon})$ , while for general convex ones, it is  $\tilde{O}(\frac{\sqrt{d}}{\sqrt[4]{n}\epsilon})$ . This is the same as in the case of pointwise loss functions [276].

### Online Private Pairwise Learning with Finite Buffer

In the previous section, we consider the online private pairwise learning and proposed Private Pairwise GIGA algorithm. However, from Algorithm 3.3.9 and Algorithm 3.3.10 we can see that in each iteration, this requires one to memorize and store all the previous data records, which is computationally/storagewise expensive. Thus, for the online pairwise loss function, it has been studied in additional the finite-buffer setting [360, 175]. Thus in this section, we study online private pairwise learning with finite buffer.

We assume that the buffer updates in stream oblivious. More specifically, we require the buffer update rule to decide upon the inclusion of a particular point  $z_i$  in the buffer based only on its stream index  $i \in [n]$ . Such as Reservoir Sampling [290] and FIFO. Stream oblivious policy allow us to decouple buffer construction randomness from training sample randomness which makes analysis easier [175]. We also assume that the adversary cannot get the status of the buffer. Next we give some definitions related to the finite-buffer online learning.

We consider a buffer  $B$  with capacity  $s$ , and we denote it is  $B_t$  in the  $t$ -th iteration, which stores a sketch of the stream. Now at each step after receiving the data record  $z_t$ . The

penalty becomes

$$\hat{L}_t^{\text{buf}}(w_{t-1}, B_t) = \frac{1}{|B_t|} \sum_{z \in B_t} \ell(w_{t-1}; z_t, z). \quad (3.111)$$

An online learning algorithm  $\mathcal{A}$  will be said to have a finite-buffer regret bound  $\mathcal{R}_{\mathcal{A}}^{\text{buf}}(n)$  if its presents an ensemble  $w_1, w_2, \dots, w_{n-1}$  such that

$$\sum_{t=2}^n \hat{L}_t^{\text{buf}}(w_{t-1}, B_t) - \min_{w \in \mathcal{C}} \sum_{t=2}^n \hat{L}_t^{\text{buf}}(w, B_t) \leq \mathcal{R}_{\mathcal{A}}^{\text{buf}}(n). \quad (3.112)$$

## Algorithms and main results

There are several buffer updating, such as RS algorithm in [360], RS-x, RS- $x^2$ . In this paper, we use RS-x as the buffer updating. See Algorithm 3.3.11 for detail. Combing with Algorithm 3.3.11 and our previous Pairwise Private GIGA, we can get efficient algorithms.

---

### Algorithm 3.3.11 RS-x: Stream Subsampling with Replacement [175]

---

```

1: Input: Buffer  $B$ , new data record  $z_t$ , buffer size  $s$ , timestep  $t$ .
2: if  $|B| < s$  then
3:    $\text{TMP} = B \cup \{z_t\}$ 
4: else
5:   if  $t = s + 1$  then
6:      $\text{TMP} = B \cup \{z_t\}$ 
7:     Repopulate  $B$  with  $s$  points sampled uniformly with replacement from  $\text{TMP}$ 
8:   else
9:     Independently, replace each point of  $B$  with  $z_t$  with probability  $\frac{1}{t}$ 
10:  end if
11: end if
```

---

We have the following differential privacy guarantee and upper bounds of finite-buffer regret (3.112).

**Theorem 3.3.4.** For any  $0 < \epsilon, \delta < 1$ , Algorithm 3.3.12 and 3.3.13 are  $(\epsilon, \delta)$ -DP. Moreover, for strongly convex loss function under Assumption 3.3.1 and if  $\frac{\epsilon}{\log 1/\delta} \ll 1$ , we have the following finite buffer regret for the outputs  $\{w_1, \dots, w_{n-1}\}$ :

$$\frac{\mathcal{R}_{\mathcal{A}}^{\text{buf}}(n)}{n-1} \leq O\left(\frac{G^2 L^2 \|\mathcal{C}\|_2 \sqrt{d} \log^{1.5} \frac{n}{\zeta} \sqrt{\log 1/\delta}}{\sqrt{n} \alpha^2 \epsilon}\right). \quad (3.113)$$

---

**Algorithm 3.3.12** Finite-buffer Online Pairwise Private GIGA-Strongly Convex (FBOOnPairStrC)

---

- 1: **Input:** Privacy parameters  $\epsilon$  and  $\delta$ , sequence of data record  $\{z_1, z_2, \dots, z_n\}$ , convex set  $\mathcal{C}$ , pairwise loss function  $\ell(\cdot; \cdot, \cdot)$ . Buffer  $B$  with size  $s$ .
  - 2: **Parameters:**  $\ell$  is  $G$ -Lipschitz,  $L$ -smooth and  $\alpha$ -strongly convex . Step time  $T_1 = \max\{\lceil \frac{16L^2}{\alpha^2} \rceil, 7\}$ .
  - 3: Randomly sample  $w_0 \in \mathcal{C}$
  - 4: **for**  $t = 1, 2, \dots, T_1$  **do**
  - 5:     Receive the data record  $z_t$  and we get the penalty  $\hat{L}_t^{\text{buf}}(w_{t-1}, B_t)$  when  $t \geq 2$ .
  - 6:     Update buffer  $B_{t+1}$  by using Algorithm 3.3.11 with  $(B_t, z_t, s, t)$ .
  - 7:     Randomly choose a parameter  $w_t \in \mathcal{C}$ .
  - 8: **end for**
  - 9: **for**  $t = T_1 + 1, \dots, n$  **do**
  - 10:     Receive the data record  $z_t$  and we get the penalty  $\hat{L}_t^{\text{buf}}(w_{t-1}, B_t)$
  - 11:     Set step size  $\eta_t = \frac{t-1}{t-2} \frac{2}{\alpha t}$
  - 12:      $w_t = \Pi_{\mathcal{C}}[w_{t-1} - \eta_t \nabla \hat{L}_t^{\text{buf}}(w_{t-1}, B_t)]$ , where  $\Pi_{\mathcal{C}}$  is the projection onto the set  $\mathcal{C}$ .
  - 13:     Compute  $\rho$  which satisfies  $\rho + 2\sqrt{\rho \log(\frac{1}{\delta})} = \epsilon$ . Then set  $\sigma_t^2 = \frac{32G^2(n-T_1)}{\alpha^2 t^2 \rho}$ . Let  $\tilde{w}_t = w_t + n_t$ , where  $n_t \sim \mathcal{N}(0, \sigma_t^2 \mathbb{I}_d)$ .
  - 14:     Output  $w_t = \arg \min_{w \in \mathcal{C}} \|w - \tilde{w}_t\|_2^2$ .
  - 15:     Update buffer  $B_{t+1}$  by using Algorithm 3.3.11 with  $(B_t, z_t, s, t)$ .
  - 16: **end for**
- 

---

**Algorithm 3.3.13** Finite-buffer Online Pairwise Private GIGA-Convex (FBOOnPairC)

---

- 1: **Input:** Privacy parameters  $\epsilon$  and  $\delta$ , sequence of data record  $\{z_1, z_2, \dots, z_n\}$ , convex set  $\mathcal{C}$ , pairwise loss function  $\ell(\cdot, \cdot, \cdot)$ . Buffer  $B$  with size  $s$ .  $\alpha$  is a parameter will be specified later.
  - 2: **Parameters:**  $\ell$  is  $G$ -Lipschitz,  $L$ -smooth and convex.
  - 3: Randomly select a point  $w_0 \in \mathcal{C}$ . Let  $\tilde{\ell}(w; z, z') = \ell(w; z, z') + \frac{\alpha}{2} \|w - w_0\|_2^2$ .
  - 4: Run Algorithm 3.3.12 with loss function  $\tilde{\ell}$ , which is  $\tilde{G} = G + \alpha \|\mathcal{C}\|_2$ -Lipschitz,  $\tilde{L} = L + \alpha$ -smooth and  $\alpha$ -strongly convex.
-

The following upper bound for generalized regret.

$$\begin{aligned} \frac{\mathcal{R}_{\mathcal{P}, \mathcal{A}}(n)}{n-1} &\leq O\left(\frac{G^2 \|\mathcal{C}\|_2 L^2 \sqrt{d} \log^{1.5} \frac{n}{\zeta} \sqrt{\log 1/\delta}}{\alpha^2 \epsilon \sqrt{n}}\right. \\ &\quad \left.+ \frac{C_d}{\sqrt{s}} + G \|\mathcal{C}\|_2 \sqrt{\frac{\log \frac{n}{\zeta}}{s}}\right). \end{aligned} \quad (3.114)$$

and the following regret bound:

$$\frac{\mathcal{R}_{\mathcal{A}}(n, D)}{n-1} \leq \frac{G^2 \|\mathcal{C}\|_2 L^2 \sqrt{d} \log^{1.5} \frac{n}{\zeta} \sqrt{\log 1/\delta}}{\alpha^2 \epsilon \sqrt{n}} + C_d \sqrt{\frac{\log \frac{n}{\zeta}}{s}}. \quad (3.115)$$

Where  $C_d$  is the dependence of dimensionality in the Rademacher Average  $\mathcal{R}_n(\ell \circ \mathcal{C})$  in (3.103).

For the general convex loss function, we just replace the term of  $O\left(\frac{G^2 \|\mathcal{C}\|_2 L^2 \sqrt{d} \log^{1.5} \frac{n}{\zeta} \sqrt{\log 1/\delta}}{\alpha^2 \epsilon \sqrt{n}}\right)$  above to  $O\left(\frac{L^2 G^2 \|\mathcal{C}\|_2^2 \sqrt{d} \log^{1.5} \frac{n}{\zeta} \sqrt{\log 1/\delta}}{\epsilon n^{\frac{1}{4}}}\right)$ . when we set  $\alpha = O\left(\frac{1}{\sqrt[4]{n}}\right)$  in Algorithm 3.3.13.

From the above Theorem 3.3.4 we can see that, unlike the infinite-buffer case, *i.e.*, Algorithms 3.3.9 and 3.3.10, the buffer capacity  $s$  plays an important role on the generalization performance. Only if  $s = \omega(n)$ , then these bounds are asymptotically the same as in the infinite-buffer case for strongly convex function, while it only needs to be  $\omega(\sqrt{n})$  for the general convex case.

### 3.3.3 Offline Private Pairwise Learning

In this section, we study differentially private pairwise learning in offline settings. As shown in Definition 3.3.1, we always assume that each  $z_i$  is sampled from some unknown distribution  $\mathcal{P}$ .

#### Generalization error induced by generalized regret

We first observe that Algorithm 3.3.9 and 3.3.10 preserve  $(\epsilon, \delta)$ -DP in the offline settings. Also, as discussed in (3.106) and (3.107), if we get the generalized regret for the output

$\{w_1, w_2, \dots, w_{n-1}\}$ , we can easily obtain a generalization error by (3.107). By a theorem in [175], we can have the following generalization bounds for  $\bar{w} = \frac{w_1 + \dots + w_{n-1}}{n-1}$  of Algorithm 3.3.9 and 3.3.10.

**Theorem 3.3.5.** Under Assumption 3.3.1, the parameter  $\bar{w} = \frac{w_1 + \dots + w_{n-1}}{n-1}$  satisfies the following generalization error for loss function  $\ell$  with probability at least  $1 - 2\zeta$  if  $\frac{\epsilon}{\log \frac{1}{\delta}} \ll 1$ , where  $w_1, w_2, \dots, w_{n-1}$  are the outputs of Algorithm 3.3.10 (Algorithm 3.3.9 for strongly convex loss functions),

$$\text{Err}_{\mathcal{P}}(\bar{w}) \leq O\left(\frac{\sum_{t=2}^n \mathcal{R}_{t-1}(\ell \circ \mathcal{C})}{n-1} + \frac{L^2 G^2 \|\mathcal{C}\|_2^2 \sqrt{d} \log^{1.5} \frac{n}{\zeta} \sqrt{\log \frac{1}{\delta}}}{\epsilon \sqrt[4]{n}}\right). \quad (3.116)$$

Moreover, if the loss is  $\alpha$ -strongly convex, then we have:

$$\text{Err}_{\mathcal{P}}(\bar{w}) \leq O\left(\frac{1}{n-1} \sum_{t=2}^n \mathcal{R}_{t-1}(\ell \circ \mathcal{C}) + \frac{G^2 L^2 \|\mathcal{C}\|_2 \sqrt{d} \log^{1.5} \frac{n}{\zeta} \sqrt{\log \frac{1}{\delta}}}{\alpha^2 \epsilon \sqrt{n}}\right). \quad (3.117)$$

**Remark 3.3.2.** Thus, for Example 1, the generalization error is  $\tilde{O}\left(\frac{d}{\epsilon \sqrt[4]{n}}\right)$  for logistic loss function while it is  $\tilde{O}\left(\frac{d}{\epsilon \sqrt{n}}\right)$  if adding an additional Frobenious regularization to the loss functions<sup>8</sup>. Similar result holds for Example 2, where the generalization error for logistic loss is  $\tilde{O}\left(\frac{\sqrt{d}}{\epsilon \sqrt[4]{n}}\right)$  while it is  $\tilde{O}\left(\frac{\sqrt{d}}{\epsilon \sqrt{n}}\right)$  in the case of with additional  $\ell_2$ -norm regularization.

### 3.3.4 Improved Upper Bounds for Offline Setting

Inspired by the sensitivity of Pairwise GIGA in Lemma 3.3.1 and Theorem 3.3.5, we propose an offline DP algorithm which has better upper bounds compared to (3.116) and (3.117). The basic idea is to use output perturbation. More specifically, we first run Pairwise GIGA in the offline settings and then add some Gaussian noises to  $\tilde{w} = \frac{w_1 + \dots + w_n}{n}$  to keep the algorithm  $(\epsilon, \delta)$ -DP, since the sensitivity of  $\tilde{w}$  is based on each  $w_i$ , which can be obtained by Lemma 3.3.1. For the general convex loss functions, we can still use the perturbation idea,

---

<sup>8</sup>Note that for Example 1 since the parameter is a positive matrix, the dimensionality will be  $O(d^2)$ .

which is the same as in Algorithm 3.3.10. See Algorithm 3.3.14 and 3.3.15 for details.

The reason that we can improve the generalization error is due to the following fact.

From Algorithms 3.3.9 and 3.3.10, we can see that the output sequences  $\{w_1, w_2, \dots, w_{n-1}\}$  satisfy the conditions of  $(\epsilon, \delta)$ -DP in each iteration. However, in the offline setting, we only need to ensure that the final output is DP. Thus, instead of adding noise in each iteration, we can add noises only once to the final output, which means that we can add a smaller scale of noises compared to the online ones.

---

**Algorithm 3.3.14 Offline Pairwise Private GIGA-Strongly Convex (OffPairStrC)**


---

- 1: **Input:** Privacy parameters  $\epsilon$  and  $\delta$ , sequence of data record  $\{z_1, z_2, \dots, z_n\}$ , constrained convex set  $\mathcal{C}$ , pairwise loss function  $\ell(\cdot; \cdot, \cdot)$ , and step number  $T_1 = \max\{\lceil \frac{16L^2}{\alpha^2} \rceil, 7\}$ .
  - 2: **Parameters:**  $\ell$  is  $G$ -Lipschitz,  $L$ -smooth and  $\alpha$ -strongly convex over  $w$ .
  - 3: Randomly sample  $w_1, \dots, w_{T_1} \in \mathcal{C}$ .
  - 4: **for**  $t = T_1 + 1, \dots, n$  **do**
  - 5:     Set step size  $\eta_t = \frac{t-1}{t-2} \frac{2}{\alpha t}$ .
  - 6:      $w_t = \arg \min_{w \in \mathcal{C}} \|w - (w_{t-1} - \eta_t \nabla \hat{L}_t(w_{t-1}, D_t))\|_2^2$ , i.e., projecting  $w_{t-1} - \eta_t \nabla \hat{L}_t(w_{t-1}, D_t)$  onto the convex set  $\mathcal{C}$ .
  - 7: **end for**
  - 8: Let  $\tilde{w} = \frac{w_1 + \dots + w_n}{n}$ .
  - 9: Let  $\bar{w} = \tilde{w} + \sigma$ , where  $\sigma \sim \mathcal{N}(0, \frac{128G^2 \log^2 n \log(1.25/\delta)}{\alpha^2 n^2 \epsilon^2} I_d)$ .
  - 10: Return  $\hat{w} = \arg \min_{w \in \mathcal{C}} \|w - \bar{w}\|_2^2$ .
- 

---

**Algorithm 3.3.15 Pairwise Private GIGA-Convex (OffPairC)**


---

- 1: **Input:** Privacy parameters  $\epsilon$  and  $\delta$ , sequence of data record  $\{z_1, z_2, \dots, z_n\}$ , constrained convex set  $\mathcal{C}$ , pairwise loss function  $\ell(\cdot; \cdot, \cdot)$ , and a parameter  $\alpha$  to be defined later.
  - 2: **Parameters:**  $\ell$  is  $G$ -Lipschitz,  $L$ -smooth and convex over  $w$ .
  - 3: Let  $\tilde{\ell}(w; z, z') = \ell(w; z, z') + \frac{\alpha}{2} \|w - w_0\|_2^2$ ,  $w_0$  is any point in  $\mathcal{C}$ .
  - 4: Run Algorithm 3.3.14 with loss function  $\tilde{\ell}$ , which is  $\tilde{G} = G + \alpha \|\mathcal{C}\|_2$ -Lipschitz,  $\tilde{L} = L + \alpha$ -smooth and  $\alpha$ -strongly convex.
- 

**Theorem 3.3.6.** For any  $0 < \epsilon, \delta \leq 1$ , Algorithm 3.3.14 is  $(\epsilon, \delta)$ -DP for any  $\alpha$ -strongly convex loss functions satisfying Assumption 3.3.1. Moreover, if  $\frac{\epsilon}{\log \frac{1}{\delta}} \ll 1$ , then with probability at least  $1 - 2\zeta$ , the output  $\hat{w}$  satisfies:

$$\text{Err}_{\mathcal{P}}(\hat{w}) \leq O\left(\frac{\sqrt{d}G^2 \|\mathcal{C}\|_2 \log \frac{n}{\zeta} \sqrt{\log \frac{1}{\delta}}}{\alpha n \epsilon} + \frac{1}{n} \sum_{t=1}^n \mathcal{R}_t(\ell \circ \mathcal{C})\right). \quad (3.118)$$

Algorithm 3.3.15 is  $(\epsilon, \delta)$ -DP for any convex loss functions satisfying Assumption 3.3.1 if  $\alpha = O(\frac{1}{\sqrt{n}})$ . Moreover, if  $\frac{\epsilon}{\log \frac{1}{\delta}} \ll 1$ , then with probability at least  $1 - 2\zeta$ , the output  $\hat{w}$  satisfies:

$$\text{Err}_{\mathcal{P}}(\hat{w}) \leq O\left(\frac{\sqrt{d}G^2\|\mathcal{C}\|_2^2 \log \frac{n}{\zeta} \sqrt{\log \frac{1}{\delta} \log n}}{\sqrt{n}\epsilon} + \frac{1}{n} \sum_{t=1}^n \mathcal{R}_t(\ell \circ \mathcal{C})\right). \quad (3.119)$$

From Theorem 3.3.6, we can see that for strongly and general convex loss functions, the bounds in (3.119) and (3.118) are respectively lower than those in (3.116) and (3.117). Specifically, for general convex loss functions, we can improve the upper bound from  $\tilde{O}(\frac{\sqrt{d}}{\epsilon \sqrt[4]{n}})$  to  $\tilde{O}(\frac{\sqrt{d}}{\epsilon \sqrt{n}})$ .

### 3.3.5 Experiments

In this section, we empirically evaluate the performance of the proposed differentially private algorithms on real-world datasets. We take two popular pairwise learning tasks, i.e., AUC maximization and metric learning, as examples. All of the experiments in this paper are conducted over 20 runs of different random permutations for each adopted dataset, and we report the averaged results.

#### Experimental setup

**Datasets.** We use six real-world datasets that are widely adopted in pairwise learning tasks. These datasets are the **Diabetes** dataset, the **Diabetic Retinopathy** dataset, the **Hepatitis** dataset, the **Parkinson Speech** dataset, the **Auto Riskness**<sup>9</sup> and the **Cancer** dataset [94]. The statistical information of them is described in Table 3.4.

**Performance measures.** To evaluate the performance of the proposed algorithms, we use the following measures:

1. **AUC:** For AUC maximization task, we report the AUC measurement [360] for each of

---

<sup>9</sup><http://www.gagolewski.com/resources/data/ordinal-regression/>

Table 3.4: The statistics of the adopted datasets.

<b>Dataset</b>	<b>Size</b>	<b>Dimension</b>
Diabetes	768	20
Hepatitis	155	19
Cancer	699	10
Diabetic Retinopathy	1, 151	20
Parkinson Speech	1, 040	27
Auto Riskness	160	16

the proposed algorithms over every adopted dataset. A larger AUC value means that the corresponding AUC maximization algorithm can generate more accurate results.

2. *Classification Accuracy*: For metric learning task, we calculate the classification accuracy that is defined as the percentage of the correctly classified samples in the test set. The less the classification accuracy, the worse the performance of the proposed algorithm. In this paper, the KNN classifier is adopted to assign labels to the test samples. For the KNN classifier, we set  $K$  to be 3.
3. *Objective function value*: For both metric learning task and AUC maximization task, we also report the objective function value of the proposed differentially private algorithms. A smaller objective function value means that the original pairwise learning model is less perturbed.

**Baselines.** Since there is no existing work that addresses the privacy issue in pairwise learning, in experiments, we take the original pairwise learning algorithms that do not take any actions to protect the private information as the baselines. We denote the baseline methods as **NonPrivate**, which is the GIGA for pairwise loss functions [175].

### Experiments for AUC maximization

We first evaluate the performance of the proposed differentially private pairwise learning algorithms (i.e., OnPairStrC, OnPairC, OffPairStrC and OffPairC) for AUC maximization task (see Example 2 for the problem formulation). We add additional  $\ell_2$  regularization

$\frac{\lambda}{2} \|w\|_2^2$  with  $\lambda = 10^{-3}$  to loss function for the strongly convex case.

We study the effect of the training size  $n$  and the privacy parameter  $\epsilon$  on the performance of the proposed OnPairStrC, OnPairC, OffPairStrC and OffPairC algorithms. Here we fix  $\delta = \frac{1}{n}$  and consider three cases where the value of parameter  $\epsilon$  is set to be 0.5, 1.5 and 2.5, respectively. For OnPairStrC and OffPairStrC, we vary the training size from 40 to 90 and conduct the experiment on the Hepatitis, Auto Riskness and Cancer datasets. For OnPairC and OffPairC, the experiment is conducted on the Diabetes, Parkinson Speech and Diabetic Retinopathy datasets and we vary the training size from 50 to 350. In Figure 3.9 and Figure 3.10, we respectively report the objective values of OnPairStrC and OnPairC. The experimental results show that the larger the value of the training size  $n$ , the smaller the objective value. Additionally, when  $n$  is fixed, the smaller the value of  $\epsilon$ , the larger the objective value is. The performance of the proposed algorithms are comparable with that of the baseline, which can be observed from Figure 3.10. The results for OffPairStrC and OffPairC are shown in Figure 3.11 and Figure 3.12, respectively. Figure 3.11 shows the objective value of OffPairStrC when the training size varies and Figure 3.12 reports the AUC measurement of OffPairC. The results in the two figures also show that the larger the training size is or privacy parameter  $\epsilon$  is, the higher the AUC measurement value is, which means that the proposed algorithm is less perturbed and more accurate. These experimental results verify that the proposed online differential private algorithms can achieve good utility while guarantee strong privacy protection when they are applied to the AUC maximization task.

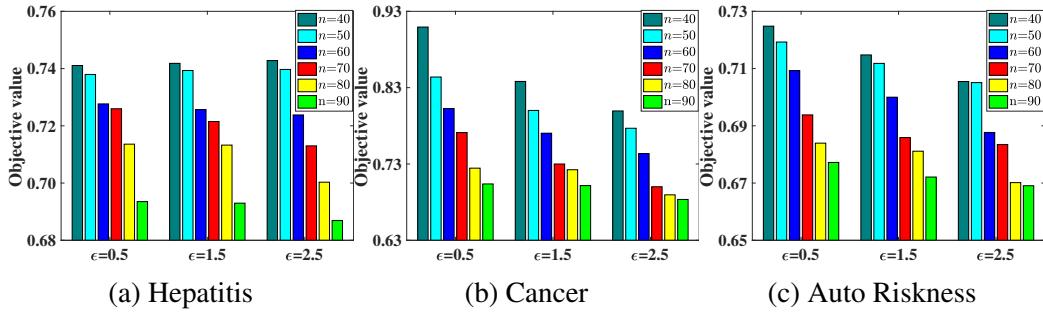


Figure 3.9: The objective value of OnPairStrC for AUC maximization.

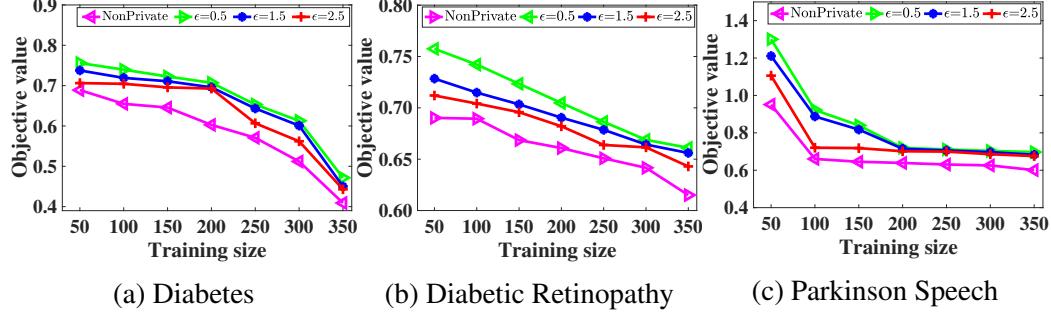


Figure 3.10: The objective value of OnPairC for AUC maximization.

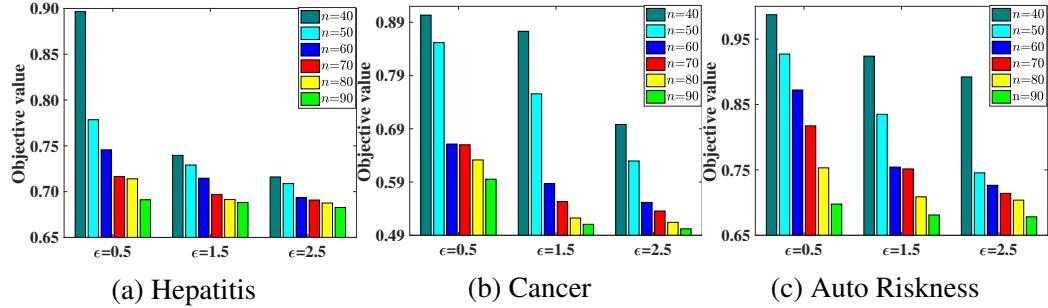


Figure 3.11: The objective value of OffPairStrC for AUC maximization.

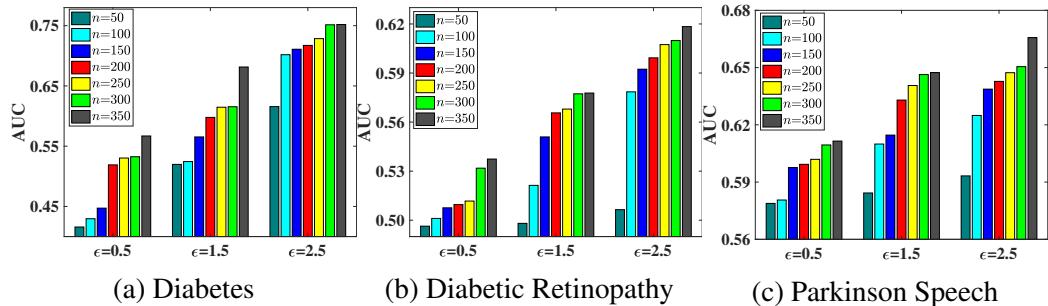


Figure 3.12: The AUC measurement of OffPairC.

# Experiments for Metric Learning

Next, we evaluate the performance of the proposed differentially private pairwise learning algorithms for the metric learning task (see Example 1 for the problem formulation). Similar to the experiments for AUC maximization, we evaluate the effect of the privacy parameter  $\epsilon$  and the training size  $n$ . In this section, we only report the experimental results for general convex pairwise learning algorithms, i.e., OnPairC and OffPairC.

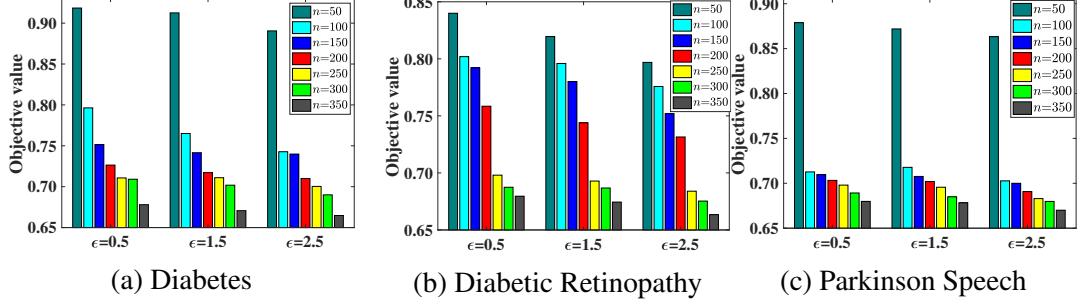


Figure 3.13: The objective value of OnPairC for metric learning task under different training sizes.

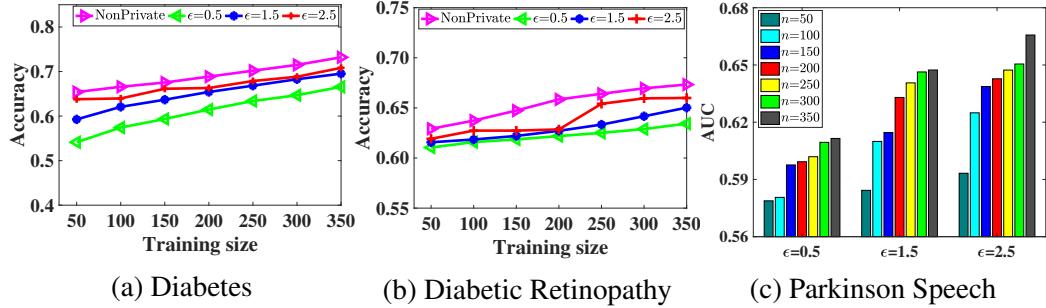


Figure 3.14: The classification accuracy of OffPairC for metric learning task under different training sizes.

In these experiments, the value of  $\delta$  is fixed as  $\frac{1}{n}$ , and we consider three cases where the parameter  $\epsilon$  is set to be 0.5, 1.5 and 2.5, respectively. We first calculate the objective value of OnPairC when the training size varies from 50 to 350, and the results on the Diabetes, Parkinson Speech and Diabetic Retinopathy datasets are shown in Figure 3.13. As for the offline algorithm OffPairC, we report the classification accuracy in Figure 3.14. As we can see, the derived experimental results are similar to that for AUC maximization. The proposed algorithms perform competitively with the baseline when we vary the values of  $n$  and  $\epsilon$ .

### 3.3.6 Omitted Proofs

#### Proof of Lemma 3.3.1

For the sake of convenience, we call the non-private version of Algorithm 3.3.9 as Pairwise GIGA and denote by  $w_t = \mathcal{A}(D)$ ,  $w'_t = \mathcal{A}(D')$ . Also, we let  $D_t = \{z_1, \dots, z_t\}$ .

We will show that the sensitivity of the  $t$ -th iteration in Pairwise GIGA is at most  $\frac{8G}{\alpha t}$ .

We prove it by induction.

We first consider the case  $1 \leq t \leq T_1$ . Since  $w_1, \dots, w_{T_1}$  are selected randomly, their values do not depend on the underlying dataset. Thus, we have  $w_t = w'_t$  for all  $1 \leq t \leq T_1$ .

Next, we consider  $t > T_1$ . There are two cases, *i.e.*,  $D - D' = \{z_t, z'_t\}$  and  $D - D' = \{z_i, z'_i\}$ , where  $i < t$ .

For the first case, since  $D - D' = \{z_t, z'_t\}$ , we have  $w_{t-1} = w'_{t-1}$ . Thus

$$\begin{aligned} \|w_t - w'_t\|_2 &\leq \|w_{t-1} - \eta_t \nabla \hat{L}_t(w_{t-1}, D_t) - w'_{t-1} + \eta_t \nabla \hat{L}_t(w_{t-1}, D'_t)\|_2 \\ &= \eta_t \|\nabla \hat{L}_t(w_{t-1}, D_t) - \nabla \hat{L}_t(w_{t-1}, D'_t)\|_2 \\ &\leq \frac{t-1}{t-2} \frac{2G}{\alpha t} \leq \frac{4G}{\alpha t}, \end{aligned}$$

where the last inequality is due to the G-Lipschitz assumption on  $\ell$  and the assumption of  $t \geq 3$ .

For the second case, we have the following

$$\|w_t - w'_t\|_2^2 \leq \|(w_{t-1} - \eta_t \nabla \hat{L}_t(w_{t-1}, D_t)) - (w'_{t-1} - \eta_t \nabla \hat{L}_t(w'_{t-1}, D'_t))\|_2^2 \quad (3.120)$$

$$\begin{aligned} &\leq \|w_{t-1} - w'_{t-1}\|_2^2 + \eta_t^2 \|\nabla \hat{L}_t(w_{t-1}, D_t) - \nabla \hat{L}_t(w'_{t-1}, D'_t)\|_2^2 \\ &\quad - 2\eta_t (w_{t-1} - w'_{t-1})^T (\nabla \hat{L}_t(w_{t-1}, D_t) - \nabla \hat{L}_t(w'_{t-1}, D'_t)). \end{aligned} \quad (3.121)$$

For the term  $\|\nabla \hat{L}_t(w_{t-1}, D_t) - \nabla \hat{L}_t(w'_{t-1}, D'_t)\|_2^2$ , we have

$$\begin{aligned}
& \|\nabla \hat{L}_t(w_{t-1}, D_t) - \nabla \hat{L}_t(w'_{t-1}, D'_t)\|_2^2 \\
&= \left\| \frac{1}{t-1} \sum_{j \neq i} [\nabla \ell(w_{t-1}; z_t, z_j) - \nabla \ell(w'_{t-1}; z_t, z_j)] \right. \\
&\quad \left. + \frac{1}{t-1} [\nabla \ell(w_{t-1}; z_t, z_i) - \nabla \ell(w'_{t-1}; z_t, z'_i)] \right\|_2^2 \\
&\leq 2 \left\| \frac{1}{t-1} \sum_{j \neq i} [\nabla \ell(w_{t-1}; z_t, z_j) - \nabla \ell(w'_{t-1}; z_t, z_j)] \right\|_2^2 \\
&\quad + 2 \left\| \frac{1}{t-1} [\nabla \ell(w_{t-1}; z_t, z_i) - \nabla \ell(w'_{t-1}; z_t, z'_i)] \right\|_2^2 \\
&\leq 2L^2 \left( \frac{t-2}{t-1} \right)^2 \|w_{t-1} - w'_{t-1}\|_2^2 + \frac{8G^2}{(t-1)^2}, \tag{3.122}
\end{aligned}$$

where the last inequality is due to the  $L$ -smoothness and  $G$ -Lipschitz of the loss function  $\ell$ .

For the term  $(w_{t-1} - w'_{t-1})^T (\nabla \hat{L}_t(w_{t-1}, D_t) - \nabla \hat{L}_t(w'_{t-1}, D'_t))$ , we have:

$$\begin{aligned}
& (w_{t-1} - w'_{t-1})^T (\nabla \hat{L}_t(w_{t-1}, D_t) - \nabla \hat{L}_t(w'_{t-1}, D'_t)) \\
&= (w_{t-1} - w'_{t-1})^T \left[ \frac{1}{t-1} \sum_{j \neq i} [\nabla \ell(w_{t-1}; z_t, z_j) - \nabla \ell(w'_{t-1}; z_t, z_j)] \right. \\
&\quad \left. + \frac{1}{t-1} [\nabla \ell(w_{t-1}; z_t, z_i) - \nabla \ell(w'_{t-1}; z_t, z'_i)] \right]. \tag{3.123}
\end{aligned}$$

By the  $\alpha$ -strongly convexity of the loss function, we have

$$(w_{t-1} - w'_{t-1})^T \left[ \frac{1}{t-1} \sum_{j \neq i} [\nabla \ell(w_{t-1}; z_t, z_j) - \nabla \ell(w'_{t-1}; z_t, z_j)] \right] \geq \alpha \frac{t-2}{t-1} \|w_{t-1} - w'_{t-1}\|_2^2. \tag{3.124}$$

Also due to the  $G$ -Lipschitz, we have

$$|(w_{t-1} - w'_{t-1})^T \left[ \frac{1}{t-1} [\nabla \ell(w_{t-1}; z_t, z_i) - \nabla \ell(w'_{t-1}; z_t, z'_i)] \right]| \leq \frac{2G \|w_{t-1} - w'_{t-1}\|_2}{t-1}. \tag{3.125}$$

Plugging (3.124) and (3.125) into (3.123), we have

$$(w_{t-1} - w'_{t-1})^T (\nabla \hat{L}_t(w_{t-1}, D_t) - \nabla \hat{L}_t(w'_{t-1}, D'_t)) \quad (3.126)$$

$$\geq \alpha \frac{t-2}{t-1} \|w_{t-1} - w'_{t-1}\|_2^2 - \frac{2G\|w_{t-1} - w'_{t-1}\|_2}{t-1}. \quad (3.127)$$

Plugging (3.126) and (3.122) into (3.121), we get

$$\begin{aligned} \|w_t - w'_t\|_2^2 &\leq (1 + 2L^2\eta_t^2(\frac{t-2}{t-1})^2 - 2\eta_t\alpha\frac{t-2}{t-1})\|w_{t-1} - w'_{t-1}\|_2^2 \\ &\quad + \frac{8G^2\eta_t^2}{(t-1)^2} + \frac{4\eta_t G\|w_{t-1} - w'_{t-1}\|_2}{t-1}. \end{aligned} \quad (3.128)$$

Now taking  $\eta_t = \frac{t-1}{t-2}\frac{2}{\alpha t}$  and  $\|w_{t-1} - w'_{t-1}\|_2 \leq \frac{8G}{\alpha(t-1)}$ , we have

$$\begin{aligned} \|w_t - w'_t\|_2^2 &\leq (1 + \frac{8L^2}{\alpha^2 t^2} - \frac{4}{t})\frac{64G^2}{\alpha^2(t-1)^2} + \frac{32G^2}{\alpha^2 t^2(t-2)^2} + \frac{64G^2}{\alpha^2 t(t-1)(t-2)} \\ &\leq (1 + \frac{8L^2}{\alpha^2 t^2} - \frac{4}{t} + \frac{1}{2(t-2)^2} + \frac{1}{(t-2)})\frac{64G^2}{\alpha^2(t-1)^2}. \end{aligned} \quad (3.129)$$

What we still need to prove is

$$(1 + \frac{8L^2}{\alpha^2 t^2} - \frac{4}{t} + \frac{1}{2(t-2)^2} + \frac{1}{(t-2)})\frac{64G^2}{\alpha^2(t-1)^2} \leq \frac{64G^2}{\alpha^2 t^2}. \quad (3.130)$$

After simplifying both sides we now need to show

$$\frac{8L^2}{\alpha^2} + \frac{t^2}{2(t-2)^2} + \frac{t^2}{t-2} \leq 2t + 1. \quad (3.131)$$

By the assumption on  $t \geq T_1 = \max\{\frac{16L^2}{\alpha^2}, 7\}$ , we have  $\frac{t}{2} \geq \frac{8L^2}{\alpha^2}$ ,  $\frac{3}{2}t \geq \frac{t^2}{t-2}$  and  $1 \geq \frac{t^2}{2(t-2)^2}$ .

Thus, (3.131) is true, and we have

$$\|w_t - w'_t\|_2^2 \leq \frac{64G^2}{\alpha^2 t^2}.$$

This completes the proof.

### Proof of Theorem 3.3.1

By Lemma 3.3.1, we know the  $\ell_2$  norm sensitivity in the  $t$ -th iteration is upper bounded by  $\frac{8G}{\alpha t}$ . Now, by the Gaussian mechanism we can get that each iteration of Algorithm 3.3.9 is  $\frac{\rho}{n-T_1}$ -zCDP for  $T_1 < t \leq n$ . Then by Lemma composition theorem of zCDP we can see that Algorithm 3.3.9 is  $\rho$ -zCDP. Thus it is  $(\epsilon, \delta)$ -DP.

### Proof of Theorem 3.3.2

For the sake of convenience, we call the non-private version of Algorithm 3.3.9 as Pairwise GIGA and denote by  $w_t = \mathcal{A}(D)$ ,  $w'_t = \mathcal{A}(D')$ . Also, we let  $D_t = \{z_1, \dots, z_t\}$ . As we said earlier, in the case of  $\frac{\epsilon}{\log \frac{1}{\delta}} \ll 1$  we can see  $\sigma_t^2 = c \frac{\log \frac{1}{\delta} G^2(n-T_1)}{\alpha^2 \epsilon^2 t^2} = O(\frac{\log \frac{1}{\delta} G^2 n}{\alpha^2 \epsilon^2 t^2})$  for  $c = 128$ .

We first prove the following lemma:

**Lemma 3.3.2.** Let  $\mathcal{R}_{GIGA}(n, D)$  be the regret of (non-private) Pairwise GIGA on the stream  $\{z_1, z_2, \dots, z_n\}$ , then the outputs  $w_1, \dots, w_{n-1}$  satisfies

$$\begin{aligned} & \sum_{t=2}^n \hat{L}_t(w_{t-1}, D_t) - \min_{w \in \mathcal{C}} \sum_{t=2}^n \hat{L}_t(w, D_t) \\ & \leq \mathcal{R}_{GIGA}(n, D) + G \sum_{t=T_1+1}^T \|n_{t-1}\|_2 + GT_1 \|\mathcal{C}\|_2. \end{aligned} \tag{3.132}$$

*Proof of Lemma 3.3.2.* We denote the output of Pairwise GIGA as  $\tilde{w}_1, \dots, \tilde{w}_{n-1}$ . Then, by

the G-Lipschitz property of  $\ell$  and  $\hat{L}_t$ , we get

$$\begin{aligned}
& \sum_{t=2}^n \hat{L}_t(w_{t-1}, D_t) - \min_{w \in \mathcal{C}} \sum_{t=2}^n \hat{L}_t(w, D_t) \\
& \leq \sum_{t=2}^n \hat{L}_t(w_{t-1}, D_t) - \sum_{t=2}^n \hat{L}_t(\tilde{w}_{t-1}, D_t) + \mathcal{R}_{GIGA}(n, D) \\
& \leq G \sum_{t=2}^n \|w_{t-1} - \tilde{w}_{t-1}\|_2 + \mathcal{R}_{GIGA}(n, D) \\
& = \mathcal{R}_{GIGA}(n, D) + G \sum_{t=T_1+1}^n \|n_{t-1}\|_2 + GT_1 \|\mathcal{C}\|_2.
\end{aligned}$$

□

Next we bound the term of  $\sum_{t=T_1+1}^T \|n_{t-1}\|_2$ . For a Gaussian distribution  $x \sim \mathcal{N}(0, \sigma^2 I_d)$ , with probability at least  $1 - \zeta$  we have  $\|x - \sigma\|_2 \leq \sigma\sqrt{d}\sqrt{2\log 2/\zeta}$ . Thus, by the above concentration bound and taking the union, we have the following with probability at least  $1 - \zeta$

$$\begin{aligned}
\sum_{t=T_1+1}^T \|n_{t-1}\|_2 & \leq O\left(\sum_{t=T_1}^n \frac{\sqrt{d}\sqrt{\log \frac{n}{\zeta}}G\sqrt{n-T_1}\sqrt{\log 1/\delta}}{\alpha\epsilon t}\right) \\
& \leq O\left(\frac{G\sqrt{d}\log^{1.5}\frac{n}{\zeta}\sqrt{n}\sqrt{\log 1/\delta}}{\alpha\epsilon}\right).
\end{aligned} \tag{3.133}$$

Combining this with Lemma 3.3.2 and (3.133), we can get the following with probability at least  $1 - \zeta$

$$\begin{aligned}
& \sum_{t=2}^n \hat{L}_t(w_{t-1}, D_t) - \min_{w \in \mathcal{C}} \sum_{t=2}^n \hat{L}_t(w, D_t) \\
& \leq \mathcal{R}_{GIGA}(n) + O\left(\frac{G^2\sqrt{d}\log^{1.5}\frac{n}{\zeta}\sqrt{n}\sqrt{\log 1/\delta}}{\alpha\epsilon} + \frac{GL^2}{\alpha^2} \|\mathcal{C}\|_2\right).
\end{aligned}$$

Using the regret bound analysis of GIGA in [371, 144] on strongly convex functions  $\{\hat{L}_t\}_{t=1}^n$  and by the fact that they are  $\alpha$ -strongly convex, we can get

$$\mathcal{R}_{GIGA}(n, D) \leq \frac{G^2(1 + \log n)}{2\alpha}.$$

Thus, in total we have

$$\begin{aligned} & \sum_{t=2}^n \hat{L}_t(w_{t-1}, D_t) - \min_{w \in \mathcal{C}} \sum_{t=2}^n \hat{L}_t(w, D_t) \\ & \leq O\left(\frac{G^2 \sqrt{d} \log^{1.5} \frac{n}{\zeta} \sqrt{n} \sqrt{\log 1/\delta}}{\alpha \epsilon} + \frac{GL^2}{\alpha^2} \|\mathcal{C}\|_2 + \frac{G^2(1 + \log n)}{\alpha}\right). \end{aligned} \quad (3.134)$$

For the expected regret, we only need to get an upper bound on  $\mathbb{E} \sum_{t=T_1+1}^n \|n_t\|_2$ . We can follow the techniques in [160] and show that

$$\mathbb{E} \sum_{t=T_1+1}^n \|n_t\|_2 \leq O\left(\sqrt{d} \frac{G}{\alpha} \sqrt{n} \frac{\log n \sqrt{\log 1/\delta}}{\epsilon}\right). \quad (3.135)$$

### Proof of Theorem 3.3.3

By the perturbation strategy in Algorithm 3.3.10 and Theorem 3.3.1 we can get the following for the loss function after perturbation  $\tilde{\ell} = \ell + \frac{\alpha}{2} \|w - w_0\|^2$

$$\begin{aligned} \mathcal{R}_{\mathcal{A}, \tilde{\ell}}(n, D) & \leq O\left(\frac{(G + \alpha \|\mathcal{C}\|_2)^2 \sqrt{d} \log^{1.5} \frac{n}{\zeta} \sqrt{n} \sqrt{\log 1/\delta}}{\alpha \epsilon}\right. \\ & \quad \left. + \frac{(G + \alpha \|\mathcal{C}\|_2)(L + \alpha)^2}{\alpha^2} \|\mathcal{C}\|_2 + \frac{(G + \alpha \|\mathcal{C}\|_2)^2(1 + \log n)}{\alpha}\right). \end{aligned}$$

Since  $\mathcal{R}_{\mathcal{A}, \ell}(n, D) \leq \mathcal{R}_{\mathcal{A}, \tilde{\ell}}(n, D) + n\alpha \|\mathcal{C}\|_2^2$ , we have

$$\begin{aligned} \mathcal{R}_{\mathcal{A}, \ell}(n, D) & \leq O\left(\frac{(G + \alpha \|\mathcal{C}\|_2)^2 \sqrt{d} \log^{1.5} \frac{n}{\zeta} \sqrt{n} \sqrt{\log 1/\delta}}{\alpha \epsilon}\right. \\ & \quad \left. + \frac{(G + \alpha \|\mathcal{C}\|_2)(L + \alpha)^2}{\alpha^2} \|\mathcal{C}\|_2 + \frac{(G + \alpha \|\mathcal{C}\|_2)^2(1 + \log n)}{\alpha} + n\alpha \|\mathcal{C}\|_2^2\right). \end{aligned}$$

Taking  $\alpha = O(\frac{1}{\sqrt[4]{n}})$ , we get

$$\mathcal{R}_{\mathcal{A}, \ell}(n, D) \leq O\left(\frac{L^2 G^2 \|\mathcal{C}\|_2^2 \sqrt{d} \log^{1.5} \frac{n}{\zeta} n^{\frac{3}{4}} \sqrt{\log 1/\delta}}{\epsilon}\right).$$

### Proof of Theorem 3.3.4

For the  $(\epsilon, \delta)$ -DP, we can follow the proof of Theorem 3.3.1. What we only need to show is the bound, we have the following lemma:

**Lemma 3.3.3** (Theorem 6 in [175]). Let  $w_1, \dots, w_{n-1}$  be an ensemble of parameters generated by an online algorithm working with a finite buffer of capacity  $s$  and a  $B$ -bounded loss function  $\ell$ . Moreover, suppose that the algorithm guarantees a finite-buffer regret bound of  $\mathcal{R}_{\mathcal{A}}^{\text{buf}}(n)$ . Then, for any  $\delta > 0$ , we have the following with probability at least  $1 - \zeta$ :

$$\mathcal{R}_{\mathcal{A}}(n) \leq \frac{\mathcal{R}_{\mathcal{A}}^{\text{buf}}(n)}{n-1} + O\left(\frac{C_d}{\sqrt{s}} + B\sqrt{\frac{\log \frac{n}{\zeta}}{s}}\right).$$

**Lemma 3.3.4** (Lemma 26 in [175]). Suppose we have an online algorithm  $\mathcal{A}$  that incurs finite-buffer penalties based on a buffer  $B$  of size  $s$  that is updated using RS-x. Suppose further that the learning algorithm generates  $\{w_1, \dots, w_{n-1}\}$ , then with probability at least  $1 - \delta$ , we have

$$\mathcal{R}_{\mathcal{P}, \mathcal{A}}(n) \leq \mathcal{R}_{\mathcal{A}}^{\text{buf}}(n) + O(C_d n \sqrt{\frac{\log \frac{n}{\zeta}}{s}}).$$

Now we denote the regret of the non-private version of Algorithm 3.3.12 as  $\mathcal{R}_{GIGA}^{\text{buf}}(n)$ . Thus, by the same proof as in Theorem 3.3.2 we have the following

$$\mathcal{R}_{\mathcal{A}}^{\text{buf}}(n) \leq \mathcal{R}_{GIGA}^{\text{buf}}(n) + O\left(\frac{G^2 \sqrt{d} \log^{1.5} \frac{n}{\zeta} \sqrt{n} \sqrt{\log 1/\delta}}{\alpha \epsilon} + \frac{GL^2}{\alpha^2} \|\mathcal{C}\|_2\right).$$

For the term  $\mathcal{R}_{GIGA}^{\text{buf}}(n)$ , since  $\{\hat{L}_t^{\text{buf}}\}_{t=1}^n$  are all  $\alpha$ -strongly convex, then we can use the GIGA algorithm in [144] with the functions  $\{\hat{L}_t^{\text{buf}}\}_{t=1}^n$  and get

$$\mathcal{R}_{GIGA}^{\text{buf}}(n) \leq \frac{2G^2(1 + \log n)}{\alpha}.$$

Thus we get the proofs.

For the general convex loss function, we have the same trick as in the proofs of Theorem

3.3.2 and 3.3.3, that is, replacing  $G = G + \alpha \|\mathcal{C}\|_2$ ,  $L = L + \alpha \|\mathcal{C}\|_2$  and take  $\alpha = O(\frac{1}{\sqrt[4]{n}})$ .

### Proof of Theorem 3.3.5

We first rephrase a lemma in [175].

**Lemma 3.3.5** (Theorem 3 in [175]). Let  $w_1, \dots, w_{n-1}$  be an ensemble of parameters generated by an online learning algorithm working with a  $B$ -bounded pairwise loss function  $\ell$  that guarantees a regret bound of  $\mathcal{R}(n)$ . Then for any  $\delta > 0$ , we have the following with probability at least  $1 - \delta$ ,

$$\begin{aligned} L_{\mathcal{P}}(\bar{w}) &\leq \frac{1}{n-1} \sum_{t=2}^n L_{\mathcal{P}}(w_{t-1}) \\ &\leq \min_{w \in \mathcal{C}} L_{\mathcal{P}}(w) + \frac{4}{n-1} \sum_{t=2}^n \mathcal{R}_{t-1}(\ell \circ \mathcal{C}) + \frac{\mathcal{R}(n)}{n-1} + 6B \sqrt{\frac{\log \frac{n}{\delta}}{n-1}}, \end{aligned} \quad (3.136)$$

where  $\mathcal{R}(\ell \circ \mathcal{C})$  is the Rademacher average for the class of functions  $\ell \circ \mathcal{C}$  in (3).

For the strongly convex loss functions, by Lemma 3.3.5, we can get with probability at least  $1 - 2\zeta$ ,

$$L_{\mathcal{P}}(\bar{w}) - \min_{w \in \mathcal{C}} L_{\mathcal{P}}(w) \leq O\left(\frac{1}{n-1} \sum_{t=2}^n \mathcal{R}_{t-1}(\ell \circ \mathcal{C}) + \frac{G^2 L^2 \|\mathcal{C}\|_2 \sqrt{d} \log^{1.5} \frac{n}{\zeta} \sqrt{\log 1/\delta}}{\alpha^2 \epsilon \sqrt{n}}\right). \quad (3.137)$$

For the general convex ones, we have with probability at least  $1 - 2\zeta$

$$\begin{aligned} L_{\mathcal{P}}(\bar{w}) - \min_{w \in \mathcal{C}} L_{\mathcal{P}}(w) &\leq O\left(\frac{1}{n-1} \sum_{t=2}^n \mathcal{R}_{t-1}(\ell \circ \mathcal{C}) + G \|\mathcal{C}\|_2 \sqrt{\frac{\log \frac{n}{\zeta}}{n-1}} + \frac{L^2 G^2 \|\mathcal{C}\|_2^2 \sqrt{d} \log^{1.5} \frac{n}{\zeta} n^{\frac{3}{4}} \sqrt{\log 1/\delta}}{\epsilon(n-1)}\right) \\ &= O\left(\frac{1}{n-1} \sum_{t=2}^n \mathcal{R}_{t-1}(\ell \circ \mathcal{C}) + \frac{L^2 G^2 \|\mathcal{C}\|_2^2 \sqrt{d} \log^{1.5} \frac{n}{\zeta} \sqrt{\log 1/\delta}}{\epsilon \sqrt[4]{n}}\right). \end{aligned}$$

### Proof of Theorem 3.3.6

We first prove Algorithm 3.3.14 is  $(\epsilon, \delta)$  differentially private. What we only need to show is the sensitivity of  $\tilde{w}$  is  $\frac{8G \log n}{n\alpha}$ . Since by Lemma 3.3.1, we know  $\|w_t - w'_t\|_2 \leq \frac{8G}{\alpha t}$ , thus

$$\|\bar{w} - \bar{w}'\|_2 \leq \frac{\sum_{t=1}^n \frac{8G}{\alpha t}}{n} \leq \frac{8G \log n}{n\alpha}. \quad (3.138)$$

Thus by Gaussian mechanism we can show it is  $(\epsilon, \delta)$ -differentially private.

Next we analyze the generalization error, we have the following with probability  $1 - \zeta$ :

$$\begin{aligned} & L_{\mathcal{P}}(\hat{w}) - \min_{w \in \mathcal{C}} L_{\mathcal{P}}(w) \\ & \leq L_{\mathcal{P}}(\hat{w}) - L_{\mathcal{P}}(\tilde{w}) + L_{\mathcal{P}}(\tilde{w}) - \min_{w \in \mathcal{C}} L_{\mathcal{P}}(w) \end{aligned} \quad (3.139)$$

$$\begin{aligned} & \leq G\|\hat{w} - \tilde{w}\|_2 + \frac{4}{n-1} \sum_{t=2}^n \mathcal{R}_{t-1}(\ell \circ \mathcal{C}) + \frac{\mathcal{R}_{GIGA}(n, D)}{n} + 6G\|\mathcal{C}\|_2 \sqrt{\frac{\log \frac{n}{\zeta}}{n}}, \\ & \leq G\|\hat{w} - \bar{w}\|_2 + G\|\bar{w} - \tilde{w}\|_2 + \frac{4}{n-1} \sum_{t=2}^n \mathcal{R}_{t-1}(\ell \circ \mathcal{C}) + \frac{\mathcal{R}_{GIGA}(n, D)}{n} + 6G\|\mathcal{C}\|_2 \sqrt{\frac{\log \frac{n}{\zeta}}{n}}, \end{aligned} \quad (3.140)$$

where  $\mathcal{R}_{GIGA}(n, D)$  is the regret of Pairwise GIGA on the strongly convex loss function  $\{\hat{L}_t\}_{t=1}^n$ . The last inequality is by the  $G$ -Lipschitz property and Lemma 3.3.5.

Also, by [144], the regret of Pairwise GIGA on the strongly convex loss function  $\{\hat{L}_t\}_{t=1}^n$  is  $\mathcal{R}_{GIGA}(n, D) \leq \frac{2G^2(1+\log n)}{\alpha}$ . For the term  $\|\hat{w} - \bar{w}\|_2$ , by definition of  $\hat{w}$ , we have

$$\|\hat{w} - \bar{w}\|_2 \leq \|\tilde{w} - \bar{w}\|_2.$$

For the term  $\|\bar{w} - \tilde{w}\| = \|\sigma\|$ , we have with probability at least  $1 - \zeta$ ,

$$\|\sigma\|_2 \leq \frac{8G\sqrt{d}\sqrt{2}\sqrt{\log 1/\zeta \log 1.25/\delta} \log n}{\alpha n \epsilon}.$$

Thus in total we have:

$$\begin{aligned} L_{\mathcal{P}}(\bar{w}) - \min_{w \in \mathcal{C}} L_{\mathcal{P}}(w) &\leq O\left(\frac{\sqrt{d}G^2\sqrt{\log 1/\zeta \log 1/\delta} \log n}{\alpha n \epsilon}\right. \\ &\quad \left.+ \frac{1}{n} \sum_{t=1}^n \mathcal{R}_t(\ell \circ \mathcal{C}) + \frac{G^2 \log n}{\alpha n} + G\|\mathcal{C}\|_2 \sqrt{\frac{\log \frac{n}{\zeta}}{n}}\right). \end{aligned}$$

For the convex loss function, as the same as above, we have

$$\begin{aligned} L_{\mathcal{P}}(\bar{w}) - \min_{w \in \mathcal{C}} L_{\mathcal{P}}(w) &\leq O\left(\frac{\sqrt{d}(G + \alpha\|\mathcal{C}\|_2)^2 \sqrt{\log 1/\zeta \log 1/\delta} \log n}{\alpha n \epsilon}\right. \\ &\quad \left.+ \frac{1}{n} \sum_{t=1}^n \mathcal{R}_t(\ell \circ \mathcal{C}) + \frac{(G + \alpha\|\mathcal{C}\|_2)^2 \log n}{\alpha n} + (G + \alpha\|\mathcal{C}\|_2)\|\mathcal{C}\|_2 \sqrt{\frac{\log \frac{n}{\zeta}}{n}} + \alpha\|\mathcal{C}\|_2^2\right). \end{aligned}$$

When we take  $\alpha = O(\frac{1}{\sqrt{n}})$ , we have

$$L_{\mathcal{P}}(\bar{w}) - \min_{w \in \mathcal{C}} L_{\mathcal{P}}(w) \leq O\left(\frac{\sqrt{d}G^2\|\mathcal{C}\|_2^2 \log \frac{n}{\zeta} \sqrt{\log 1/\delta} \log n}{\sqrt{n}\epsilon} + \frac{1}{n} \sum_{t=1}^n \mathcal{R}_t(\ell \circ \mathcal{C})\right). \tag{3.141}$$

# **Chapter 4**

## **Empirical Risk Minimization with Non-Convex Loss Functions in Differential Privacy Model**

In Chapter 3 we studied different settings of DP-ERM with convex loss functions. However, several empirical studies have revealed that non-convex loss functions can achieve better classification accuracy than convex loss functions [231], and recent developments in Deep Neural Networks [133] have further suggested that the loss functions are more likely to be non-convex in real world applications. Thus, there is an urgent need for the research community to shift its focus from convex to non-convex loss functions. However, due to the fact that finding the global minimum for non-convex functions is NP-hard, which implies that measuring the utility by the expected excess empirical risk may not always be a good choice. Thus, to study the problem, one possible is to change the measurement of error for our private estimator. So far, only a few papers [356, 324] have considered the utility of DP-ERM with non-convex loss functions, but all of them measure the utility by  $\ell_2$  norm of the gradient, instead of the expected excess empirical risk. In the following three sections, we will study the theoretical behaviors under three different types of measurements, *i.e.*,

first order stationary measurement, excess empirical (population) risk and second order stationary measurement, respectively. To make each chapter independent and self-contained, we will review the definition of DP-ERM in each Chapter. We also note that the notations of loss function, constraint set and parameter space may be different across different chapters. We first review some definitions in convex optimization that will be used throughout the whole chapter.

**Definition 4.0.1** (Lipschitz Function). A loss function  $f : \mathcal{C} \times \mathcal{X} \rightarrow \mathbb{R}$  is G-Lipschitz (under  $\ell_2$ -norm) over  $\theta$ , if for any  $z \in \mathcal{X}$  and  $\theta_1, \theta_2 \in \mathcal{C}$ , we have  $|f(\theta_1, z) - f(\theta_2, z)| \leq G\|\theta_1 - \theta_2\|_2$ .

**Definition 4.0.2** (L-smooth Function). A loss function  $f : \mathcal{C} \times \mathcal{X} \rightarrow \mathbb{R}$  is L-smooth over  $\theta$  with respect to the norm  $\|\cdot\|$  if for any  $z \in \mathcal{X}$  and  $\theta_1, \theta_2 \in \mathcal{C}$ , we have

$$\|\nabla f(\theta_1, z) - \nabla f(\theta_2, z)\|_* \leq L\|\theta_1 - \theta_2\|,$$

where  $\|\cdot\|_*$  is the dual norm of  $\|\cdot\|$ . If  $f$  is differentiable, this yields

$$f(\theta_1, z) \leq f(\theta_2, z) + \langle \nabla f(\theta_2, z), \theta_1 - \theta_2 \rangle + \frac{L}{2}\|\theta_1 - \theta_2\|^2.$$

**Definition 4.0.3** ( $\rho$ -Hessian Lipschitz). A twice-differentiable loss function  $\ell : \mathcal{C} \times \mathcal{X} \rightarrow \mathbb{R}$  is called  $\rho$ -Hessian Lipschitz if for any  $z \in \mathcal{X}$  and  $\theta_1, \theta_2 \in \mathcal{C}$  we have

$$\|\nabla^2 \ell(\theta_1, z) - \nabla^2 \ell(\theta_2, z)\|_2 \leq \rho\|\theta_1 - \theta_2\|_2.$$

## 4.1 First Order Stationary View

Before going into details, we first review the definition of DP-ERM with convex loss functions.

**Definition 4.1.1** (DP-ERM). Given a dataset  $D = \{z_1, \dots, z_n\}$  from a data universe  $\mathcal{X}$  and a closed convex set  $\mathcal{C} \subseteq \mathbb{R}^p$ , DP-ERM is to find  $x^{\text{priv}} \in \mathcal{C}$  so as to minimize the empirical risk, *i.e.*  $F^r(x, D) = \frac{1}{n} \sum_{i=1}^n f(x, z_i) + r(x)$ , with the guarantee of being differentially private, where  $f$  is the loss function and  $r$  is some simple (non-)smooth convex function called **regularizer**. When the inputs are drawn i.i.d from an unknown underlying distribution  $\mathcal{P}$  on  $\mathcal{X}$ , we also consider the population risk  $\mathbb{E}_{z \sim \mathcal{P}}[f(x, z)]$ . If the loss function is convex, the utility of the algorithm is measured by the expected excess empirical risk, *i.e.*  $\mathbb{E}_{\mathcal{A}}[F^r(x^{\text{priv}}, D)] - \min_{x \in \mathcal{C}} F^r(x, D)$ , or the expected excess population risk (generalization error), *i.e.*  $\mathbb{E}_{z \sim \mathcal{P}, \mathcal{A}}[f(x^{\text{priv}}, z)] - \min_{x \in \mathcal{C}} \mathbb{E}_{z \sim \mathcal{P}}[f(x, z)]$ , where the expectation of  $\mathcal{A}$  is taking over all the randomness of the algorithm.

As we mentioned in previously, to study DP-ERM with non-convex loss functions, one way is to change the measurement of error. Since we know that for any function, its global minimum satisfies that its  $\ell_2$ -norm of the gradient of the function is zero, thus, we can use the  $\ell_2$ -norm of the gradient of the function as a measurement. So far, only a few papers [356, 324] have considered the utility of DP-ERM with non-convex loss functions. Despite the aforementioned progress on this problem, there are still quite a few remaining issues. 1) Previous work has obtained the error bounds for the smooth loss functions with smooth regularizer; it is not clear whether they can be extended to non-smooth regularizer, such as  $\ell_1$  norm. 2) Even though existing work has considered the error bound measured by empirical risk, it is not clear what is the generalization property of the problem. Particularly, it is unknown what is the error bound measured by population risk for non-convex loss functions and its difference with the convex ones [29]. 3) Existing work mainly focuses on the low dimensional case, where  $n \gg p$ . It is still unknown what can be done for the high dimensional case. In this paper, we will address the above issues. Our main results are listed in **Table 4.1**. Below is a more detailed description of our contributions.

1. For low dimensional space, we consider the general case for DP-ERM with non-convex loss function and non-smooth regularizer. For this case (see **Algorithm 4.1.16**), we

generalize the approaches in [356, 324], which consider only smooth regularizer and unconstrained domain, *i.e.*  $\mathcal{C} = \mathbb{R}^p$ ). Particularly, we use as the utility the  $\ell_2$  norm of the projected gradient, while [356, 324] use the  $\ell_2$  norm of the gradient. Then, we resolve some practical issues in [356, 324] by using zero Concentrated Differential Privacy. Finally, we study the generalization property of the private estimator. By using  $\ell_2$  norm of the gradient in the empirical risk, we show an upper bound of the population risk with non-convex loss functions at the point  $\theta^{\text{priv}}$  based on **the expected  $\ell_2$ -norm of the gradient**, *i.e.*  $\mathbb{E}_{\mathcal{A}} \|\mathbb{E}_{z \sim \mathcal{P}} [\nabla f(x^{\text{priv}}, z)]\|_2$ .

2. For high dimensional space (*i.e.*  $p \gg n$ ), we first show that by using the differentially private version of Frank-Wolfe method, it is possible to measure the utility by Frank-Wolfe gap (see **Algorithm 4.1.17**), and the utility upper bound depends only on the Gaussian Width of the constraint set  $\mathcal{C}$  (see Definition 4.1.3), instead of the dimensionality  $p$  of the underlying space. Then, we improve the robustness of the above approach for non-smooth regularizer, while still maintain the same utility upper bound (see **Algorithm 4.1.18**) for the case of  $\|\mathcal{C}\|_2 \leq 1$  by using the  $\ell_2$  norm of the projected gradient. Finally, we consider a special case where  $\mathcal{C}$  is a polytope and the loss function is  $\ell_1$ -Lipschitz, which has been studied in [270] for the convex case. For this case (see **Algorithm 4.1.19**), we present a method which uses Frank-Wolfe gap to measure the utility and achieves an upper bound depending only on  $\log p$ , instead of the Gaussian Width or the dimensionality of the underlying space.

### 4.1.1 Low Dimension Case

#### Extending to Non-Smooth Regularizer

In this section, we consider DP-ERM with non-convex loss function and non-smooth convex regularizer, *i.e.*,

Method	Assumption	Utility Upper Bd	Non-smooth Regularizer	Measurement
[356]	Smooth, $\ell_2$ -norm Lipschitz	$O(\frac{\sqrt[4]{p \ln(\frac{1}{\delta})} \ln(\frac{n}{\delta})}{\sqrt{n\epsilon}})$	No	$\ell_2$ norm of gradient
[327]	Smooth, $\ell_2$ -norm Lipschitz	$O(\frac{\sqrt[4]{p \ln(\frac{1}{\delta})}}{\sqrt{n\epsilon}})$	No	$\ell_2$ norm of gradient
<b>Algorithm 4.1.16</b>	Smooth, $\ell_2$ -norm Lipschitz	$O(\frac{\sqrt[4]{p \ln(\frac{1}{\delta})}}{\sqrt{n\epsilon}})$	Yes	$\ell_2$ norm of projected gradient
<b>Algorithm 4.1.17</b>	Smooth, $\ell_2$ -norm Lipschitz, $\mathcal{C}$ bounded	$O(\frac{\sqrt[4]{(\ \mathcal{C}\ _2^2 + G_{\mathcal{C}}^2) \ln(\frac{1}{\delta})}}{\sqrt{n\epsilon}})$	No	Frank-Wolfe gap
<b>Algorithm 4.1.18</b>	Smooth, $\ell_2$ -norm Lipschitz, $\mathcal{C}$ bounded	$O(\frac{\sqrt[4]{(\ \mathcal{C}\ _2^2 + G_{\mathcal{C}}^2) \ln(\frac{1}{\delta})}}{\sqrt{n\epsilon}})$	Yes	$\ell_2$ norm of projected gradient
<b>Algorithm 4.1.19</b>	Smooth, $\ell_1$ -norm Lipschitz, $\mathcal{C}$ is $\ell_1$ norm ball (or polytope)	$O(\frac{\sqrt[4]{\ln(\frac{1}{\delta})} \sqrt{\ln(np)}}{\sqrt{n\epsilon}})$	No	Frank-Wolfe gap

Table 4.1: Comparisons with previous  $(\epsilon, \delta)$ -DP algorithms for DP-ERM with non-convex loss function. We assume that the Lipschitz and smooth parameters are 1, and  $\|\mathcal{C}\|_2 \leq 1$ .

$$\min_{x \in \mathcal{C}} F^r(x, D) = \frac{1}{n} \sum_{i=1}^n f(x, z_i) + r(x). \quad (4.1)$$

For convenience, we let  $F(x) = \frac{1}{n} \sum_{i=1}^n f(x, z_i)$  and  $F^r(x) = F^r(x, D)$ .

**Assumption 4.1.1.**  $F(x)$  is assumed to be differentiable and  $L$ -smooth over  $x$  with respect to  $\ell_2$  norm. Also, the loss function  $f(\cdot, z)$  is assumed to be  $G$ -Lipschitz over  $x$  with respect to  $\ell_2$ -norm for all  $z \in \mathcal{X}$ .

In order to measure the utility for (4.1), we define the **generalized projected gradient** as  $\mathcal{P}_{\mathcal{C}}(x, g, \gamma) = \frac{1}{\gamma}(x - x^+)$ , where

$$x^+ = \arg \min_{u \in \mathcal{C}} \{ \langle g, u \rangle + \frac{1}{2\gamma} \|x - u\|_2^2 + r(u) \}. \quad (4.2)$$

Note that this measurement has been widely used in the optimization community for studying the convergence and non-stationarity, such as [127, 128]. Actually, if  $\mathcal{C} = \mathbb{R}^p$  and  $r(x) \equiv 0$ , we have  $\mathcal{P}_{\mathcal{C}}(x, \nabla F(x), \gamma) = \nabla F(x) = \nabla F^r(x)$ .

Based on the Projected Gradient Descent, we have the following algorithm for DP-ERM with non-convex loss function and non-smooth convex regularizer.

---

**Algorithm 4.1.16** DP-PGD( $F, x_1, T, \sigma, \{\gamma_k\}_{k=1}^T$ )

---

**Input:**  $T$  is the maximum number of iterations,  $x_1$  is the initial point, and  $\{\gamma_k\}_{k=1}^T$  is the step size.  $\epsilon$  and  $\delta$  are privacy parameters.

- 1: **for**  $k = 1, \dots, T$  **do**
  - 2:     Compute  $x_{k+1} = \arg \min_{u \in \mathcal{C}} \{ \langle \nabla F(x_k) + \epsilon_k, u \rangle + \frac{1}{2\gamma_k} \|u - x_k\|_2^2 + r(u) \}$ , where  $\epsilon_k \sim N(0, \sigma^2 I_p)$ , here  $\sigma$  can be chosen by Theorem 1 or as the following:
  - 3:     Compute  $\rho$  which satisfies  $\rho + 2\sqrt{\rho \log(\frac{1}{\delta})} = \epsilon$ . Then set  $\sigma^2 = \frac{2L^2T}{n^2\rho^2}$ .
  - 4: **end for**
  - 5: **return**  $x_R \in \{x_1, \dots, x_T\}$  such that  $R$  is uniformly sampled from  $\{1, 2, \dots, T\}$ .
- 

**Theorem 4.1.1.** There exist constants  $c, c_1$ , such that for any  $0 < \epsilon < c_1 T, 0 < \delta < 1$ ,

**DP-PGD** (Algorithm 4.1.16) is  $(\epsilon, \delta)$ -differentially private if

$$\sigma^2 = c \frac{G^2 T \ln(\frac{1}{\delta})}{n^2 \epsilon^2}. \quad (4.3)$$

**Theorem 4.1.2.** Under **Assumption 4.1.1**, if we take  $\sigma^2$  as in (4.3),  $\{\gamma\}_{k=1}^T = \frac{1}{2L}$  and  $T = O(\frac{n\epsilon}{\sqrt{p \ln(\frac{1}{\delta})}})$  in Algorithm 4.1.16, the following inequality holds,

$$\mathbb{E} \|g_{\mathcal{C},R}\|_2 \leq O\left(\frac{\sqrt[4]{p \ln(\frac{1}{\delta})}}{\sqrt{n\epsilon}}\right), \quad (4.4)$$

where  $g_{\mathcal{C},R} = \frac{1}{\gamma_k}(x_R - x_{R+1})$ .

**Remark 4.1.1.** Note that if we remove the non-smoothness restriction on the regularizer and assume that  $\mathcal{C} = \mathbb{R}^p$ , the upper bound in Theorem 4.1.2 becomes the same as in [324]. Thus Theorem 4.1.2 can be viewed as a generalization of theirs.

Also it is worth noting that if we use the output in classical non-convex optimization algorithm directly, such as the one on Page 26 in [230], *i.e.*  $\|g_{\mathcal{C},R}\|_2 = \min_{1 \leq k \leq T} \|g_{\mathcal{C},k}\|_2$ , the algorithm will not be differentially private. Thus, here we use another randomizer on  $R$ . This is a main difference between our algorithm and those optimization algorithms.

It is notable that the variance of noise (4.3) in Theorem 4.1.1, which is based on Moment Accountant (Lemma 2.1.7), just states the existence of such constant  $c$  without specifying it.

Actually we can follow the way in [1] which is based on grid search for finding this hidden constant. However, this procedure is costly and complex, here we propose a more practical approach by transforming zero Concentrated Different Privacy (zCDP) [52] to  $(\epsilon, \delta)$ -DP, which corresponds to the step 3 in Algorithm 4.1.16<sup>1</sup>.

The idea is that we first make the algorithm to be  $\rho$ -zCDP and then transfer to  $(\epsilon, \delta)$ -DP, *i.e.* we first compute the number  $\rho$  which satisfies  $\rho + 2\sqrt{\rho \log(\frac{1}{\delta})} = \epsilon$ . Then we perform Algorithm 1 for  $T$  iterations. We can easily get in this case the variance satisfies  $\sigma_2^2 = \frac{2L^2T}{n^2(\sqrt{\log(1/\delta)} + \epsilon - \sqrt{\log(1/\delta)})^2}$ . When  $\frac{\epsilon}{\log(1/\delta)} \ll 1$  (this case will always holds since in practice we select  $\epsilon = 0.1 - 0.5$  and  $\delta = \frac{1}{n}$ ), by expanding Taylor series of  $\sqrt{1+x}$ , we have  $(\sqrt{\log(1/\delta)} + \epsilon - \sqrt{\log(1/\delta)})^2 \simeq \frac{\epsilon^2}{4\log(1/\delta)}$ , so  $\sigma_2^2 \simeq \frac{8L^2T \log(1/\delta)}{n^2 \epsilon^2}$ . We can see that compared with moment accountant method, our method is much more practical and simpler, compared with advanced composition theorem, it adds less noise in each iteration (see Experiment section for details).

## Extension to Population Risk

An important problem in machine learning is to use population risk to measure the performance of an estimator. It indicates how well the estimator performs on unseen examples from the same distribution. Based on the idea of measuring the utility of  $\theta^{\text{priv}}$  by the  $\ell_2$  norm of the gradient of the empirical risk, in this section, we show an upper bound of  $\theta^{\text{priv}}$  on the population risk based on the  $\ell_2$  norm of the gradient for non-convex loss functions, *i.e.*  $\mathbb{E}_{\mathcal{A}} \|\mathbb{E}_{z \sim \mathcal{P}} [\nabla f(x^{\text{priv}}, z)]\|_2$ , where  $\mathcal{A}$  is the randomized algorithm which outputs the private estimator  $x^{\text{priv}}$ .

Due to the hardness of the problem even in non-private settings, we need to make some assumptions. Below, we only consider the non-regularizer case.

**Assumption 4.1.2.** The gradient of the loss function is  $\tau$ -sub-Gaussian. That is, for any  $\lambda \in \mathbb{R}^p$  and  $x \in \mathbb{R}^p$ , we have  $\mathbb{E}\{\exp(\langle \lambda, \nabla f(x, z) - \mathbb{E}[\nabla f(x, z)] \rangle)\} \leq \exp(\frac{\tau^2 \|\lambda\|^2}{2})$ .

---

<sup>1</sup>Recently, [192] also proposed a similar way of reducing the noise in DP-GD based on zCDP. However, here we do not compare with it since there is no theoretical guarantee in their paper.

**Assumption 4.1.3.** The Hessian of the population risk is bounded. That is, there exists an  $H$  such that  $\|\nabla^2 \mathbb{E}_{z \sim \mathcal{P}}[f(x_0, z)]\|_2 \leq H$  for all  $x_0 \in \mathbb{R}^p$ . Also, the Hessian of the loss function is  $L$ -Lipschitz. That is, for every  $z$  and  $x_1, x_2 \in \mathbb{R}^p$ , we have  $\frac{\|\nabla^2 f(x_1, z) - \nabla^2 f(x_2, z)\|_2}{\|x_1 - x_2\|_2} \leq L$ , where the  $\ell_2$  norm of the Hessian is the operator norm. Furthermore, we assume that the constant  $H, L$  cannot be too large with respect to  $\tau$  and  $p$ . This means that there exists a constant  $c$  such that  $H \leq \tau^2 p^c$  and  $L \leq \tau^3 p^c$ .

Note that the first assumption is quite standard for analyzing the population risk [74]. The second assumption is very common in many non-convex loss functions, such as robust regression and binary classification. The examples can be found in [218]. Based on recent results on non-convex learning, we now have the following theorem.

**Theorem 4.1.3.** Under Assumption 4.1.1, 4.1.2 and 4.1.3, if  $n \geq \Omega(p \log(p))$ , then for any  $0 < \epsilon, \delta, \beta \leq 1$ , there is an  $(\epsilon, \delta)$ -differentially private algorithm  $\mathcal{A}$  which outputs  $x_R$  satisfying the following with probability at least  $1 - \beta$ ,

$$\begin{aligned} \mathbb{E}_{\mathcal{A}} \|\mathbb{E}_{z \sim \mathcal{P}}[\nabla f(x_R, z)]\|_2 &\leq O\left(\sqrt{\frac{\tau^2 p \log(\frac{\tau}{\beta}) \log n}{n}} + \frac{\sqrt[4]{p \log(\frac{1}{\delta})}}{\sqrt{n\epsilon}}\right) \\ &= O\left(\tau \sqrt{\frac{p \log(\frac{\tau}{\beta}) \log n \sqrt{\log \frac{1}{\delta}}}{n\epsilon}}\right). \end{aligned} \quad (4.5)$$

**Remark 4.1.2.** As we can see from above theorem, compared with the uniform convergence error, *i.e.* the first term in the right side of (4.5), the error due to differential privacy, *i.e.* the second term in the right side of (4.5), is less when we consider  $\epsilon, \delta$  as constants. Thus, the effect of differential privacy on the convergence error is just making the efficient sample complexity  $n$  become  $n\epsilon$ . This is quite different from the population risk in convex loss functions under differential privacy, where the error caused by privacy plays a much more important role, *i.e.* there is additional factor of  $\sqrt{p}$  in the population risk under differential privacy compared with non-private case. For details, please refer to Appendix F in [29]. An open problem is that whether this bound is tight, or whether we can deal with the high

dimensional case, we left these for future works.

### 4.1.2 High Dimension Case

#### Error Bounded by Frank-Wolfe Gap

The utility bound in (4.4) depends on the dimensionality  $p$ . In high dimensional (*i.e.*,  $p \gg n$ ) space, such a dependence may no longer be desirable. For convex loss functions, [269] showed that it is possible to make the utility bound (using the expected excess empirical risk as the measurement) depend only on the Gaussian Width of the constrained set  $\mathcal{C}$ , which could be considerably smaller than  $O(\sqrt{p})$  when  $\mathcal{C}$  is a bounded closed centrally symmetric convex set  $\mathcal{C} \subseteq \mathbb{R}^p$  (such as  $l_1$ -norm ball). Thus, a natural question is whether such an improvement can also be achievable for non-convex loss functions. Below we give an affirmative answer by showing that this is indeed possible for non-convex loss function (without considering the non-smoothness constraint on the regularizer, *i.e.*,  $r(x) \equiv 0$ ).

We start our discussion with some definitions and lemmas which will be used in this and next section.

**Definition 4.1.2** (Minkowski Norm). The Minkowski norm (denoted by  $\|\cdot\|_{\mathcal{C}}$ ) with respect to a centrally symmetric convex set  $\mathcal{C} \subseteq \mathbb{R}^p$  is defined as follows. For any vector  $v \in \mathbb{R}^p$ ,  $\|\cdot\|_{\mathcal{C}} = \min\{r \in \mathbb{R}^+ : v \in r\mathcal{C}\}$ . The dual norm of  $\|\cdot\|_{\mathcal{C}}$  is denoted as  $\|\cdot\|_{\mathcal{C}^*}$ ; for any vector  $v \in \mathbb{R}^p$ ,  $\|v\|_{\mathcal{C}^*} = \max_{w \in \mathcal{C}} |\langle w, v \rangle|$ .

**Definition 4.1.3** (Gaussian Width). Let  $b \sim \mathcal{N}(0, I_p)$  be a Gaussian random vector in  $\mathbb{R}^p$ . The Gaussian width for a set  $\mathcal{C}$  is defined as  $G_{\mathcal{C}} = \mathbb{E}_b[\sup_{w \in \mathcal{C}} \langle b, w \rangle]$ .

Compared with the dimensionality  $p$ , Gaussian Width of a convex set  $\mathcal{C} \subseteq \mathbb{R}^p$  could be much smaller. For example, when  $\mathcal{C}$  is  $l_1$ -norm unit ball,  $G_{\mathcal{C}} = O(\sqrt{\log p})$ ; when  $\mathcal{C}$  is the set of all unit  $s$ -sparse vectors on  $\mathbb{R}^p$ ,  $G_{\mathcal{C}} = O(\sqrt{s \log(p/s)})$ .

**Lemma 4.1.1.** [269] For  $W = (\max_{w \in \mathcal{C}} \langle w, v \rangle)^2$ , where  $v \sim \mathcal{N}(0, I_p)$ , we have  $\mathbb{E}_v[W] = O(G_{\mathcal{C}}^2 + \|\mathcal{C}\|_2^2)$ .

For simplicity, we let  $\|\cdot\|$  denote  $\|\cdot\|_{\mathcal{C}}$  and  $\|\cdot\|_*$  denote  $\|\cdot\|_{\mathcal{C}^*}$  in this section.

Our algorithm is based on the Frank-Wolfe method, where a differentially private version of Frank-Wolfe has been studied in [270] for LASSO. Frank-Wolfe method can be viewed as a greedy algorithm which moves towards the optimum solution in the first order approximation. It reduces the problem to solving a minimization problem of linear function, which exploits the geometric property of the constrained set  $\mathcal{C}$ . It also provides a new measurement of the non-stationarity, called Frank-Wolfe gap, for the utility, which has already been used in [187, 247]. Formally, the Frank-Wolfe gap at a point  $x$  of the function  $F$  is defined as:  $\mathcal{G}(x) = \max_{v \in \mathcal{C}} \langle v - x, -\nabla F(x) \rangle$ ,  $x \in \mathcal{C}$ . Since the gap  $\mathcal{G}(x) = 0$  if and only if  $x$  is a stationary point, it could provide of stationarity for a point. Our following algorithm uses the Frank-Wolfe gap as a measurement for DP-ERM with non-convex smooth loss functions.

---

**Algorithm 4.1.17** DP-FW-L2( $F, x_1, T, \sigma, \{\gamma_t\}_{t=1}^T$ )

---

**Input:**  $T$  is the maximum of iterations,  $x_1$  is the initial point, and  $\{\gamma_t\}_{t=1}^T$  is the step size.  
**for**  $k = 1, \dots, T$  **do**  
    Compute  $v_t = \arg \max_{v \in \mathcal{C}} \langle v, -(\nabla F(x) + \epsilon_t) \rangle$ , where  $\epsilon_k \sim N(0, \sigma^2 I_p)$ .  
     $x_{t+1} = x_t + \gamma_t(v_t - x_t)$ .  
**end for**  
**return**  $x_R \in \{x_1, \dots, x_T\}$  such that  $R$  is uniformly sampled from  $\{1, \dots, T\}$ .

---

**Theorem 4.1.4.** Let  $\mathcal{C}$  be a bounded, closed, centrally symmetric convex set. Then, there exist constants  $c, c_1$ , under **Assumption 4.1.1** and for any  $0 < \epsilon < c_1 T$ ,  $0 < \delta < 1$ , **DP-FW-L2** (Algorithm 4.1.17) is  $(\epsilon, \delta)$ -differentially private if  $\sigma^2$  is chosen as in (4.3). Moreover, if taking  $\{\gamma_t\}_{t=1}^T = O\left(\frac{\sqrt[4]{(\|\mathcal{C}\|_2^2 + G_{\mathcal{C}}^2) \ln \frac{1}{\delta}}}{\|\mathcal{C}\|_2 \sqrt{n\epsilon}}\right)$  and  $T = O\left(\frac{n\epsilon}{\sqrt{(\|\mathcal{C}\|_2^2 + G_{\mathcal{C}}^2) \ln \frac{1}{\delta}}}\right)$ , the following holds,

$$\mathbb{E}[\mathcal{G}_R] \leq O\left(\frac{\|\mathcal{C}\|_2 \sqrt[4]{(\|\mathcal{C}\|_2^2 + G_{\mathcal{C}}^2) \ln \frac{1}{\delta}}}{\sqrt{n\epsilon}}\right), \quad (4.6)$$

where  $\mathcal{G}_t = \max_{v \in \mathcal{C}} \langle -\nabla F(x_t), v - x_t \rangle$ .

### 4.1.3 Error Bounded by Norm of Gradient

So far we have presented two methods for the general non-convex case and the high dimension case, respectively. Theorem 4.1.4 enables us to bound the utility using Gaussian Width, but has some robustness issue with non-smooth regularizer. Contrarily, Theorem 4.1.2 can handle non-smooth regularizer, but its utility depends on the dimensionality of the space. Below we show in **Algorithm 4.1.18** that it is actually possible to combine the advantages of both methods by using Mirror Descent.

**Definition 4.1.4.** A function  $w : \mathcal{C} \rightarrow \mathbb{R}$  is said to be a distance generating function with modulus  $\alpha > 0$  (w.r.t.  $\|\cdot\|$  norm), if  $w$  is continuously differentiable and strongly convex satisfying the following inequality for any  $x, z \in \mathcal{C}$ ,  $\langle x - z, \nabla w(x) - \nabla w(z) \rangle \geq \alpha \|x - z\|^2$ . The Bregman Divergence associated with  $w$  is defined as  $V(x, z) = w(x) - w(z) - \langle \nabla w(z), x - z \rangle$ .

Similar to (4.2), we define the generalized projected gradient as  $\mathcal{P}_{\mathcal{C}}(x, g, \gamma) = \frac{1}{\gamma}(x - x^+)$ , where  $x^+ = \arg \min_{u \in \mathcal{C}} \{\langle g, u \rangle + \frac{1}{\gamma}V(u, x) + r(u)\}$ . Note that (4.2) is a special case in which  $w(x) = \frac{1}{2}\|x\|_2^2$ .

---

**Algorithm 4.1.18 DP-PMD( $F, x_1, T, \sigma, \{\gamma_k\}_{k=1}^T, w(\cdot)$ )**


---

**Input:**  $T$  is the maximum number of iterations,  $x_1$  is the initial point,  $w : \mathcal{C} \rightarrow \mathbb{R}$  is a distance generating function with modulus 1 (w.r.t.  $\|\cdot\|$  norm) and  $V(\cdot, \cdot)$  is its Bregman Divergence,  $\{\gamma_k\}_{k=1}^T$  is the step size.

- 1: **for**  $k = 1, \dots, T$  **do**
  - 2:     Compute  $x_{k+1} = \arg \min_{u \in \mathcal{C}} \{\langle \nabla F(x_k) + \epsilon_k, u \rangle + \frac{1}{\gamma_k}V(u, x_k) + r(u)\}$ , where  $\epsilon_k \sim N(0, \sigma^2 I_p)$ .
  - 3: **end for**
  - 4: **return**  $x_R \in \{x_1, \dots, x_T\}$  where  $R$  is uniformly sampled from  $\{1, \dots, T\}$ .
- 

**Theorem 4.1.5.** Let  $\mathcal{C}$  be a bounded closed centrally symmetric convex set. Then, under

**Assumption 4.1.1** and for any  $0 < \epsilon < c_2 T, \delta > 0$ , **DP-PMD** (Algorithm 4.1.18) is  $(\epsilon, \delta)$ -differentially private if  $\sigma^2$  is chosen as in (4.3). Moreover, if taking  $\{\gamma\}_{k=1}^T = \frac{1}{2L\|\mathcal{C}\|_2^2}$  and

$T = O\left(\frac{n\epsilon\|\mathcal{C}\|_2}{\sqrt{(\|\mathcal{C}\|_2^2 + G_{\mathcal{C}}^2)\ln(\frac{1}{\delta})}}\right)$ , the following holds

$$\mathbb{E}\|g_{\mathcal{C},R}\|_2 \leq O\left(\frac{\|\mathcal{C}\|_2^{\frac{3}{2}}\sqrt[4]{(\|\mathcal{C}\|_2^2 + G_{\mathcal{C}}^2)\ln(\frac{1}{\delta})}}{\sqrt{n\epsilon}}\right), \quad (4.7)$$

where  $g_{\mathcal{C},k} = \frac{1}{\gamma_k}(x_k - x_{k+1})$ .

**Remark 4.1.3.** If  $\|\mathcal{C}\|_2 \leq 1$ ,  $\mathcal{G}_{\mathcal{C}} = o(\sqrt{p})$ , from Theorems 4.1.5 and 4.1.2, we can see that the utility bound of (4.7) is always less than (4.4). One of the main reasons for us to have Theorem 4.1.5 is the fact that we can exploit the geometric structure of the problem (by Remark 2 and the Mirror Descent). Moreover, when we ignore the terms related to  $\mathcal{C}$ , the upper bounds in Theorem 4.1.5 and 4.1.4 actually achieve the same upper bound, although the utilities are measured quite differently.

#### 4.1.4 Further Reducing the Utility

Theorem 4.1.5 allows us to bound the utility quite well for the general non-convex case. However, as shown in [270, 181], the utility can be further reduced for some convex loss functions to a level depending only on  $\log(p)$ , rather than  $G_{\mathcal{C}}$  or  $p$ . This inspires us to ask whether there is any special case for non-convex loss functions to achieve the same. In this section, we give an affirmative answer to this by showing (in **Algorithm 4.1.19**) that there is indeed a case where the Frank-Wolf gap depends only on  $\log(p)$ . We consider problem (4.1) without the regularizer term.

**Assumption 4.1.4.**  $F(x)$  is assumed to be differentiable and  $L$ -smooth over  $x$  w.r.t  $\ell_2$ -norm, and  $f(\cdot, z)$  is assumed to be  $G$ -Lipschitz over  $x$  with respect to  $\ell_1$ -norm for all  $z \in \mathcal{X}$ .  $\mathcal{C} \subseteq \mathbb{R}^p$  is assumed to be a closed convex set. Furthermore,  $\mathcal{C}$  is assumed to be the convex hull of some finite set  $A$ , i.e.,  $\mathcal{C} = \text{Conv}(A)$  and bounded. (For example,  $\mathcal{C}$  could be a polytope.)

---

**Algorithm 4.1.19** DP-FW-L1( $F, x_1, T, \sigma, \{\gamma_t\}_{t=1}^T$ )

---

**Input:**  $T$  is the iteration number and  $x_1$  is the initial point.  $\{\gamma_t\}_{t=1}^T$  is the step size.  $\mathcal{C} \subseteq \mathbb{R}^p$  be the convex hull of a compact set  $A \subseteq \mathbb{R}^p$ .

- 1: **for**  $t = 1, \dots, T$  **do**
  - 2:   Use exponential mechanism  $\mathcal{M}(D, u, \mathcal{R})$ , where  $\mathcal{R} = A$ ,  $u(D, s) = -\langle s, \nabla F(x_t, D) \rangle$ , to ensure  $(\frac{\epsilon}{\sqrt{8T \ln(\frac{1}{\delta})}}, 0)$ -differentially private. Denote the output as  $\tilde{x}_t$ .
  - 3:   Compute  $x_{t+1} = (1 - \gamma_t)x_t + \gamma_t \tilde{x}_t$ .
  - 4: **end for**
  - 5: **return**  $x_R \in \{x_1, \dots, x_T\}$  where  $R$  is uniformly sampled from  $\{1, 2, \dots, T\}$ .
- 

**Theorem 4.1.6.** Assume  $A$  is a finite set. Then, for any  $\epsilon, \delta > 0$ , **DP-FW-L1** (Algorithm 4.1.19) ensures  $(\epsilon, \delta)$ -differentially private. Furthermore, if we set  $T = O(\frac{n\epsilon}{\sqrt{\ln(\frac{1}{\delta})} \ln(|A|n/\eta)})$  and  $\{\gamma_t\}_{t=1}^T = \sqrt{\frac{2}{T\|\mathcal{C}\|_2^2}}$ . Then with probability at least  $1 - \eta$ , the following holds

$$\mathbb{E}[\mathcal{G}_R] \leq O\left(\frac{\|\mathcal{C}\|_1 \sqrt[4]{\ln(\frac{1}{\delta})} \sqrt{\ln \frac{n|A|}{\eta}}}{\sqrt{n\epsilon}}\right), \quad (4.8)$$

where  $\mathcal{G}_t = \max_{v \in \mathcal{C}} \langle -\nabla F(x_t), v - x_t \rangle$ .

**Corollary 4.1.1.** If  $\mathcal{C}$  is an  $\ell_1$ -norm ball or a simplex in  $\mathbb{R}^p$ , then we can see that  $A$  is the set of the vertices of  $\mathcal{C}$ , in this case, the Frank-Wolf gap in (4.8) is  $\mathbb{E}\mathcal{G}_R = O(\frac{\sqrt[4]{\ln(\frac{1}{\delta})} \sqrt{\ln(np)}}{\sqrt{n\epsilon}})$ .

Note that since  $A$  in step 2 of Algorithm 4.1.19 is finite and  $u$  is a linear function, it could run in  $O(|A|p)$  time; also we can use Report-Noisy-Max in [104] instead of the exponential mechanism, see [205] for details. The above bound could be the smallest among all the results presented so far. For example, when  $\mathcal{C}$  contains the unit Euclidean ball,  $G_C = \Omega(\sqrt{p})$ . Thus, all the previous results depend on  $p$  while (4.8) depends only on  $\log(p)$ .

## 4.1.5 Experimental Results

In this section, we study the performance of differentially private gradient descent method with non-convex loss functions. Particularly, we consider the case where the sigmoid

function is the loss and  $\ell_1$ -norm is the regularizer, *i.e.*

$$\min_{\theta \in \mathbb{R}^p} F^r(\theta, D) = \frac{1}{n} \sum_{i=1}^n \frac{1}{1 + \exp(-y_i \langle \theta, x_i \rangle)} + \frac{\lambda}{2} \|\theta\|_1,$$

where  $\{x_i\}_{i=1}^n$  are the feature vectors and  $\{y_i\}_{i=1}^n$  are the corresponding labels.

## Experiment Settings

Due to the hardness of computing the Frank-Wolfe gap, we test Algorithm 4.1.19 and measure the  $\ell_2$ -norm of the generalized projected gradient. We set  $\lambda = 10^{-4}$ , and evaluate our algorithms on both synthetic and real world datasets. The synthetic dataset is generated in the same way as in logistic regression. The size of the synthetic dataset is  $(5 \times 10^4, 50)$ . For the real-world datasets, we use the same datasets as in the convex case.

For the differential privacy parameters, we choose  $\epsilon$  from  $\{0.1, 0.5, 2, 5\}$ , and a fixed  $\delta = 10^{-4}$ . For the optimization algorithms, the initial vector is selected randomly. Also, since the step size does not affect differential privacy, we use the the same way as in <http://cvxr.com/tfocs/> to choose the step size, where the initial step size is 0.1. All the experiments are performed on MATLAB.

## Experiments Results

Figure 4.1, 4.2 and 4.3 are the results on synthetic, Covertype, and IJCNN datasets, respectively, with varying parameters. Figure 4.4 shows the results of different differentially private composition methods on different datasets.

Results in (a) of Figures 4.1, 4.2 and 4.3 show the effect of the iteration number  $T$  on the norm of gradient. From these figures, we can see that although different  $T$  values can cause the magnitude of the added noise change in each iteration, it has less effect on the norm of projected gradient than  $\epsilon$ . This is due to the fact that all the upper bounds in our theoretical analysis are independent of  $T$ , which makes the norm of projected gradient stable after some

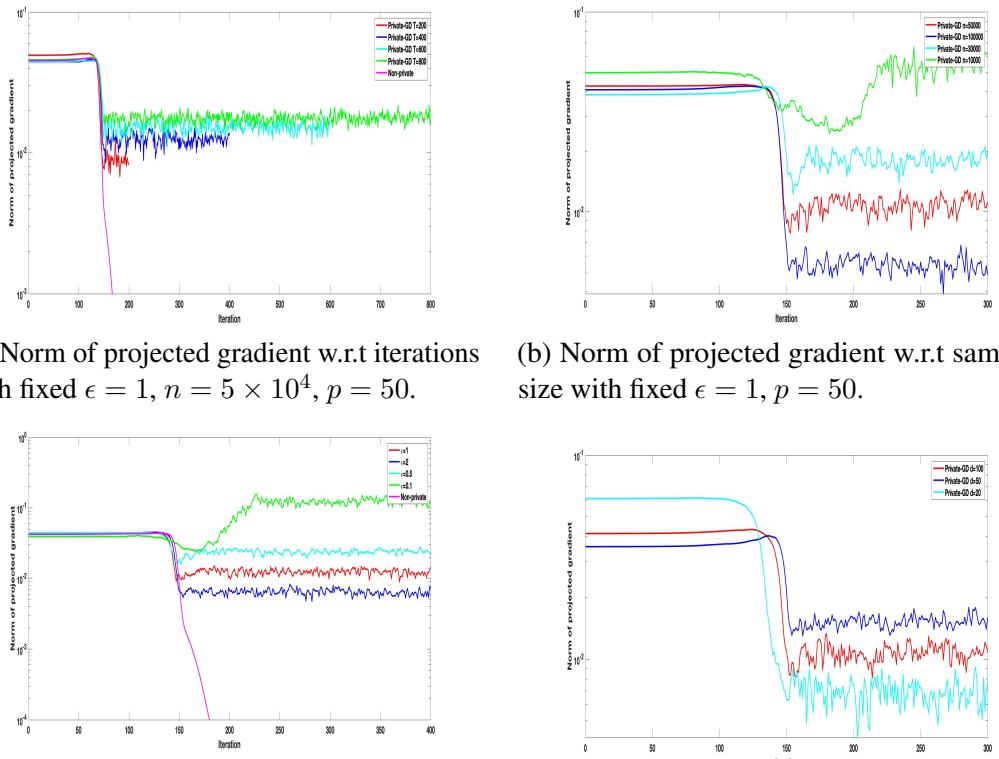


Figure 4.1: Experimental results on synthetic datasets for nonconvex case.

iterations.

Results in (b) of Figures 4.1, 4.2 and 4.3 depict the effect of sample size  $n$ . From these figures, we can observe that the norm of gradient is reverse proportional to the sample size, which is consistent with our theoretical analysis.

The effect of the privacy parameter  $\epsilon$  is plotted in (c) of Figures 4.1, 4.2 and 4.3. These figures show that a larger  $\epsilon$ , which means less privacy, leads to a smaller error, i.e. the norm of projected gradient is smaller. This is consistent with our theoretical analysis.

The effect of dimensionality is depicted in (d) of Figures 4.1, 4.2 and 4.3. These figures indicate that there is a positive correlation between the dimensionality and the norm of gradient, i.e. the higher the dimensionality, the larger the norm of gradient.

Figure 4.4 shows the comparisons of our method with some existing ones. The figure suggests that on all the datasets our method has less error compared to the advanced

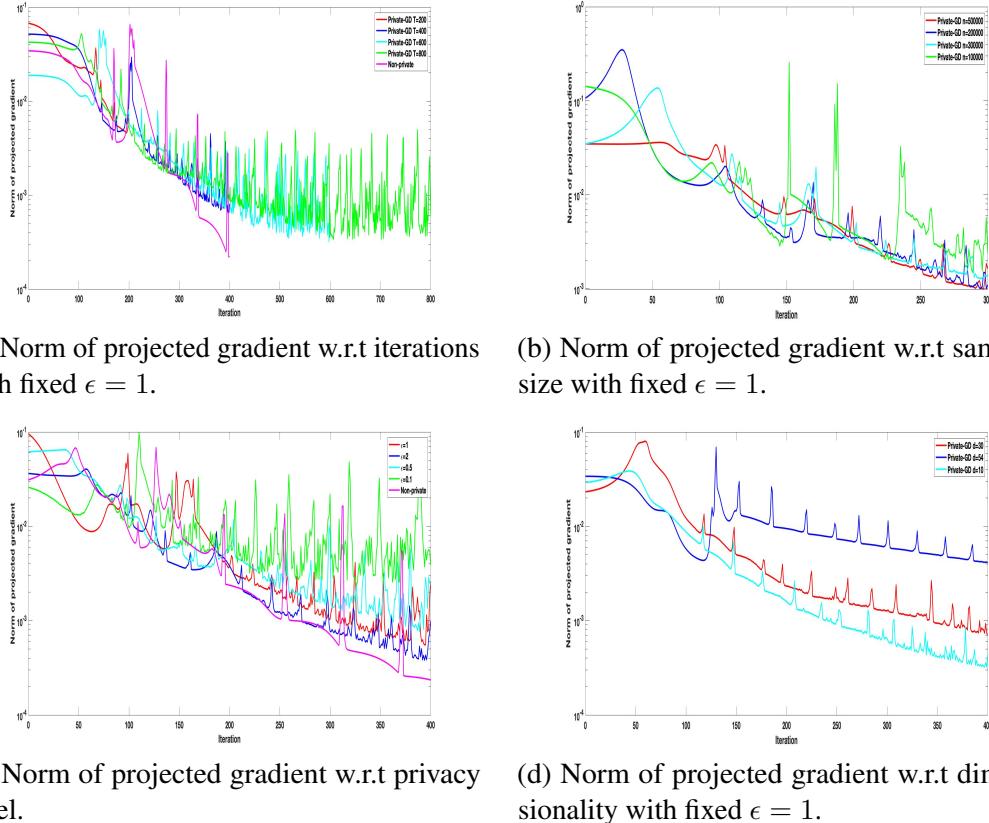


Figure 4.2: Experimental results on Covertype dataset for nonconvex case.

composition theorem and moment accountant methods. This confirms our previous analysis.

#### 4.1.6 Omitted Proofs

##### Proof of Theorem 4.1.1

In order to proof Theorem 4.1.1, we have the following lemma.

**Lemma 4.1.2.** Let  $x^+ = \arg \min_{u \in \mathcal{C}} \{\langle \nabla F(x) + \epsilon, u \rangle + \frac{1}{2\gamma} \|u - x\|_2^2 + r(u)\}$ . Then the following is true.

$$\langle \nabla F(x), x - x^+ \rangle \geq \frac{1}{\gamma} \|x^+ - x\|_2^2 + r(x^+) - r(x) + \langle \epsilon, x^+ - x \rangle.$$

Since Lemma 4.1.2 is a special case of Lemma 4.1.5 (*i.e.*,  $w = \frac{1}{2} \|x\|_2^2$ ), we will only prove Lemma 4.1.5.

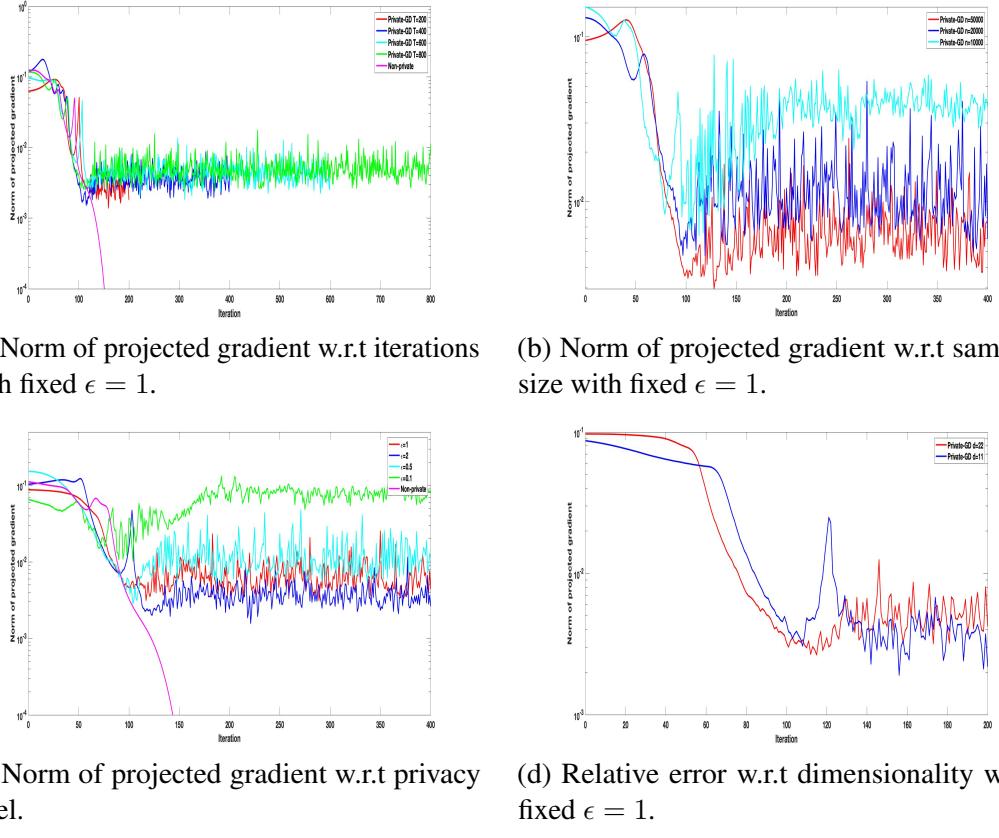


Figure 4.3: Experimental results on IJCNN dataset for nonconvex case.

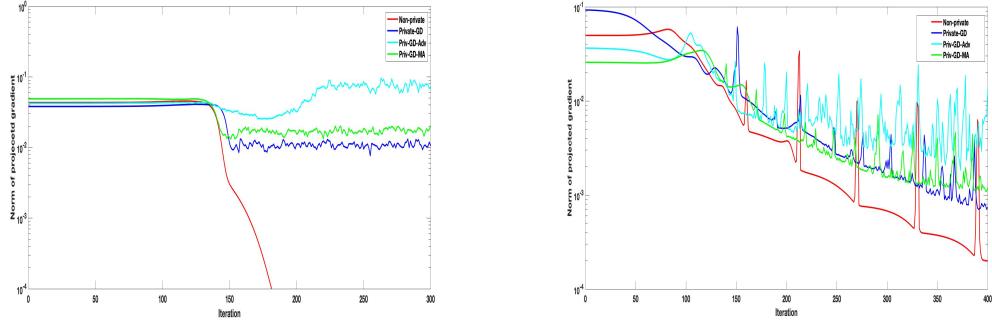
By the L-smooth property of  $F(x)$  and  $g_{C,k} = \frac{1}{\gamma_k}(x_k - x_{k+1})$ , we have

$$F(x_{k+1}) \leq F(x_k) - \gamma_k \langle \nabla F(x_k), g_{C,k} \rangle + \frac{L\gamma_k^2}{2} \|g_{C,k}\|_2^2.$$

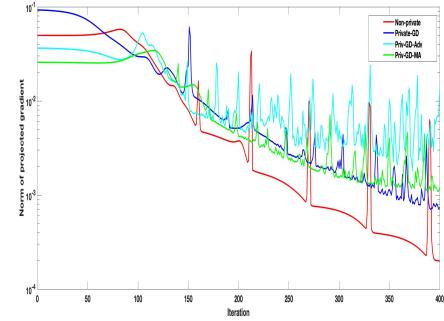
In Lemma 4.1.2, taking  $x^+ = x_{k+1}$ ,  $x = x_k$ ,  $\gamma = \gamma_k$  and  $x - x^+ = \gamma_k g_{C,k}$ , we have

$$\begin{aligned} F^r(x_{k+1}) &\leq F^r(x_k) - (\gamma_k - \frac{L}{2}\gamma_k^2) \|g_{C,k}\|_2^2 - \gamma_k \langle \epsilon_k, g_{C,k} \rangle \\ &\leq F^r(x_k) - (\gamma_k - \frac{L}{2}\gamma_k^2 - \frac{L}{2}\gamma_k^2) \|g_{C,k}\|_2^2 + \frac{\|\epsilon_k\|_2^2}{2L}, \end{aligned}$$

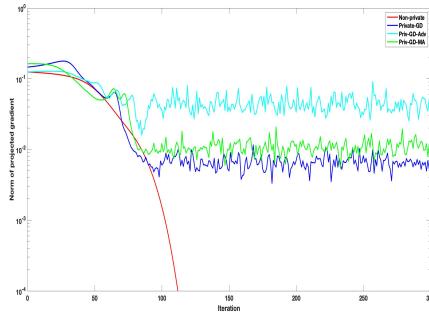
where the last inequality comes from Cauchy Inequality. Summing this over  $k = 1 \dots T$



(a) Norm of projected gradient w.r.t different methods on synthetic dataset with fixed  $\epsilon = 1$ ,  $n = 5 \times 10^4$  and  $p = 50$ .



(b) Norm of projected gradient w.r.t different methods on Covertype dataset with fixed  $\epsilon = 1$ .



(c) Norm of projected gradient w.r.t different methods on IJCNN dataset with fixed  $\epsilon = 1$ .

Figure 4.4: Experimental results on the norm of projected gradient w.r.t different methods. The left one is for synthetic dataset, the middle one is for Covertype dataset, and the right one is for IJCNN dataset.

and taking the expectation with  $\{\epsilon\}_{k=1}^T$ , we have

$$\sum_{k=1}^T (\gamma_k - L\gamma_k^2) \mathbb{E}\|g_{C,k}\|_2^2 \leq F^r(x_1) - F^r(x^*) + \frac{Tp\sigma^2}{2L}.$$

By the definition of  $g_{C,R}$ , we have

$$\mathbb{E}\|g_{C,R}\|_2^2 = \frac{1}{T} \sum_{k=1}^T \mathbb{E}\|g_{C,k}\|_2^2.$$

Taking  $\{\gamma_k\}_{k=1}^T = \frac{1}{2L}$  and  $\sigma^2$  as in Theorem 8, we obtain

$$\mathbb{E}\|g_{C,R}\|_2^2 \leq \frac{4L(F^r(x_1) - F^r(x^*))}{T} + O\left(\frac{pG^2T \ln(\frac{1}{\delta})}{n^2\epsilon^2}\right).$$

Setting  $T = O\left(\frac{n\epsilon\sqrt{L}}{G\sqrt{p \ln(\frac{1}{\delta})}}\right)$ , and since  $\mathbb{E}Z \leq \sqrt{\mathbb{E}Z^2}$ , we get the result.

### Proof of Theorem 4.1.3

The proof is based on the following theorem in [218]:

**Lemma 4.1.3** (Theorem 1 in [218]). Under Assumption 2 and 3, there exists a universal constant  $C_0$ , such that if letting  $C = C_0 \max\{c_h, \log(r\tau/\delta), 1\}$ , the following holds: The sample gradient converges uniformly to the population gradient in Euclidean norm, namely, if  $n \geq Cp \log p$ , we have with probability at least  $1 - \delta$ ,

$$\sup_{\theta \in \mathbb{B}^p(r)} \|\nabla F_n(\theta) - \nabla \hat{F}(\theta)\|_2 \leq \tau \sqrt{\frac{Cp \log n}{n}},$$

where  $F_n(\theta) = \frac{1}{n} \sum_{i=1}^n f(\theta, x_i)$ ,  $\hat{F}(\theta) = \mathbb{E}_{x \sim \mathcal{P}}[f(\theta, x)]$ .

Actually, we can extend the theorem from the restriction  $\theta \in \mathbb{B}^p(r)$  to any ball with radius  $r$ , that is  $\theta \in \mathbb{B}^p(x_0, r)$ . From Theorem 4.1.1, we have  $\mathbb{E}_{\mathcal{A}} \|\nabla F(x_R, D)\|_2 \leq O\left(\frac{\sqrt{p \ln(\frac{1}{\delta})}}{\sqrt{n\epsilon}}\right)$ . By Lemma 4.1.3, we know that for each  $x_R$ , we have  $\|\nabla F(x_R, D) - \nabla \hat{F}(x_R)\|_2 \leq \tau \sqrt{\frac{Cp \log n}{n}}$ . Thus,  $\mathbb{E}_{\mathcal{A}} \|\nabla F(x_R, D) - \nabla \hat{F}(x_R)\|_2 \leq \tau \sqrt{\frac{Cp \log n}{n}}$ . Consequently, we have  $\mathbb{E}_{\mathcal{A}} \|\nabla \mathbb{E}_{x \sim \mathcal{P}}[f(x_R, x)]\|_2 \leq O\left(\frac{\sqrt{p \ln(\frac{1}{\delta})}}{\sqrt{n\epsilon}} + \tau \sqrt{\frac{Cp \log n}{n}}\right)$ .

### Proof of Theorem 4.1.4

We first need the following lemma:

**Lemma 4.1.4.** For any vector  $v$ , we have  $\|v\|_2 \leq \|\mathcal{C}\|_2 \|v\|_{\mathcal{C}}$ , where  $\|\mathcal{C}\|_2$  is the  $\ell_2$ -diameter and  $\|\mathcal{C}\|_2 = \sup_{x,y \in \mathcal{C}} \|x - y\|_2$ .

Lemma 4.1.4 implies that any smooth convex function  $F(\theta)$ , which is L-smooth with respect to  $\ell_2$  norm, is  $L\|\mathcal{C}\|_2^2$ -smooth with respect to  $\|\cdot\|_{\mathcal{C}}$  norm, which is the motivation of our algorithm.

*Proof.* If  $v = 0$ , this is trivially true. Otherwise, we will show that  $\frac{\|v\|_2}{\|\mathcal{C}\|_2} \leq \|v\|_{\mathcal{C}}$ . This is equivalent to show that  $v \notin \frac{\|v\|_2}{\|\mathcal{C}\|_2} \mathcal{C}$ . Taking any  $y \in \mathcal{C}$ , since  $\|\frac{\|v\|_2}{\|\mathcal{C}\|_2} y\|_2 = \frac{\|v\|_2}{\|\mathcal{C}\|_2} \|y\|_2$ , we know that  $\|y\|_2 < \|\mathcal{C}\|_2$ . Thus,  $\|\frac{\|v\|_2}{\|\mathcal{C}\|_2} y\|_2 < \|v\|_2$ . We get  $v \notin \frac{\|v\|_2}{\|\mathcal{C}\|_2} \mathcal{C}$ .  $\square$

Let  $\tilde{L}$  denote  $L\|\mathcal{C}\|_2^2$ , and  $D$  denote the diameter of  $\mathcal{C}$  w.r.t.  $\|\cdot\|$  norm. By the  $L$ -smooth property and Lemma 4.1.4, we have

$$F(x_{t+1}) \leq F(x_t) + \gamma_t \langle \nabla F(x_t), v_t - x_t \rangle + \frac{\tilde{L}\gamma_t^2}{2} \|v_t - x_t\|^2. \quad (4.9)$$

Let  $\hat{v}_t = \arg \max_{v \in \mathcal{C}} \langle v, -\nabla F(x_t) \rangle$ . By the optimality of  $v_t$ , we have

$$\langle v_t, -\nabla F(x_t) - \epsilon_t \rangle \geq \langle \hat{v}_t, -\nabla F(x_t) - \epsilon_t \rangle,$$

This implies

$$\langle v_t - \hat{v}_t, \nabla F(x_t) \rangle \leq \langle v_t - \hat{v}_t, -\epsilon_t \rangle. \quad (4.10)$$

From (4.9), we get

$$F(x_{t+1}) \leq F(x_t) + \gamma_t \langle \nabla F(x_t), v_t - \hat{v}_t \rangle + \gamma_t \langle \nabla F(x_t), \hat{v}_t - x_t \rangle + \frac{\gamma_t^2 \tilde{L}}{2} \|v_t - x_t\|^2.$$

Plugging (4.10) into (4.9) and by the fact that  $\langle \nabla F(x_t), \hat{v}_t - x_t \rangle = -\mathcal{G}_t$  (from the definition of  $\hat{v}_t$ ), we obtain

$$\begin{aligned} \gamma_t \mathcal{G}_t &\leq F(x_t) - F(x_{t+1}) + \gamma_t \langle v_t - \hat{v}_t, -\epsilon_t \rangle + \frac{\tilde{L}\gamma_t^2}{2} D^2 \\ &\leq F(x_t) - F(x_{t+1}) + \frac{\gamma_t^2 \tilde{L} \|v_t - \hat{v}_t\|^2}{2} + \frac{\|\epsilon_t\|_*^2}{2\tilde{L}} + \frac{\tilde{L}\gamma_t^2}{2} D^2 \\ &\leq F(x_t) - F(x_{t+1}) + \frac{\|\epsilon_t\|_*^2}{2\tilde{L}} + \tilde{L}\gamma_t^2 D^2, \end{aligned}$$

where the second inequality is due to Cauchy Inequality. By the definition of  $\mathcal{G}_R$ , we have

$\mathbb{E}[\mathcal{G}_R] = \frac{1}{T} \sum_{t=1}^T \mathbb{E}[\mathcal{G}_t]$ . Since  $\{\gamma_t\}_{t=1}^T = \gamma$ , summing the above over  $t = 1 \dots, T$  and

taking the expectation, also from Lemma 4.1.1, we have

$$\mathbb{E}G_R \leq \frac{F(x_1) - F(x^*)}{\gamma T} + \tilde{L}\gamma D^2 + \frac{1}{\gamma}O((\|\mathcal{C}\|_2^2 + G_{\mathcal{C}}^2)\frac{G^2 T \ln(\frac{1}{\delta})}{n^2 \epsilon^2}).$$

Taking  $\gamma = O(\frac{\sqrt[4]{G^2(\|\mathcal{C}\|_2^2 + G_{\mathcal{C}}^2)\ln\frac{1}{\delta}}}{\sqrt{\tilde{L}D\sqrt{n\epsilon}}})$  and  $T = O(\frac{n\epsilon}{\sqrt{(\|\mathcal{C}\|_2^2 + G_{\mathcal{C}}^2)G^2\ln\frac{1}{\delta}}})$ , by definition of  $\|\cdot\|$ , and  $D \leq O(1)$ , we obtain the result.

### Proof of Theorem 4.1.5

We first proof the following lemma:

**Lemma 4.1.5.** Let  $w$  be a distance generating function with modulus  $\alpha$  w.r.t.  $\|\cdot\|$  norm, and  $x^+ = \arg \min_{u \in \mathcal{C}} \{\langle \nabla F(x) + \epsilon, u \rangle + \frac{1}{\gamma}V(u, x) + h(u)\}$ . Then the following is true

$$\langle \nabla F(x), x - x^+ \rangle \geq \frac{\alpha}{\gamma} \|x^+ - x\|^2 + r(x^+) - r(x) + \langle \epsilon, x^+ - x \rangle.$$

*Proof.* By the optimality of  $x^+$ , we know that there exits a  $p \in \partial r(x^+)$  such that

$$\langle \nabla F(x) + \epsilon + \frac{1}{\gamma}[\nabla w(x^+) - \nabla w(x)] + p, u - x^+ \rangle \geq 0, \forall x \in \mathcal{C}. \quad (4.11)$$

Letting  $u = x$  in above inequality, we have

$$\langle \nabla F(x), x - x^+ \rangle \geq \frac{1}{\gamma} \langle \nabla w(x^+) - \nabla w(x), x^+ - x \rangle + \langle p + \epsilon, x^+ - x \rangle.$$

By the strongly convexity of  $w$  and  $\langle p, x^+ - x \rangle \geq r(x^+) - r(x)$ , we get the proof.  $\square$

Since  $F(\theta)$  is  $L$ -smooth w.r.t  $\ell_2$  norm, we know that it is  $L\|\mathcal{C}\|_2^2$ -smooth w.r.t  $\|\cdot\|$  norm. Let  $\tilde{L} = L\|\mathcal{C}\|_2^2$ . We have

$$F(x_{k+1}) \leq F(x_k) - \gamma_k \langle \nabla F(x_k), g_{\mathcal{C}, k} \rangle + \frac{\tilde{L}\gamma_k^2}{2} \|g_{\mathcal{C}, k}\|^2.$$

In Lemma 4.1.5, taking  $x^+ = x_{k+1}, x = x_k, \gamma = \gamma_k, x - x^+ = \gamma_k g_{\mathcal{C},k}$ , we have

$$\begin{aligned} F^r(x_{k+1}) &\leq F^r(x_k) - (\gamma_k - \frac{\tilde{L}}{2}\gamma_k^2)\|g_{\mathcal{C},k}\|^2 - \gamma_k\langle\epsilon_k, g_{\mathcal{C},k}\rangle \\ &\leq F^r(x_k) - (\gamma_k - \frac{\tilde{L}}{2}\gamma_k^2 - \frac{\tilde{L}}{2}\gamma_k^2)\|g_{\mathcal{C},k}\|^2 + \frac{\|\epsilon_k\|_*^2}{2\tilde{L}}, \end{aligned}$$

where the last inequality follows from Cauchy Inequality. Summing this over  $k = 1 \dots T$  and taking the expectation with  $\{\epsilon\}_{k=1}^T$  and by Lemma 4.1.1, we have

$$\sum_{k=1}^T (\gamma_k - \tilde{L}\gamma_k^2)\mathbb{E}\|g_{\mathcal{C},k}\|^2 \leq F^r(x_1) - F^r(x^*) + \frac{TO(\|\mathcal{C}\|_2^2 + G_{\mathcal{C}}^2)\sigma^2}{2\tilde{L}}.$$

By the definition of  $R$ , we have  $\mathbb{E}\|g_{\mathcal{C},R}\|^2 = \frac{1}{T} \sum_{k=1}^T \mathbb{E}\|g_{\mathcal{C},k}\|^2$ . Taking  $\{\gamma_k\}_{k=1}^T = \frac{1}{2\tilde{L}}$  and  $\sigma^2$ , we get

$$\mathbb{E}\|g_{\mathcal{C},R}\|^2 \leq \frac{4\tilde{L}(F^r(x_1) - F^r(x^*))}{T} + O\left(\frac{(\|\mathcal{C}\|_2^2 + G_{\mathcal{C}}^2)G^2T\ln(\frac{1}{\delta})}{n^2\epsilon^2}\right).$$

Setting  $T = O(\frac{n\epsilon\sqrt{\tilde{L}}}{G\sqrt{(\|\mathcal{C}\|_2^2 + G_{\mathcal{C}}^2)\ln(\frac{1}{\delta})}})$ , we have  $\mathbb{E}\|g_{\mathcal{C},R}\|^2 \leq O(\frac{G\sqrt{\tilde{L}}\sqrt{(\|\mathcal{C}\|_2^2 + G_{\mathcal{C}}^2)\ln(\frac{1}{\delta})}}{n\epsilon})$ . Also by Lemma 4.1.4, we have the result.

### Proof of Theorem 4.1.6

By exponential mechanism and advanced composition theorem, we can see that it is  $(\epsilon, \delta)$ -differentially private. By the  $G$ -Lipschitz (w.r.t  $\ell_1$ -norm) property of the loss function, we know that  $\Delta u \leq O(\frac{\|\mathcal{C}\|_1 G}{n\epsilon})$ . Let  $\beta = O(\frac{G\|\mathcal{C}\|_1\sqrt{8T\ln(\frac{1}{\delta})}\ln(\frac{|A|T}{\eta})}{n\epsilon})$ . By the utility bound of exponential mechanism, we know that in each iteration, with probability  $1 - \frac{\eta}{T}$ , the following holds

$$\langle\tilde{\theta}_t, \nabla F(\theta_t)\rangle \leq \min_{v \in A} \langle v, \nabla F(\theta_t)\rangle + \beta. \quad (4.12)$$

Let  $s_t = \arg \min_{u \in \mathcal{A}} \langle u, \nabla F(x_t) \rangle$ . By the  $L$ -smooth property and (4.12), we have

$$\begin{aligned} \frac{L}{2} \|x_{t+1} - x_t\|_2^2 &\geq F(x_{t+1}) - F(x_t) - \langle F(x_t), x_{t+1} - x_t \rangle \\ &= F(x_{t+1}) - F(x_t) - \gamma_t \langle \nabla F(x_t), \tilde{x}_t - x_t \rangle \\ &\geq F(x_{t+1}) - F(x_t) - \gamma_t (\langle \nabla F(x_t), s_t - x_t \rangle + \beta). \end{aligned}$$

Note that  $\min_{u \in \mathcal{C}} \langle u - x_t, \nabla F(x_t) \rangle = \min_{u \in \mathcal{A}} \langle u - x_t, \nabla F(x_t) \rangle = \langle s_t - x_t, \nabla F(x_t) \rangle = -\mathcal{G}_t$ .

Thus, we have

$$F(x_{t+1}) - F(x_t) + \gamma_t \mathcal{G}_t \leq \gamma_t \beta + \frac{L \gamma_t^2}{2} \|\mathcal{C}\|_2^2. \quad (4.13)$$

Summing over  $t = 1, \dots, T$ , we get with probability  $1 - \eta$ ,

$$\left( \sum_{t=1}^T \gamma_t \right) \mathcal{G}_R \leq F(x_1) - F(x^*) + \left( \sum_{t=1}^T \gamma_t \right) \beta + \left( \sum_{t=1}^T \gamma_t^2 \right) \|\mathcal{C}\|_2^2.$$

Taking  $\{\gamma_t\}_{t=1}^T = \gamma$ , we have

$$\mathcal{G}_R \leq \frac{F(x_1) - F(x^*)}{\gamma T} + \frac{\gamma \|\mathcal{C}\|_2^2 L}{2} + O\left(\frac{G \|\mathcal{C}\|_1 \sqrt{T \ln(\frac{1}{\delta})} \ln(\frac{|A|T}{\eta})}{n\epsilon}\right).$$

Taking  $T = O\left(\frac{n\epsilon \|\mathcal{C}\|_2}{\|\mathcal{C}\|_1 G \sqrt{\ln(\frac{1}{\delta}) \ln(|A|n)}}\right)$  and  $\gamma = \sqrt{\frac{2}{T \|\mathcal{C}\|_2^2 L}}$ , by the relation  $\|\mathcal{C}\|_2 \leq \|\mathcal{C}\|_1$  we get the result.

## 4.2 Global Minimum View

In the previous section, we study the approach of using first order stationary measurement to measure the error of private estimation. Despite some obvious advantages with such an approach, it also endows a few limitations: 1) although [356, 328, 309] showed that the gradient norm tends to 0 as  $n$  goes to infinity, there is no guarantee that such an estimator will be close to any non-degenerate local minimum [5]; 2) the gradient-norm estimator is

not always consistent with the excess empirical (population) risk of the loss function, *i.e.*,  $\hat{L}(w^{\text{priv}}) - \hat{L}(w^*)$ , where  $w^*$  is the optimal solution [29, 67]. Thus, it is difficult to compare the obtained solution with either the global or local minima. This propels us to the following interesting question.

**Can the excess empirical (population) risk be used to measure the error of non-convex loss functions under differential privacy?**

In the section, we consider the  $\ell_2$ -norm regularized DP-ERM with non-convex loss functions and propose an  $(\epsilon, \delta)$ -DP algorithm, named DP-GLD (Algorithm 4.2.20), and prove that its excess empirical (or population) risk is upper bounded by  $\tilde{O}\left(\frac{d \log(1/\delta)}{\log n \epsilon^2}\right)$  when  $\log n \geq O(d)$ , where  $n$  is the data size and  $d$  is the dimensionality of the space. Our technique is based on some recent developments in Bayesian learning and (stochastic) Gradient Langevin Dynamics [243, 70, 348, 284]. Interestingly, we show that the  $\frac{1}{\log n}$  term in the empirical risk bound can be further improved to  $\frac{1}{n^{\Omega(1)}}$  by a highly non-trivial analysis on the time-average error of a dynamic system.

Next, we consider upper bounding the excess population risk. Instead of determining the optimal bound, we show how to improve the bounds for some specific problems. Particularly, we focus on the generalized linear model with non-convex loss functions and the robust regressions problem with additional assumptions, and present an  $(\epsilon, \delta)$ -DP algorithm for them with population risk  $O\left(\frac{\sqrt[4]{d}}{\sqrt{n}\epsilon}\right)$ .

**Related Work:** Previous works on the DP version of SGLD have focused on Bayesian learning, such as [335, 195], which differ from our work considerably. Firstly, our work mainly focuses on achieving  $(\epsilon, \delta)$ -DP for ERM with non-convex loss functions, and on measuring the error of a private estimator with respect to the global or local minima. Secondly, existing works assume that the temperature-parameter  $\beta$  in the gradient Langevin dynamics is one or some constant, while  $\beta$  in our problem is not even a constant, making the analysis significantly more challenging in our work than in previous ones (see Remark

4.2.2 for more details).

### 4.2.1 Preliminaries

**Problem Setting** Given a dataset  $D = \{z_1 = (x_1, y_1), z_2 = (x_2, y_2) \dots, z_n = (x_n, y_n)\}$  from a data universe  $\mathcal{Z}$  and a closed convex set  $\mathcal{C} \subseteq \mathbb{R}^d$ , where  $\{x_i\}_{i=1}^n$  are feature vectors and  $\{y_i\}_{i=1}^n$  are labels or responses. DP-ERM is to find  $w^{\text{priv}} \in \mathcal{C}$  by minimizing the empirical risk defined as  $\hat{L}^r(w, D) \triangleq \hat{L}(w, D) + r(w) \triangleq \frac{1}{n} \sum_{i=1}^n \ell(w, z_i) + r(w)$ , with the guarantee of being differentially private (defined below). Here  $\ell$  is the loss function; and  $r(\cdot)$  is some simple (non)-smooth convex regularizer. The utility of an algorithm is measured by the **expected excess empirical risk** (which we call *empirical risk*), *i.e.*,

$$\text{Err}_D^r(w^{\text{priv}}) = \mathbb{E}[\hat{L}^r(w^{\text{priv}}, D)] - \min_{w \in \mathcal{C}} \hat{L}^r(w, D),$$

where the expectation is taking over the randomness of the algorithm. When the data are drawn i.i.d. from an unknown underlying distribution  $\mathcal{P}$  on  $\mathcal{Z}$ , we also seek to minimize the population risk, defined as  $L_{\mathcal{P}}^r(w) = \mathbb{E}_{z \in \mathcal{P}}[\ell(w, z)] + r(w)$ . The **expected excess population risk** (which we call *population risk*) becomes<sup>2</sup>

$$\text{Err}_{\mathcal{P}}^r(w^{\text{priv}}) = \mathbb{E}[L_{\mathcal{P}}^r(w^{\text{priv}})] - \min_{w \in \mathcal{C}} L_{\mathcal{P}}^r(w).$$

#### Markov semigroups and Infinitesimal Generator

In order to be self-contained, in this section we introduce the background and some preliminaries of Markov diffusion process. We refer the reader to [243, 18, 70] for more details.

**Definition 4.2.1.** For two Borel measures  $\mu, \nu$  on  $\mathbb{R}^d$  with finite second moments, the 2-Wasserstein distance,  $\mathcal{W}_2(\mu, \nu)$ , is defined as:  $\mathcal{W}_2(\mu, \nu) := \inf\{(\mathbb{E}\|V - W\|_2^2)^{\frac{1}{2}} : \mu =$

---

<sup>2</sup>If there is no regularizer, we will simply denote as  $\text{Err}_{\mathcal{P}}$ .

$\mathcal{L}(V), \nu = \mathcal{L}(W)\}$ , where the infinitum is taken over all the random couples  $(V, W)$  whose values are taken in  $\mathbb{R}^d \times \mathbb{R}^d$  with marginals  $V \sim \mu$  and  $W \sim \nu$ .  $\mathcal{L}(V)$  means the probability law of the random vector  $V$ .

Let  $\{W_t\}_{t \geq 0}$  be a continuous-time homogeneous Markov process with values in  $\mathbb{R}^d$ , and  $P = \{P_t\}_{t \geq 0}$  be the corresponding Markov semigroup. That is

$$P_s g(W_t) = \mathbb{E}[g(W_{s+t})|W_t]$$

for all  $s, t \geq 0$  and all bounded measurable functions  $g : \mathbb{R}^d \mapsto \mathbb{R}$ . A Borel probability measure  $\pi$  is called stationary or invariant if  $\int_{\mathbb{R}^d} P_t g d\pi = \int_{\mathbb{R}^d} g d\pi$  for all  $g$  and  $t$ . Each of  $P_t$  can be extended to a bounded linear operator on  $L^2(\pi)$ , such that  $P_t g \geq 0$  whenever  $g \geq 0$  and  $P_t 1 = 1$  for all  $t$ . The infinitesimal generator of the semigroup is a linear operator  $\mathcal{L}$  defined on a dense subspace  $\mathcal{D}(\mathcal{L})$  of  $L^2(\pi)$  such that for any  $g \in \mathcal{D}(\mathcal{L})$ , we have  $\partial_t P_t g = \mathcal{L} P_t g$ . Also,  $\mathcal{L}$  can be defined as

$$\mathcal{L}g(W_t) := \lim_{h \rightarrow 0} \frac{P_h g(W_t) - g(W_t)}{h}.$$

The infinitesimal generator  $\mathcal{L}$  defines the Dirichlet form

$$\mathcal{E}(g) := - \int_{\mathbb{R}^d} g \mathcal{L} g d\pi.$$

Let  $P$  be a Markov semigroup with the unique invariant distribution  $\pi$  and the Dirichlet form  $\mathcal{E}$ . We say that  $\pi$  satisfies a Poincaré inequality with constant  $c$  if for all probability measures  $\mu \ll \pi$ , we have

$$\chi^2(\mu||\pi) \leq c \mathcal{E}\left(\sqrt{\frac{d\mu}{d\pi}}\right),$$

where  $\chi^2(\mu||\pi) := \|\frac{d\mu}{d\pi} - 1\|_{L^2(\pi)}^2$  is the  $\chi^2$  divergence, and  $\frac{1}{c} \leq \lambda$  with  $\lambda$  being the spectral

gap

$$\lambda := \inf\left\{\frac{\mathcal{E}g}{\int_{\mathbb{R}^d} g^2 d\pi} : g \in C^2, g \neq 0, \int_{\mathbb{R}^d} g = 0\right\}.$$

We say that  $\pi$  satisfies a Logarithmic Sobolev inequality with constant  $c$  if, for all  $\mu \ll \pi$ ,

$$D(\mu||\pi) \leq 2c\mathcal{E}\left(\sqrt{\frac{d\mu}{d\pi}}\right),$$

where  $D(\mu||\pi) = \int d\mu \log \frac{d\mu}{d\pi}$  is the KL-divergence.

Consider a Markov process  $\{W_t\}_{t \geq 0}$  with a unique invariant distribution  $\pi$  and the Dirichlet form  $\mathcal{E}$  such that  $\pi$  satisfies a Logarithmic Sobolev inequality with constant  $c$ . Then, we have the following [18]:

1. Let  $\mu_t := \mathcal{L}(W_t)$ , then we have  $D(\mu_t||\pi) \leq D(\mu_0||\pi)e^{-\frac{2t}{c}}$ .
2. If  $\mathcal{E}g = \alpha \int \|\nabla g\|^2 d\pi$  for some  $\alpha > 0$ , then, for any  $\mu \ll \pi$ ,  $\mathcal{W}_2(\mu, \pi) \leq \sqrt{2c\alpha D(\mu||\pi)}$ .

Given a data set  $D \in \mathcal{Z}^n$  with Langevin Monte Carlo Dynamic:

$$dW_t = -\nabla F(W_t, D)dt + \sqrt{2}dB_t. \quad (4.14)$$

If  $\nabla F(\cdot, D)$  is Lipschitz, then the Gibbs measure  $\pi_D(dw) \propto e^{-\beta F(w; D)}$  is the unique invariant measure of the underlying Markov semigroup. Its infinitesimal generator is

$$\mathcal{L}g(W_t) = (-\nabla F(W_t; D) \cdot \nabla + \Delta^2)g(W_t).$$

The corresponding Dirichlet form is

$$\mathcal{E}g = \int_{\mathbb{R}^d} \|\nabla g\|^2 d\pi.$$

Under some assumptions about the loss function, [243] shows that the Gibbs measure satisfy logarithmic Sobolev inequality.

**Lemma 4.2.1.** [Proposition 3.2 and Appendix B in [243]] For some  $\beta \geq O(1)$ , all of the Gibbs measures  $\pi$  satisfy a logarithmic Sobolev inequality with constant

$$c_{LS} \leq O\left(\frac{1}{\lambda^*}(d + \beta)\right),$$

where  $\lambda_*$  is the uniform spectral gap

$$\lambda_* := \inf_{D \in \mathcal{Z}^n} \inf \left\{ \frac{\int_{\mathbb{R}^d} \|\nabla g\|^2 d\pi_D}{\int_{\mathbb{R}^d} g^2 d\pi_D} : g \in C^1(\mathbb{R}^d) \cap L^2(\pi_D), g \neq 0, \int_{\mathbb{R}} g d\pi_D = 0 \right\}.$$

which satisfies:

$$\frac{1}{\lambda_*} \leq O\left(\frac{d + \beta}{\beta} \exp(O(\beta + d))\right).$$

Moreover, exponential dependence of  $\frac{1}{\lambda}$  on  $\beta$  is unavoidable in the presence of multiple local minima and saddle points.

The following shows a connection between time average of the diffusion and the corresponding Poisson equation.

The Poisson equation is an elliptic PDE on the basis of the infinitesimal generator associated with the Langevin dynamics. For the generator  $\mathcal{L}$  corresponding to the underlying Markov semigroup, we define the Poisson equation as

$$\mathcal{L}\psi = \phi - \bar{\phi},$$

where  $\phi$  is the test function, and  $\bar{\phi} := \int \phi(x) \pi(dx)$ .

## 4.2.2 Excess Risk of DP-ERM with Non-convex Loss Functions

We make the following assumptions in this section unless specified otherwise.

**Assumption 4.2.1.** 1. The hypothesis space  $\mathcal{C} = \mathbb{R}^d$ , regularizer is  $\ell_2$  norm, e.g.,  $r(\cdot) = \frac{\lambda}{2} \|\cdot\|^2$  for some  $\lambda > 0$ .

2. For any  $z \in \mathcal{Z}$ ,  $\ell(\cdot, z)$  is  $L$ -Lipschitz, and  $\ell(0, z) \leq A$ .
3. For each  $z \in \mathcal{Z}$ ,  $\ell(\cdot, z)$  is twice differentiable and is  $M$ -smooth.

These assumptions are quite standard in the DP-ERM literature with convex loss functions [67, 66]. For some non-convex loss functions such as the sigmoid function, it is easy to see that these assumptions are satisfied. For convenience, we assume that  $A, \lambda, L, M$  are all constants, which will be omitted in the big  $O$  notation. Also the big  $\tilde{O}$  terms omit the log terms.

We first review Gradient Langevin Dynamics (GLD), a popular generalization of the gradient descent algorithm. For ERM, the GLD algorithm executes the following recursion for  $w$  at iteration  $k$ :

$$w_k = w_{k-1} - \eta_{k-1} \nabla \hat{L}^r(w_{k-1}, D) + \sqrt{\frac{2\eta_{k-1}}{\beta}} \xi_{k-1}, \quad (4.15)$$

where  $\xi_{k-1}$  is a standard  $d$ -dimensional Gaussian random vector,  $\eta_{k-1}$  is the step size and  $\beta > 0$  is the inverse temperature parameter. Actually, GLD can be viewed as a discrete-time approximation of a continuous-time Langevin diffusion, described by the following stochastic differential equation (SDE):

$$dW_t = -\nabla \hat{L}^r(W_t, D) dt + \sqrt{2\beta^{-1}} dB_t, \quad (4.16)$$

where  $(B_t)_{t \geq 0}$  is a standard Brownian motion. It has been shown that the distribution of diffusion process in (4.16) converges to its stationary distribution, *i.e.* the Gibbs measure  $\pi(dw) \propto \exp(-\beta \hat{L}^r(w, D))$  [75]. Moreover, when  $\beta \rightarrow \infty$ , the distribution concentrates around the minimizer of  $\hat{L}^r(w, D)$ . By choosing the step size  $\eta$  properly, GLD can maintain differential privacy, as described in Algorithm 4.2.20.

It can be shown that Algorithm 4.2.20 ensures DP under certain conditions, as stated in Theorem 4.2.1.

---

**Algorithm 4.2.20** DP-GLD

---

**Input:** T is the iteration number.  $\epsilon, \delta$  are privacy parameters.

- 1: Choose an arbitrary point  $w_0$  from distribution density  $p_0(w)$  or fix the initial point  $w_0$ .
  - 2: Denote  $\eta = \frac{cn^2\epsilon^2}{L^2\beta T \log(1/\delta)}$ , where  $c = \frac{1}{c_2^2}$  is from Lemma 2.1.7.
  - 3: **for**  $k = 1, 2, \dots, T$  **do**
  - 4:      $w_k = w_{k-1} - \eta \nabla \hat{L}(w_{k-1}, D) + \sqrt{\frac{2\eta}{\beta}} \xi_{k-1}$ , where  $\xi_{k-1} \sim \mathcal{N}(0, I_d)$
  - 5: **end for**
  - 6:
  - 7: Return  $w_T$  or randomly sample  $j \in [T]$  and return  $w_j$ .
- 

**Theorem 4.2.1.** There exist constant numbers  $c_1$  and  $c_2$ , such that for any  $0 < \epsilon < c_1 T$  and  $0 < \delta < 1$ , Algorithm 4.2.20 is  $(\epsilon, \delta)$ -differentially private.

Our idea for proving an upper bound of the excess risk of ERM is based on the analysis of the convergence rate of GLD as in [83, 84, 243]. Let  $\mu_k$  be the probability law of  $w_k$  in (4.15), and  $\nu_{k\eta}$  the law of  $W_{k\eta}$  in (4.16). Our main step is to analyze the 2-Wasserstein distance  $\mathcal{W}_2(\mu_k, \pi)$ , which can be decomposed into  $\mathcal{W}_2(\mu_k, \nu_{k\eta})$  and  $\mathcal{W}_2(\nu_{k\eta}, \pi)$ . The key observation of our analysis is that in DP-GLD with  $k = T$ ,  $\eta T$  is a fixed number according to Algorithm 4.2.20, *i.e.*,  $\eta T = \Theta(\frac{n^2\epsilon^2}{\beta \log(1/\delta)L^2})$ . This means that the term  $\mathcal{W}_2(\nu_{T\eta}, \pi)$  is always fixed, no matter how large  $T$  is. For the  $\mathcal{W}_2(\mu_T, \nu_{T\eta})$  term, since  $w_T$  is a discretized version of  $W_{T\eta}$ , when  $\eta$  approaches 0,  $\mathcal{W}_2(\mu_T, \nu_{T\eta})$  will also approach 0. Thus, it appears that the best of what we can do for bounding  $\mathcal{W}_2(\mu_T, \pi)$  in DP-GLD is to bound it by  $\mathcal{W}_2(\mu_T, \nu_{T\eta})$ , *i.e.*,

$$\lim_{T \rightarrow \infty} \mathcal{W}_2(\mu_T, \pi) \leq \mathcal{W}_2(\nu_{T\eta}, \pi).$$

The above distance can be bounded as shown in a recent work by using Logarithmic Sobolev inequality [243]. For our problem, Theorem 4.2.2 extends the results by adapting recent non-asymptotic GLD theory [243, 348] and giving an upper bound of the excess risk for some initial points.

**Theorem 4.2.2.** Under the conditions of Theorem 4.2.1, if take  $T \geq \Theta(\frac{(M+\lambda)^2 n^2 \epsilon^2}{\lambda \beta \log(1/\delta) L^2})$  and  $\beta \geq \max\{\frac{4}{\lambda}, d\}$  in Algorithm 4.2.20, and assume that the probability law,  $\mu_0$ , of the initial

hypothesis  $w_0$  has a bounded and strictly positive density function w.r.t. Lebesgue measure on  $\mathbb{R}^d$ , and  $k_0 = \log \int_{\mathbb{R}^d} e^{\|w\|^2} p_0(w) dw < \infty$ , then the population risk at  $w_T$  is bounded by

$$\begin{aligned} \text{Err}_{\mathcal{P}}^r(w_T) &\leq O\left(\frac{n^{\frac{5}{2}}\epsilon^{\frac{5}{2}}}{\beta^{\frac{3}{4}}T^{\frac{1}{4}}\log(1/\delta)} + \exp(O(\beta))\exp\left[-\frac{n^2\epsilon^2}{\beta\log(1/\delta)\exp(O(\beta))}\right]\right. \\ &\quad \left. + \frac{\exp(O(\beta))}{n} + \frac{d\log(\beta)}{\beta}\right). \end{aligned} \quad (4.17)$$

The above bound implies that  $\lim_{T \rightarrow \infty} \text{Err}_{\mathcal{P}}^r(w_T) \leq O\left(\frac{\exp(O(\beta))\log(1/\delta)}{n^2\epsilon^2} + \frac{\exp(O(\beta))}{n} + \frac{d\log(\beta)}{\beta}\right)$  by the constraint on  $T$ .

For the empirical risk, we have

$$\text{Err}_D^r(w_T) \leq O\left(\exp(O(\beta))\exp\left[-\frac{n^2\epsilon^2}{\beta\log(1/\delta)\exp(O(\beta))}\right] + \frac{n^{\frac{5}{2}}\epsilon^{\frac{5}{2}}}{\beta^{\frac{3}{4}}T^{\frac{1}{4}}\log(1/\delta)} + \frac{d\log(\beta)}{\beta}\right). \quad (4.18)$$

**Remark 4.2.1.** We can see from Theorem 4.2.2 that the excess risk is only meaningful when  $\beta \geq O(d)$ . If set  $\beta = O(\log n)$ , or equivalently  $\log n \geq O(d)$ , both the excess *population and empirical risks* are bounded by  $\tilde{O}\left(\frac{\log(1/\delta)}{n\epsilon^2} + \frac{d}{\log(n)}\right) = \tilde{O}\left(\frac{\log(1/\delta)d}{\log(n)\epsilon^2}\right)$  when  $T \rightarrow \infty$ . These bounds are larger than the ones for convex loss functions, which are  $O\left(\frac{\sqrt{d\log(1/\delta)}}{n\epsilon}\right)$  and  $O\left(\frac{d\log(1/\delta)}{n^2\epsilon^2}\right)$  for population and empirical risks, respectively [29].

Next, we improve the bounds in Theorem 4.2.2 by using a finer analysis of the time-average error for the SDE (4.16). We show that Algorithm 4.2.20 achieves a lower error bound in term of  $n$ , i.e.,  $O\left(\frac{1}{n^{\Omega(1)}}\right)$  instead of  $O\left(\frac{1}{\log n}\right)$  for the empirical risk when fixing the initial point for  $w$ .

**Theorem 4.2.3.** With the same assumption as in Theorem 4.2.2 and a fixed initial point for  $w$ , if we return  $w_j$  in Algorithm 4.2.20 instead of  $w_T$ , where  $j$  is uniformly sampled from  $\{1, \dots, T\}$ , then the empirical risk is bounded, for sufficiently large  $T$ , by:

$$\text{Err}_D^r(w_j) \leq O\left(C\left[\frac{\beta^2\log(1/\delta)}{n^2\epsilon^2} + \frac{n^2\epsilon^2}{T\beta^2\log(1/\delta)}\right] + \frac{d}{\beta}\log(\beta)\right), \quad (4.19)$$

where  $C = C(d, \beta)$  is a function of  $d, \beta$ . Moreover, the bound is polynomially depending on  $\beta$  (assuming that  $d$  is a constant) with degree independent of  $d$ . In other words, if  $\beta$  satisfies  $\Theta(C \frac{\beta^2 \log(1/\delta)}{n^2 \epsilon^2}) = \Theta(\frac{d}{\beta} \log(\beta))$  in (4.19), then there exists a constant  $0 < \tau < 1$ , such that

$$\lim_{T \rightarrow \infty} \text{Err}_D^r(w_T) \leq \tilde{O}\left(\frac{C_0(d) \log(1/\delta)}{n^\tau \epsilon^\tau}\right), \quad (4.20)$$

where  $C_0(d)$  is a function of  $d$ .

**Remark 4.2.2.** Theorem 4.2.3 is a significant improvement over Theorem 4.2.2, which is derived based on a novel and non-trivial analysis on the time-average error of SDEs. Specifically, three points are worth emphasizing: 1) Although the time-average-error analysis of an SDE has been studied, for example in [291, 70], the non-asymptotic bounds in those results cannot be applied directly to our problem. This is because  $\beta$  in those results is assumed to be a constant. However, in our problem  $\beta$  is not even a constant, as it can be seen from (4.17) and (4.18). Furthermore, those results are based on the boundedness assumption on the solution of a Poisson equation (*e.g.*, Assumption 1 in [70] and Theorem 9 in [291]), which is too strong for our problem. Note that if  $\beta$  were a parameter, the hidden constant  $C$  in the bounds of [291, 70] would depend on  $\beta$ . Fortunately, through a rather non-trivial analysis, we are able to show that the constant is at most polynomially depending on  $\beta$ . Even though the exact degree of the polynomial is unknown, it is independent of  $d$ . 2) Our result is significant in the sense that it provides new bounds for diffusion-based Bayesian sampling such as [291, 70], where the dependency on  $d$  in their error bounds can be quantified, a key missing piece in previous results. 3) We reveal in Theorem 4.2.3 that if a random  $w_j$ , instead of the final  $w_T$ , is returned, one can improve the term related to  $n$  in the empirical risk bound from  $1/\log n$  to  $n^{-\tau}$ . It can be seen from (4.19) that the relationships between  $\beta$ ,  $d$ , and the constant  $C$  play an important role in proving the bound of the empirical risk. Since we are mainly targeting at the rate in terms of  $n$ , it suffices to consider only the relationship between  $\beta$  and  $C$ . We leave as an open problem to determine

whether it is possible to obtain an even tighter or explicit bound for the empirical risk. The ideal scenario is that  $C$  is independent of  $\beta$ . In this case, a better and more accurate bound of  $\tilde{O}(C_1(d)/(n\epsilon)^{\frac{2}{3}})$  can be obtained, where  $C_1(d)$  is a function of  $d$ .

From Theorem 4.2.2 and 4.2.3, we can see that the error bound for the *excess population risk* in terms of  $n$  is  $\frac{1}{\log n}$  (see Remark 4.2.1), while for the empirical risk it is  $\frac{1}{n^\tau}$ , where ideally  $\tau \leq \frac{2}{3}$  (see Remark 4.2.2). A natural question is thus to determine whether these bounds are tight. In the following, we first show that for loss functions satisfying Assumption 1, there is an  $\epsilon$ -DP algorithm whose error bound of the empirical risk is  $\tilde{O}(\frac{d}{n\epsilon})$  (and whose time complexity is exponential).

**Theorem 4.2.4.** For any  $\beta < 1$ , there is an  $\epsilon$ -differentially private algorithm, whose output  $w^{\text{priv}}$  satisfies, with probability at least  $1 - \beta$ ,  $\hat{L}^r(w^{\text{priv}}, D) - \hat{L}^r(w^*, D) \leq \tilde{O}(\frac{d}{n\epsilon})$ . The time complexity is  $O((1 + \frac{2Ln\epsilon}{\lambda d})^d n)$ .

Note that since  $\Theta(\frac{d}{n\epsilon})$  is the optimal bound for general convex functions [29], our empirical-risk bound of  $\tilde{O}(\frac{d}{n\epsilon})$  is thus near optimal.

In general, we can use an  $\alpha$ -net and the exponential mechanism to obtain a private estimator, which has an upper bound of  $\tilde{O}(\max\{\frac{d}{n\epsilon}, \alpha\})$  for the empirical risk with a time complexity of  $O((1 + \frac{2L}{\lambda\alpha})^d n)$ . Now consider the case that  $d$  is a constant. We can see that for the exponential mechanism, the bound in (4.20) can be obtained if we take  $\frac{1}{\alpha} = O(n^\tau)$ . However, in this case, the running time of exponential mechanism is  $O(n^{\tau d+1})$  compared to  $\tilde{O}(\text{Poly}(n, d))$  with Algorithm 4.2.20. Alternatively, the running time for achieving error  $\gamma$  in Algorithm 4.2.20 is polynomial in  $\frac{1}{\gamma}$ , while it is  $O((\frac{1}{\gamma}))^d$  with an exponential mechanism (for sufficient large  $n$ ). This means that Algorithm 4.2.20 is much more efficient when  $d$  is large.

Next, we consider upper bounding the excess population risk. Instead of determining the optimal bound, we show how to improve the bounds for some specific problems. Particularly, we focus on the generalized linear model with non-convex loss functions and the robust

regressions problem with additional assumptions, and present an  $(\epsilon, \delta)$ -DP algorithm for them with population risk  $O(\frac{\sqrt{d}}{\sqrt{n}\epsilon})$ . Note that these problems have been extensively studied in literature related to non-convex learning theory, such as [218, 115, 202, 207]. Here, we adopt the same assumptions as in [115].

**Generalized Linear Model** We consider the problem of learning a generalized linear model (GLM) with squared loss. We assume that  $\mathcal{X} = \{x \in \mathbb{R}^d | \|x\|_2 \leq 1\}$ ,  $\mathcal{C} = \{w \in \mathbb{R}^d | \|w\|_2 \leq 1\}$  and  $\mathcal{Y} = \{0, 1\}$ ,  $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ . With a link function  $\sigma$ , GLM endows a loss function:  $\ell(w, (x, y)) = (\sigma(\langle w, x \rangle) - y)^2$ . We further make the following assumptions on the link function, which includes the sigmoid and probit functions<sup>3</sup>.

**Assumption 4.2.2.** Let  $\mathcal{S} = [-1, 1]$ , we assume that

1.  $\exists$  constant  $C_\sigma \geq 1$  s.t.  $\max\{\sigma'(s), \sigma''(s)\} \leq C_\sigma$ , for  $\forall s \in \mathcal{S}$ .
2.  $\exists$  constant  $c_\sigma > 0$  s.t.  $\sigma'(s) \geq c_\sigma$ , for  $\forall s \in \mathcal{S}$ .
3. There exists some  $\|w^*\|_2 \leq 1$  such that  $\mathbb{E}[y|x] = \sigma(\langle w^*, x \rangle)$ .
4.  $|\sigma(s)| \leq B$  for some constant  $B > 0$ , for  $\forall s \in \mathcal{S}$ .

**Robust Regression** Let  $\mathcal{Z}$  and  $\mathcal{C}$  be the same as in GLM, and  $\mathcal{Y} = [-Y, Y]$  for some constant  $Y$ . For a non-convex positive loss function  $\psi$ , the loss of robust regression is defined as  $\ell(w, (x, y)) = \psi(\langle x, w \rangle - y)$ . We make the following assumptions on  $\psi$ , which includes the biweight loss function<sup>4</sup> [202].

**Assumption 4.2.3.** Let  $\mathcal{S} = [-(1 + Y), (1 + Y)]$ .

1.  $\exists C_\psi \geq 1$ , s.t.  $\max\{\psi'(s), \psi''(s)\} \leq C_\psi$ , for  $\forall s \in \mathcal{S}$ .

---

<sup>3</sup>The probit function is  $\sigma(s) = \Phi(s)$ , where  $\Phi$  is the Gaussian cumulative distribution function.

<sup>4</sup>For a fixed parameter  $c > 0$ , the biweight loss is defined as  $\psi(s) = \frac{c^2}{6} \cdot \begin{cases} 1 - (1 - (\frac{s}{c})^2)^3, & |t| \leq c \\ 1, & |t| \geq c. \end{cases}$

2.  $\psi'(\cdot)$  is odd with  $\psi'(s) > 0$ , for  $\forall s > 0$ ; and  $h(s) := \mathbb{E}_\xi[\psi'(s + \xi)]$  satisfies  $h'(0) > c_\psi$ , where  $c_\psi > 0$ .

3. There is  $w^* \in \mathcal{C}$  such that  $y = \langle w^*, x \rangle + \xi$ , where  $\xi$  is symmetric noise with a zero-mean given  $x$ .

**Algorithm 4.2.21 DP-FW-L2**

**Input:**  $T$  is the number of iterations,  $w_1$  is the initial point, and  $\{\gamma_t\}_{t=1}^T$  is the step size.  $\epsilon$  and  $\delta$  are privacy parameters.

- 1: **for**  $t = 1, \dots, T$  **do**
- 2:     Compute  $v_t = \arg \max_{v \in \mathcal{C}} \langle v, -(\nabla \hat{L}(w_t, D) + \epsilon_t) \rangle$ , where  $\epsilon_t \sim N(0, \sigma^2 I_d)$  for some  $\sigma$ .
- 3:      $w_{t+1} = w_t + \gamma_t(v_t - w_t)$ .
- 4: **end for**
- 5:
- 6: Return  $w_R \in \{w_1, \dots, w_T\}$  such that  $R$  is uniformly sampled from  $\{1, \dots, T\}$ .

Algorithm 4.2.21 solves both problems and is motivated by the fact that the population risk satisfies the inequality,  $L_{\mathcal{P}}(w) - L_{\mathcal{P}}(w^*) \leq \mu \langle \nabla L_{\mathcal{P}}(w), w - w^* \rangle$ , for some constant  $\mu > 0$  and  $\forall w \in \mathcal{C}$ . Thus, it suffices to get an upper bound of  $\langle \nabla L_{\mathcal{P}}(w), w - w^* \rangle$ . It turns out that this can be obtained via a DP version of the Frank-wolfe method.

**Theorem 4.2.5.** For the general linear model with Assumption 4.2.2, there exist constants  $c_1$  and  $c_2 > 0$  such that for any  $0 < \epsilon < c_1 T$  and  $0 < \delta < 1$ , Algorithm 4.2.21 is  $(\epsilon, \delta)$ -DP when  $\sigma^2 = c_2 \frac{C_\sigma^2 (B+1)^2 T \log \frac{1}{\delta}}{n^2 \epsilon^2}$ . Moreover, if taking  $\gamma_t = O\left(\frac{\sqrt[4]{d \ln \frac{1}{\delta}}}{\sqrt{n \epsilon}}\right)$  for all  $t \in [1, \dots, T]$  with  $T = O\left(\frac{n \epsilon}{\sqrt{d \ln \frac{1}{\delta}}}\right)$ , we have  $\text{Err}_{\mathcal{P}}(w_R) \leq O\left(\frac{\sqrt[4]{d \ln \frac{1}{\delta}}}{\sqrt{n \epsilon}}\right)$ , where the big- $O$  notations omit other terms.

For the case of robust regression with Assumption 4.2.3, if we take  $\sigma^2 = c_2 \frac{C_\psi^2 T \log \frac{1}{\delta}}{n^2 \epsilon^2}$ , the algorithm is  $(\epsilon, \delta)$ -DP. Moreover, with the same conditions on  $T, \{\gamma_t\}_{t=1}^T$  as above, it can be derived that  $\text{Err}_{\mathcal{P}}(w_R) \leq O\left(\frac{\sqrt[4]{d \ln \frac{1}{\delta}}}{\sqrt{n \epsilon}}\right)$ , where the big- $O$  notations omit other terms.

Motivated by Algorithm 4.2.21, under the conditions of  $\mathcal{X} = \{x \in \mathbb{R}^d | \|x\|_\infty \leq 1\}$  and  $\mathcal{C} = \{w \in \mathbb{R}^d | \|w\|_1 \leq 1\}$ , we can actually derive an upper bound of the population risk

that depends only logarithmically on  $d$  (*i.e.*,  $\log d$ ), indicating that it is suitable for high dimensional applications. Note that the conditions on  $\mathcal{X}$  and  $\mathcal{C}$  have been considered in linear regression [270]. We adopt them to our problem and extend their DP-Frank-Wolfe algorithm to Algorithm 4.2.22.

---

**Algorithm 4.2.22** DP-FW-L1

---

**Input:**  $T$  is the iteration number and  $w_1$  is the initial point.  $\{\gamma_t\}_{t=1}^T$  is the step size.  $A$  is the set of vertices of  $\mathcal{C}$ .  $\epsilon$  and  $\delta$  are privacy parameters.

- 1: **for**  $t = 1, \dots, T$  **do**
  - 2:   Use exponential mechanism  $\mathcal{M}(D, u, \mathcal{R})$ , where  $\mathcal{R} = A$ ,  $u(D, s) = -\langle s, \nabla \hat{L}(w_t, D) \rangle$ , to ensure  $(\frac{\epsilon}{\sqrt{8T \ln(\frac{1}{\delta})}}, 0)$ -differential privacy. Denote the output as  $\tilde{w}_t$ .
  - 3:   Compute  $w_{t+1} = (1 - \gamma_t)w_t + \gamma_t \tilde{w}_t$ .
  - 4: **end for**
  - 5:
  - 6: Return  $w_R \in \{w_1, \dots, w_T\}$ , where  $R$  is uniformly sampled from  $\{1, 2, \dots, T\}$ .
- 

**Theorem 4.2.6.** Let  $\mathcal{X} = \{x \in \mathbb{R}^d | \|x\|_\infty \leq 1\}$  and  $\mathcal{C} = \{w \in \mathbb{R}^d | \|w\|_1 \leq 1\}$ . For the GLM and robust regression problems, Algorithm 4.2.22 is  $(\epsilon, \delta)$ -DP with sensitivities  $\Delta = O(\frac{C_\sigma(B+1)}{n})$  and  $\Delta = O(\frac{C_\psi}{n})$ , respectively. Furthermore, if we set  $T = O(\frac{n\epsilon}{\sqrt{\ln(\frac{1}{\delta}) \ln(dn/\eta)}})$  and  $\{\gamma_t\}_{t=1}^T = O(\sqrt{\frac{2}{T}})$ , then with probability at least  $1 - \eta$ , we have  $\text{Err}_P(w_R) \leq O(\frac{4\sqrt{\ln(\frac{1}{\delta})} \sqrt{\ln \frac{nd}{\eta}}}{\sqrt{n\epsilon}})$ . Here the big- $O$  notations omit other terms.

### 4.2.3 Omitted Proofs

#### Proof of Theorem 4.2.1

Firstly, we can see that if in each iteration

$$w_k = w_{k-1} + \eta(\nabla \hat{L}^r(w_{k-1}, D) + \xi_1) + \frac{\sqrt{\eta}}{\sqrt{\beta}} \xi_2,$$

where  $\xi_1 \sim \mathcal{N}(0, \frac{L^2 c_2^2 \log(1/\delta) T}{n^2 \epsilon^2} I_d)$  and  $\xi_2 \sim \mathcal{N}(0, I_d)$ , then by moment account (Lemma 2.1.7), we can see that it is  $(\epsilon, \delta)$ -differentially private for  $\epsilon < c_1 T$  and  $0 < \delta < 1$ . Furthermore, if

$\eta^2 \frac{L^2 c_2^2 \log(1/\delta) T}{n^2 \epsilon^2} = \frac{\eta}{\beta}$  or  $\eta = \frac{n^2 \epsilon^2}{T L^2 c_2^2 \beta \log(1/\delta)}$ , then it is equivalent to the updating in Algorithm 4.2.20. This completes the proof.

### Proof of Theorem 4.2.2

The proof follows the framework of the proof in [243].

**Notations** For a given dataset  $D$ , we denote the corresponding Gibbs measure as  $\pi_D \propto e^{-\beta \hat{L}^r(w, D)}$ . Also, let  $\mu_{T,D} = \mathcal{L}(w_T|D)$  and  $\nu_{t,D} = \mathcal{L}(W_t|D)$ .

Firstly, we show that our assumptions about the loss function and  $w_0$  meet the assumptions in [243]. Actually, our setting implies that  $f(w, z) = \ell(w, z) + \frac{\lambda}{2} \|w\|^2$  in [243]. It is easy to see that  $A = A$ ,  $B = L$ ,  $M = M + \lambda$  in [243]. Also, when  $\ell(\cdot, z)$  is  $L$ -Lipschitz, we know that  $f(w, z) = \ell(w, z) + \frac{\lambda}{2} \|w\|^2$  is  $(m = \frac{\lambda}{2}, b = \frac{L^2}{2\lambda})$ -dissipative (that is,  $\langle w, \nabla f(w, z) \rangle \geq \frac{\lambda}{2} \|w\|^2 - \frac{L^2}{2\lambda}$ ), which satisfies assumption A.3 in [243]. For A.4, we can see that Algorithm 4.2.20 is just the non-stochastic version. Hence,  $\delta = 0$ . Thus, most of the analysis in [243] can also be applied here. For self-completeness, we will rephrase them so that they fit our differentially private context.

Now, we briefly introduce the proof in [243]. Let  $\hat{w}^*$  be the output of the Gibbs algorithm under which the conditional distribution of  $\hat{w}^*$  is equal to  $\pi_D$ . Then, we decompose the population risk into the following

$$\mathbb{E} L_{\mathcal{P}}^r(w_T) - L_{\mathcal{P}}^r(w^*) = \mathbb{E} L_{\mathcal{P}}^r(w_T) - \mathbb{E} L_{\mathcal{P}}^r(\hat{w}^*) + \mathbb{E} L_{\mathcal{P}}^r(\hat{w}^*) - \mathbb{E} \hat{L}^r(\hat{w}^*, D) + \mathbb{E} \hat{L}^r(\hat{w}^*, D) - L_{\mathcal{P}}^r(w^*).$$

For the second term, by Proposition 3.5 in [243] we have

$$\mathbb{E} L_{\mathcal{P}}^r(\hat{w}^*) - \mathbb{E} \hat{L}^r(\hat{w}^*, D) \leq O\left(\frac{(\beta + d)c_{LS}}{n}\right) = O\left(\frac{\exp(O(\beta + d))}{n}\right) = O\left(\frac{\exp(O(\beta))}{n}\right), \quad (4.21)$$

where the big  $O$  notation hides the parameters of  $M, b, B$  (that is  $L, M, \lambda$  in our setting) by the assumption of  $\beta > d$ .

For the third term, we have the following theorem:

**Lemma 4.2.2** ([243]). For any  $\beta \geq \frac{2}{m}$ ,

$$\mathbb{E}\hat{L}^r(\hat{w}^*, D) - L_{\mathcal{P}}^r(w^*) \leq O\left(\frac{d}{\beta} \log(\beta)\right),$$

where the big  $O$  notation omits the factor of  $M, m$ .

In order to estimate the term of  $\mathbb{E}L_{\mathcal{P}}^r(w_T) - \mathbb{E}L_{\mathcal{P}}^r(\hat{w}^*)$ , we have to estimate  $\mathbb{E}\hat{L}_D^r(w_T) - \mathbb{E}\hat{L}_D^r(\hat{w}^*)$  for each  $D \in \mathcal{Z}^n$ . The goal is to get an upper bound for  $\mathcal{W}_2(\mu_{T,D}, \pi_D) \leq \mathcal{W}_2(\mu_T, \nu_{T,\eta,D}) + \mathcal{W}_2(\nu_{T,\eta,D}, \pi_D)$  for all dataset  $D$ .

For the term  $\mathcal{W}_2(\nu_{T,\eta,D}, \pi_D)$ , since  $\nu$  is related to the continuous-time Langevin diffusion (4.14), and  $T\eta$  is a fixed value, which is independent of  $\eta$ , we have (see Section 3.4 in [243]):

$$\mathcal{W}_2(\nu_{T,\eta,D}, \pi_D) \leq O\left(\sqrt{(d+\beta)c_{LS}}e^{-\frac{T\eta}{\beta c_{LS}}}\right) = O\left(\exp(O(\beta))\exp\left(-\frac{T\eta}{O(\exp(\beta))}\right)\right). \quad (4.22)$$

Note that  $T\eta = \frac{n^2\epsilon^2}{L^2c_2^2\beta\log(1/\delta)}$ .

Our final goal is to estimate  $\mathcal{W}_2(\mu_{T,D}, \nu_{T,\eta,D})$ . The proof is the same as in [243]. However, we can see that  $\frac{m}{4M^2} = \frac{\lambda}{8(M+\lambda)^2} \leq 1$ . This means that in order to use the result in [243], we have to ensure that  $\eta \leq O(\frac{m}{M^2}) = O(\frac{\lambda}{(M+\lambda)^2})$ . That is,  $T \geq C\frac{n^2\epsilon^2(M+\lambda)^2}{\beta L^2\log(1/\delta)\lambda}$ .

We can easily get (see Proposition 3.1 in [243]):

$$\mathcal{W}_2^2(\mu_{T,D}, \nu_{T,\eta,D}) \leq O(\beta\sqrt{\eta}(T\eta)^2). \quad (4.23)$$

Thus, we have

$$\mathcal{W}_2(\mu_{T,D}, \pi_D) \leq O\left(\frac{(n\epsilon)^{\frac{5}{2}}}{\beta^{\frac{3}{4}}\log(1/\delta)T^{\frac{1}{4}}} + \sqrt{(d+\beta)c_{LS}}e^{-\frac{T\eta}{\beta c_{LS}}}\right). \quad (4.24)$$

For all  $D \in \mathcal{Z}^n$ , we have

$$\begin{aligned} \int \hat{L}^r(w, D) \mu_{T,D}(dw) - \int \hat{L}^r(w, D) \pi_D(dw) &\leq O\left(\frac{(n\epsilon)^{\frac{5}{2}}}{\beta^{\frac{3}{4}} \log(1/\delta) T^{\frac{1}{4}}} \right. \\ &\quad \left. + \exp(O(\beta)) \exp\left(-\frac{n^2 \epsilon^2}{\log(1/\delta) O(\exp(\beta))}\right)\right), \end{aligned} \quad (4.25)$$

where  $O$  is independent of  $\beta, T, n, \epsilon, \delta$ .

Combining this with Lemmas 4.2.2, (4.25) and (4.21), we have the proof.

The result of the limit comes from the fact that  $\exp(-x) \leq \frac{1}{x}$ .

### Proof of Theorem 4.2.3

For convenience, we let  $F(w)$  denote  $\hat{L}^r(w, D)$ . Then, the updating becomes

$$w_{t+1} = w_t - \eta \nabla F(w_t) + \sqrt{\frac{2\eta}{\beta}} \zeta_t. \quad (4.26)$$

By scaling  $\eta' = \frac{\eta}{\beta}$  and  $F' = \beta F$ , we have

$$w_{t+1} = w_t - \eta' \nabla F'(w_t) + \sqrt{2\eta'} \zeta_t. \quad (4.27)$$

Note that the technique of rescaling is commonly used in other papers, *e.g.*, [83, 348].

The continuous Langevin dynamic corresponding to (4.27) is

$$dW(t) = -\nabla F'(W(t))dt + \sqrt{2}dB(t). \quad (4.28)$$

$$\mathcal{L}g = -\nabla g \cdot \nabla F' + \Delta^2 g \quad (4.29)$$

Also the invariant distribution is  $\pi(dw) \propto e^{-F'(W)}$ , and the Poisson equation is

$$\mathcal{L}\psi = \phi - \bar{\phi}, \quad (4.30)$$

where  $\bar{\phi} = \int \phi(w)\pi(dw)$  and  $\phi$  is the testing function.

A seemingly straightforward way to prove the result is to use the theorem on finite time sample average error of SGLD, such as Theorem 2 in [70] or (55) in [291] to our equation (4.27). However, both papers consider only the case of  $\beta = 1$ , and their assumptions are quite strong compared to ours. This means that the hidden constants in their bounds may depend on  $\beta$  and the dimensionality  $d$ . However, as can be seen from above,  $\beta$  cannot be assumed as a constant in our problem. Thus we cannot directly apply their results.

Next, we will use some of the ideas in the proof of Theorem 9 in [291] to show that the constants depend only polynomially on  $\beta$  and the degree of the polynomial is independent of  $d$ . We refer the reader to Section 9 in [291].

For convenience, we assume that the test function  $\phi = F$ . Now consider the solution  $\psi$  to the Possion equation (4.30) for  $\phi$  (note that the existence will be shown later for a class  $\phi$  of functions). Also, for  $\psi(w_{t+1})$ , we use Taylor expansion at  $w_t$ ; that is (note that since we now only need to estimate the bias, we just expand it to the third order, which is different from the one in [291]),

$$\begin{aligned} \psi(w_{t+1}) &= \psi(w_t) + \nabla\psi(w_t)(w_{t+1} - w_t) + \frac{1}{2}(w_{t+1} - w_t)^T \nabla^2\psi(w_t)(w_{t+1} - w_t) + \mathcal{R}_t \\ \end{aligned} \tag{4.31}$$

$$\begin{aligned} &= \psi(w_t) + \nabla\psi(w_t)(-\eta' \nabla F'(w_t) + \sqrt{2\eta'} \zeta_t) + \\ &\quad \frac{1}{2}\eta'^2 \nabla F'(w_t) \nabla'^2\psi(w_t) \nabla F'(w_t) - \sqrt{2\eta'} \eta' \nabla F'(w_t) \zeta_t + \eta' \zeta_t^T \nabla^2\psi(w_t) \zeta_t + \mathcal{R}_t, \end{aligned} \tag{4.32}$$

where  $\mathcal{R}_t = \frac{1}{6} \int_0^1 s^2 \psi^{(3)}(sw_t + (1-s)w_{t+1})(w_{t+1} - w_t, w_{t+1} - w_t, w_{t+1} - w_t) ds$  and (4.32) comes from (4.27).

Taking the expectation on  $\psi(w_{t+1})$ , we have

$$\begin{aligned}\mathbb{E}\psi(w_{t+1}) - \mathbb{E}\psi(w_t) &= -\eta' \nabla \psi(w_t) \nabla F'(w_t) \\ &\quad + \eta' \Delta^2 \psi(w_t) + \frac{1}{2} \eta'^2 \nabla F'(w_t) \nabla'^2 \psi(w_t) \nabla F'(w_t) + \mathbb{E}\mathcal{R}_t.\end{aligned}\quad (4.33)$$

By (4.29), we have

$$\eta' \mathbb{E}\mathcal{L}(\psi)(w_t) = \mathbb{E}\psi(w_{t+1}) - \mathbb{E}\psi(w_t) - \frac{1}{2} \eta'^2 \nabla F'(w_t) \nabla'^2 \psi(w_t) \nabla F'(w_t) - \mathbb{E}\mathcal{R}_t. \quad (4.34)$$

Summing over all  $t$  for  $t = 1, \dots, T$  and dividing  $\eta'T$  on both sides, by Poisson equation (4.30) we get:

$$\begin{aligned}\mathbb{E}\left(\frac{\sum_{t=1}^T \phi(w_t)}{T} - \bar{\phi}\right) &= \frac{1}{\eta'T} \mathbb{E}[\psi(w_{T+1}) - \psi(w_1)] - \frac{1}{\eta'T} \mathbb{E} \sum_{t=1}^T \mathcal{R}_t \\ &\quad - \frac{1}{2} \frac{\eta'}{T} \sum_{t=1}^T \nabla F'(w_t) \nabla'^2 \psi(w_t) \nabla F'(w_t).\end{aligned}\quad (4.35)$$

What we need to prove are the following inequalities.

$$\sup_t \mathbb{E}\psi(w_t) \leq C_1, \quad (4.36)$$

$$\sup_t \mathbb{E}\mathcal{R}_t \leq \eta'^2 C_2 \quad (4.37)$$

$$\sup_t \mathbb{E} \nabla F'(w_t) \nabla'^2 \psi(w_t) \nabla F'(w_t) \leq C_3, \quad (4.38)$$

where  $C_1, C_2, C_3$  are independent of  $\eta$  and at most polynomially depending on  $\beta$  with their degrees independent of  $d$  (note that they may depend on  $d$ , but we only care about  $\beta$ ). If we can show these, then we have the proof.

To prove these inequalities, we want to show for the testing function  $\phi$  and its corre-

sponding  $\psi$  the following

$$\|\psi^{(i)}\| \leq C_i^1 f, \forall i = \{0, 1, 2, 3\}, \quad (4.39)$$

where  $\{C_i^1\}$  are constants that are at most polynomially depending on  $\beta$  (with degrees independent of  $d$ ) and  $f(x) = 1 + \|x\|_2^2$  is the quadratic function.

Also, we want to show for every  $m \in \mathbb{N}$ ,

$$\sup_t \mathbb{E}\|w_t\|_2^m \leq C_m^2 < \infty, \quad (4.40)$$

where  $\{C_m^2\}$  are constants that also are at most polynomially depending on  $\beta$  (with degrees independent of  $d$ ).

It is easy to see that if the above inequalities (*i.e.*, (4.39)(4.40)) can be proven, then for (4.36) we have  $\sup_t \mathbb{E}\psi(w_t) \leq O(C_0^1 C_2^2)$ ; for (4.37), we have

$$\begin{aligned} \sup_t \mathbb{E}\mathcal{R}_t &\leq O(C_3^1(1 + \|w_t\|^2)[\eta'^3\|\nabla F'(w_t)\|^3 + \eta'^2\|\nabla F'(w_t)\|]) \\ &\leq O(\eta'^2 C_3^1 f[\|\beta\nabla F(w_t)\|^3 + \|\beta\nabla F(w_t)\|]). \end{aligned} \quad (4.41)$$

Since  $F$  is smooth, we have  $\|\nabla F(w)\| \leq \frac{M_0}{2}(1 + \|w\|)$  for some  $M_0$  independent of  $\beta$ . Thus, by (4.41), we have  $\sup_t \mathbb{E}\mathcal{R}_t \leq O(\eta'^2 C)$ , where  $C$  is at most polynomially depending on  $\beta$ . Similarly, we can show for (4.38).

Thus, our goal is now to prove (4.39) and (4.40). For (4.40), we have the following theorem:

**Theorem 4.2.7.** For every  $m$ , if  $\beta > d$  and sufficiently small  $\eta$  in (4.26)

$$\sup_t \mathbb{E}\|w_t\|_2^{2m} \leq C_m^2 < \infty, \quad (4.42)$$

where  $w_t$  is in (4.26) and  $C_m^2$  is independent of  $\beta$ .

**Proof of Theorem 4.2.7.** For  $m=1$ , it has been shown in Lemma 3.2 in [243] that  $C_1^2 = O(\frac{d}{\beta}) = O(1)$ , which satisfies our requirements. Actually, for any  $m$ , we can follow the proof of Lemma 3.2 in [243], to show that there is a sufficiently small  $\eta$  which makes the Theorem hold.

For example, when  $m = 2$ ,  $\mathbb{E}\|w_t\|_2^4 = \mathbb{E}\|w_{t-1} - \eta\nabla F(w_{t-1}) + \sqrt{2\eta/\beta}\zeta_{t-1}\|^4 \leq O(\mathbb{E}\|w_{t-1} - \eta\nabla F(w_{t-1})\|^4 + \eta^4)$ , also for  $\mathbb{E}\|w_{t-1} - \eta\nabla F(w_{t-1})\|^4 \leq \mathbb{E}\|w_{t-1}\|^4(1 - \Theta(\eta) + \Theta(\eta^2) - \Theta(\eta^3) + \Theta(\eta^4)) + O(\eta)$ . The constants in the big- $\Theta$  and big- $O$  notations are independent of  $\beta$ . Thus, if we take a sufficiently small  $\eta$ , which makes  $(1 - \Theta(\eta) + \Theta(\eta^2) - \Theta(\eta^3) + \Theta(\eta^4)) < 1$ , then we can get an upper bound that is independent of  $\beta$  for  $\beta \geq \max\{1, d\}$ . The same argument goes for all  $m \in \mathbb{N}$ . Thus we have the proof.  $\square$

**Proof of (4.40)** Now by Theorem 4.2.7, we have for every  $m$ ,  $\sup_t \mathbb{E}\|w_t\|^m < C_m$ , where  $C_m$  is at most polynomially depending on (actually is independent of)  $\beta$ , since by Jensen's Inequality we have  $\sup_t \mathbb{E}\|w_t\|_2^m \leq \sqrt{\mathbb{E}\|w_t\|^{2m}}$ . This proves (4.40).

**Proof of (4.39)** For (4.39), the key point is that our testing function is bounded by a quadratic function, due to the L-smoothness of our assumption. We have the following theorem due to Theorem 1 and 2 in [241] (corresponding to the case of  $\alpha = 1 > 0$ ,  $b(x) = F'(x) = \beta L^r(w, D)$  and  $r_0 = \infty$  in the Has'minski's assumption) and Theorem 13 in [291].

**Theorem 4.2.8** (Theorem 1 and 2 in [241]). Consider the Poisson equation in  $\mathbb{R}^d$ ,

$$\mathcal{L}u(x) = -f(x), \quad (4.43)$$

where  $\mathcal{L}$  is the infinitesimal generator of the diffusion process (4.28). We further assume that  $\int f(x)\pi(dx) = 0$ , where  $\pi$  is the invariant measure of the diffusion process. If  $\|f(x)\| \leq C_1 + C_2\|x\|^s$  for some  $s > 0$  and some constants  $C_1, C_2$ . Then (4.43) defines a continuous function  $u(x)$  which belongs to the Sobolev class  $W_{p,\text{loc}}^2$  for any  $p > 1$ , and

satisfies the following properties,

1. There exists a constant  $C'$  such that

$$|u(x)| \leq C'(1 + \|x\|^s), \quad (4.44)$$

where  $C'$  is determined only by  $C_1, C_2$  and  $C_m$ , and  $C_m$  is determined only by the constants in equations (4)-(6) in [241] for  $m > s + 2$ .

2. Moreover, there is a constant  $C$  such that

$$\|\nabla u(x)\| \leq C(1 + \|x\|^s), \quad (4.45)$$

where  $C$  is determined only by  $C_1, C_2$  and  $C_m$ , and  $C_m$  is determined only by the constants in equations (4)-(6) in [241] for  $m > s + 2$ .

Now by the proofs of Theorem 1 and 2 in [241], we have the following theorem:

**Theorem 4.2.9.** For our test function  $\phi$ , if fixing  $m = 6$  in Theorem 4.2.8, then  $C'$  in (4.44) is polynomially depending on  $C_m, C_1, C_2$ , and the same for  $C$  in (4.45).

*Proof.* The proof of  $C'$  depending polynomial on  $C_1, C_2, C_m$  can be easily found in the proof of Theorem 1 and Theorem 2 in [241]. Since for our test function  $\phi$ ,  $s = 2$  by the  $M$ -smooth property. Thus, we need  $m > \beta + 2$  and  $m > 2k + 2$  for some  $k > 0$  (See Proposition 1 in [241]). This means that choosing  $m = 6$  can satisfy the condition in Theorem 1 of [241].

For  $C$ , we follows the proof of Theorem 1 in [241]. In [241], the proof is by (4.44), Sobolev embedding theorem and Theorem 9.11 and (9.40) in [129]. From the proof of Theorem 9.11 in [129], we can see that the hidden constant behind is only polynomially depending on the upper bounds of the coefficients of the second order PDE, which means only polynomially depending on  $\beta$  in our problem. Also by Sobolev embedding theorem,

we can see that the polynomial dependence on  $\beta$  will be unchanged. Thus, we have the proof for  $C$ .  $\square$

Next, we show that  $C_1, C_2, C_m$  are at most polynomially depending on  $\beta$ .

**Theorem 4.2.10.** For a fixed number  $m$  in (4)-(6) in [241] related to the diffusion process (4.28),  $C_1, C_2$  and the constants in (4)-(6) in [241] are at most polynomially depending on  $\beta$  (which is  $r$  in assumption  $A_b$  in [241]). Thus,  $C_m$  is at most polynomially depending on  $\beta$ .

*Proof.* For  $C_1, C_2$ , since  $f = \bar{\phi} - \phi$ , which corresponds to (4.43) in Theorem 4.2.8, where  $\phi = \hat{L}^r(\cdot, D)$ , hence we have  $\|f(x)\| \leq \|\hat{L}^r(x, D)\| + \|\bar{\phi}\|$ . For the term of  $\hat{L}^r(x, D)$ , since it is  $(M + \lambda)$ -smooth, thus  $\hat{L}^r(x, D) \leq \frac{M_0}{2}(1 + \|x\|^2)$  for some  $M_0$  which is independent of  $\beta$ . For the term  $\bar{\phi} = \int \hat{L}^r(w, D)\pi(dw)$ , by Proposition 3.4 of [243], we know that if  $\beta \geq \frac{2}{m} = \frac{4}{\lambda}$ , then  $\bar{\phi} \leq O(\frac{d}{\beta} \log(\beta + d) + \min \phi)$ , which is at most polynomially depending on  $\beta$ . Thus,  $C_1, C_2$  are at most polynomially depending on  $\beta$  with their degrees independent of the dimensionality  $d$ .

Now, let us consider  $C_m$ . Actually, by the proof of Theorem 1 in [241], we can see that  $C_m$  only depends polynomially on the constants of (4)-(6) in proposition 1 in [241]. Thus, it suffices to show that constants of (4)-(6) in proposition 1 in [241] depends only polynomially on  $\beta$ .

To show this, we can see that  $\beta^{\frac{\lambda}{2}}$  corresponds to  $r$  and  $\alpha = 1$  in [241]. The proof of proposition 1 in [241] comes from Lemma 1-Lemma 8 in [287]. From the proof in [287], we know that all the constants of (4)-(6) in proposition 1 in [241] are polynomially depending on  $r$ , i.e.  $\beta$  and their degrees are independent of  $d$ .

Thus  $C_1, C_2, C_m$  are all at most polynomially depending on  $\beta$  with their degrees independent of  $d$ .  $\square$

To summarize, we have the following theorem:

**Theorem 4.2.11.** The constant  $C$  and  $C'$  in (4.44), (4.45) are at most polynomially depending on  $\beta$ , moreover, the degree of the polynomial is independent on  $d$ .

Upto now, we have showed that  $\|\psi^i\| \leq C_i^1 f$  for  $i = \{0, 1\}$ , where  $f$  is a quadratic function and  $C_i^1$  are polynomially depending on  $\beta$ .

What is still left is for  $i = \{2, 3\}$ . To prove this, our idea is to use the trick in [291] (see A.9-A.11 and Lemma 15 in [291]). That is, we note that the derivatives of  $\psi$  can be expressed as the solution to different Poisson equations. Also, by iterating Theorem 4.2.8, 4.2.9, 4.2.10, 4.2.11, we can get all the constant  $C_i^1$  depending at most polynomially on  $\beta$ .

Putting all these together, we have showed that

$$\mathbb{E}\left(\frac{\sum_{t=1}^T \phi(w_t)}{T} - \bar{\phi}\right) \leq C\left(\frac{1}{\eta' T} + \eta'\right), \quad (4.46)$$

where  $C$  is at most polynomially depending on  $\beta$  whose degree is independent of  $d$  (we omit other terms and consider only  $\beta$ ). Taking  $\eta' = \frac{\eta}{\beta}$  and  $\eta$  in Algorithm 1, also noting that  $\mathbb{E}\hat{L}^r(w_j, D) = \mathbb{E}\frac{\sum_{i=1}^T \hat{L}^r(w_i, D)}{T}$  and  $\phi = \hat{L}^r(\cdot, D)$ , by Proposition 3.4 in [243], we can get the proof.

### Proof of Theorem 4.2.4

**Lemma 4.2.3.** [104] For the exponential mechanism  $\mathcal{M}(D, u, \mathcal{R})$ , we have

$$\Pr\{u(\mathcal{M}(D, u, \mathcal{R})) \leq OPT_u(x) - \frac{2\Delta u}{\epsilon}(\ln |\mathcal{R}| + t)\} \leq e^{-t}.$$

where  $OPT_u(x)$  is the highest score in the range  $\mathcal{R}$ , i.e.  $\max_{r \in \mathcal{R}} u(D, r)$ .

We first show that the optimal value  $w^* = \arg \min_{w \in \mathbb{R}^d} \hat{L}^r(w, D)$  contained in the ball  $\mathbb{B}^d(\frac{L}{\lambda})$ . This is because under our assumption,  $\hat{L}^r(w, D)$  is  $(\frac{\lambda}{2}, \frac{L^2}{2\lambda})$ -dissipative. That is,  $\forall w \in \mathbb{R}^d, \langle w, \nabla \hat{L}^r(w, D) \rangle \geq \frac{\lambda}{2}\|w\|^2 - \frac{L^2}{2\lambda}$ . Thus,  $w^* = \arg \min_{w \in \mathbb{B}^d(\frac{L}{\lambda})} \hat{L}^r(w, D)$ .

For any  $\alpha > 0$ , by a simple volume argument (Lemma 5.2 in [289]) we can see that there

exists an  $\alpha$ -net  $\mathcal{N}_\alpha$  whose size is at most  $(1 + \frac{2L}{\lambda\alpha})^d$ . Then, by the property that  $\hat{L}^r(w, D)$  is  $O(L)$ -Lipschitz, we have the following:

$$\min_{w \in \mathcal{N}_\alpha} \hat{L}^r(w, D) - \hat{L}^r(w^*, D) \leq O(L\alpha).$$

Now consider the following  $\epsilon$ -DP algorithm. We set the score function  $u(D, w) = -(\hat{L}^r(w, D) - \hat{L}^r(w_0, D))$ , where  $w_0 \in \mathbb{B}^d(\frac{L}{\lambda})$  is an arbitrary point; the range space  $\mathcal{R} = \mathcal{N}_\alpha$ . Since  $\hat{L}^r(w, D)$  is  $O(L)$ -Lipschitz in  $\mathbb{B}^d(\frac{L}{\lambda})$ , the sensitivity is at most  $O(\frac{L}{n})$ . Thus by Lemma 4.2.3 after running exponential mechanism, we have with probability at least  $1 - \beta$ ,

$$\hat{L}^r(w^{\text{priv}}, D) - \min_{w \in \mathcal{N}_\alpha} \hat{L}^r(w, D) \leq O\left(\frac{d \ln \frac{1}{\alpha\beta}}{n\epsilon}\right).$$

Thus, from the above and taking  $\alpha = \frac{d}{n\epsilon}$ , we have

$$\hat{L}^r(w^{\text{priv}}, D) - \hat{L}^r(w^*, D) \leq \tilde{O}\left(\frac{d}{n\epsilon}\right).$$

Actually, this is the lower bound for ERM under general convex functions with the constrained set  $\mathcal{C} = \mathcal{B}^d(r)$  under  $\epsilon$  differential privacy, see Theorem 5.2 in [29]. By this, we can easily get a lower bound for non-convex loss functions under our assumptions. We thus have the following theorem:

**Theorem 4.2.12.** Consider DP-ERM problem with  $\hat{L}^r(w, D) = \frac{1}{n} \sum_{i=1}^n \ell(w, z_i) + r(w)$ , where  $r(w) = \frac{\lambda}{2} \|w\|^2$ ,  $\ell(w, z) = -\langle w, z \rangle - \frac{\lambda}{2} \|w\|^2$ ,  $\mathcal{C} = \mathbb{B}^d(r)$  for some constant  $r$ . Then for every  $\epsilon$ -differentially private algorithm, there is a dataset  $D = \{z_1, \dots, z_n\} \subseteq \{-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\}^d$  such that, with probability at least  $1/2$ , we must have:

$$\hat{L}^r(w^{\text{priv}}, D) - \min_{w \in \mathcal{C}} \hat{L}^r(w, D) \geq \Omega\left(\min\left\{1, \frac{d}{n\epsilon}\right\}\right).$$

On the other hand, under our assumptions about  $\mathcal{C} = \mathbb{B}^d(r)$ , there is an  $\epsilon$ -differentially

private algorithm, whose output  $w^{\text{priv}}$  satisfies with probability at least  $1 - \beta$ ,

$$\hat{L}^r(w^{\text{priv}}, D) - \hat{L}^r(w^*, D) \leq \tilde{O}\left(\frac{d}{n\epsilon}\right).$$

The time complexity is  $O((1 + \frac{2Ln\epsilon}{d\lambda})^d n)$ .

Thus we can get an near optimal bound for general non-convex loss functions under  $\epsilon$ -differential privacy.

### Proof of Theorem 4.2.5

Before showing the proof, we first give an upper bound on the Frank-Wolfe gap of the output in Algorithm 4.2.23 for general smooth and Lipschitz loss functions with general convex set  $\mathcal{C}$ . We start with the definition of Gaussian Width:

**Definition 4.2.2** (Minkowski Norm). The Minkowski norm (denoted by  $\|\cdot\|_{\mathcal{C}}$ ) with respect to a centrally symmetric convex set  $\mathcal{C} \subseteq \mathbb{R}^d$  is defined as follows. For any vector  $v \in \mathbb{R}^d$ ,  $\|\cdot\|_{\mathcal{C}} = \min\{r \in \mathbb{R}^+ : v \in r\mathcal{C}\}$ . The dual norm of  $\|\cdot\|_{\mathcal{C}}$  is denoted as  $\|\cdot\|_{\mathcal{C}^*}$ ; for any vector  $v \in \mathbb{R}^d$ ,  $\|v\|_{\mathcal{C}^*} = \max_{w \in \mathcal{C}} |\langle w, v \rangle|$ .

**Definition 4.2.3** (Gaussian Width). Let  $b \sim \mathcal{N}(0, I_d)$  be a Gaussian random vector in  $\mathbb{R}^d$ .

The Gaussian width for a set  $\mathcal{C}$  is defined as  $G_{\mathcal{C}} = \mathbb{E}_b[\sup_{w \in \mathcal{C}} \langle b, w \rangle]$ .

### Algorithm 4.2.23 DP-FW-L2

**Input:**  $T$  is the maximum of iterations,  $w_1$  is the initial point, and  $\{\gamma_t\}_{t=1}^T$  is the step size.  $\epsilon$  and  $\delta$  are privacy parameters.

```

for  $t = 1, \dots, T$  do
    Compute  $v_t = \arg \max_{v \in \mathcal{C}} \langle v, -(\nabla L(w_t, D) + \epsilon_t) \rangle$ , where  $\epsilon_t \sim N(0, \sigma^2 I_d)$ .
     $w_{t+1} = w_t + \gamma_t(v_t - w_t)$ .
end for
```

Return  $w_R \in \{w_1, \dots, w_T\}$  such that  $R$  is uniformly sampled from  $\{1, \dots, T\}$ .

---

**Theorem 4.2.13.** Let  $\mathcal{C}$  be a bounded, closed, centrally symmetric convex set. Assume that  $\hat{L}(w, D)$  is differentiable and  $M$ -smooth over  $w$  with respect to  $\ell_2$  norm, and the loss function  $\ell(\cdot, z)$  is  $L$ -Lipschitz over  $x$  with respect to  $\ell_2$ -norm for all  $z \in \mathcal{Z}$ . Then, there are constants  $c_1, c_2 > 0$  such that for any  $0 < \epsilon < c_1 T, 0 < \delta < 1$ , **DP-FW-L2** (Algorithm 4.2.23) is  $(\epsilon, \delta)$ -differentially private if  $\sigma^2 = c_2 \frac{L^2 T \ln(\frac{1}{\delta})}{n^2 \epsilon^2}$ . Moreover, if take  $\{\gamma_t\}_{t=1}^T = O\left(\frac{\sqrt{(\|\mathcal{C}\|_2^2 + G_{\mathcal{C}}^2) \ln \frac{1}{\delta}}}{\|\mathcal{C}\|_2 \sqrt{n\epsilon}}\right)$  and  $T = O\left(\frac{n\epsilon}{\sqrt{(\|\mathcal{C}\|_2^2 + G_{\mathcal{C}}^2) \ln \frac{1}{\delta}}}\right)$ , the following holds,

$$\mathbb{E}[\mathcal{G}_R] \leq O\left(\frac{\|\mathcal{C}\|_2 \sqrt[4]{(\|\mathcal{C}\|_2^2 + G_{\mathcal{C}}^2) \ln \frac{1}{\delta}}}{\sqrt{n\epsilon}}\right), \quad (4.47)$$

where  $\mathcal{G}_t = \max_{v \in \mathcal{C}} \langle -\nabla \hat{L}(w_t, D), v - w_t \rangle$ .

*Proof.* To prove Theorem 4.2.13, we need the following lemmas.

**Lemma 4.2.4.** For any vector  $v$ , we have  $\|v\|_2 \leq \|\mathcal{C}\|_2 \|v\|_{\mathcal{C}}$ , where  $\|\mathcal{C}\|_2$  is the  $\ell_2$ -diameter and  $\|\mathcal{C}\|_2 = \sup_{x,y \in \mathcal{C}} \|x - y\|_2$ .

Lemma 4.2.4 implies that any smooth convex function  $F(\theta)$ , which is  $M$ -smooth with respect to  $\ell_2$  norm, is  $M\|\mathcal{C}\|_2^2$ -smooth with respect to  $\|\cdot\|_{\mathcal{C}}$  norm, which is the motivation of our algorithm.

*Proof.* If  $v = 0$ , this is trivially true. Otherwise, we will show that  $\frac{\|v\|_2}{\|\mathcal{C}\|_2} \leq \|v\|_{\mathcal{C}}$ . This is equivalent to show that  $v \notin \frac{\|v\|_2}{\|\mathcal{C}\|_2} \mathcal{C}$ . Taking any  $y \in \mathcal{C}$ , since  $\|\frac{\|v\|_2}{\|\mathcal{C}\|_2} y\|_2 = \frac{\|v\|_2}{\|\mathcal{C}\|_2} \|y\|_2$ , we know that  $\|y\|_2 < \|\mathcal{C}\|_2$ . Thus,  $\|\frac{\|v\|_2}{\|\mathcal{C}\|_2} y\|_2 < \|v\|_2$ . We have  $v \notin \frac{\|v\|_2}{\|\mathcal{C}\|_2} \mathcal{C}$ .  $\square$

**Proof of Theorem 4.2.13.** For convenience, we let the norm  $\|\cdot\| = \|\cdot\|_{\mathcal{C}}$ , and  $F(w) = \hat{L}(w, D)$ . Let  $\tilde{M}$  denote  $M\|\mathcal{C}\|_2^2$ , and  $D$  denote the diameter of  $\mathcal{C}$  w.r.t.  $\|\cdot\|$  norm. By the  $M$ -smoothness property and Lemma 4.2.4, we have

$$F(w_{t+1}) \leq F(w_t) + \gamma_t \langle \nabla F(w_t), v_t - w_t \rangle + \frac{\tilde{M} \gamma_t^2}{2} \|v_t - w_t\|^2. \quad (4.48)$$

Let  $\hat{v}_t = \arg \max_{v \in \mathcal{C}} \langle v, -\nabla F(w_t) \rangle$ . By the optimality of  $v_t$ , we have

$$\langle v_t, -\nabla F(w_t) - \epsilon_t \rangle \geq \langle \hat{v}_t, -\nabla F(w_t) - \epsilon_t \rangle.$$

This implies that

$$\langle v_t - \hat{v}_t, \nabla F(w_t) \rangle \leq \langle v_t - \hat{v}_t, -\epsilon_t \rangle. \quad (4.49)$$

From (4.48), we get

$$F(w_{t+1}) \leq F(w_t) + \gamma_t \langle \nabla F(w_t), v_t - \hat{v}_t \rangle + \gamma_t \langle \nabla F(w_t), \hat{v}_t - w_t \rangle + \frac{\gamma_t^2 \tilde{M}}{2} \|v_t - w_t\|^2.$$

Plugging (4.49) into (4.48) and by the fact that  $\langle \nabla F(w_t), \hat{v}_t - w_t \rangle = -\mathcal{G}_t$  (from the definition of  $\hat{v}_t$ ), we obtain

$$\begin{aligned} \gamma_t \mathcal{G}_t &\leq F(w_t) - F(w_{t+1}) + \gamma_t \langle v_t - \hat{v}_t, -\epsilon_t \rangle + \frac{\tilde{M} \gamma_t^2}{2} D^2 \\ &\leq F(w_t) - F(w_{t+1}) + \frac{\gamma_t^2 \tilde{M} \|v_t - \hat{v}_t\|^2}{2} + \frac{\|\epsilon_t\|_*^2}{2\tilde{M}} + \frac{\tilde{M} \gamma_t^2}{2} D^2 \\ &\leq F(w_t) - F(w_{t+1}) + \frac{\|\epsilon_t\|_*^2}{2\tilde{M}} + \tilde{M} \gamma_t^2 D^2, \end{aligned}$$

where the second inequality is due to Cauchy Inequality. By the definition of  $\mathcal{G}_R$ , we have  $\mathbb{E}[\mathcal{G}_R] = \frac{1}{T} \sum_{t=1}^T \mathbb{E}[\mathcal{G}_t]$ . Since  $\{\gamma_t\}_{t=1}^T = \gamma$ , summing the above over  $t = 1 \dots, T$  and taking the expectation, we have

$$\mathbb{E}\mathcal{G}_R \leq \frac{F(w_1) - F(w^*)}{\gamma T} + \tilde{M} \gamma D^2 + \frac{1}{\gamma} O((\|\mathcal{C}\|_2^2 + G_{\mathcal{C}}^2) \frac{L^2 T \ln(\frac{1}{\delta})}{n^2 \epsilon^2}).$$

Taking  $\gamma = O(\frac{\sqrt{G^2(\|\mathcal{C}\|_2^2 + G_{\mathcal{C}}^2) \ln \frac{1}{\delta}}}{\sqrt{\tilde{M}D\sqrt{n\epsilon}}})$  and  $T = O(\frac{n\epsilon}{\sqrt{(\|\mathcal{C}\|_2^2 + G_{\mathcal{C}}^2)L^2 \ln \frac{1}{\delta}}})$ , and by definition of  $\|\cdot\|$  and the fact that  $D \leq O(1)$ , we have the proof.  $\square$

We first consider the Generalized Linear Model. The following inequality has been proved in [115]. We rephrase it here to make the proof self-complete. Denote by  $L_{\mathcal{P}}(w) =$

$\mathbb{E}_{(x,y) \sim \mathcal{Z}} \ell(w; x, y)$  and  $\hat{L}(w, D) = \frac{1}{n} \sum_{i=1}^n \ell(w; x_i, y_i)$ .

### Generalized Linear Model

**Lemma 4.2.5.** For a fixed  $w$ ,  $L_{\mathcal{P}}(w) - L_{\mathcal{P}}(w^*) \leq \frac{C_{\sigma}}{2c_{\sigma}} \langle \nabla L_{\mathcal{P}}(w), w - w^* \rangle$ .

*Proof.* Let  $w \in \mathcal{C}$  be fixed. Then, we have

$$\begin{aligned} \langle \nabla L_{\mathcal{P}}(w), w - w^* \rangle &= 2\mathbb{E}_{(x,y)}[\sigma(\langle w, x \rangle) - y)\sigma'(\langle w, x \rangle)\langle w - w^*, x \rangle] \\ &= 2\mathbb{E}_x[(\sigma(\langle w, x \rangle) - \sigma(\langle w^*, x \rangle))\sigma'(\langle w, x \rangle)\langle w - w^*, x \rangle]. \end{aligned}$$

By Assumption 4.2.2, we have

$$\begin{aligned} L_{\mathcal{P}}(w) - L_{\mathcal{P}}(w^*) &= \mathbb{E}(\sigma(\langle w, x \rangle) - \sigma(\langle w^*, x \rangle))^2 \\ &\leq \frac{C_{\sigma}}{2c_{\sigma}} \langle \nabla L_{\mathcal{P}}(w), w - w^* \rangle. \end{aligned}$$

□

By Lemma 4.2.5 and Theorem 4.2.13, we only need to bound  $\langle \nabla L_{\mathcal{P}}(w), w - w^* \rangle$ .

Before doing that, we show that the empirical risk is Lipschitz and smooth, which satisfies the assumption in Theorem 4.2.13. It is due to:

$$\|\nabla \hat{L}(w, D)\|_2 = \left\| \frac{1}{n} \sum_{i=1}^n (\sigma(\langle w, x_i \rangle) - y_i) \sigma'(\langle w, x \rangle) x_i^T \right\|_2 \leq C_{\sigma}(B+1), \quad (4.50)$$

and

$$\begin{aligned} \|\nabla^2 \hat{L}(w, D)\|_2 &= \left\| \frac{1}{n} \sum_{i=1}^n [(\sigma'(\langle w, x_i \rangle))^2 + \sigma''(\langle w, x_i \rangle)(\sigma(\langle w, x_i \rangle) - y_i)] x_i x_i^T \right\|_2 \\ &\leq C_{\sigma}^2 + C_{\sigma}(B+1). \end{aligned}$$

Thus,  $\hat{L}(w, D)$  is  $(C_{\sigma}(B+1))$ -Lipschitz and  $C_{\sigma}^2 + C_{\sigma}(B+1)$ -smooth. Also, since  $\mathcal{C}$  is the

unit  $\ell_2$  norm, we have  $G_{\mathcal{C}} = O(\sqrt{d})$  and  $\|\mathcal{C}\|_2 = 1$ . Thus, we get  $\mathbb{E}[\mathcal{G}_R] \leq O\left(\frac{\sqrt[4]{d \ln \frac{1}{\delta}}}{\sqrt{n\epsilon}}\right)$  by Theorem 4.2.13.

By the definition of  $\mathcal{G}_R$ , we know that  $\mathbb{E}[\mathcal{G}_R] \geq \mathbb{E}\langle \nabla \hat{L}(w_R, D), w_R - w^* \rangle$ . Taking the expectation w.r.t  $\{(x_1, y_1), \dots, (x_n, y_n)\}$ , we then have  $\mathbb{E}[\mathcal{G}_R] \geq \mathbb{E}\langle \nabla L_{\mathcal{P}}(w_R), w_R - w^* \rangle$ . Combing it with Lemma 4.2.5, we get

$$\mathbb{E}L_{\mathcal{P}}(x_R) - L_{\mathcal{P}}(w^*) \leq O\left(\frac{C_\sigma}{2c_\sigma} \frac{\sqrt[4]{d \ln \frac{1}{\delta}}}{\sqrt{n\epsilon}}\right).$$

**Robust Regression** We now consider robust regression. We begin with showing a similar result as in Lemma 4.2.5. First, the smoothness of  $\psi$  implies that for any  $s, s^* \in \mathcal{S}$ , we have

$$\psi(s) - \psi(s^*) \leq \psi'(s^*)(s - s^*) + \frac{C_\psi}{2}(s - s^*)^2.$$

Taking  $s = \langle w, x \rangle$  and  $s^* = \langle w^*, x \rangle$ , and then taking expectation w.r.t.  $\{(x_1, y_1), \dots, (x_n, y_n)\}$ , we get

$$\begin{aligned} L_{\mathcal{P}}(w) - L_{\mathcal{P}}(w^*) &\leq \mathbb{E}_{x,y}[\psi'(\langle w^*, x \rangle - y)\langle w - w^*, x \rangle] + \frac{C_\psi}{2}\mathbb{E}\langle w - w^*, x \rangle^2 \\ &= \langle \nabla L_{\mathcal{P}}(w^*), w - w^* \rangle + \frac{C_\psi}{2}\mathbb{E}\langle w - w^*, x \rangle^2. \end{aligned} \quad (4.51)$$

By Assumption 4.2.3, we have

$$\nabla L_{\mathcal{P}}(w^*) = \mathbb{E}_{x,\xi}[\psi'(-\xi)x] = 0.$$

Thus, we get  $L_{\mathcal{P}}(w) - L_{\mathcal{P}}(w^*) \leq \frac{C_\psi}{2}\mathbb{E}\langle w - w^*, x \rangle^2$ . On the other hand, using gradient we have

$$\begin{aligned} \langle \nabla L_{\mathcal{P}}(w), w - w^* \rangle &= \mathbb{E}_x[\mathbb{E}_\xi \psi'(\langle w - w^*, x \rangle - \xi)\langle w - w^*, x \rangle] \\ &= \mathbb{E}_x[h(\langle w - w^*, x \rangle)\langle w - w^*, x \rangle]. \end{aligned}$$

By the assumption on function  $h(\cdot)$ , we get

$$h(\langle w - w^*, x \rangle) \langle w - w^*, x \rangle = \frac{h(\langle w - w^*, x \rangle)}{\langle w - w^*, x \rangle} \langle w - w^*, x \rangle^2 \geq c_\psi \langle w - w^*, x \rangle^2,$$

where the inequality is due to the fact that  $h(0) = 0$  and  $h'(0) \geq c_\psi$ .

Taking the expectation, we have

$$\langle \nabla L_{\mathcal{P}}(w), w - w^* \rangle \geq c_\psi \mathbb{E}_x \langle w - w^*, x \rangle^2.$$

Thus, we have the following lemma.

**Lemma 4.2.6.**

$$L_{\mathcal{P}}(w) - L_{\mathcal{P}}(w^*) \leq \frac{C_\psi}{2c_\psi} \langle \nabla L_{\mathcal{P}}(w), w - w^* \rangle.$$

It is easily to get that the loss function  $\ell(w, (x, y)) = \psi(\langle w, x \rangle - y)$  is  $C_\psi$ -Lipschitz and  $C_\psi$ -smooth. Using the same argument as in the proof for the case of Generalized Linear model, we get the proof.  $\square$

#### 4.2.4 Proof of Theorem 4.2.6

We first give an upper bound on the Frank-Wolfe gap of general  $\ell_1$ -norm Lipschitz and smooth loss functions.

**Definition 4.2.4.** The loss function  $\ell$  is L-Lipschitz under  $\ell_1$ -norm over  $w$ , if for any  $z \in \mathcal{Z}$  and  $w_1, w_2 \in \mathcal{C}$ ,  $|\ell(w_1, z) - \ell(w_2, z)| \leq L \|w_1 - w_2\|_1$  holds.

**Definition 4.2.5.** A loss function  $\ell : \mathcal{C} \times \mathcal{Z} \mapsto \mathbb{R}$  is M-smooth over  $w$  with respect to the  $\|\cdot\|_1$  norm if for any  $z \in \mathcal{Z}$  and  $w_1, w_2 \in \mathcal{C}$ , the following holds

$$\|\nabla \ell(w_1, z) - \nabla \ell(w_2, z)\|_\infty \leq M \|w_1 - w_2\|_1.$$

If  $f$  is differentiable, this yields  $\ell(w_1, z) \leq \ell(w_2, z) + \langle \nabla \ell(w_2, z), w_1 - w_2 \rangle + \frac{M}{2} \|w_1 - w_2\|_1^2$ .

**Assumption 4.2.4.**  $\hat{L}(w, D)$  is assumed to be differentiable and  $M$ -smooth over  $x$  w.r.t  $\ell_1$ -norm, and  $\ell(\cdot, z)$  is assumed to be  $L$ -Lipschitz over  $x$  with respect to  $\ell_1$ -norm for all  $z \in \mathcal{X}$ .  $\mathcal{C} \subseteq \mathbb{R}^d$  is assumed to be a closed convex set. Furthermore,  $\mathcal{C}$  is assumed to be the convex hull of some finite set  $A$ , i.e.,  $\mathcal{C} = \text{Conv}(A)$  and bounded. (For example,  $\mathcal{C}$  could be a polytope.)

---

**Algorithm 4.2.24 DP-FW-L1**


---

**Input:**  $T$  is the iteration number and  $x_1$  is the initial point.  $\{\gamma_t\}_{t=1}^T$  is the step size.  $\mathcal{C} \subseteq \mathbb{R}^p$  is the convex hull of a compact set  $A \subseteq \mathbb{R}^d$ .  $\epsilon$  and  $\delta$  are privacy parameters.

- 1: **for**  $t = 1, \dots, T$  **do**
  - 2:   Use exponential mechanism  $\mathcal{M}(D, u, \mathcal{R})$ , where  $\mathcal{R} = A$ ,  $u(D, s) = -\langle s, \nabla \hat{L}(w_t, D) \rangle$ , to ensure  $(\frac{\epsilon}{\sqrt{8T \ln(\frac{1}{\delta})}}, 0)$ -differential privacy. Denote the output as  $\tilde{w}_t$ .
  - 3:   Compute  $w_{t+1} = (1 - \gamma_t)w_t + \gamma_t \tilde{w}_t$ .
  - 4: **end for**
  - 5:
  - 6: Return  $w_R \in \{w_1, \dots, w_T\}$ , where  $R$  is uniformly sampled from  $\{1, 2, \dots, T\}$ .
- 

**Theorem 4.2.14.** Under Assumption 4.2.4 and assuming that  $A$  is a finite set, then for any  $\epsilon, \delta > 0$ , **DP-FW-L1** (Algorithm 4.2.24) ensures  $(\epsilon, \delta)$ -differentially private. Furthermore, if set  $T = O(\frac{n\epsilon}{\sqrt{\ln(\frac{1}{\delta}) \ln(|A|n/\eta)}})$  and  $\{\gamma_t\}_{t=1}^T = \sqrt{\frac{2}{MT\|\mathcal{C}\|_1^2}}$ , then with probability at least  $1 - \eta$ , the following holds

$$\mathbb{E}[\mathcal{G}_R] \leq O\left(\frac{\|\mathcal{C}\|_1 \sqrt{\ln(\frac{1}{\delta})} \sqrt{\ln \frac{n|A|}{\eta}}}{\sqrt{n\epsilon}}\right), \quad (4.52)$$

where  $\mathcal{G}_t = \max_{v \in \mathcal{C}} \langle -\nabla \hat{L}(w_t, D), v - w_t \rangle$ .

**Proof of Theorem 4.2.14.** For convenience, we let  $F(w) = \hat{L}(w, D)$ . By exponential mechanism and advanced composition theorem, we can see that it is  $(\epsilon, \delta)$ -differentially private. By the  $L$ -Lipschitz (w.r.t  $\ell_1$ -norm) property of the loss function, we know that  $\Delta u \leq O(\frac{\|\mathcal{C}\|_1 L}{n})$ . Let  $\beta = O(\frac{L\|\mathcal{C}\|_1 \sqrt{8T \ln(\frac{1}{\delta})} \ln(\frac{|A|T}{\eta})}{n\epsilon})$ . By the utility bound of exponential mechanism (Lemma 4.2.3), we know that in each iteration, with probability  $1 - \frac{\eta}{T}$ , the following holds

$$\langle \tilde{w}_t, \nabla F(w_t) \rangle \leq \min_{v \in A} \langle v, \nabla F(w_t) \rangle + \beta. \quad (4.53)$$

Let  $s_t = \arg \min_{u \in \mathcal{A}} \langle u, \nabla F(w_t) \rangle$ . By the  $M$ -smooth property and (4.53), we have

$$\begin{aligned} \frac{M}{2} \|w_{t+1} - w_t\|_1^2 &\geq F(w_{t+1}) - F(w_t) - \langle F(w_t), w_{t+1} - w_t \rangle \\ &= F(w_{t+1}) - F(w_t) - \gamma_t \langle \nabla F(w_t), \tilde{w}_t - w_t \rangle \\ &\geq F(w_{t+1}) - F(w_t) - \gamma_t (\langle \nabla F(w_t), s_t - w_t \rangle + \beta). \end{aligned}$$

Note that  $\min_{u \in \mathcal{C}} \langle u - w_t, \nabla F(w_t) \rangle = \min_{u \in \mathcal{A}} \langle u - w_t, \nabla F(w_t) \rangle = \langle s_t - w_t, \nabla F(w_t) \rangle = -\mathcal{G}_t$ . Thus, we have

$$F(w_{t+1}) - F(w_t) + \gamma_t \mathcal{G}_t \leq \gamma_t \beta + \frac{M \gamma_t^2}{2} \|\mathcal{C}\|_1^2. \quad (4.54)$$

Summing over  $t = 1, \dots, T$ , we get with probability  $1 - \eta$ ,

$$\left( \sum_{t=1}^T \gamma_t \right) \mathcal{G}_R \leq F(w_1) - F(w^*) + \left( \sum_{t=1}^T \gamma_t \right) \beta + \frac{M}{2} \left( \sum_{t=1}^T \gamma_t^2 \right) \|\mathcal{C}\|_1^2.$$

Taking  $\{\gamma_t\}_{t=1}^T = \gamma$ , we have

$$\mathcal{G}_R \leq \frac{F(w_1) - F(w^*)}{\gamma T} + \frac{\gamma \|\mathcal{C}\|_1^2 M}{2} + O\left(\frac{L \|\mathcal{C}\|_1 \sqrt{T \ln(\frac{1}{\delta})} \ln(\frac{|A|T}{n})}{n\epsilon}\right).$$

Taking  $T = O\left(\frac{n\epsilon}{L\sqrt{\ln(\frac{1}{\delta}) \ln(|A|n)}}\right)$  and  $\gamma = \sqrt{\frac{2}{T\|\mathcal{C}\|_1^2 M}}$ , we get the result.  $\square$

**Generalized Linear Model** We first show the Lipschitz and Smooth properties w.r.t  $\ell_1$ -norm. Since  $\nabla \ell(w, x, y) = \sigma(\langle w, x \rangle - y) \sigma'(\langle w, x \rangle) x^T$ , by the Lipschitzness and the assumption, we have

$$\|(\sigma(\langle w, x \rangle) - y) \sigma'(\langle w, x \rangle) x^T\|_\infty \leq C_\sigma (B + 1).$$

Let  $w_1, w_2 \in \mathcal{C}$ , we have

$$\begin{aligned}
& \|(\sigma(\langle w_1, x \rangle) - y)\sigma'(\langle w_1, x \rangle)x^T - (\sigma(\langle w_2, x \rangle) - y)\sigma'(\langle w_2, x \rangle)x^T\|_\infty \\
& \leq |(\sigma(\langle w_1, x \rangle) - y)\sigma'(\langle w_1, x \rangle) - (\sigma(\langle w_2, x \rangle) - y)\sigma'(\langle w_2, x \rangle)| \\
& \leq |\sigma(\langle w_1, x \rangle)\sigma'(\langle w_1, x \rangle) - \sigma(\langle w_2, x \rangle)\sigma'(\langle w_2, x \rangle)| + |\sigma'(\langle w_1, x \rangle) - \sigma'(\langle w_2, x \rangle)| \\
& \leq (C_\sigma^2 + (B+1)C_\sigma)|\langle w_1 - w_2, x \rangle| \\
& \leq (C_\sigma^2 + (B+1)C_\sigma)\|w_1 - w_2\|_1.
\end{aligned}$$

Thus, by Theorem 4.2.14, we know  $\mathbb{E}[\mathcal{G}_R] \leq O\left(\frac{\sqrt[4]{\ln(\frac{1}{\delta})}\sqrt{\ln \frac{np}{\eta}}}{\sqrt{ne}}\right)$ . The remaining part of the proof is by Lemma 4.2.5 and is the same as in the proof of Theorem 4.2.5.

**Robust Regression** For the case of linear regression, it is almost the same as in the case of generalized linear model, we omit it here.

### 4.3 Local Minimum/Second Order Stationary View

In Chapter 4.1 we study using first order stationary measurement to measure the error of private estimation, despite some obvious advantages with such an approach (such as the sample complexity is relatively low), it also endows an oblivious limitation: although [356, 328, 309] showed that the gradient norm tends to 0 as  $n$  goes to infinity, there is no guarantee that such an estimator will be close to any non-degenerate local minimum [5]. To solve this issue, in Chapter 4.2 we study the using the global error measurement (*i.e.*, the empirical (population) risk) to measure the private estimator. However, we showed that the sample complexity may be exponential to the dimensionality. Thus, our question, is there any other measurement which could guarantee that our private estimator close to some local minimum while also keep the sample complexity to be small? In this section, we will provide an affirm answer.

Recent research on deep neural network training [125, 179] and many other machine

learning problems [126, 124, 123] has shifted their attentions to obtaining local minima. It has been shown that fast convergence to a local minimum is actually sufficient for such tasks, but convergence to critical points (*i.e.*, points with vanished gradients) is often not acceptable. This motivates us to investigate efficient techniques for finding local minima. However, as shown in [14], computing a local minimum could be quite challenging as it is actually NP-hard for non-convex functions. Fortunately, many non-convex functions in machine learning are known to be strict saddle [126], meaning that a second-order stationary point (or approximate local minimum) is sufficient to obtain a close enough point to some local minimum.

To find (approximate) local minima, [126] have recently proposed an elegant approach using a noisy version of gradient descent. Their method adds some scaled Gaussian noise in each iteration to the gradient before updating, rather than directly using SGD. Such a way of finding local minima resembles the idea used by the DP community for achieving differential privacy for SGD [29, 328, 309]. In DP-SGD, some Gaussian noise is also added to the gradient in each iteration to make it  $(\epsilon, \delta)$ -DP. Although these two algorithms focus on different perspectives (one for escaping saddle points while the other for making the algorithm DP), they both inject random Gaussian noise to the gradients in each iteration. This naturally leads us to another question:

**Can we find some approximate local minimum which escapes saddle points, while keeping the algorithm  $(\epsilon, \delta)$ -differentially private?**

In this section, We first show that when the data size  $n$  is large enough, there exist polynomial-time<sup>5</sup>  $(\epsilon, \delta)$ -DP algorithms that can find an  $\alpha$ -approximate local minimum of the empirical risk in both constrained and non-constrained settings. To the best of our knowledge, this is the first result that reveals a connection between differential privacy and saddle-point escaping.

However, this method has several issues, which hamper its applications in big data.

---

<sup>5</sup>For the constrained case, polynomial-time solutions are only for some specified sets, see Remark 4.3.1 for details.

Firstly, the sample complexity (or equivalently error bound) is relatively high. It is not clear whether it can be improved. Secondly, this method needs to calculate the gradient and Hessian matrix of the whole objective function in each iteration, which is prohibitive in large scale datasets.

To address the aforementioned theoretical and practical issues, we then propose a new method called Differentially Private Trust Region (DP-TR) which is capable of escaping saddle points privately. Particularly, we first show that our algorithm can output an  $\alpha$ -SOSP with high probability and less sample complexity. To make our method scalable, we then present a stochastic version of DP-TR called Differentially Private Stochastic Trust Region (DP-STR) with the same functionality. We show that DP-STR is much faster and has asymptotically the same sample complexity as DP-TR. Finally, we provide comprehensive experimental studies on the practical performance of our methods in escaping saddle point under differential privacy model. We first impose the following assumption on the loss function considered in this section.

**Assumption 4.3.1.** The loss function is  $L$ -Lipschitz,  $M$ -smooth and  $\rho$ -Hessian Lipschitz. We further assume that the empirical risk  $L(w, D)$  is bounded by a constant  $B$ <sup>6</sup>. If  $\mathcal{C}$  is closed, we denote the diameter of  $\mathcal{C}$  as  $D = \max_{x, x' \in \mathcal{C}} \|x - x'\|_2$ .

### 4.3.1 Finding Approximate Local Minimum Privately Using DP-GD

#### Unconstrained Case

**Definition 4.3.1.**  $w$  is called a second-order stationary point (SOSP) of a twice differentiable function  $F$  if

$$\|\nabla F(w)\|_2 = 0 \text{ and } \lambda_{\min}(\nabla^2 F(w)) \geq 0 ,$$

where  $\lambda_{\min}$  denotes its smallest eigenvalue.

---

<sup>6</sup>Note that if the empirical risk is not bounded, we can still use the same proof after replacing the constant by the term  $L(w_1, D) - L(w^*, D)$ . We make such an assumption for convenience.

Since it is extremely challenging to find an exact SOSP [126], we turn to its approximation. The following defintion of  $\alpha$ -approximate SOSP relaxes the first- and second-order optimality conditions.

**Definition 4.3.2** ([5]).  $w$  is an  $\alpha$ -second-order stationary point ( $\alpha$ -SOSP) or  $\alpha$ -approximate local minimum of a twice differentiable function  $F$ , if <sup>7</sup>

$$\|\nabla F(w)\|_2 \leq \alpha \text{ and } \lambda_{\min}(\nabla^2 F(w)) \geq -\sqrt{\rho\alpha}. \quad (4.55)$$

**Definition 4.3.3** (DP-SOSP). Given  $\alpha, \epsilon, \delta > 0$ , DP-SOSP is to identify the smallest sample complexity  $n(\alpha, p, \epsilon, \delta)$  such that when  $n \geq n(\alpha, p, \epsilon, \delta)$ , for any dataset  $D$  of size  $n$ , there is an  $(\epsilon, \delta)$ -DP algorithm which outputs an  $\alpha$ -SOSP of the empirical risk function  $L(w, D)$  with high probability.

To find an  $\alpha$ -SOSP privately, we present Algorithm 4.3.25. Comparing with the first-order noisy gradient descent methods, such as those in [126, 163, 349, 164], the main difference is that the noises added should be in the scale of  $O(\frac{\sqrt{T}}{n\epsilon})$ , which depends on the iteration number  $T$ . This dependency makes Algorithm 4.3.25 more complex than previous related algorithms.

---

#### Algorithm 4.3.25 DP-GD

**Input:**  $T$  is the iteration number and  $w_1$  is the initial point.  $\{\gamma_t\}_{t=1}^T$  is the step size.  $\epsilon$  and  $\delta$  are privacy parameters.

- 1: **for**  $t = 1, \dots, T$  **do**
  - 2:     Compute  $w_{t+1} = w_t - \eta_t(\nabla L(w_t, D) + \epsilon_t)$ , where  $\epsilon_t \sim N(0, \sigma^2 I_d)$  for some  $\sigma$ .
  - 3: **end for**
  - 4:
  - 5: **Return**  $\{w_1, \dots, w_{T+1}\}$ .
- 

To prove that Algorithm 4.3.25 has the ability of escaping saddle points, we first show that the iteration number satisfies  $T = \tilde{O}(\frac{MB}{\alpha^2})$  when the magnitude of the noise is small

---

<sup>7</sup>This is a special version of  $(\epsilon, \gamma)$ -SOSP [126]. Our results can be easily extended to the general definition. The same applies to the constrained case.

enough (*i.e.*, when  $n$  is large enough). Based on this fact, we then prove that Algorithm 4.3.25 can find an  $\alpha$ -SOSP with high probability. Our results are summarized in the following theorem.

**Theorem 4.3.1.** Under Assumption 4.3.1, there exist constants  $c_1, c_2$ , such that for any  $0 < \epsilon < c_1 T$ , Algorithm 4.3.25 is  $(\epsilon, \delta)$ -DP if  $\sigma^2 = c_2 \frac{L^2 \log \frac{1}{\delta} T}{n^2 \epsilon^2}$ . Moreover, if the data size  $n$  is large enough such that

$$n \geq \tilde{\Omega}\left(\frac{\sqrt{MB} \sqrt{\log \frac{1}{\delta} d} \log \frac{1}{\xi} L}{\epsilon \alpha^2}\right), \quad (4.56)$$

and choose  $T = \tilde{O}\left(\frac{MB}{\alpha^2}\right)$ ,  $\{\eta_t\}_{t=1}^T = \frac{1}{M}$ , then with probability  $1 - \zeta$ , one of the outputs is an  $\alpha$ -SOSP of the empirical risk  $L(\cdot, D)$ . Here the  $\tilde{O}$  and  $\tilde{\Omega}$  terms omit other log factors.

Recently, [309, 328] show that there are  $(\epsilon, \delta)$ -DP algorithms satisfying  $\|\nabla L(w^{\text{priv}}, D)\|_2 \leq O\left(\frac{4\sqrt{d \log \frac{1}{\delta}}}{\sqrt{n\epsilon}}\right)$ . Thus, to achieve an  $\epsilon$ -first-order stationary point, the size  $n$  should satisfy the condition of  $n \geq \Omega\left(\frac{\sqrt{d \log \frac{1}{\delta}}}{\epsilon \alpha^2}\right)$ . Comparing to the sample complexity in (4.56) for  $\epsilon$ -SOSP, we can see that they are actually asymptotically almost the same (up to some log factors).

Theorem 4.3.1 ensures the existence of an approximate SOSP among  $\{w_1, \dots, w_{T+1}\}$ . To find such a SOSP with high probability, we propose Algorithm 4.3.26, which incurs an additional  $O(\sqrt{d})$  factor in the sample size  $n$  in (4.56).

**Theorem 4.3.2.** There exist constants  $c_1, c_2$  such that when  $\sigma_1^2 = c_1 \frac{\log \frac{1}{\delta} TL^2}{n^2 \epsilon^2}$  and  $\sigma_2^2 = c_2 \frac{\log \frac{1}{\delta} M^2 d T}{n^2 \epsilon^2}$ , Algorithm 4.3.26 is  $(\epsilon, \delta)$ -DP. Furthermore, with probability at least  $1 - \xi - \frac{T}{p^C}$  for some sufficiently large  $C > 0$ , the output is an  $\alpha$ -SOSP when the sample size satisfies

$$n \geq \tilde{\Omega}\left(\frac{Md\sqrt{MB} \sqrt{\log \frac{1}{\delta} \log \frac{1}{\xi} L}}{\rho \epsilon \alpha^2}\right).$$

---

**Algorithm 4.3.26** Selecting SOSP

---

- 1: Run Algorithm 4.3.25 to ensure  $(\frac{\epsilon}{2}, \frac{\delta}{2})$ -differential privacy on finding an  $\frac{\alpha}{2}$ -SOSP with probability at least  $1 - \frac{\xi}{2}$ . Let the output be  $\{w_1, \dots, w_{T+1}\}$ .
  - 2: **for**  $t = 1, \dots, T + 1$  **do**
  - 3:     Let  $g_t = \nabla L(w_t, D) + \epsilon_t$ , where  $\epsilon_t \sim \mathcal{N}(0, \sigma_1^2 I_d)$ .  $\tilde{H}_t = \nabla^2 L(w_t, D) + H_t$ , where  $H_t$  a symmetric matrix with its upper triangle (including the diagonal) being i.i.d samples from  $\mathcal{N}(0, \sigma_2^2)$  and each lower triangle entry is copied from its upper triangle counterpart.
  - 4:     **if**  $\|g_t\|_2 \leq \alpha$  and  $\lambda_{\min}(\tilde{H}_t) \geq -\sqrt{\rho\alpha}$  **then**
  - 5:         Return  $w_t$ .
  - 6:     **end if**
  - 7: **end for**
- 

### Constrained Case

In this section we consider a constrained-version of SOSP studied in last section (see Definition 4.3.4).

**Definition 4.3.4** ([222]). For a twice differentiable function  $F$  and a closed convex set  $\mathcal{C}$ ,  $w^*$  is an  $\alpha$ -second-order stationary point in the constraint set  $\mathcal{C}$  if: 1)  $\nabla F(w^*)^T(w - w^*) \geq -\alpha$ , for  $\forall w \in \mathcal{C}$ , and 2)  $(w - w^*)^T \nabla^2 F(w^*)(w - w^*) \geq -\sqrt{\rho\alpha}$ , for  $\forall w \in \mathcal{C}$ , s.t.  $\nabla F(w^*)^T(w - w^*) = 0$ .

Recently, [222] proposed an algorithm for escaping the saddle points in the above constrained case. Motivated by their algorithm and the ideas in the proof of Theorem 4.3.1, we propose Algorithm 4.3.27 as a DP-version of the problem with a theoretical guarantee presented in Theorem 4.3.3.

**Theorem 4.3.3.** There exist constants  $c_1, c_2, c_3$  and sufficiently large  $C$  such that for any  $0 < \epsilon < c_1 T, 0 < \delta < 1$ , if  $\sigma_1^2 = c_2 \frac{\log \frac{1}{\delta} L^2 T}{n^2 \epsilon^2}$  and  $\sigma_2^2 = c_3 \frac{\log \frac{1}{\delta} d M^2 T}{n^2 \epsilon^2}$ , Algorithm 4.3.27 is  $(\epsilon, \delta)$ -DP. Moreover, taking  $T = O(\max\{\frac{D^2 MB}{\alpha^2}, \frac{B\rho^{1/2} D^6}{\Phi^3 \alpha^{3/2}}\}) = O(\frac{BM\rho^{1/2} D^6}{\Phi^3 \alpha^2})$ ,  $\theta = \frac{\Phi\alpha}{2D^3}$ ,  $0 < \Phi \leq \frac{9}{5}$ ,  $\{\eta_t\}_{t=1}^T = \frac{\alpha}{2D^2 M}$  and  $r = \frac{\Phi^2 \Phi \alpha}{72\rho D^3}$ , we have that for any  $0 < \xi < 1$ , with probability at least  $1 - \xi - \frac{T}{p^C}$ , Algorithm 4.3.27 outputs  $w_t$ , which is an  $\alpha$ -SOSP of the

---

**Algorithm 4.3.27 DP-GD-SO**


---

**Input:**  $T$  is the iteration number and  $x_1$  is the initial point.  $\{\gamma_t\}_{t=1}^T$  is the step size.  $\epsilon$  and  $\delta$  are privacy parameters.  $\theta, \sigma_1, \sigma_2$  are parameters to be specified later.

```

1: for  $t = 1, \dots, T$  do
2:   Compute  $g_t = \nabla L(w_t, D) + \epsilon_t$ , where  $\epsilon_t \sim \mathcal{N}(0, \sigma_1^2 I_d)$  for some  $\sigma_1$ .
3:   Compute  $v_t = \arg \max_{v \in \mathcal{C}} \{-g_t^T v\}$ .
4:   if  $g_t^T(v_t - w_t) < -\frac{\alpha}{2}$  then
5:     Compute  $w_{t+1} = (1 - \eta_t)w_t + \eta_t v_t$ .
6:   else
7:     Let  $\tilde{H}_t = \nabla^2 L(w_t, D) + H_t$ , where  $H_t$  is a symmetric matrix with its upper triangle (including the diagonal) being i.i.d samples from  $\mathcal{N}(0, \sigma_2^2)$  and each lower triangle entry is copied from its upper triangle counterpart.
8:     Find  $u_t$ , a  $\Phi$ -approximate solution of

$$\min_u q(u) = (u - w_t)^T \tilde{H}_t (u - w_t)$$


$$\text{s.t. } u \in \mathcal{C}, g_t^T(u - w_t) \leq r$$

9:     if  $q(u_t) \leq \frac{-\Phi\sqrt{\rho\alpha}}{2}$  then
10:       Compute  $w_{t+1} = (1 - \theta)w_t + \theta u_t$ .
11:     else
12:       Return  $w_t$ .
13:     end if
14:   end if
15: end for
16:

```

---

empirical risk  $L(\cdot, D)$ , if the sample size  $n$  satisfies:

$$n \geq \tilde{\Omega}\left(\max\left\{\frac{LD^7\sqrt{dMB\log\frac{1}{\delta}}\log\frac{1}{\xi}\rho^{1/4}}{\epsilon\alpha^2}, \frac{\sqrt{\log\frac{1}{\delta}dBMLD^4}\log\frac{1}{\xi}\rho^{1/4}}{\epsilon\alpha^2}, \frac{d\sqrt{BM^3\log\frac{1}{\delta}}D^5\log\frac{1}{\xi}}{\rho^{1/4}\alpha^{3/2}\epsilon}\right\}\right).$$

Here the  $\tilde{\Omega}$ -notation omits  $\Phi$  and other log terms.

**Remark 4.3.1.** Firstly, we note that when omitting other terms in the bound in Theorem 4.3.3 such as  $L, B, \Phi, D, G, \rho$ , the sample complexity for escaping saddle points in the constrained case is  $\tilde{\Omega}\left(\frac{d}{\epsilon\alpha^2}\right)$ . Compared with the unconstrained case in Theorem 4.3.2, they are asymptotically the same. Secondly, a quadratic programming problem needs to be

solved in step 8 of Algorithm 4.3.27. For a general constraint set  $\mathcal{C}$ , solving the quadratic problem is NP-hard. However, for some specified sets such as intersection of ellipsoids or balls, an approximate solution can be obtained in polynomial time. See [222] for more details.

### 4.3.2 Improved Sample Complexity via DP-TR Method

Our ideas are derived from the trust region method proposed in [78], we now briefly introduce the trust region method. In each step of the trust region method for a function  $F(\cdot)$ , it solves a Quadratic Constraint Quadratic Program (QCQP):

$$h^k = \arg \min_{h \in \mathbb{R}^d, \|h\|_2 \leq r} \langle \nabla F(w^k), h \rangle + \frac{1}{2} \langle \nabla^2 F(w^k)h, h \rangle, \quad (4.57)$$

where  $r$  is called the *trust-region radius*. Then, it updates in the following way

$$w^{k+1} = w^k + h^k.$$

Since the function  $F(w)$  is non-convex, this indicates that the sub-problem (4.57) is non-convex. However, its global minimum can be characterized by the following lemma.

**Lemma 4.3.1** (Corollary 7.2.2 in [78]). Any global minimum of the problem (4.57) should satisfy

$$(\nabla^2 F(w^k) + \lambda I)h^k = -\nabla F(w^k), \quad (4.58)$$

where the dual variable  $\lambda \geq 0$  should satisfies the conditions of  $\nabla^2 F(x^k) + \lambda I \succ 0$  and  $\lambda(\|h^k\|_2 - r) = 0$ .

It is worth noting that in practice sub-problem (4.57) can be solved by the Lanczos method efficiently (see [134] for details). For the dual variable  $\lambda$  in Lemma 4.3.1, it can be solved by almost any QCQP solver such as CVX [135].

## Differentially Private Trust Region Method

The key idea of our DP-TR is the following. In each iteration, instead of using the gradient and Hessian of the empirical risk directly to the sub-problem (4.57), we use their perturbed versions to ensure DP. That is, we use  $\tilde{\nabla}L(w^k, D) = \nabla L(w^k, D) + \epsilon_k$  and  $\tilde{\nabla}^2L(w^k, D) = \nabla^2L(w^k, D) + H_k$ , where  $\epsilon_t$  is a Gaussian vector and  $H_t$  is a randomized symmetric Gaussian matrix (since a Hessian matrix is symmetric, we need to add a symmetric random matrix). The main steps of DP-TR are given in Algorithm 4.3.28.

For the stopping criteria, we use the dual variable  $\lambda^k$  and see whether the value is greater or less than some threshold. This criteria enable the last-term convergence analysis in Theorem 4.3.5.

The following theorem shows that Algorithm 4.3.28 is  $(\epsilon, \delta)$ -DP.

**Theorem 4.3.4.** For any  $\epsilon, \delta > 0$ , Algorithm 4.3.28 is  $(\epsilon, \delta)$ -differentially private under Assumption 4.3.1.

---

### Algorithm 4.3.28 DP-TR

**Input:** Privacy parameters  $\epsilon, \delta$ , trust-region radius  $r$ , iteration number  $T$  (to be specified later), initial vector  $w^0$  and error term  $\alpha$

- 1: Let  $\phi = (\sqrt{\epsilon + \ln \frac{1}{\delta}} - \sqrt{\ln \frac{1}{\delta}})^2$ .
- 2: **for**  $k = 0, \dots, T - 1$  **do**
- 3:     Denote  $\tilde{\nabla}L(w^k, D) = \nabla L(w^k, D) + \epsilon_k$ , where  $\epsilon_t \sim \mathcal{N}(0, \sigma^2 I_d)$  with  $\sigma^2 = \frac{4G^2T}{n^2\phi}$ .
- 4:     Denote  $\tilde{\nabla}^2L(w^k, D) = \nabla^2L(w^k, D) + H_k$ , where  $H_t$  is a symmetric matrix with its upper triangle (including the diagonal) being i.i.d samples from  $\mathcal{N}(0, \sigma_2^2)$ ,  $\sigma_2^2 = \frac{4pM^2T}{n^2\phi}$ , and each lower triangle entry is copied from its upper triangle counterpart.
- 5:     Solve the following QCQP and get  $h^k$  and dual variable  $\lambda^k$ ,

$$h^k = \arg \min_{h \in \mathbb{R}^d, \|h\|_2 \leq r} \langle \tilde{\nabla}L(w^k, D), h \rangle + \frac{1}{2} \langle \tilde{\nabla}^2L(w^k, D)h, h \rangle,$$

- 6:     Let  $w^{k+1} = w^k + h^k$ .
  - 7:     **if**  $\lambda^k \leq \sqrt{\alpha\phi}$  **then**
  - 8:         Output  $w_\alpha = w^{k+1}$ .
  - 9:     **end if**
  - 10: **end for**
-

The following theorem shows that when the data size  $n$  is large enough, then with high probability the output of Algorithm 4.3.28 will be an  $\alpha$ -SOSP.

**Theorem 4.3.5.** Under Assumption 4.3.1, for any given  $\alpha$ , if we take  $r = \sqrt{\frac{\alpha}{\rho}}$ ,  $T = \frac{6\sqrt{\rho}\Delta}{\alpha^{1.5}}$ , then with probability at least  $1 - \zeta - \frac{T}{p^c}$  for some universal constant  $c > 0$  and  $\zeta > 0$ , the algorithm outputs a point which is an  $O(\alpha)$ -SOSP if  $n$  satisfies

$$n \geq \Omega\left(\frac{p \ln \frac{1}{\zeta} \sqrt{\ln \frac{1}{\delta}}}{\alpha^{1.75} \epsilon}\right), \quad (4.59)$$

where the Big- $\Omega$  notation omits the terms of  $G, M, \rho, \Delta, \ln \frac{1}{\alpha}$ .

**Remark 4.3.2.** We note that in Theorem 4.3.2 to output an  $O(\alpha)$ -SOSP with high probability, the data size  $n$  needs to satisfy  $n \geq \Omega\left(\frac{p\sqrt{\ln \frac{1}{\delta}}}{\alpha^2 \epsilon}\right)$ , while the dependency on  $\alpha$  in (4.59) is  $\frac{1}{\alpha^{1.75}}$ . Thus, we improve the sample size by a factor of  $O\left(\frac{1}{\alpha^{0.25}}\right)$ . Equivalently, if we fix  $n$ , Theorem 4.3.5 ensures that Algorithm 4.3.28 outputs a point which is  $O\left(\left(\frac{p\sqrt{\ln \frac{1}{\delta}}}{n\epsilon}\right)^{\frac{4}{7}}\right)$ -SOSP, while Theorem 4.3.2 outputs a point which is  $O\left(\left(\frac{p\sqrt{\ln \frac{1}{\delta}}}{n\epsilon}\right)^{\frac{1}{2}}\right)$ -SOSP. We can see that our algorithm yields better approximate SOSP than the previous one. We leave as open problems to determine whether the sample complexity in (4.59) can be further improved and what is the optimal bound of the sample complexity.

Also, in Theorem 4.3.2 the number of iterations is  $T = \tilde{O}\left(\frac{1}{\alpha^2}\right)$ , while Algorithm 4.3.28 needs only  $O\left(\frac{1}{\alpha^{1.5}}\right)$  iterations. This means that the running time of Algorithm 4.3.28 is  $O\left(\frac{n\text{Poly}(p)}{\alpha^{1.5}}\right)$ , while it is  $O\left(\frac{n\text{Poly}(p)}{\alpha^2}\right)$  in Theorem 4.3.2. Thus, our algorithm has an improved time complexity for the term of  $\frac{1}{\alpha}$  compared with the previous one. Moreover, as we will see in the experiment section, our algorithms is indeed faster than the previous one.

Theorem 4.3.5 shows the explicit step size control of the DP-TR method: Since the dual variable satisfies  $\lambda^k > \sqrt{\alpha\rho}$  for all but the last iteration. Thus we can always find a solution to the trust-region sub-problem (4.57) in the boundary, *i.e.*,  $\|h^k\|_2 = r$ , according to Lemma 4.3.1.

## Differentially Private Stochastic Trust Region Method

In the previous section we show that our method DP-TR needs less samples and is faster than DP-GD (Algorithm 4.3.25). However, as mentioned in Remark 4.3.2, the time complexities of both algorithms are linearly dependent on the sample size  $n$ , which is prohibitive in large scale datasets. Thus, a natural question is to determine whether it is possible to design an algorithm that shares the advantages of DP-TR and meanwhile is scalable. In this section we give an affirmative answer to this question by providing a stochastic version of DP-TR called Differentially Private Stochastic Trust Region method (DP-STR).

The key idea of DP-STR is that, instead of evaluating the gradient and Hessian matrix of the whole function  $L(w, D)$  in each iteration, we will uniformly sub-sample two sets of indices  $\mathcal{S}, \mathcal{T} \subseteq [n]$  and calculate the gradients and Hessian matrix of the loss function with the samples corresponding to the set  $\mathcal{S}$  and  $\mathcal{T}$ , respectively. That is

$$\nabla L(w^k, \mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \nabla \ell(w^k, x_i), \quad (4.60)$$

$$\nabla^2 L(w^k, \mathcal{T}) = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \nabla^2 \ell(w^k, x_i). \quad (4.61)$$

Then, similar to DP-TR, we add some Gaussian noise and random Gaussian matrix to  $\nabla L(w^k, \mathcal{S})$  and  $\nabla^2 L(w^k, \mathcal{T})$ , respectively, to ensure  $(\epsilon, \delta)$ -DP. See Algorithm 4.3.29 for details. Note that since zCDP can not be guaranteed by sub-sampling, we use the traditional advanced composition theorem Lemma 2.1.5 and sub-sampling property Lemma 2.1.2 to guarantee  $(\epsilon, \delta)$ -DP.

**Theorem 4.3.6.** For any  $0 < \epsilon, \delta < 1$ , Algorithm 4.3.29 is  $(\epsilon, \delta)$ -differentially private.

**Theorem 4.3.7.** Under Assumption 4.3.1, for a given  $\alpha$ , if we take  $r = \sqrt{\frac{\alpha}{\rho}}$ ,  $T = \frac{6\sqrt{\rho}\Delta}{\alpha^{1.5}}$ ,  $|\mathcal{S}| \geq \Omega(\frac{L^2 \ln \frac{p}{\zeta}}{\alpha^2})$  and  $|\mathcal{T}| \geq \Omega(\frac{M^2 \ln \frac{p}{\zeta}}{\alpha\rho})$  in Algorithm 4.3.29, then with probability at least  $1 - 3\zeta - \frac{T}{p^c}$  for some universal constant  $c > 0$  and  $\zeta > 0$ , the algorithm outputs a point that is an  $O(\alpha)$ -SOSP if  $n$  satisfies (4.59), which is the same as in Theorem 4.3.5.

---

**Algorithm 4.3.29** DP-STR

---

**Input:** Privacy parameters  $\epsilon, \delta$ , trust-region radius  $r$ , iteration number  $T$ , sub-sampling size  $|\mathcal{S}|, |\mathcal{T}|$  (to be specified later), initial vector  $w^0$  and error term  $\alpha$ .

- 1: **for**  $k = 0, \dots, T - 1$  **do**
  - 2:     Uniformly sub-sample two independent indices sets  $\mathcal{S}, \mathcal{T} \subseteq [n]$  with size  $|\mathcal{S}|$  and  $|\mathcal{T}|$ , respectively.
  - 3:     Denote  $\tilde{\nabla}L(w^k, \mathcal{S}) = \nabla L(w^k, \mathcal{S}) + \epsilon_k$ , where  $\epsilon_t \sim \mathcal{N}(0, \sigma^2 I_d)$  with  $\sigma^2 = \frac{256G^2 \ln \frac{5T}{\delta} \ln \frac{2}{\delta}}{n^2 \epsilon^2}$  and  $\nabla L(w^k, \mathcal{S})$  is given in (4.60)
  - 4:     Denote  $\tilde{\nabla}^2 L(w^k, \mathcal{T}) = \nabla^2 L(w^k, \mathcal{T}) + H_k$ , where  $H_t$  is a symmetric matrix with its upper triangle (including the diagonal) being i.i.d samples from  $\mathcal{N}(0, \sigma_2^2)$ ,  $\sigma_2^2 = \frac{256pM^2T \ln \frac{2}{\delta} \ln \frac{5T}{\delta}}{n^2 \epsilon^2}$ , and each lower triangle entry is copied from its upper triangle counterpart.  $\nabla^2 L(w^k, \mathcal{T})$  is given in (4.61).
  - 5:     Solve the following QCQP and get  $h^k$  and dual variable  $\lambda^k$ ,
$$h^k = \arg \min_{h \in \mathbb{R}^d, \|h\|_2 \leq r} \langle \tilde{\nabla}L(w^k, \mathcal{S}), h \rangle + \frac{1}{2} \langle \tilde{\nabla}^2 L(w^k, \mathcal{T})h, h \rangle,$$
  - 6:     Let  $w^{k+1} = w^k + h^k$ .
  - 7:     **if**  $\lambda^k \leq \sqrt{\alpha\rho}$  **then**
  - 8:         Output  $w_\alpha = w^{k+1}$ .
  - 9:     **end if**
  - 10: **end for**
- 

Comparing with Theorem 4.3.5, we can see that the sample complexity of Theorem 4.3.7 is the same while the time complexity of Algorithm 4.3.29 is  $O(T(|\mathcal{S}| + |\mathcal{T}|)\text{Poly}(p)) = O(\frac{\text{Poly}(p)}{\alpha^{3.5}})$ , which is independent of the sample size  $n$ . This means that DP-STR is faster and scalable to large scale datasets.

**Remark 4.3.3.** We note that it is unknown whether the DP-GD (Algorithm 4.3.25) can be extended to a stochastic version whose time complexity is independent of the size  $n$ . Algorithm 4.3.25 consists of two routines, one is the Differentially Private Gradient Descent method and the other one is the procedure of selecting an  $\alpha$ -SOSP. The first one can be easily extend to a stochastic version, which is similar as the one in [356]. However, for the second one, it needs to calculate the whole Hessian matrix and verify some conditions as stopping criteria, but it is unknown whether we can extend it to a stochastic version. Compared with Algorithm 4.3.25, in Algorithm 4.3.28 we use the Hessian matrix for Trust-Region sub-problem and use the dual variable  $\lambda^k$  as our stopping criteria. Thus, this is why we can

extend Algorithm 4.3.28 to a stochastic version.

Note that in Algorithm 4.3.29 we use the basic subsampling technique for DP-STR to improve the time complexity. In [328], the authors proposed the Stochastic Variance Reduction Gradient method to improve the gradient complexity for DP-ERM with convex less functions and show it is superior to the DP-SGD method. Thus, it is unknown whether we can use the same idea to our problem to further improve the time complexity or gradient complexity. Moreover, in both of Algorithm 4.3.28 and 4.3.29, we assume that we can exactly solve the Trust-Region sub-problem (4.57). However, in most cases, exactly solving the problem is quite hard and costly. Thus whether we can relax this assumption is still an open problem. We leave these as further research.

### 4.3.3 Experiments

In this section, we present numerical experiments for different non-convex Empirical Risk Minimization problems on different datasets to demonstrate the our DP-GD, DP-TR and DP-STR algorithms in finding SOSP under differential privacy.

#### Experimental Settings

**Baselines** We will use our methods (DP-GD, DP-TR and DP-STR) after carefully tuning the algorithms for a fair comparison. For the QCQP sub-problem in Algorithm 4.3.28 and 4.3.29, we use the CVX package [135] to solve it.

**Datasets** We evaluate the algorithms on real-world datasets with  $n \gg p$ . Specifically, we use the datasets, Covertype and IJCNN, which are commonly used in the study of DP-ERM such as [328, 319, 311]. More information about these datasets is listed in Table 4.2. We normalize each row of the datasets as preprocessing.

Table 4.2: Summary of Datasets used in the experiments.

Dataset	Sample size $n$	dimension $p$
Covertype	581,012	54
IJCNN	35,000	22

**Evaluated Problems** For the loss functions we will follow the studies in [182, 367, 309].

The first non-convex problem that will be investigated is logistic regression with a non-convex regularizer  $r(w) = \sum_{i=1}^p \frac{\lambda w_i^2}{1+w_i^2}$ . Specifically, suppose that we are given training data  $\{(x_i, y_i)\}_{i=1}^n$ , where  $x_i \in \mathbb{R}^p$  and  $y \in \{-1, 1\}$  are, respectively, the feature vector and label of the  $i$ -th data record. The corresponding ERM is

$$\min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \log(1 + \exp(-y_i \langle x_i, w \rangle)) + r(w).$$

In the experiment, we set  $\lambda = 10^{-3}$ .

The second problem that will be considered is the sigmoid regression with  $\ell_2$  norm regularizer. Given training dataset  $\{(x_i, y_i)\}_{i=1}^n$  where  $x_i \in \mathbb{R}^p$  and  $y \in \{-1, 1\}$  are, respectively, the feature vector and label of the  $i$ -th data record. Then, minimization problem is

$$\min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \frac{1}{1 + \exp(-y_i \langle x_i, w \rangle)} + \frac{\lambda}{2} \|w\|_2^2.$$

In the experiment, we set  $\lambda = 10^{-3}$ .

**Measurements** We first study how the optimality gap, *i.e.*,  $L(w_\alpha, D) - \min_{w \in \mathbb{R}^p} L(w, D)$ , changes w.r.t the privacy level  $\epsilon$  or time (second). For the optimal solution of the problem  $\min_{w \in \mathbb{R}^p} L(w, D)$ , we obtain it through multiple runs of the classical trust region method and taking the best one. Besides the expected excess empirical risk, we also use the gradient norm, *i.e.*,  $\|\nabla L(w_\alpha, D)\|_2$ , to measure the utility. For logistic regression, we also consider its classification accuracy w.r.t privacy level, where the non-private case is obtained by running the trust region method and taking the best. For each experiment, we run 10 times

and take the average as the final output. In all experiments, we set  $\delta = \frac{1}{n}$  and  $\alpha = 10^{-1}$ .

## Experimental Results

Figure 4.5 shows the classification accuracy of the private classifier given by the sigmoid regression on the Covertype and IJCNN datasets w.r.t different privacy levels. We can see that the accuracy increases when  $\epsilon$  becomes larger, which means that the algorithm will be non-private. From Remark 4.3.2 we can see that this is due to the fact that when  $\epsilon$  is larger, we can output an SOSP which is closer to the local minimum. Also, the accuracy of the non-private case is 86% and 95% for Covertype and IJCNN dataset, respectively. This indicates that the accuracy is comparable to the non-private case when  $\epsilon \geq 1.5$ .

The first and second subfigures of Figure 4.6, 4.7, 4.8 and 4.9 depict the optimality gap and the gradient norm w.r.t different privacy level  $\epsilon$  of the two non-convex problems on Covertype and IJCNN datasets. For Covertype, we set the batchsize as 50000, while for IJCNN we set it as 5000. From the figures, we can see that compared with DP-GD, our DP-TR method has better performance on both the optimality gap and the gradient norm. This is due to the fact that DP-TR has improved the bound of SOSP (see Remark 4.3.2). However, the results of DP-STR are worse than that of DP-GD and DP-TR. We attribute this to the fact that the noise level of DP-STR added in each iteration (steps 2 and 3) is higher than that of DP-TR and DP-GD. For example, in Step 2 of Algorithm 4.3.29 we add a Gaussian noise with variance  $\sigma^2 = \frac{256L^2 \ln \frac{5T}{\delta} \ln \frac{2}{\delta}}{n^2 \epsilon^2}$  to each coordinate, while in step 3 of Algorithm 4.3.28 we only need to add a Gaussian noise with variance  $\sigma^2 = \frac{4L^2 T}{n^2 \phi} \approx \frac{64L^2 T \log \frac{1}{\delta}}{n^2 \epsilon^2}$ . Equivalently, the sub-optimality of DP-STR is due to the higher level of noise that needs to be added, which is required by the Advanced Composition Theorem to ensure  $(\epsilon, \delta)$ -DP. We leave it as an open problem to determine how to improve the practical performance of DP-STR.

The third subfigures of Figure 4.6, 4.7, 4.8 and 4.9 show the results on the optimality gap w.r.t time of the two non-convex problems on the datasets of Covertype and IJCNN. Here we fix  $\epsilon$  to be 1 in all the experiments. We can see that although the gap of DP-STR is

worse than that of DP-GD and DP-TR, its running time is the least one. This is due to the fact that DP-STR needs only to evaluate a subset of the gradient and Hessian matrix, instead of the full ones as in DP-TR and DP-GD.

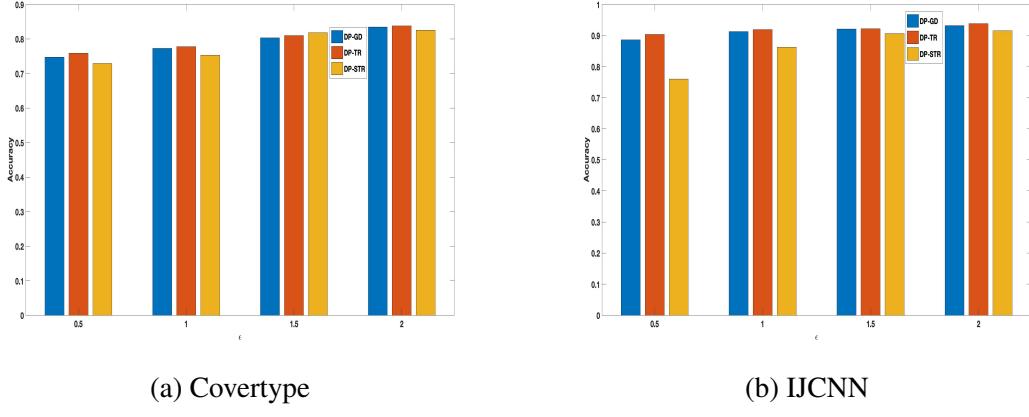


Figure 4.5: Accuracy w.r.t privacy level on Covertype and IJCNN datasets

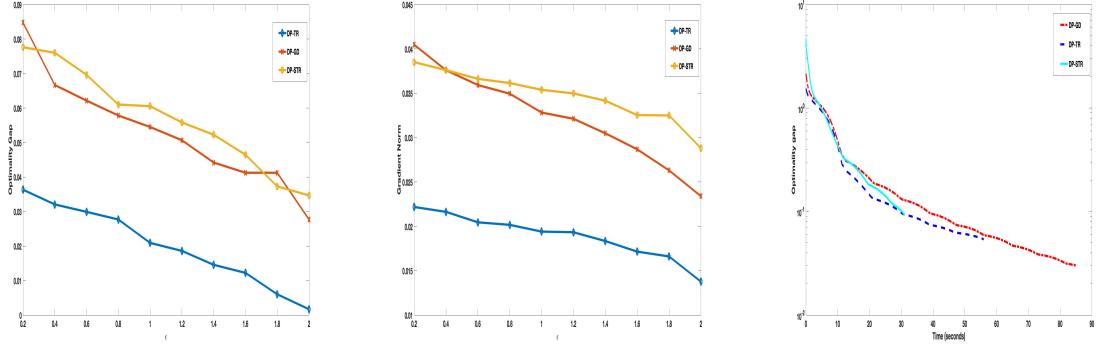


Figure 4.6: Results of logistic regression with non-convex regularizer on Covertype dataset

### 4.3.4 Omitted Proofs

#### Proof of Theorem 4.3.1

The guarantee of  $(\epsilon, \delta)$ -DP comes from the Moment Accountant in Lemma 2.1.7. Below we show that one of  $\{w_1, w_2, \dots, w_T\}$  is  $\alpha$ -SOSP with high probability.

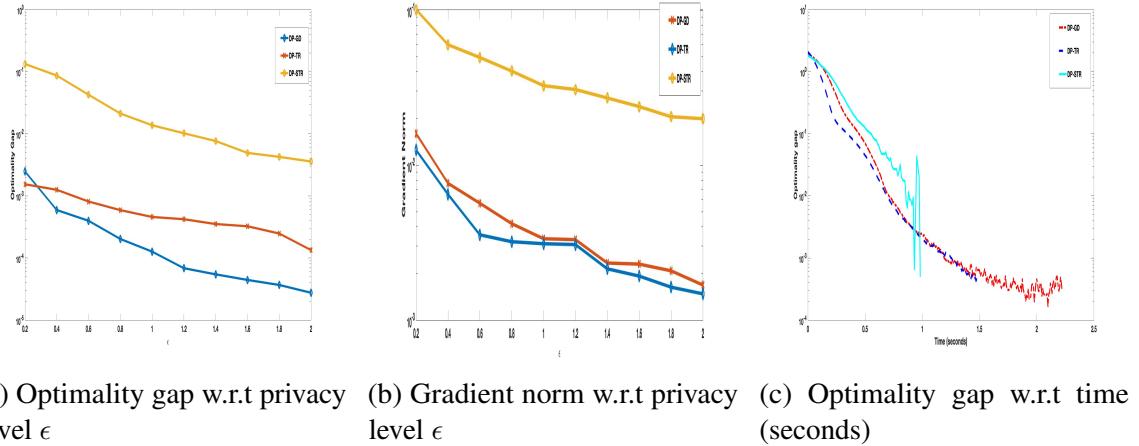


Figure 4.7: Results of logistic regression with non-convex regularizer on IJCNN dataset

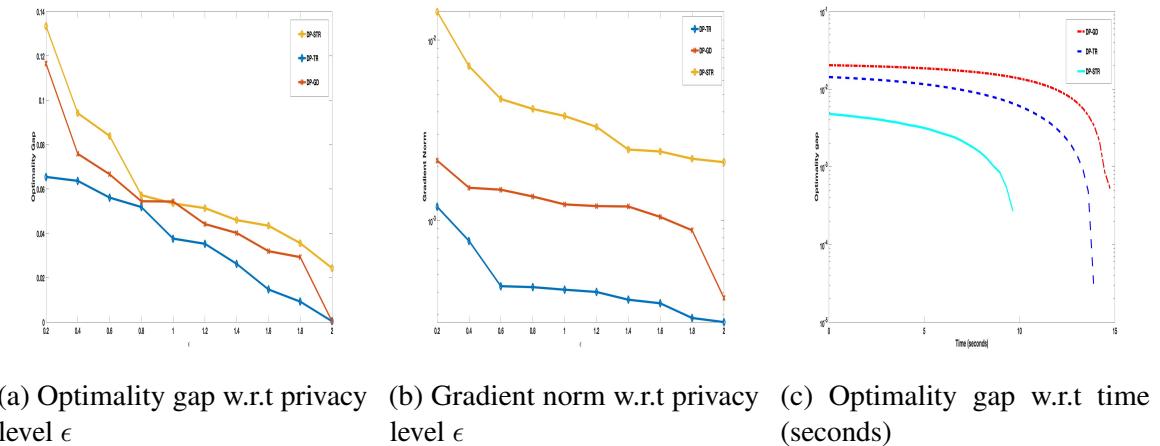


Figure 4.8: Results of sigmoid regression with  $\ell_2$  norm regularizer on Covertype dataset

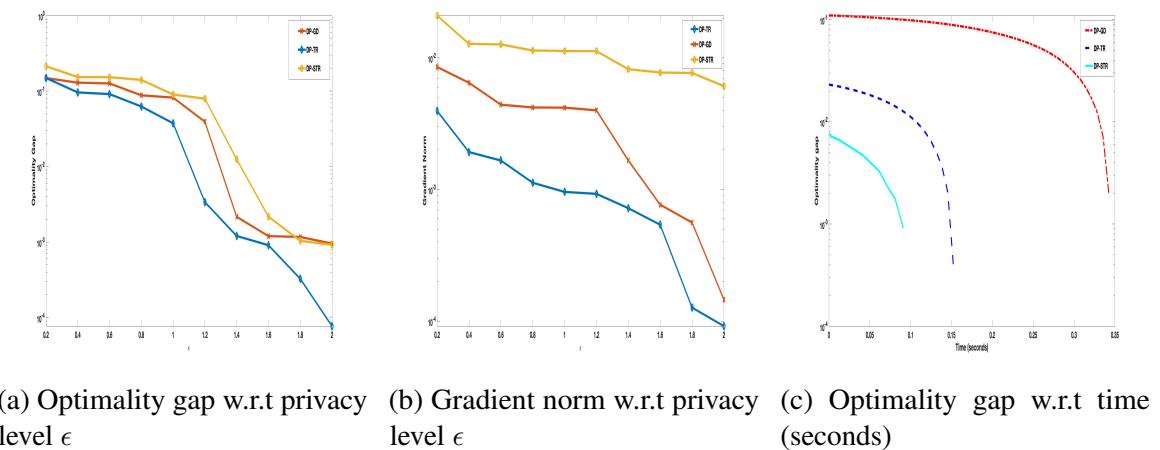


Figure 4.9: Results of sigmoid regression with  $\ell_2$  norm regularizer on IJCNN dataset

For convenience, we use the following notations  $F(w) = L(w, D)$ ,  $\{\eta_t\} = \eta = \frac{1}{M}$ ,  $\Phi = \sqrt{\frac{\alpha^3}{\rho}}\chi^{-3}c^{-5}$ ,  $r = \alpha\chi^{-3}c^{-6}$ ,  $\Gamma = \frac{\chi c}{\eta\sqrt{\rho\alpha}}$  and  $\chi = \max\{1, C_1 \log \frac{dMB}{\rho\alpha\xi}\}$  for some constant  $C_1$  and enough large constant  $c$ .

By the concentration inequality of Gaussian distribution, we have the following lemma.

**Lemma 4.3.2.** With probability at least  $1 - \frac{\xi}{2}$ , for all  $i \in [T]$ ,

$$\|\epsilon_i\|_2 \leq \frac{\sqrt{2c_2 \log \frac{1}{\delta} T d L \log \frac{4T}{\xi}}}{n\epsilon} \leq r = \alpha\chi^{-3}c^{-6}.$$

Below we assume that the event in Lemma 4.3.2 happens. Next, we show the following.

**Lemma 4.3.3.** If  $\|F(w_t)\|_2 \geq \alpha$ , then we have

$$F(w_{t+1}) - F(w_t) \leq -\eta \frac{\alpha^2}{4}.$$

*Proof of Lemma 4.3.3.* By the  $M$ -smoothness and taking  $\eta = \frac{1}{M}$ , we have

$$\begin{aligned} F(w_{t+1}) &\leq F(w_t) + \langle \nabla F(w_t), w_{t+1} - w_t \rangle + \frac{M}{2} \|w_{t+1} - w_t\|_2^2 \\ &\leq F(w_t) - \eta \|\nabla F(w_t)\|_2^2 + \eta \|\nabla F(w_t)\|_2 \|\epsilon_t\|_2 \\ &\quad + \frac{\eta^2 M}{2} [\|\nabla F(w_t)\|_2^2 + 2\|\nabla F(w_t)\|_2 \|\epsilon_t\|_2 + \|\epsilon_t\|_2^2] \\ &= F(w_t) - \eta \|\nabla F(w_t)\|_2 \left[ \frac{1}{2} \|\nabla F(w_t)\|_2 - 2\|\epsilon_t\|_2 \right] + \frac{\eta}{2} \|\epsilon_t\|_2^2 \\ &\leq F(w_t) - \frac{\eta\alpha^2}{4}, \end{aligned}$$

where the last inequality is due to the following: by the assumption on  $n$ , we have  $\|\epsilon_t\| \leq \alpha\xi^{-3}c^{-6} \leq \frac{\alpha}{20}$  for sufficiently large  $c$  and  $\|\nabla F(w_t)\|_2 \geq \alpha$ .  $\square$

Next, we prove the following key lemma:

**Lemma 4.3.4.** If  $\|\nabla F(w_t)\| \leq \alpha$  and  $\lambda_{\min}(\nabla^2 F(w_t)) \leq -\sqrt{\rho\alpha}$ , then in Algorithm 4, with probability  $1 - \xi$ , we have  $F(w_{t+\Gamma}) - F(w_t) \leq -\Phi$ .

*Proof of Lemma 4.3.4.* To prove this lemma, we need the following lemmas.

**Lemma 4.3.5.**

$$F(w_{t+1}) - F(w_t) \leq -\frac{\eta}{4} \|\nabla F(w_t)\|_2^2 + 5\eta \|\epsilon_t\|_2^2.$$

*Proof of Lemma 4.3.5.* By the  $M$ -smoothness, we have

$$\begin{aligned} F(w_{t+1}) &\leq F(w_t) + \langle \nabla F(w_t), w_{t+1} - w_t \rangle + \frac{M}{2} \|w_{t+1} - w_t\|_2^2 \\ &\leq F(w_t) - \eta \langle F(w_t), F(w_t) + \epsilon_t \rangle + \frac{\eta^2 M}{2} (\|\nabla F(w_t)\|_2^2 + 2\|\nabla F(w_t)\|_2 \|\epsilon_t\|_2 + \|\epsilon_t\|_2^2) \\ &\leq F(w_t) - \frac{\eta}{2} \|\nabla F(w_t)\|_2^2 + 2\eta \|\nabla F(w_t)\|_2 \|\epsilon_t\|_2 + \frac{\eta}{2} \|\epsilon_t\|_2^2 \\ &\leq F(w_t) - \frac{\eta}{4} \|\nabla F(w_t)\|_2^2 + 5\eta \|\epsilon_t\|_2^2. \end{aligned}$$

□

**Lemma 4.3.6.** For all  $t+1 \leq T$ , we have

$$\|w_{t+1} - w_1\|_2^2 \leq 8\eta T (F(w_1) - F(x_{T+1})) + 50\eta^2 T \sum_{t=1}^T \|\epsilon_t\|_2^2.$$

*Proof of Lemma 4.3.6.* For any  $t \leq T-1$ , by Lemma 4.3.5, we have

$$\begin{aligned} \|w_{t+1} - w_t\|_2^2 &\leq \eta^2 \|\nabla F(w_t) + \epsilon_t\|_2^2 \leq 2\eta^2 \|\nabla F(w_t)\|_2^2 + 2\eta^2 \|\epsilon_t\|_2^2 \\ &\leq 8\eta (F(w_t) - F(w_{t+1})) + 50\eta^2 \|\epsilon_t\|_2^2. \end{aligned}$$

Thus we have

$$\sum_{t=1}^T \|w_{t+1} - w_t\|_2^2 \leq 8\eta (F(w_1) - F(w_{T+1})) + 50\eta^2 \sum_{t=1}^T \|\epsilon_t\|_2^2.$$

In total, we get

$$\|w_{t+1} - w_1\|_2^2 \leq (\sum_{i=1}^t \|w_{i+1} - w_i\|_2)^2 \leq t \sum_{i=1}^t \|w_{i+1} - w_i\|_2^2 \leq T \sum_{t=1}^T \|w_{t+1} - w_t\|_2^2.$$

□

Let  $\text{DPGD}_\epsilon^{(t)}(x)$  be our algorithm that updates  $t$ -times with perturbations  $\{\epsilon_1, \dots, \epsilon_t\}$  fixed and begins with  $x$ . Define the stuck region as:

$$\mathcal{X}^\epsilon(\tilde{w}) = \{w | w \in \mathbb{B}_{\tilde{w}}(\eta r), \text{ and } \Pr(F(\text{DPGD}_\epsilon^{(\Gamma)}(w)) - F(\tilde{w}) \geq -\Phi) \geq \sqrt{\xi}\}. \quad (4.62)$$

Intuitively, the later perturbations of the coupling sequence are the same while the very first perturbation is used to escape the saddle points.

**Lemma 4.3.7.** There exists a large enough constant  $c$  such that if  $\|\nabla F(\tilde{w})\|_2 \leq \alpha$  and  $\lambda_{\min}(\nabla^2 F(\tilde{w})) \leq -\sqrt{\rho\epsilon}$ , then the width of  $\mathcal{X}^\epsilon(\tilde{w})$  along the minimum eigenvector of  $\tilde{w}$  is at most  $\xi\eta r\sqrt{\frac{2\pi}{d}}$ .

*Proof of Lemma 4.3.7.* To prove this lemma, we let  $e_{\min}$  be the minimum eigenvector of  $\nabla^2 F(\tilde{w})$ . It suffice to show that for any  $w_1, w'_1 \in \mathbb{B}_{\tilde{w}}(\eta r)$  satisfying the condition of  $w_1 - w'_1 = \lambda e_{\min}$ , where  $|\lambda| \geq \xi\eta r\sqrt{\frac{2\pi}{d}}$ ,  $w_1 \notin \mathcal{X}^\epsilon(\tilde{w})$  or  $w'_1 \notin \mathcal{X}^\epsilon(\tilde{w})$ .

Let  $w_{\Gamma+1} = \text{DPGD}_\epsilon^{(\Gamma)}(w_1)$  and  $w'_{\Gamma+1} = \text{DPGD}_\epsilon^{(\Gamma)}(w'_1)$ , where the two sequences are independent. To show that  $w_1 \notin \mathcal{X}^\epsilon(\tilde{w})$  or  $w'_1 \notin \mathcal{X}^\epsilon(\tilde{w})$ , it is sufficient to demonstrate that with probability at least  $1 - \xi$

$$\min\{F(w_{\Gamma+1}) - F(\tilde{w}), F(w'_{\Gamma+1}) - F(\tilde{w})\} \leq -\Phi. \quad (4.63)$$

That is due to the fact that if  $w_1, w'_1 \in \mathcal{X}^\epsilon(\tilde{w})$ , we have, with probability at least  $\xi$ , that  $F(w_{\Gamma+1}) - F(\tilde{w}) \geq -\Phi$  and  $F(w'_{\Gamma+1}) - F(\tilde{w}) \geq -\Phi$ . This will mean that with probability

at most  $1 - \xi$ ,

$$\min\{F(w_{\Gamma+1}) - F(\tilde{w}), F(w'_{\Gamma+1}) - F(\tilde{w})\} \leq -\Phi, \quad (4.64)$$

which contradicts (4.63).

To prove that (4.63) holds with probability at least  $1 - \xi$ , we need to show that

1.  $\max\{F(w_1) - F(\tilde{w}), F(w'_1) - F(\tilde{w})\} \leq \Phi,$
2. with probability at least  $1 - \delta$ ,  $\min\{F(w_{\Gamma+1}) - F(w_1), F(w'_{\Gamma+1}) - F(w'_1)\} \leq -2\Phi$ .

For (1), we have, by the definition of  $w_1 \in \mathbb{B}_{\tilde{w}}(\eta t)$  and the  $M$ -smoothness, that

$$F(w_1) - F(\tilde{w}) \leq \alpha\eta r + \frac{M}{2}(\eta r)^2 = O\left(\frac{\alpha^2}{M}\chi^{-3}c^{-6}\right) \leq \Phi$$

for sufficiently large  $c$ . Similarly, we have the same for  $F(w'_1) - F(\tilde{w})$ .

To prove (2), we first assume that it is not true, *i.e.*,

$$\min\{F(w_{\Gamma+1}) - F(w_1), F(w'_{\Gamma+1}) - F(w'_1)\} \geq -2\Phi.$$

Then, by Lemmas 4.3.2 and 4.3.6, we have  $\forall t \in [\Gamma + 1]$  that for sufficiently large  $c > 0$ ,

$$\begin{aligned} & \max\{\|w_t - \tilde{w}\|_2, \|w'_t - \tilde{w}\|_2\} \\ & \leq \max\{\|w_t - w_1\|_2 + \|w_1 - \tilde{w}\|_2, \max\{\|w'_t - w'_1\|_2 + \|w'_1 - \tilde{w}\|_2\}\} \\ & \leq \sqrt{16\eta\Gamma\Phi + 50\eta^2\Gamma^2r^2} + \eta r \\ & \leq \sqrt{16\eta\Gamma\Phi + 50\eta^2\Gamma^2\alpha^2\chi^{-4}c^{-12}} + \eta\alpha\chi^{-3}c^{-6} \\ & \leq 4\left(\sqrt{\frac{\alpha}{\rho}}\chi^{-1}c^{-2}\right) = R, \end{aligned}$$

where the last inequality is due to the fact that  $M \geq \sqrt{\rho\alpha}$ . This means that both sequences  $\{w_t\}_{t=1}^{\Gamma+1}$  and  $\{w'_t\}_{t=1}^{\Gamma+1}$  do not leave the ball with radius  $R$  around  $\tilde{w}$ . Let  $H = \nabla^2 F(\tilde{w})$  and

$x_t := w_t - w'_t$ . We have

$$\begin{aligned} x_{t+1} &= x_t - \eta[\nabla F(w_t) - \nabla F(w'_t)] = (I - \eta H)x_t - \eta\Delta_t x_t \\ &\leq (I - \eta H)^t x_1 - \eta \sum_{\tau=1}^t (I - \eta H)^{t-\tau}(\Delta_\tau x_\tau), \end{aligned}$$

where  $\Delta_t = \int_0^1 [\nabla F(w'_t + \theta(w_t - w'_t)) - H]d\theta$ . By Hessian Lipschitz, we have  $\Delta_t \leq \rho \max\{\|w_t - \tilde{w}\|_2, \|w'_t - \tilde{w}\|_2\} \leq \rho R$ . We now show the following by induction:

$$\left\| \eta \sum_{\tau=1}^t (I - \eta H)^{t-\tau}(\Delta_\tau x_\tau) \right\| \leq \frac{1}{2} \|(I - \eta H)^t x_1\|_2. \quad (4.65)$$

For the base case of  $t = 1$ , we can easily verify it using the fact that  $\eta\rho R \leq \frac{1}{2}$  for sufficiently large  $c$ . Suppose that it holds for all  $t' \leq t$ . This gives us  $\|x_{t'}\|_2 \leq 2\|(I - \eta H)^{t'} x_1\|_2$ . Let  $\gamma = \lambda_{\min}(\nabla^2 F(\tilde{w}))$ . For the case of  $t + 1 \leq \Gamma + 1$ , we have

$$\begin{aligned} \left\| \eta \sum_{\tau=1}^t (I - \eta H)^{t-\tau}(\Delta_\tau x_\tau) \right\| &\leq \eta\rho R \left\| \sum_{\tau=1}^t (I - \eta H)^{t-\tau} \right\|_2 \|x_\tau\|_2 \leq \eta\rho R \Gamma (1 + \eta\gamma)^t \|x_1\|_2 \\ &\leq \frac{1}{4} \|(I - \eta H)^t x_1\|_2, \end{aligned}$$

where the third inequality uses the fact that  $x_0$  is along the direction of the minimum eigenvector of  $H$ , and the last one is due to the fact that  $\eta\rho R \Gamma = 4c^{-1} \leq \frac{1}{4}$  for large enough constant  $c$ .

Thus, in total we have

$$\begin{aligned}
\|x_{\Gamma+1}\| &\geq \|(I - \eta H)^\Gamma x_1\|_2 - \|\eta \sum_{\tau=1}^{\Gamma} (I - \eta H)^{t-\tau} (\Delta_\tau x_\tau)\|_2 \\
&\geq \frac{1}{2} \|(I - \eta H)^\Gamma x_1\|_2 = \frac{1}{2} (1 - \eta\gamma)^\Gamma \|x_1\|_2 \\
&\geq \frac{1}{2} (1 + \eta\sqrt{\rho\epsilon})^\Gamma \|x_1\|_2 \\
&\geq \frac{1}{2} (1 + \eta\sqrt{\rho\epsilon})^\Gamma \xi \eta r \sqrt{\frac{2\pi}{d}} \\
&\geq (1 + \eta\sqrt{\rho\alpha})^\Gamma \frac{\xi\alpha\chi^{-3}c^{-6}}{2M} \\
&\geq 2^{\eta\sqrt{\rho\alpha}\Gamma} \frac{\xi\alpha\chi^{-3}c^{-6}}{2M} \\
&\geq 8\sqrt{\frac{\alpha}{\rho}}\chi^{-1}c^{-2} = 2R,
\end{aligned}$$

where the last inequality is due to the fact that  $\Gamma = \frac{\chi c}{\eta\sqrt{\rho\alpha}}$  and  $\chi = \max\{1, \log \frac{\sqrt{d}M}{\xi\sqrt{\rho\alpha}}\}$ . From the above, we can see that when  $c$  is sufficiently large, the above inequalities hold. Thus, we have  $\|x_{\Gamma+1}\|_2 \geq 2R$ . This contradicts the fact that  $\max\{\|w_t - \tilde{w}\|_2, \|w'_t - \tilde{w}\|_2\} \leq R$ . This completes the proof.  $\square$

We now return to the proof of Lemma 4.3.4. Let  $r_0 = \xi r \sqrt{\frac{2\pi}{d}}$ . By Lemma 4.3.7, we know that  $\mathcal{X}^\epsilon(w_t)$  has width at most  $\eta r_0$  in the direction of the minimum eigenvector of  $\nabla^2 F(w_t)$ . Thus, we have

$$\text{Vol}(\mathcal{X}^\epsilon(w_t)) \leq \text{Vol}(\mathbb{B}_0^{(d-1)}(\eta r)) \cdot \eta r_0, \quad (4.66)$$

which gives us

$$\frac{\text{Vol}(\mathcal{X}^\epsilon(w_t))}{\text{Vol}(\mathbb{B}_{w_t}^d(\eta t))} \leq \frac{\text{Vol}(\mathbb{B}_0^{(d-1)}(\eta r)) \cdot \eta r_0}{\text{Vol}(\mathbb{B}_{w_t}^d(\eta t))} = \frac{r_0}{r\sqrt{\pi}} \frac{\Gamma(\frac{d}{2} + 1)}{\Gamma(\frac{d}{2} + \frac{1}{2})} \leq \frac{r_0}{r\sqrt{\pi}} \sqrt{\frac{d+1}{2}} \leq 2\xi.$$

Hence, with probability at least  $1 - 2\xi$ , the perturbation lands in  $\mathbb{B}_{w_t}^d(\eta t) \setminus \mathcal{X}^\epsilon(w_t)$ . That is,

with probability at least  $1 - \sqrt{\xi}$ , the following holds

$$F(DPGD_\epsilon^{(\Gamma)}(w_t)) - F(w_t) \leq -\Phi.$$

Thus, we have the above inequality with probability at least  $(1 - \xi)(1 - 2\xi)(1 - \sqrt{\xi}) \geq 1 - 3\sqrt{\xi}$ . Reparametrizing  $\xi' = 3\sqrt{\xi}$  only affects the factors in  $\chi$ .  $\square$

Now, we prove Theorem 4.3.1.

*Proof of Theorem 4.3.1.* By Lemmas 4.3.3 and 4.3.4, we have, with probability at least  $1 - \frac{2B}{\Phi}\xi$ , that the algorithm will find an  $\alpha$ -SOSP in the following number of iterations

$$O\left(\frac{B}{\eta\alpha^2} + \frac{B\Gamma}{\Phi}\right) = O\left(\frac{B\chi^4}{\eta\alpha^2}\right).$$

What we need is that

$$\frac{\sqrt{2c_2 \log \frac{1}{\delta} T p L} \log \frac{4T}{\xi}}{n\epsilon} \leq r = \alpha\chi^{-3}c^{-6},$$

which means  $n \geq \tilde{\Omega}\left(\frac{\sqrt{MB}\xi^5 c^6 \sqrt{2c_2 \log \frac{1}{\delta} p L}}{\epsilon\alpha^2}\right)$ . Taking  $\xi = \frac{2B}{\Phi}\xi$  only affects the log term.

This completes the proof.  $\square$

### Proof of Theorem 4.3.2

By a similar argument given in the proof of Theorem 4.3.1, we know that there exist  $c_1, c_2$  that make step 2 to step 6  $(\frac{\epsilon}{2}, \frac{\delta}{2})$ -DP. Thus, the whole algorithm is  $(\epsilon, \delta)$ -DP.

By Lemma 4.3.8, we know that with probability at least  $1 - \xi - \frac{T}{p^C}$

$$\begin{aligned} \|\epsilon_t\|_2 &\leq \frac{\sqrt{c_2 \log \frac{1}{\delta} T d L} \log \frac{8T}{\xi}}{n\epsilon} = Err_1 \\ \|H_t\|_2 &\leq \frac{C \sqrt{c_3 \log \frac{1}{\delta} T M d}}{n\epsilon} = Err_2. \end{aligned}$$

Now, we assume that the above event happens. By Theorem 4.3.1, we know that with probability at least  $1 - \frac{\xi}{2}$ , there exists  $\|\nabla F(w_t)\|_2 \leq \frac{\alpha}{2}$  and  $\lambda_{\min}(\nabla^2 F(w_t)) \geq -\sqrt{\frac{\rho\alpha}{2}}$ . Thus, for this  $t$ , we have

$$\begin{aligned}\|g_t\|_2 &\leq Err_1 + \frac{\alpha}{2} \leq \alpha \\ \lambda_{\min}(\tilde{H}_t) &\geq \lambda_{\min}(\nabla^2 F(w_t)) - Err_2 \geq -\sqrt{\rho\alpha}.\end{aligned}$$

These inequalities hold when  $Err_1 \leq \frac{\alpha}{2}$  and  $Err_2 \leq (1 - \sqrt{\frac{1}{2}})\sqrt{\rho\alpha}$ . Thus, the size  $n$  should satisfy

$$n \geq \tilde{\Omega}\left(\max\left\{\frac{\sqrt{BM \log \frac{1}{\delta}}dL \log \frac{1}{\xi}}{\epsilon\alpha^2}, \frac{\sqrt{\log \frac{1}{\delta} BMM d \log \frac{1}{\xi}}}{\rho\epsilon\alpha^2}\right\}\right).$$

Combining this with Theorem 4.3.1, we get the theorem.

### Proof of Theorem 4.3.3

First, we show the guarantee of  $(\epsilon, \delta)$ -DP. By Lemma 2.1.7, we know that  $\sigma_1^2 = \frac{16c_2 \log \frac{2}{\delta} L^2 T}{n^2 \epsilon^2}$ , where  $c_2$  is the constant in Lemma 2.1.7. Hence, it is  $(\frac{\epsilon}{2}, \frac{\delta}{\epsilon})$ -DP. Due to the L-smoothness, we have that for any pair of neighboring datasets  $D, D'$ ,  $\|\nabla^2 L(w, D) - \nabla^2 L(w, D')\|_2 \leq 2L$ , which means that  $\|\nabla^2 L(w, D) - \nabla^2 L(w, D')\|_F \leq 2\sqrt{d}L$ . This implies that if we view the Hessian matrix as a vector, the  $\ell_2$ -sensitivity is  $2\sqrt{d}L$ . Also, due to symmetric structure, adding symmetric Gaussian matrix with each entry sampled from  $\mathcal{N}(0, \sigma_2^2)$ , where  $\sigma_2^2 = \frac{c_3 T \log \frac{1}{\delta} M^2 d}{n^2 \epsilon^2}$ , will ensure  $(\frac{\epsilon}{2}, \frac{\delta}{2})$ -DP. Thus, the algorithm is  $(\epsilon, \delta)$ -DP.

Then, we show the ability of escaping saddle points. For simplicity, we let  $F(\cdot) = L(\cdot, D)$ ,  $\sqrt{\rho\alpha} = \gamma$ ,  $\gamma' = \frac{\gamma}{2}$ ,  $\alpha' = \frac{\alpha}{2}$ , and  $r = \frac{\Phi^2 \gamma'^2}{18\rho D^3}$ .

We first show the following lemma by using the concentration of Gaussian distribution and the spectrum of symmetric Gaussian noise [275].

**Lemma 4.3.8.** For any  $0 < \xi < 1$ , there exists a constant  $C, c_3, c_2 > 0$  such that with

probability at least  $1 - \xi - \frac{T}{p^C}$ , for any  $t \in [T]$ ,

$$\|\epsilon_t\|_2 \leq \frac{\sqrt{c_2 \log \frac{1}{\delta} T d L \log \frac{8T}{\xi}}}{n\epsilon} = Err_1 \quad (4.67)$$

$$\|H_t\|_2 \leq \frac{C \sqrt{c_3 \log \frac{1}{\delta} T M d}}{n\epsilon} = Err_2. \quad (4.68)$$

In the remaining analysis, we assume that the events in Lemma 4.3.8 happen, and the data size  $n$  is large enough such that

$$Err_1 \leq \min\left\{\frac{\Phi^2 \gamma'^2}{18\rho D^4}, \frac{\alpha'}{4D}\right\} \quad (4.69)$$

$$Err_2 \leq \frac{\Phi \gamma'}{9D^2}. \quad (4.70)$$

Thus,  $n$  should be

$$n \geq \max\left\{\frac{18\sqrt{c_2 \log \frac{1}{\delta} T d L \log \frac{8T}{\xi}} \rho D^4}{\epsilon \Phi^2 \gamma'^2}, \frac{4\sqrt{c_2 \log \frac{1}{\delta} T d L \log \frac{8T}{\xi}} D}{\epsilon \alpha'}, \frac{9\sqrt{c_3 \log \frac{1}{\delta} T D^2 M d}}{\Phi \gamma' \epsilon}\right\}. \quad (4.71)$$

We now show the iteration complexity of Algorithm 5. First, we consider the case of  $g_t^T(v_t - w_t) \leq -\alpha'$ .

**Lemma 4.3.9.** For  $w_t$ , if  $g_t^T(v_t - w_t) \leq -\alpha'$ , then we have

$$F(w_{t+1}) \leq F(w_t) - \frac{\alpha'^2}{4D^2 M}. \quad (4.72)$$

*Proof of Lemma 4.3.9.* By the  $M$ -smoothness of  $F(\cdot)$ , we have

$$\begin{aligned} F(w_{t+1}) &\leq F(w_t) + \langle \nabla F(w_t), w_{t+1} - w_t \rangle + \frac{M}{2} \|w_{t+1} - w_t\|_2^2 \\ &\leq F(w_t) + \eta \langle g_t, v_t - w_t \rangle + \eta \langle \epsilon_t, w_t - v_t \rangle + \frac{\eta^2 M D^2}{2} \\ &\leq F(w_t) - \eta \alpha' + \eta D Err_1 + \frac{\eta^2 M D^2}{2}. \end{aligned}$$

Taking  $\eta = \frac{\alpha'}{D^2 M}$ , since  $Err_1 \leq \frac{\alpha'}{4D}$ , we have

$$F(w_{t+1}) \leq F(w_t) - \frac{\alpha'^2}{4D^2 M}.$$

□

**Lemma 4.3.10.** For a given  $w_t$ , if  $g_t(v_t - w_t) \geq -\alpha'$  and  $q(u_t) \leq -\Phi\gamma'$ , then we have

$$F(w_{t+1}) \leq F(w_t) - \frac{\Phi^3 \gamma'^3}{6\rho^2 D^6}. \quad (4.73)$$

*Proof of Lemma 4.3.10.*

$$\begin{aligned} & F(w_{t+1}) \\ & \leq F(w_t) + \langle \nabla F(w_t), w_{t+1} - w_t \rangle + \frac{1}{2}(w_{t+1} - w_t)^T \nabla^2 F(w_t)(w_{t+1} - w_t) + \frac{\rho}{6} \|w_{t+1} - w_t\|^2 \\ & \leq F(w_t) + \theta \langle \nabla F(w_t), u_t - w_t \rangle + \frac{\theta^2}{2}(u_t - w_t)^T \nabla^2 F(w_t)(u_t - w_t) + \frac{\theta^3 \rho D^3}{6} \\ & \leq F(w_t) + \theta \langle g_t, u_t - w_t \rangle + \theta \langle \epsilon, w_t - u_t \rangle + \frac{\theta^2}{2}(u_t - w_t)^T \tilde{H}_t(u_t - w_t) \\ & \quad - \frac{\theta^2}{2}(u_t - w_t)^T H_t(u_t - w_t) + \frac{\theta^3 \rho D^3}{6} \\ & \leq F(w_t) + \theta r + \theta D Err_1 - \frac{\theta^2 \Phi \gamma'}{2} + \frac{\theta^2 D^2 Err_2}{2} + \frac{\theta^3 \rho D^3}{6}. \end{aligned}$$

Taking  $\theta = \frac{\Phi \gamma'}{\rho D^3}$  and by the inequalities  $Err_1 \leq \frac{\Phi^2 \gamma'^2}{18D^4 \rho}$ ,  $Err_2 \leq \frac{\Phi \gamma'}{9D^2}$ , and  $r = \frac{\Phi^2 \gamma'^2}{18\rho D^3}$ , we get the lemma. □

By Lemmas 4.3.9 and 4.3.10, we know that under the events of Lemma 4.3.8, the algorithm terminates in  $T = O(\max\{\frac{D^2 MB}{\alpha'^2}, \frac{B\rho^2 D^6}{\Phi^3 \gamma'^3}\})$  iterations.

Next, we will show that under the events of Lemma 4.3.8, the output  $w_t$  is an  $(\alpha, \gamma)$ -SOSP.

From the theorem, we know that  $w_t$  satisfies the following conditions:

$$g_t^T(v - w_t) \geq -\alpha', \forall v \in \mathcal{C},$$

$$(u - w_t)^T \tilde{H}_t(u - w_t) \geq -\Phi\gamma', \forall u \in \mathcal{C}, g_t^T(u - w_t) \leq r.$$

We will first show that  $w_t$  satisfies the first order condition, that is

$$\max_{u \in \mathcal{C}} \langle \nabla F(w_t), u - w_t \rangle \geq \min_{u \in \mathcal{C}} (\langle g_t, u - w_t \rangle - D\text{Err}_1) \geq -\alpha' - D\text{Err}_1 \geq -\alpha.$$

We then show that  $w_t$  satisfies the second-order property.

Let  $\mathcal{A} = \{w | \langle \nabla F(w_t), w - w_t \rangle = 0\}$  and  $\mathcal{B} = \{w | g_t^T(w - w_t) \leq r\}$ . We can show that  $A \subseteq B$ . This is due to the following. For any  $w \in A$ ,  $\langle \nabla F(w_t), w - w_t \rangle = 0$ . Thus,

$$g_t^T(w - w_t) = \nabla F(w_t)^T(w - w_t) + \epsilon_t^T(w - w_t) \leq D \cdot \text{Err}_1 \leq r.$$

Finally, for any  $w \in \mathcal{C}$ , we have

$$\begin{aligned} (w - w_t)^T \nabla^2 F(w_t)(w - w_t) &= (w - w_t)^T H_t(w - w_t) - (w - w_t)^T \tilde{H}_t(w - w_t) \\ &\geq -\Phi\gamma' - D^2 \text{Err}_2 \\ &\geq -\frac{10}{9}\Phi\frac{\gamma}{2} \geq -\frac{5}{9}\Phi\gamma \geq -\gamma. \end{aligned}$$

Thus, for all  $w \in \mathcal{C}$  satisfying the condition of  $\langle \nabla F(w_t), w - w_t \rangle = 0$ , we have  $(w - w_t)^T \nabla^2 F(w_t)(w - w_t) \geq -\gamma$ .

Thus,  $n$  should satisfy

$$n \geq \tilde{\Omega} \left( \max \left\{ \frac{LD^7 \sqrt{dMB \log \frac{1}{\delta}} \log \frac{1}{\xi} \rho^{1/4}}{\epsilon \Phi^{7/2} \alpha^2}, \frac{\sqrt{\log \frac{1}{\delta} dBMLD^4} \log \frac{1}{\xi} \rho^{1/4}}{\epsilon \alpha^2 \Phi^{3/2}}, \frac{d \sqrt{BM^3 \log \frac{1}{\delta}} D^5 \log \frac{1}{\xi}}{\rho^{1/4} \Phi^{5/2} \alpha^{3/2} \epsilon} \right\} \right).$$

### Proof of Theorem 4.3.4

By the relation between zCDP and  $(\epsilon, \delta)$ -DP, we can see that it suffices to show that Algorithm 4.3.28 is  $\phi$ -zCDP.

To do this, we will show that each iteration is  $\frac{\phi}{T}$ -zCDP. Then, by the composition theorem we know that the whole algorithm is  $\phi$ -zCDP.

We first show that Step 3 is  $\frac{\phi}{2T}$ -zCDP. This is due to the Lipschitz condition in Assumption 4.3.1 the fact that the  $\ell_2$ -norm sensitivity of  $L(w, D)$  is bounded by  $\frac{2G}{n}$ , and Gaussian mechanism.

Next, we show that Step 4 is also  $\frac{\phi}{2T}$ -zCDP. Since the symmetric matrix can be viewed as a  $\frac{p(p+1)}{2}$  dimensional vector, by the  $M$ -smooth property in Assumption 4.3.1 we know that the  $\ell_2$ -sensitivity of  $\nabla^2 L(w^k, D)$  is bounded by

$$\begin{aligned} & \|\nabla^2 L(w^k, D) - \nabla^2 L(w^k, D')\|_F \leq \\ & \sqrt{p} \|\nabla^2 L(w^k, D) - \nabla^2 L(w^k, D')\|_2 \leq \frac{2M\sqrt{p}}{n}. \end{aligned}$$

Thus by Gaussian mechanism, we know that adding noise  $H_k$  ensures that it is  $\frac{\phi}{2T}$ -zCDP.

Thus, by the composition theorem of zCDP we know that each iteration is  $\frac{\phi}{T}$ -zCDP.

### Proof of Theorem 4.3.5

Before giving the proof, we first introduce the following lemmas which show the concentration bounds of Gaussian distribution and Gaussian random matrices.

**Lemma 4.3.11** ([275]). For  $x \sim \mathcal{N}(0, \sigma^2 I_p)$ , with probability at least  $1 - \zeta$  for any  $1 > \zeta > 0$ ,

$$\|x\|_2 \leq \sqrt{2p}\sigma \log \frac{1}{\zeta}.$$

Let  $Z$  be a symmetric matrix whose upper triangle entries, including the diagonal, are i.i.d samples from  $\mathcal{N}(0, \sigma^2)$ . Then, we have, with probability at least  $1 - \frac{1}{p^c}$ , that  $\|Z\|_2 \leq C\sqrt{p}\sigma$ ,

where  $c, C$  are universal constants.

By Lemma 4.3.11 and the assumption on  $n$  we know that with probability at least  $1 - \xi - \frac{T}{p^c}$  we have for all  $k \in \{0, 1, \dots, T-1\}$ ,

$$\|\epsilon_t\|_2 \leq \sqrt{2p} \log \frac{T}{\zeta} \frac{2G\sqrt{T}}{n\sqrt{\phi}} \leq \frac{\alpha}{6} \quad (4.74)$$

$$\|H_t\|_2 \leq C \frac{2pM\sqrt{T}}{n\sqrt{\phi}} \leq \frac{\sqrt{\alpha\rho}}{3}. \quad (4.75)$$

In the following, we will assume that the above events (4.74) and (4.75) occur. From Assumption 4.3.1 we have

$$L(w^{k+1}, D) \leq L(w^k, D) + \langle \nabla L(w^k, D), h^k \rangle, \frac{1}{2} \langle \nabla^2 L(w^k, D) h^k, h^k \rangle + \frac{\rho}{6} \|h^k\|_2^3. \quad (4.76)$$

Plugging  $\tilde{\nabla} L(w^k, D)$  and  $\tilde{\nabla}^2 L(w^k, D)$  into (4.76) and by Cauchy-Schwartz inequality, we obtain

$$\begin{aligned} L(w^{k+1}, D) &\leq L(w^k, D) + \langle \tilde{\nabla} L(w^k, D), h^k \rangle + \|\epsilon_k\|_2 \|h^k\|_2 \\ &+ \frac{1}{2} \langle \tilde{\nabla}^2 L(w^k, D) h^k, h^k \rangle + \frac{\rho}{6} \|h^k\|_2^3 + \frac{1}{2} \|H_k\|_2 \|h^k\|_2^2. \end{aligned} \quad (4.77)$$

By (4.74), (4.75) and the fact that  $\|h^k\|_2 \leq r = \sqrt{\frac{\alpha}{\rho}}$ , we have

$$\|\epsilon_k\|_2 \|h^k\|_2 + \frac{1}{2} \|H_k\|_2 \|h^k\|_2^2 \leq \frac{1}{3} \frac{\alpha^{1.5}}{\sqrt{\rho}}. \quad (4.78)$$

By Lemma 4.3.1, we know that the optimality of  $h^k$  indicates that there exists dual variable

$\lambda^k \geq 0$  so that

$$\tilde{\nabla}L(w^k, D) + \tilde{\nabla}^2L(w^k, D)h^k + \lambda^k h^k = 0 \quad (4.79)$$

$$\tilde{\nabla}^2L(w^k, D) + \lambda^k I_p \succ 0 \quad (4.80)$$

$$\lambda^k(\|h^k\|_2 - r) = 0. \quad (4.81)$$

Thus by (4.79) we obtain

$$\langle \tilde{\nabla}L(w^k, D) + \tilde{\nabla}^2L(w^k, D)h^k + \lambda^k h^k, h^k \rangle = 0. \quad (4.82)$$

Also, by (4.80) we have

$$\langle \tilde{\nabla}^2L(w^k, D)h^k + \lambda^k h^k, h^k \rangle \geq 0. \quad (4.83)$$

Thus, from (4.82) and (4.83) we get

$$\langle \tilde{\nabla}L(w^k, D), h^k \rangle \leq 0. \quad (4.84)$$

Moreover, (4.81) indicates that  $\|h^k\| = r = \sqrt{\frac{\alpha}{\rho}}$  since we have  $\lambda^k > \sqrt{\alpha\rho} > 0$ .

Combining (4.84), (4.80), (4.78) with (4.77), we have

$$L(w^{k+1}, D) \leq L(w^k, D) - \frac{\lambda^k \alpha}{2\rho} + \frac{1}{3} \frac{\alpha^{1.5}}{\sqrt{\rho}}. \quad (4.85)$$

Thus, if  $\lambda^k > \sqrt{\alpha\rho}$  then we have

$$L(w^{k+1}, D) \leq L(w^k, D) - \frac{1}{6\sqrt{\rho}} \alpha^{1.5}.$$

Hence, we can see that  $\lambda^k \leq \sqrt{\alpha\rho}$  in no more than  $T = \frac{6\sqrt{\rho}\Delta}{\alpha^{1.5}}$  iterations.

We now show that when  $\lambda^k \leq \sqrt{\alpha\rho}$ ,  $w^{k+1}$  is an  $O(\alpha)$ -SOSP. The reason is the following.

From (4.79), we have

$$\|\tilde{\nabla}L(w^k, D) + \tilde{\nabla}^2L(w^k, D)h^k\|_2 = \lambda^k \sqrt{\frac{\alpha}{\rho}} \leq \alpha. \quad (4.86)$$

Thus, by events (4.74) and (4.75) we have

$$\begin{aligned} \|\nabla L(w^k, D) + \nabla^2 L(w^k, D)h^k\| &\leq \|\epsilon_k\|_2 + \|H_k h^k\|_2 + \|\tilde{\nabla}L(w^k, D) + \tilde{\nabla}^2L(w^k, D)h^k\| \\ &\leq \frac{\alpha}{6} + \frac{\alpha}{3} + \alpha = 1.5\alpha. \end{aligned}$$

On the other side, by using the property of  $\rho$ -Hessian Lipschitz we have

$$\|\nabla L(w^{k+1}, D) - \nabla L(w^k, D) - \nabla^2 L(w^k, D)h^k\|_2 \leq \frac{\rho}{2}\|h^k\|^2 = \frac{\alpha}{2}.$$

Thus, combining the above two inequalities we get  $\nabla L(w^{k+1}, D) \leq 2\alpha$ .

Moreover, by using the Hessian Lipschitz property and (4.80) we have

$$\begin{aligned} \nabla^2 L(w^{k+1}, D) &\succ \nabla^2 L(w^k, D) - \rho\|h^k\|_2 I_p \\ &\succ -H_k - \lambda^k I_p - \sqrt{\alpha\rho}I_p \\ &\succ -\frac{7}{3}\sqrt{\alpha\rho}I_p, \end{aligned}$$

which means that  $w^{k+1}$  is  $9\alpha$ -SOSP.

Thus, to satisfy (4.74) and (4.75),  $n$  only needs to satisfy

$$n \geq \Omega(\max\{\frac{\sqrt{p}\sqrt{T}G\ln\frac{1}{\zeta}}{\alpha\sqrt{\phi}}, \frac{pM\sqrt{T}}{\sqrt{\alpha\rho\phi}}\}).$$

Using Taylor series, we have  $\sqrt{\phi} = \sqrt{\epsilon + \ln\frac{1}{\delta}} - \sqrt{\ln\frac{1}{\delta}} = O(\frac{\epsilon}{\sqrt{\ln\frac{1}{\delta}}})$  [311]. Also since  $T = O(\frac{\sqrt{\rho}}{\alpha^{1.5}})$ , we get the proof.

### Proof of Theorem 4.3.6

By the Advanced Composition Theorem (Lemma 2.1.5), it is sufficient to show that each iteration is  $(\frac{\epsilon}{2\sqrt{2T \ln(2/\delta)}}, \frac{\delta}{2T})$ -DP.

We first show that Step 3 is  $(\epsilon', \delta') = (\frac{\epsilon}{4\sqrt{2T \ln(2/\delta)}}, \frac{\delta}{4T})$ -DP. To show this, we consider the mechanism  $\mathcal{A}(w, D) = \sum_{i=1}^n \nabla \ell(w, x_i) + \epsilon_1$ , where  $\epsilon_1 \sim \mathcal{N}(0, \frac{8L^2 \ln \frac{1.25}{\delta'}}{(\frac{n}{2|\mathcal{S}|}\epsilon')^2})$ . By the definition of Gaussian mechanism we can see that  $\mathcal{A}(w, D)$  is  $(\frac{n}{2|\mathcal{S}|}\epsilon', \delta')$ -DP. Thus, by the Sub-sampling Property (Lemma 2.1.2), we know that the sub-sampling version of  $\mathcal{A}$ , i.e.  $\mathcal{A}(w, \mathcal{S}) = \sum_{i \in \mathcal{S}} \nabla \ell(w, x_i) + \epsilon_1$ , is  $(2\frac{|\mathcal{S}|}{n} \cdot \frac{n}{2|\mathcal{S}|}\epsilon', \delta')$ -DP. Also, we note that  $\tilde{\nabla} L(w^k, \mathcal{S}) = \frac{1}{|\mathcal{S}|} \mathcal{A}(w^k, \mathcal{S})$ . Using the values of  $\epsilon'$  and  $\delta'$ , we can see that Step 2 is  $(\frac{\epsilon}{4\sqrt{2T \ln(2/\delta)}}, \frac{\delta}{4T})$ -DP.

Following the above argument and the ideas in the proof of Theorem 1, we can also show that Step 4 of Algorithm 4.3.29 is  $(\epsilon', \delta') = (\frac{\epsilon}{4\sqrt{2T \ln(2/\delta)}}, \frac{\delta}{4T})$ -DP. Thus, in total, we know that Algorithm 4.3.29 is  $(\epsilon, \delta)$ -DP.

### Proof of Theorem 4.3.7

Before giving the proof, we first recall the following lemma which shows the error bound of (6) and (7) w.r.t the full gradient and Hessian matrix, respectively.

**Lemma 4.3.12** (Theorem 7 and 8 in [182]). With probability at least  $1 - \zeta$ ,

$$\|\nabla L(w^k, \mathcal{S}) - \nabla L(w^k, D)\|_2 \leq O(G \sqrt{\frac{\ln(\frac{p}{\zeta})}{|\mathcal{S}|}}).$$

With probability at least  $1 - \zeta$ ,

$$\|\nabla^2 L(w^k, \mathcal{T}) - \nabla^2 L(w^k, D)\|_2 \leq O(M \sqrt{\frac{\ln \frac{p}{\zeta}}{|\mathcal{T}|}}).$$

The proof is almost the same as that of Theorem 4.3.5. By Lemmas 4.3.11 and 4.3.12 and the assumptions of  $n, |\mathcal{S}|, |\mathcal{T}|$ , we have, with probability at least  $1 - 3\zeta - \frac{T}{p^c}$ , that for

all  $k \in \{0, 1, \dots, T-1\}$ ,

$$\|\tilde{\nabla}L(w^k, \mathcal{S}) - \nabla L(w^k, D)\|_2 \leq O\left(\frac{L\sqrt{pT}\ln\frac{T}{\delta}}{n\epsilon} + G\sqrt{\frac{\ln(\frac{pT}{\zeta})}{|\mathcal{S}|}}\right) \leq \frac{\alpha}{6} \quad (4.87)$$

$$\|\tilde{\nabla}^2L(w^k, \mathcal{T}) - \nabla^2L(w^k, D)\|_2 \leq O\left(\frac{pM\sqrt{T}\ln\frac{T}{\delta}}{n\epsilon} + M\sqrt{\frac{\ln\frac{pT}{\zeta}}{|\mathcal{T}|}}\right) \leq \frac{\sqrt{\alpha\rho}}{3}. \quad (4.88)$$

In the following, we will assume that the events of (4.87) and (4.88) occur for all  $k$ . Let  $s_k = \tilde{\nabla}L(w^k, \mathcal{S}) - \nabla L(w^k, D)$  and  $t_k = \tilde{\nabla}^2L(w^k, \mathcal{T}) - \nabla^2L(w^k, D)$ .

From Assumption 4.3.1 we have

$$\begin{aligned} L(w^{k+1}, D) &\leq L(w^k, D) + \langle \nabla L(w^k, D), h^k \rangle \\ &\quad + \frac{1}{2} \langle \nabla^2 L(w^k, D) h^k, h^k \rangle + \frac{\rho}{6} \|h^k\|_2^3. \end{aligned} \quad (4.89)$$

Plugging  $\tilde{\nabla}L(w^k, D)$  and  $\tilde{\nabla}^2L(w^k, D)$  into (4.89) and by Cauchy-Schwartz inequality, we obtain

$$\begin{aligned} L(w^{k+1}, D) &\leq L(w^k, D) + \langle \tilde{\nabla}L(w^k, \mathcal{S}), h^k \rangle + \|s_k\|_2 \|h^k\|_2 \\ &\quad + \frac{1}{2} \langle \tilde{\nabla}^2L(w^k, \mathcal{T}) h^k, h^k \rangle + \frac{\rho}{6} \|h^k\|_2^3 + \frac{1}{2} \|t_k\|_2 \|h^k\|_2^2. \end{aligned} \quad (4.90)$$

By (4.87), (4.88) and  $\|h^k\|_2 \leq r = \sqrt{\frac{\alpha}{\rho}}$ , we have

$$\|s_k\|_2 \|h^k\|_2 + \frac{1}{2} \|t_k\|_2 \|h^k\|_2^2 \leq \frac{1}{3} \frac{\alpha^{1.5}}{\sqrt{\rho}}. \quad (4.91)$$

By Lemma 4.3.1 we know that the optimality of  $h^k$  indicates that there exists dual variable

$\lambda^k \geq 0$  so that

$$\tilde{\nabla}L(w^k, \mathcal{S}) + \tilde{\nabla}^2L(w^k, \mathcal{T})h^k + \lambda^k h^k = 0 \quad (4.92)$$

$$\tilde{\nabla}^2L(w^k, \mathcal{T}) + \lambda^k I_p \succ 0 \quad (4.93)$$

$$\lambda^k(\|h^k\|_2 - r) = 0. \quad (4.94)$$

Thus, by (4.92) we obtain

$$\langle \tilde{\nabla}L(w^k, \mathcal{S}) + \tilde{\nabla}^2L(w^k, \mathcal{T})h^k + \lambda^k h^k, h^k \rangle = 0. \quad (4.95)$$

Also, by (4.93) we have

$$\langle \tilde{\nabla}^2L(w^k, \mathcal{T})h^k + \lambda^k h^k, h^k \rangle \geq 0. \quad (4.96)$$

Hence, we get

$$\langle \tilde{\nabla}L(w^k, \mathcal{S}), h^k \rangle \leq 0. \quad (4.97)$$

Moreover, (4.81) indicates that  $\|h^k\| = r = \sqrt{\frac{\alpha}{\rho}}$ , since  $\lambda^k > \sqrt{\alpha\rho} > 0$ .

Combining (4.97), (4.93), (4.91) with (4.90), we have

$$L(w^{k+1}, D) \leq L(w^k, D) - \frac{\lambda^k}{2} \frac{\alpha}{\rho} + \frac{1}{3} \frac{\alpha^{1.5}}{\sqrt{\rho}}. \quad (4.98)$$

Thus, if  $\lambda^k > \sqrt{\alpha\rho}$ , we have

$$L(w^{k+1}, D) \leq L(w^k, D) - \frac{1}{6\sqrt{\rho}} \alpha^{1.5}.$$

This means that that  $\lambda^k \leq \sqrt{\alpha\rho}$  in no more than  $T = \frac{6\sqrt{\rho}\Delta}{\alpha^{1.5}}$  iterations.

We now show that when  $\lambda^k \leq \sqrt{\alpha\rho}$ ,  $w^{k+1}$  is an  $O(\alpha)$ -SOSP. This is due to the following

reasons. From (4.92), we have

$$\|\tilde{\nabla}L(w^k, \mathcal{S}) + \tilde{\nabla}^2L(w^k, \mathcal{T})h^k\|_2 = \lambda^k \sqrt{\frac{\alpha}{\rho}} \leq \alpha. \quad (4.99)$$

Thus, by events (4.87) and (4.88) we have

$$\begin{aligned} \|\nabla L(w^k, D) + \nabla^2L(w^k, D)h^k\| &\leq \|s_k\|_2 + \|t_k h^k\|_2 \\ &+ \|\tilde{\nabla}L(w^k, \mathcal{S}) + \tilde{\nabla}^2L(w^k, \mathcal{T})h^k\| \\ &\leq \frac{\alpha}{6} + \frac{\alpha}{3} + \alpha = 1.5\alpha. \end{aligned}$$

On the other side, by using the property of  $\rho$ -Hessian Lipschitz we have

$$\begin{aligned} \|\nabla L(w^{k+1}, D) - \nabla L(w^k, D) - \nabla^2L(w^k, D)h^k\|_2 \\ \leq \frac{\rho}{2} \|h^k\|^2 = \frac{\alpha}{2}. \end{aligned}$$

Combining the above two inequalities, we have  $\nabla L(w^{k+1}, D) \leq 2\alpha$ .

Moreover, by using the Hessian Lipschitz condition and (4.93) we have

$$\begin{aligned} \nabla^2L(w^{k+1}, D) &\succ \nabla^2L(w^k, D) - \rho\|h^k\|_2 I_p \\ &\succ -t_k - \lambda^k I_p - \sqrt{\alpha\rho}I_p \\ &\succ -\frac{7}{3}\sqrt{\alpha\rho}I_p, \end{aligned}$$

which means that  $w^{k+1}$  is  $9\alpha$ -SOSP.

# **Chapter 5**

## **Empirical Risk Minimization in Local Differential Privacy Model**

In Chapter 3 and 4 we studied Empirical Risk Minimization (ERM) in the central differential privacy model. While in the central model where data are managed by a trusted centralized entity which is responsible for collecting them and for deciding which differentially private data analysis to perform and to release (a classical use case for this model is the one of census data [140]). In the local model instead, each individual manages his/her proper data and discloses them to a server through some differentially private mechanisms. The server collects the (now private) data of each individual and combines them into a resulting data analysis. A classical use case for this model is the one aiming at collecting statistics from user devices like in the case of Google's Chrome browser [110], and Apple's iOS-10 [273].

As we mentioned in Chapter 2 In the local model, there are two basic kinds of protocols: interactive and non-interactive. [27] have recently investigated the power of non-interactive differentially private protocols. Because of its simplicity and its efficiency in term of network latency, this type of protocols seems to be more appealing for real world applications. Both Google and Apple use the non-interactive model in their projects [273, 110].

Despite being used in industry, the interactive local model has been much less studied

than the central one. Part of the reason for this is that there are intrinsic limitations in what one can do in the local model. As a consequence, many basic questions, that are well studied in the central model, have not been completely understood in the local model, yet.

In this chapter, we study Empirical Risk Minimization in the Local Differential Privacy model. In Section 5.1 we study ERM in the non-interactive model and show some negative results. To alleviate the exponential sample complexity issue, we then relax the classical non-interactive model where we allow the server has additional public but unlabeled data. We study the theoretical behaviors of Generalized Linear Models and Non-linear Regression in this relaxed model. Finally, we transfer our attention to the high dimensional ERM in LDP model. Specifically, we investigate the sparse linear regression problem in the interactive model.

## 5.1 ERM in Non-interactive LDP model

In this section, we first study Differentially Private Empirical Risk Minimization in the non-interactive local model. Before presenting our contributions and showing comparisons with previous works, for convenience and to be self-contained we first review definition of ERM problem, then we will discuss our motivations.

**Problem setting [257, 178]** Given a convex, closed and bounded constraint set  $\mathcal{C} \subseteq \mathbb{R}^p$ , a data universe  $\mathcal{D}$ , and a loss function  $\ell : \mathcal{C} \times \mathcal{D} \mapsto \mathbb{R}$ , and an  $n$ -size dataset  $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\} \in \mathcal{D}^n$  with data records  $\{x_i\}_{i=1}^n \subset \mathbb{R}^p$  and labels (responses)  $\{y_i\}_{i=1}^n \subset \mathbb{R}$  defines an *empirical risk* function:  $L(w; D) = \frac{1}{n} \sum_{i=1}^n \ell(w; x_i, y_i)$  (note that in some settings, such as mean estimation, there may not be separate labels). When the inputs are drawn i.i.d from an unknown underlying distribution  $\mathcal{P}$  on  $\mathcal{D}$ , we can also define the *population risk* function:  $L_{\mathcal{P}}(w) = \mathbb{E}_{D \sim \mathcal{P}^n} [\ell(w; D)]$ .

Thus, we have the following two types of excess risk measured at a particular output

$w_{\text{priv}}$ : The empirical risk,

$$\text{Err}_D(w_{\text{priv}}) = L(w_{\text{priv}}; D) - \min_{w \in \mathcal{C}} L(w; D),$$

and the population risk,

$$\text{Err}_{\mathcal{P}}(w_{\text{priv}}) = L_{\mathcal{P}}(w_{\text{priv}}) - \min_{w \in \mathcal{C}} L_{\mathcal{P}}(w).$$

The problem considered in this section is to design non-interactive LDP protocols that find a private estimator  $w_{\text{priv}}$  to minimize the empirical and/or population excess risks. Alternatively, we can express our goal on this problem in terms of *sample complexity*: find the smallest  $n$  for which we can design protocols that achieve error at most  $\alpha$  (in the worst case over data sets, or over generating distributions, depending on how we measure risk).

[97] first considered the worst-case error bounds for LDP convex optimization. For 1-Lipchitz convex loss functions over a bounded constraint set, they gave a highly interactive SGD-based protocol with sample complexity  $n = O(p/\epsilon^2\alpha^2)$ ; moreover, they showed that no LDP protocol which interacts with each player only once can achieve asymptotically better sample complexity, even for linear losses.

[257] considered the round complexity of LDP protocols for convex optimization. They observed that known methods perform poorly when constrained to be run non-interactively. They gave new protocols that improved on the state-of-the-art but nevertheless required sample complexity exponential in  $p$ . Specifically, they showed:

**Theorem 5.1.1** ([257]). Under some assumptions on the loss functions, there is a non-interactive  $\epsilon$ -LDP algorithm such that for all distribution  $\mathcal{P}$  on  $\mathcal{D}$ , with probability  $1 - \beta$ , its population risk is upper bounded by

$$\text{Err}_{\mathcal{P}}(w_{\text{priv}}) \leq \tilde{O}\left(\left(\frac{\sqrt{p} \log^2(1/\beta)}{\epsilon^2 n}\right)^{\frac{1}{p+1}}\right). \quad (5.1)$$

A similar result holds for empirical risk  $\text{Err}_D(w_{\text{priv}})$ . Equivalently, to ensure an error no more than  $\alpha$ , the sample complexity needs to be  $n = \tilde{O}(\sqrt{p}c^p\epsilon^{-2}\alpha^{-(p+1)})$ , where  $c$  is some constant (approximately 2).

Furthermore, lower bounds on the parallel query complexity of stochastic optimization (e.g., [228, 344]) mean that, for natural classes of LDP optimization protocols (based on the measure of noisy gradients), the exponential dependence of the sample size on the dimensionality  $p$  (in the terms of  $\alpha^{-(p+1)}$  and  $c^p$ ) is, in general, unavoidable [257].

This situation is somehow undesirable: when the dimensionality  $p$  is high and the target error is low, the dependency on  $\alpha^{-(p+1)}$  could make the sample size quite large. However, several results have already shown that for some specific loss functions, the exponential dependency on the dimensionality can be avoided. For example, [257] show that, in the case of linear regression, there is a non-interactive  $(\epsilon, \delta)$ -LDP algorithm whose sample complexity for achieving error at most  $\alpha$  in the empirical risk is  $n = O(p \log(1/\delta)\epsilon^{-2}\alpha^{-2})$ .<sup>1</sup> This indicates that there is a gap between the general case and some specific loss functions. This motivates us to consider the following basic question:

*Are there natural conditions on the loss function which allow for non-interactive  $\epsilon$ -LDP algorithms with sample complexity sub-exponentially (ideally, it should be polynomially or even linearly) depending on the dimensionality  $p$  in the terms of  $\alpha$  or  $c$ ?*

To answer this question, we make two attempts to approach the problem from different perspectives. In the first attempt, we show that the exponential dependency on  $p$  in the term of  $\alpha^{-(p+1)}$  can be avoided if the loss function is sufficiently smooth. In the second attempt, we show that there exists a family of loss functions whose sample complexities is depending on  $p$ . Below is a summary of our main contributions.

---

<sup>1</sup>Note that these two results are for non-interactive  $(\epsilon, \delta)$ -LDP, and we mainly focus on non-interactive  $\epsilon$ -LDP algorithms. Thus, we omit terms related to  $\log(1/\delta)$  in this section.

## Our Contributions:

1. In our first attempt, we investigate the conditions on the loss function guaranteeing a sample complexity which depends polynomially on  $p$  in the term of  $\alpha$ . We first show that by using Bernstein polynomial approximation, it is possible to achieve a non-interactive  $\epsilon$ -LDP algorithm in constant or low dimensions with the following properties. If the loss function is  $(8, T)$ -smooth (see Definition 5.1.4), then with a sample complexity of  $n = \tilde{O}((c_0 p^{\frac{1}{4}})^p \alpha^{-(2+\frac{p}{2})} \epsilon^{-2})$ , the excess empirical risk is ensured to be  $\text{Err}_D \leq \alpha$ . If the loss function is  $(\infty, T)$ -smooth, the sample complexity can be further improved to  $n = \tilde{O}(4^{p(p+1)} D_p^2 p \epsilon^{-2} \alpha^{-4})$ , where  $D_p$  depends only on  $p$ . Note that in the first case, the sample complexity is lower than the one in [257] when  $\alpha \leq O(\frac{1}{p})$ , and in the second case, the sample complexity depends only polynomially on  $\alpha^{-1}$ , instead of the exponential dependence as in [257]. Furthermore, our algorithm does not assume convexity for the loss function and thus can be applied to non-convex loss functions.
2. Then, we address the efficiency issue, which has only been partially studied in previous works [257]. Following an approach similar to [27], we propose an algorithm for our loss functions which has only 1-bit communication cost and  $O(1)$  computation cost for each client, and achieves asymptotically the same error bound as the original one. Additionally, we present a novel analysis for the server showing that if the loss function is convex and Lipschitz and the convex set satisfies some natural conditions, then there is an algorithm which achieves the error bound of  $O(p\alpha)$  and runs in polynomial time in  $\frac{1}{\alpha}$  (instead of exponential time as in [257]) if the loss function is  $(\infty, T)$ -smooth.
3. In our second attempt, we study the conditions on the loss function guaranteeing a sample complexity which depends polynomially on  $p$  (in both terms of  $\alpha$  and  $c$ ). We show that for any 1-Lipschitz generalized linear convex loss function, *i.e.*,

$\ell(w; x, y) = f(y_i \langle w, x_i \rangle)$  for some 1-Lipschitz convex function  $f$ , there is a non-interactive  $(\epsilon, \delta)$ -LDP algorithm, whose sample complexity for achieving error  $\alpha$  in empirical risk depends only linearly, instead of exponentially, on the dimensionality  $p$ . Our idea is based on results from Approximation Theory. We first consider the case of hinge loss functions. For this class of functions, we use Bernstein polynomials to approximate their derivative functions after smoothing, and then we apply the Stochastic Inexact Gradient Descent algorithm [102]. Next we extend the result to all convex general linear functions. The key idea is to show that any 1-Lipschitz convex function in  $\mathbb{R}$  can be expressed as a linear combination of some linear functions and hinge loss functions, *i.e.*, plus functions of inner product  $[\langle w, s \rangle]_+ = \max\{0, \langle w, s \rangle\}$ . Based on this, we propose a general method which is called the polynomial of inner product approximation.

4. Finally, we show the generality of our technique by applying polynomial approximation to other problems. Specifically, we give a non-interactive LDP algorithm for answering the class of  $k$ -way marginals queries, by using Chebyshev polynomial approximation, and a non-interactive LDP algorithm for answering the class of smooth queries, by using trigonometric polynomial approximation.

Table 5.1 shows the detailed comparisons between our results and the results in [257, 362].

### 5.1.1 Related Work

[178] initiated the study of learning under local differential privacy. Specifically, they showed a general equivalence between learning in the local model and learning in the statistical query model. [32] gave the first lower bounds for the accuracy of LDP protocols, for the special case of counting queries (equivalently, binomial parameter estimation).

The general problem of LDP convex risk minimization was first studied by [97], which

Methods	Sample Complexity	Assumption on the Loss Function
[257, Claim 4]	$\tilde{O}(4^p \alpha^{-(p+2)} \epsilon^{-2})$	1-Lipschitz
[257, Theorem 10]	$\tilde{O}(2^p \alpha^{-(p+1)} \epsilon^{-2})$	1-Lipschitz and Convex
[257]	$\Theta(p \epsilon^{-2} \alpha^{-2})$	Linear Regression
[362]	$O\left(p\left(\frac{8}{\alpha}\right)^4 \log \log(8/\alpha) \left(\frac{4}{\epsilon}\right)^{2c} \log(8/\alpha) + 2\left(\frac{1}{\alpha^2 \epsilon^2}\right)\right)$	Smooth Generalized Linear
<b>This Paper</b>	$\tilde{O}\left((c_0 p^{\frac{1}{4}})^p \alpha^{-(2+\frac{p}{2})} \epsilon^{-2}\right)$	(8, T)-smooth
<b>This Paper</b>	$\tilde{O}(4^{p(p+1)} D_p^2 \epsilon^{-2} \alpha^{-4})$	$(\infty, T)$ -smooth
<b>This Paper</b>	$p \cdot \left(\frac{C}{\alpha^3}\right)^{O(1/\alpha^3)} / \epsilon^{O(\frac{1}{\alpha^3})}$	Hinge Loss
<b>This Paper</b>	$p \cdot \left(\frac{C}{\alpha^3}\right)^{O(1/\alpha^3)} / \epsilon^{O(\frac{1}{\alpha^3})}$	1-Lipschitz Convex Generalized Linear

Table 5.1: Comparisons on the sample complexities for achieving error  $\alpha$  in the empirical risk, where  $c$  is a constant. We assume that  $\|x_i\|_2, \|y_i\| \leq 1$  for every  $i \in [n]$  and the constraint set  $\|\mathcal{C}\|_2 \leq 1$ . Asymptotic statements assume  $\epsilon, \delta, \alpha \in (0, 1/2)$  and ignore dependencies on  $\log(1/\delta)$ .

provided tight upper and lower bounds for a range of settings. Subsequent work considered a range of statistical problems in the LDP setting, providing upper and lower bounds—we omit a complete list here.

[257] initiated the study of the round complexity of LDP convex optimization, connecting it to the parallel complexity of (non-private) stochastic optimization.

Convex risk minimization in the *non-interactive* LDP model received considerable recent attentions [362, 257, 299] (see Table 5.1 for details). [257] first studied the problem with general convex loss functions and showed that the exponential dependence on the dimensionality is unavoidable for a class of non-interactive algorithms. In this paper, we investigate the conditions on the loss function that allow us to avoid the issue of exponential dependence on  $p$  in the sample complexity. [82] showed that an exponential lower bound on either the term of  $\frac{1}{\alpha}$  or the dimension  $p$  on the number of samples necessary to solve the standard task of learning a large-margin linear separator in the non-interactive LDP model.

The work most related to ours (*i.e.*, the second attempt) is that of [362], which also considered some specific loss functions in high dimensions, such as sparse linear regression

and kernel ridge regression. The major differences with our results are the following. Firstly, although they studied a similar class of loss functions (*i.e.*, Smooth Generalized Linear Loss functions) and used the polynomial approximation approach, their approach needs quite a few assumptions on the loss function in addition to the smoothness condition, such as Lipschitz smoothness and boundedness on the higher order derivative functions, which are clearly not satisfied by the hinge loss functions. Contrarily, our results only assume the 1-Lipschitz convex condition on the loss function. Secondly, even though the idea in our algorithm for the hinge loss functions is similar to theirs, we also consider generalized linear loss function by using techniques from approximation theory.

[185, 359] recently studied the problem of releasing k-way marginal queries in LDP. They compared different LDP methods to release marginal statistics, but did not consider methods based on polynomial approximation.

### 5.1.2 Preliminaries

---

#### Algorithm 5.1.30 1-dim LDP-AVG

---

```

1: Input: Player  $i \in [n]$  holding data  $v_i \in [0, b]$ , privacy parameter  $\epsilon$ .
2: for Each Player  $i$  do
3:   Send  $z_i = v_i + \text{Lap}(\frac{b}{\epsilon})$ 
4: end for
5: for The Server do
6:   Output  $a = \frac{1}{n} \sum_{i=1}^n z_i$ .
7: end for

```

---

Since we only consider non-interactive LDP through the section, we will use LDP as non-interactive LDP below.

As an example that will be useful throughout the paper, the next lemma shows a property of an  $\epsilon$ -LDP algorithm for computing 1-dimensional average.

**Lemma 5.1.1.** For any  $\epsilon > 0$ , Algorithm 5.1.30 is  $\epsilon$ -LDP. Moreover, if player  $i \in [n]$  holds value  $v_i \in [0, b]$  and  $n > \log \frac{2}{\beta}$  with  $0 < \beta < 1$ , then, with probability at least  $1 - \beta$ , the

output  $a \in \mathbb{R}$  satisfies:

$$|a - \frac{1}{n} \sum_{i=1}^n v_i| \leq \frac{2b\sqrt{\log \frac{2}{\beta}}}{\sqrt{n}\epsilon}.$$

**Bernstein polynomials and approximation** We give here some basic definitions that will be used in the sequel; more details can be found in [7, 203, 219].

**Definition 5.1.1.** Let  $k$  be a positive integer. The Bernstein basis polynomials of degree  $k$  are defined as  $b_{v,k}(x) = \binom{k}{v} x^v (1-x)^{k-v}$  for  $v = 0, \dots, k$ .

**Definition 5.1.2.** Let  $f : [0, 1] \mapsto \mathbb{R}$  and  $k$  be a positive integer. Then, the Bernstein polynomial of  $f$  of degree  $k$  is defined as  $B_k(f; x) = \sum_{v=0}^k f(v/k) b_{v,k}(x)$ . We denote by  $B_k$  the Bernstein operator  $B_k(f)(x) = B_k(f, x)$ .

Bernstein polynomials can be used to approximate some smooth functions over  $[0, 1]$ .

**Definition 5.1.3** ([219]). Let  $h$  be a positive integer. The iterated Bernstein operator of order  $h$  is defined as the sequence of linear operators  $B_k^{(h)} = I - (I - B_k)^h = \sum_{i=1}^h \binom{h}{i} (-1)^{i-1} B_k^i$ , where  $I = B_k^0$  denotes the identity operator and  $B_k^i$  is defined as  $B_k^i = B_k \circ B_k^{k-1}$ . The iterated Bernstein polynomial of order  $h$  can be computed as  $B_k^{(h)}(f; x) = \sum_{v=0}^k f(\frac{v}{k}) b_{v,k}^{(h)}(x)$ , where  $b_{v,k}^{(h)}(x) = \sum_{i=1}^h \binom{h}{i} (-1)^{i-1} B_k^{i-1}(b_{v,k}; x)$ .

Iterated Bernstein operator can well-approximate multivariate  $(h, T)$ -smooth functions.

**Definition 5.1.4** ([219]). Let  $h$  be a positive integer and  $T > 0$  be a constant. A function  $f : [0, 1]^p \mapsto \mathbb{R}$  is  $(h, T)$ -smooth if it is in class  $\mathcal{C}^h([0, 1]^p)$  and its partial derivatives up to order  $h$  are all bounded by  $T$ . We say it is  $(\infty, T)$ -smooth, if for every  $h \in \mathbb{N}$  it is  $(h, T)$ -smooth.<sup>2</sup>

Note that  $(h, T)$ -smoothness is incomparable with the Lipschitz smoothness. In  $(h, T)$ -smoothness, we assume it is smooth up to the  $h$ -th order while Lipschitz smooth is only for the first order, from this view,  $(h, T)$ -smoothness is stronger than the Lipschitz smoothness.

---

<sup>2</sup> $\mathcal{C}^h([0, 1]^p)$  means the class of functions that is  $h$ -th order smooth in the interval  $[0, 1]^p$ .

However, in Lipschitz smoothness we assume the gradient norm of the function will be bounded by some constant while  $(h, T)$ -smoothness assumes that each partial derivative (or each coordinate of the gradient) is bounded by some constant, so from this view Lipschitz smoothness is stronger than  $(h, T)$ -smoothness.

**Lemma 5.1.2** ([219]). If  $f : [0, 1] \mapsto \mathbb{R}$  is a  $(2h, T)$ -smooth function, then for all positive integers  $k$  and  $y \in [0, 1]$ , we have  $|f(y) - B_k^{(h)}(f; y)| \leq TD_h k^{-h}$ , where  $D_h$  is a constant independent of  $k, f$  and  $y$ .

The above lemma is for univariate functions, which has been extended to multivariate functions in [7].

**Definition 5.1.5.** Assume  $f : [0, 1]^p \mapsto \mathbb{R}$  and let  $k_1, \dots, k_p, h$  be positive integers. The multivariate iterated Bernstein polynomial of order  $h$  at  $y = (y_1, \dots, y_p)$  is defined as:

$$B_{k_1, \dots, k_p}^{(h)}(f; y) = \sum_{j=1}^p \sum_{v_j=0}^{k_j} f\left(\frac{v_1}{k_1}, \dots, \frac{v_p}{k_p}\right) \prod_{i=1}^p b_{v_i, k_i}^{(h)}(y_i). \quad (5.2)$$

We denote  $B_k^{(h)} = B_{k_1, \dots, k_p}^{(h)}(f; y)$  if  $k = k_1 = \dots = k_p$ .

**Lemma 5.1.3** ([7]). If  $f : [0, 1]^p \mapsto \mathbb{R}$  is a  $(2h, T)$ -smooth function, then for all positive integers  $k$  and  $y \in [0, 1]^p$ , we have

$$|f(y) - B_k^{(h)}(f; y)| \leq O(pTD_h k^{-h}).$$

Where  $D_h$  is a universal constant only related to  $h$ .

In the following, we will rephrase some basic definitions and lemmas on Chebyshev polynomial approximation.

**Definition 5.1.6.** The Chebyshev polynomials  $\{\mathcal{T}(x)_n\}_{n \geq 0}$  are recursively defined as follows

$$\mathcal{T}_0(x) \equiv 1, \mathcal{T}_1(x) \equiv x \text{ and } \mathcal{T}_{n+1}(x) = 2x\mathcal{T}_n(x) - \mathcal{T}_{n-1}(x).$$

It satisfies that for any  $n \geq 0$

$$\mathcal{T}_n(x) = \begin{cases} \cos(n \arccos(x)), & \text{if } |x| \leq 1 \\ \cosh(n \operatorname{narccosh}(x)), & \text{if } x \geq 1 \\ (-1)^n \cosh(n \operatorname{narccosh}(-x)), & \text{if } x \leq -1 \end{cases}$$

**Definition 5.1.7.** For every  $\rho > 0$ , let  $\Gamma_\rho$  be the ellipse  $\Gamma$  of foci  $\pm 1$  with major radius  $1 + \rho$ .

**Definition 5.1.8.** For a function  $f$  with a domain containing in  $[-1, 1]$ , its degree- $n$  Chebyshev truncated series is denoted by  $P_n(x) = \sum_{k=0}^n a_k \mathcal{T}_k(x)$ , where the coefficient  $a_k = \frac{2-1[k=0]}{\pi} \int_{-1}^1 \frac{f(x) \mathcal{T}_k(x)}{\sqrt{1-x^2}} dx$ .

**Lemma 5.1.4** (Cheybeshev Approximation Theorem [280]). Let  $f(z)$  be a function that is analytic on  $\Gamma_\rho$  and has  $|f(z)| \leq M$  on  $\Gamma_\rho$ . Let  $P_n(x)$  be the degree- $n$  Chebyshev truncated series of  $f(x)$  on  $[-1, 1]$ . Then, we have

$$\max_{x \in [-1, 1]} |f(x) - P_n(x)| \leq \frac{2M}{\rho + \sqrt{2\rho + \rho^2}} (1 + \rho + \sqrt{2\rho + \rho^2})^{-n},$$

$$|a_0| \leq M, \text{ and } |a_k| \leq 2M(1 + \rho + \sqrt{2\rho + \rho^2})^{-k}.$$

The following theorem shows the convergence rate of the Stochastic Inexact Gradient Method [102], which will be used in our algorithm. We first give the definition of inexact oracle (see Section 5.1.8 for the algorithm and general theory of SIGM).

**Definition 5.1.9.** For an objective function  $f$ , a  $(\gamma, \beta, \sigma)$  stochastic oracle returns a tuple

$(F_{\gamma,\beta,\sigma}(w; \xi), G_{\gamma,\beta,\sigma}(w; \xi))$  ( $\xi$  means the randomness in the algorithm) such that

$$\mathbb{E}_\xi[F_{\gamma,\beta,\sigma}(w; \xi)] = f_{\gamma,\beta,\sigma}(w),$$

$$\mathbb{E}_\xi[G_{\gamma,\beta,\sigma}(w; \xi)] = g_{\gamma,\beta,\sigma}(w),$$

$$\mathbb{E}_\xi[\|G_{\gamma,\beta,\sigma}(w; \xi) - g_{\gamma,\beta,\sigma}(w)\|_2^2] \leq \sigma^2,$$

$$0 \leq f(v) - f_{\gamma,\beta,\sigma}(w) - \langle g_{\gamma,\beta,\sigma}(w), v - w \rangle \leq \frac{\beta}{2}\|v - w\|^2 + \gamma, \forall v, w \in \mathcal{C}.$$

**Lemma 5.1.5** ([102]). Assume that  $f(w)$  is endowed with a  $(\gamma, \beta, \sigma)$  stochastic oracle with  $\beta \geq O(1)$ . Then, the sequence  $w_k$  generated by SIGM algorithm satisfies the following inequality

$$\mathbb{E}[f(w_k)] - \min_{w \in \mathcal{C}} f(w) \leq \Theta\left(\frac{\beta\sigma\|\mathcal{C}\|_2^2}{\sqrt{k}} + \gamma\right).$$

### 5.1.3 LDP-ERM with Smooth Loss Functions

In this section, we will mainly focus on reducing the sample complexity of  $\frac{1}{\alpha}$ . We first show that if the loss function is  $\infty$ -smooth (with some additional assumptions), then its sample complexity can be reduced to only polynomial in  $\frac{1}{\alpha}$  instead of exponential dependency in the previous paper. Then we talk about how to reduce the communication and computation cost for each user and also provide an algorithm which can let the server solve the problem more efficient.

In this section, we impose the following assumptions on the loss function.

**Assumption 1:** We let  $x$  denote  $(x, y)$  for simplicity unless specified otherwise. We assume that there is a constraint set  $\mathcal{C} \subseteq [0, 1]^p$  and for every  $x \in \mathcal{D}$  and  $w \in \mathcal{C}$ ,  $\ell(\cdot; x)$  is well defined on  $[0, 1]^p$  and  $\ell(w; x) \in [0, 1]$ . These closed intervals can be extended to arbitrarily bounded closed intervals.

Note that our assumptions are similar to the ‘Typical Settings’ in [257], where  $\mathcal{C} \subseteq [0, 1]^p$  appears in their Theorem 10, and  $\ell(w; x) \in [0, 1]$  from their 1-Lipschitz requirement and  $\|\mathcal{C}\|_2 \leq 1$ . We note that the above assumptions on  $x_i, y_i$  and  $\mathcal{C}$  are quite common for the

studies of LDP-ERM [257, 362].

### Basic Idea

Definition 5.1.5 and Lemma 5.1.3 tell us that if the value of the empirical risk function, *i.e.* the average of the sum of loss functions, is known at each of the grid points  $(\frac{v_1}{k}, \frac{v_2}{k} \dots \frac{v_p}{k})$ , where  $(v_1, \dots, v_p) \in \mathcal{T} = \{0, 1, \dots, k\}^p$  for some large  $k$ , then the function can be well approximated. Our main observation is that this can be done in the local model by estimating the average of the sum of loss functions at each of the grid points using Algorithm 5.1.30. This is the idea of Algorithm 5.1.31.

---

#### Algorithm 5.1.31 Local Bernstein Mechanism

---

- 1: **Input:** Player  $i \in [n]$  holds a data record  $x_i \in \mathcal{D}$ , public loss function  $\ell : [0, 1]^p \times \mathcal{D} \mapsto [0, 1]$ , privacy parameter  $\epsilon > 0$ , and parameter  $k$ .
  - 2: Construct the grid  $\mathcal{T} = \{\frac{v_1}{k}, \dots, \frac{v_p}{k}\}_{\{v_1, \dots, v_p\}}$ , where  $\{v_1, \dots, v_p\} \in \{0, 1, \dots, k\}^p$ .
  - 3: **for** Each grid point  $v = (\frac{v_1}{k}, \dots, \frac{v_p}{k}) \in \mathcal{T}$  **do**
  - 4:     **for** Each Player  $i \in [n]$  **do**
  - 5:         Calculate  $\ell(v; x_i)$ .
  - 6:     **end for**
  - 7:     Run Algorithm 5.1.30 with  $\epsilon = \frac{\epsilon}{(k+1)^p}$  and  $b = 1$  and denote the output as  $\tilde{L}(v; D)$ .
  - 8: **end for**
  - 9: **for** The Server **do**
  - 10:     Construct Bernstein polynomial, as in (5.2), based on the perturbed empirical loss function values  $\{\tilde{L}(v; D)\}_{v \in \mathcal{T}}$ . Denote  $\tilde{L}(\cdot; D)$  the corresponding function.
  - 11:     Compute  $w_{\text{priv}} = \arg \min_{w \in \mathcal{C}} \tilde{L}(w; D)$ .
  - 12: **end for**
- 

**Theorem 5.1.2.** For any  $\epsilon > 0$  and  $0 < \beta < 1$ , Algorithm 5.1.31 is  $\epsilon$ -LDP.<sup>3</sup> Assume that the loss function  $\ell(\cdot; x)$  is  $(2h, T)$ -smooth for all  $x \in \mathcal{D}$ , some positive integer  $h$  and constant  $T = O(1)$ . If the sample complexity  $n$  satisfies the condition of  $n = O\left(\frac{\log \frac{1}{\beta} 4^{p(h+1)}}{\epsilon^2 D_h^2}\right)$ , then by setting  $k = O\left(\left(\frac{D_h \sqrt{pn\epsilon}}{2^{(h+1)p} \sqrt{\log \frac{1}{\beta}}}\right)^{\frac{1}{h+p}}\right)$ , with probability at least  $1 - \beta$  we have:

$$\text{Err}_D(w_{\text{priv}}) \leq \tilde{O}\left(\frac{\log^{\frac{h}{2(h+p)}}\left(\frac{1}{\beta}\right) D_h^{\frac{p}{p+h}} p^{\frac{p}{2(h+p)}} 2^{(h+1)p} \frac{h}{h+p}}{n^{\frac{h}{2(h+p)}} \epsilon^{\frac{h}{h+p}}}\right), \quad (5.3)$$

---

<sup>3</sup>Note that we can use Advanced Composition Theorem in [104] to reduce the noise. For simplicity, we omit it here; the following algorithms are also the same.

where  $\tilde{O}$  hides the log and  $T$  terms.

From (5.3) we can see that in order to achieve error  $\alpha$ , the sample complexity needs to be

$$n = \tilde{O}(\log \frac{1}{\beta} D_h^{\frac{2p}{h}} p^{\frac{p}{h}} 4^{(h+1)p} \epsilon^{-2} \alpha^{-(2+\frac{2p}{h})}). \quad (5.4)$$

This implies the following special cases.

**Corollary 5.1.1.** If the loss function  $\ell(\cdot; x)$  is  $(8, T)$ -smooth for all  $x \in \mathcal{D}$  and some constant  $T$ , and  $n, \epsilon, \beta, k$  satisfy the condition in Theorem 5.1.2 with  $h = 4$ , then with probability at least  $1 - \beta$ , the sample complexity to achieve  $\alpha$  error is

$$n = \tilde{O}(\alpha^{-(2+\frac{p}{2})} \epsilon^{-2} (4^5 \sqrt{D_4} p^{\frac{1}{4}})^p).$$

Note that the sample complexity for general convex loss functions in [257] is  $n = \tilde{O}(\alpha^{-(p+1)} \epsilon^{-2} 2^p)$ , which is considerably worse than ours when  $\alpha \leq O(\frac{1}{p})$ , that is either in the low dimensional case or with high accuracy.

**Corollary 5.1.2.** If the loss function  $\ell(\cdot; x)$  is  $(\infty, T)$ -smooth for all  $x \in \mathcal{D}$  and some constant  $T$ , and  $n, \epsilon, \beta, k$  satisfy the condition in Theorem 5.1.2 with  $h = p$ , then with probability at least  $1 - \beta$ , the output  $w_{\text{priv}}$  of Algorithm 5.1.31 satisfies:

$$\text{Err}_D(w_{\text{priv}}) \leq \tilde{O}\left(\frac{\log \frac{1}{\beta} D_p^{\frac{1}{2}} p^{\frac{1}{4}} \sqrt{2}^{(p+1)p}}{n^{\frac{1}{4}} \epsilon^{\frac{1}{2}}}\right),$$

where  $\tilde{O}$  hides the log and  $T$  terms. Thus, to achieve error  $\alpha$ , with probability at least  $1 - \beta$ , the sample complexity needs to be

$$n = \tilde{O}\left(\max\left\{4^{p(p+1)} \log\left(\frac{1}{\beta}\right) D_p^2 p \epsilon^{-2} \alpha^{-4}, \frac{\log \frac{1}{\beta} 4^{p(p+1)}}{\epsilon^2 D_p^2}\right\}\right). \quad (5.5)$$

It is worth noticing that from (5.4) we can see that when the term  $\frac{h}{p}$  grows, the term  $\alpha$  decreases. Thus, for loss functions that are  $(\infty, T)$ -smooth, we can get a smaller dependency

than the term  $\alpha^{-4}$  in (5.5). For example, if we take  $h = 2p$ , then the sample complexity is  $n = O(\max\{c_2^{p^2} \log \frac{1}{\beta} D_{2p} \sqrt{p} \epsilon^{-2} \alpha^{-3}, \frac{\log \frac{1}{\beta} c^{p^2}}{\epsilon^2 D_{2p}^2}\})$  for some constants  $c, c_2$ . When  $h \rightarrow \infty$ , the dependency on the error becomes  $\alpha^{-2}$ , which is the optimal bound, even for convex functions.

Our analysis on the empirical excess risk does not use the convexity assumption. While this gives a bound which is not optimal, even for  $p = 1$ , it also says that our result holds for non-convex loss functions and constrained domain set, as long as they are smooth enough.

From (5.5), we can see that our sample complexity is lower than the one in [257] when  $\alpha \leq O(\frac{1}{16^p})$ . It is notable that this bound is less reasonable since in practice could be very large. However, there are still many cases where the condition still holds. For example, in low dimensional space to achieve the best performance for ERM, quite often the error is set to be extremely small, e.g.,  $\alpha = 10^{-10} \sim 10^{-14}$  [166].

Using the convexity assumption of the loss function, we can also give a bound on the population excess risk. Here we will show only the case of  $(\infty, T)$ , as the general case is basically the same.

**Theorem 5.1.3.** Under the conditions in Corollary 5.1.2, if we further assume that the loss function  $\ell(\cdot; x)$  is convex and 1-Lipschitz for all  $x \in \mathcal{D}$ , then with probability at least  $1 - 2\beta$ , we have:

$$\text{Err}_{\mathcal{P}}(w_{\text{priv}}) \leq \tilde{O}\left(\frac{(\sqrt{\log 1/\beta})^{\frac{1}{4}} D_p^{\frac{1}{4}} p^{\frac{1}{8}} \sqrt[4]{2}^{p(p+1)}}{\beta n^{\frac{1}{12}} \epsilon^{\frac{1}{4}}}\right).$$

That is, if we have sample complexity

$$n = \tilde{O}\left(\max\left\{\frac{\log \frac{1}{\beta} 4^{p(p+1)}}{\epsilon^2 D_p^2}, (\sqrt{\log 1/\beta})^3 D_p^3 p^{\frac{3}{2}} 8^{p(p+1)} \epsilon^{-3} \alpha^{-12} \beta^{-12}\right\}\right),$$

then  $\text{Err}_{\mathcal{P}}(w_{\text{priv}}) \leq \alpha$ .

Corollary 5.1.2 provides a partial answer to our motivational questions. That is, for loss functions which are  $(\infty, T)$ -smooth, there is an  $\epsilon$ -LDP algorithm for the empirical and population excess risks achieving error  $\alpha$  with sample complexity which is independent

of the dimensionality  $p$  in the term of  $\alpha$ . This result does not contradict the results in [257]. Indeed, the example used to show the unavoidable dependency between the sample complexity and  $\alpha^{-\Omega(p)}$ , to achieve an  $\alpha$  error, is actually non-smooth.

## More Efficient Algorithms

Algorithm 5.1.31 has computational time and communication complexity for each player which are exponential in the dimensionality. This is clearly problematic for every realistic practical application. For this reason, in this section, we investigate more efficient algorithms. For convenience, in this section we focus only on the case of  $(\infty, T)$ -smooth loss functions, but our results can easily be extended to more general cases.

We first consider the computational issue on the users side. The following lemma, shows an  $\epsilon$ -LDP algorithm (which is different from Algorithm 5.1.30) for efficiently computing  $p$ -dimensional average (notice the extra conditions on  $n$  and  $p$  compared with Lemma 5.1.1).

**Lemma 5.1.6** ([234]). Consider player  $i \in [n]$  holding data  $v_i \in \mathbb{R}^p$  with coordinate between 0 and  $b$ . Then for  $0 < \beta < 1$ ,  $0 < \epsilon$  such that  $n \geq 8p \log(\frac{8p}{\beta})$  and  $\sqrt{n} \geq \frac{12}{\epsilon} \sqrt{\log \frac{32}{\beta}}$ , there is an  $\epsilon$ -LDP algorithm, LDP-AVG, with probability at least  $1 - \beta$ , the output  $a \in \mathbb{R}^p$  satisfying<sup>4</sup>:

$$\max_{j \in [d]} |a_j - \frac{1}{n} \sum_{i=1}^n [v_i]_j| \leq O\left(\frac{bp}{\sqrt{n}\epsilon} \sqrt{\log \frac{p}{\beta}}\right).$$

Moreover, the computational cost for each user is  $O(1)$ .

By using Lemma 5.1.6 and by discretizing the grid with some interval steps, we can design an algorithm which requires  $O(1)$  computation time and  $O(\log n)$ -bits communication per player (see [234] for details; in Section 5.1.7 we have an algorithm with  $O(\log \log n)$ -bits communication per player). However, we would like to do even better and obtain constant communication complexity.

---

<sup>4</sup>Note that here we use a weak version of their result, one can get a finer analysis. For simplicity, we will omit it in the paper.

Instead of discretizing the grid, we apply a technique, proposed first by [27], which permits us to transform any ‘sampling resilient’  $\epsilon$ -LDP protocol into a protocol with 1-bit communication complexity (at the expense of increasing the shared randomness in the protocol). Roughly speaking, a protocol is sampling resilient if its output on any dataset  $S$  can be approximated well by its output on a random subset of half of the players.

Since our algorithm only uses the LDP-AVG protocol, we can show that it is indeed sampling resilient. Inspired by this result and the algorithm behind Lemma 5.1.6, we propose Algorithm 5.1.32 and obtain the following theorem.

**Theorem 5.1.4.** For any  $0 < \epsilon \leq \ln 2$  and  $0 < \beta < 1$ , Algorithm 5.1.32 is  $\epsilon$ -LDP. If the loss function  $\ell(\cdot; x)$  is  $(\infty, T)$ -smooth for all  $x \in \mathcal{D}$  and  $n = \tilde{O}(\max\{\frac{\log \frac{1}{\beta} 4^{p(p+1)}}{\epsilon^2 D_p^2}, p(k + 1)^p \log(k + 1), \frac{1}{\epsilon^2} \log \frac{1}{\beta}\})$ , then by setting  $k = O\left(\left(\frac{D_p \sqrt{pn\epsilon}}{2^{(p+1)p} \sqrt{\log \frac{1}{\beta}}}\right)^{\frac{1}{2p}}\right)$ , the results in Corollary 5.1.2 hold with probability at least  $1 - 4\beta$ . Moreover, for each player the time complexity is  $O(1)$ , and the communication complexity is 1-bit.

Now we study the algorithm from the server’s computational complexity perspective. The polynomial construction time complexity is  $O(n)$ , where the most inefficient part is finding  $w_{\text{priv}} = \arg \min_{w \in \mathcal{C}} \tilde{L}(w; D)$ . In fact, this function may be non-convex; but unlike general non-convex functions, it can be  $\alpha$ -uniformly approximated by the empirical loss function  $L(\cdot; D)$  if the loss function is convex (by the proof of Theorem 5.1.2), although we do not have access to the empirical risk function. Thus, we can see this problem as an instance of Approximately-Convex Optimization, which has been studied recently by [249]. Before doing that, we first give the definition of the condition on the constraint set.

**Definition 5.1.10** ([249]). We say that a convex set  $\mathcal{C}$  is  $\mu$ -well conditioned for  $\mu \geq 1$ , if there exists a function  $F : \mathbb{R}^p \mapsto \mathbb{R}$  such that  $\mathcal{C} = \{x | F(x) \leq 0\}$  and for every  $x \in \partial K : \frac{\|\nabla^2 F(x)\|_2}{\|\nabla F(x)\|_2} \leq \mu$ .

**Lemma 5.1.7** (Theorem 3.2 in [249]). Let  $\epsilon, \Delta$  be two real numbers such that  $\Delta \leq \max\{\frac{\epsilon^2}{\mu\sqrt{p}}, \frac{\epsilon}{p}\} \times \frac{1}{16348}$ . Then, there exists an algorithm  $\mathcal{A}$  such that for any given  $\Delta$

---

**Algorithm 5.1.32** Player-Efficient Local Bernstein Mechanism with 1-bit communication per player

---

- 1: **Input:** Player  $i \in [n]$  holds a data record  $x_i \in \mathcal{D}$ , public loss function  $\ell : [0, 1]^p \times \mathcal{D} \mapsto [0, 1]$ , privacy parameter  $\epsilon \leq \ln 2$ , and parameter  $k$ .
- 2: **Preprocessing:**
- 3: Generate  $n$  independent public strings
- 4:  $y_1 = \text{Lap}(\frac{1}{\epsilon}), \dots, y_n = \text{Lap}(\frac{1}{\epsilon})$ .
- 5: Construct the grid  $\mathcal{T} = \{\frac{v_1}{k}, \dots, \frac{v_p}{k}\}_{\{v_1, \dots, v_p\}}$ , where  $\{v_1, \dots, v_p\} \in \{0, 1, \dots, k\}^p$ .
- 6: Partition randomly  $[n]$  into  $d = (k+1)^p$  subsets  $I_1, I_2, \dots, I_d$ , and associate each  $I_j$  to a grid point  $\mathcal{T}(j) \in \mathcal{T}$ .
- 7: **for** Each Player  $i \in [n]$  **do**
- 8:     Find  $I_l$  such that  $i \in I_l$ . Calculate  $v_i = \ell(\mathcal{T}(l); x_i)$ .
- 9:     Compute  $p_i = \frac{1}{2} \frac{\Pr[v_i + \text{Lap}(\frac{1}{\epsilon}) = y_i]}{\Pr[\text{Lap}(\frac{1}{\epsilon}) = y_i]}$
- 10:     Sample a bit  $b_i$  from Bernoulli( $p_i$ ) and send it to the server.
- 11: **end for**
- 12: **for** The Server **do**
- 13:     **for**  $i = 1 \dots n$  **do**
- 14:         Check if  $b_i = 1$ , set  $\tilde{z}_i = y_i$ , otherwise  $\tilde{z}_i = 0$ .
- 15:     **end for**
- 16:     **for** each  $l \in [d]$  **do**
- 17:         Compute  $v_\ell = \frac{n}{|I_l|} \sum_{i \in I_\ell} \tilde{z}_i$
- 18:         Denote the corresponding grid point  $(\frac{v_1}{k}, \dots, \frac{v_p}{k}) \in \mathcal{T}$  of  $I_l$ , then denote  $\hat{L}((\frac{v_1}{k}, \dots, \frac{v_p}{k}); D) = v_\ell$ .
- 19:     **end for**
- 20:     Construct Bernstein polynomial for the perturbed empirical loss  $\{\hat{L}(v; D)\}_{v \in \mathcal{T}}$  as in Algorithm 5.1.31. Denote  $\tilde{L}(\cdot; D)$  the corresponding function.
- 21:     Compute  $w_{\text{priv}} = \arg \min_{w \in \mathcal{C}} \tilde{L}(w; D)$ .
- 22: **end for**

---

approximate convex function  $\tilde{f}$  over a  $\mu$ -well-conditioned convex set  $\mathcal{C} \subseteq \mathbb{R}^p$  of diameter 1 (that is, there exists a 1-Lipschitz convex function  $f : \mathcal{C} \mapsto \mathbb{R}$  such that for every  $x \in \mathcal{C}$ ,  $|f(x) - \tilde{f}(x)| \leq \Delta$ ),  $\mathcal{A}$  returns a point  $\tilde{x} \in \mathcal{C}$  with probability at least  $1 - \delta$  in time  $\text{Poly}(p, \frac{1}{\epsilon}, \log \frac{1}{\delta})$  and with the following guarantee:  $\tilde{f}(\tilde{x}) \leq \min_{x \in \mathcal{C}} \tilde{f}(x) + \epsilon$ .

Based on Lemma 5.1.7 (for  $\tilde{L}(w; D)$ ) and Corollary 5.1.2, and taking  $\epsilon = O(p\alpha)$ , we have the following.

**Theorem 5.1.5.** Under the conditions in Corollary 5.1.2, and assuming that  $n$  satisfies  $n = \tilde{O}(4^{p(p+1)} \log(1/\beta) D_p^2 p \epsilon^{-2} \alpha^{-4})$ , that the loss function  $\ell(\cdot; x)$  is 1-Lipschitz and convex for every  $x \in \mathcal{D}$ , that the constraint set  $\mathcal{C}$  is convex and  $\|\mathcal{C}\|_2 \leq 1$ , and satisfies  $\mu$ -well-condition property (see Definition 5.1.10), if the error  $\alpha$  satisfies  $\alpha \leq C \frac{\mu}{p\sqrt{p}}$  for some universal constant  $C$ , then there is an algorithm  $\mathcal{A}$  which runs in  $\text{Poly}(n, \frac{1}{\alpha}, \log \frac{1}{\beta})$  time for the server,<sup>5</sup> and with probability  $1 - 2\beta$  the output  $\tilde{w}_{\text{priv}}$  of  $\mathcal{A}$  satisfies  $\tilde{L}(\tilde{w}_{\text{priv}}; D) \leq \min_{w \in \mathcal{C}} \tilde{L}(w; D) + O(p\alpha)$ , which means that  $\text{Err}_D(\tilde{w}_{\text{priv}}) \leq O(p\alpha)$ .

Combining Theorem 5.1.5 with Corollary 5.1.2, and taking  $\alpha = \frac{\alpha}{p}$ , we have our final result:

**Theorem 5.1.6.** Under the conditions of Corollary 5.1.2, Theorem 5.1.4 and 5.1.5, for any  $C \frac{\mu}{\sqrt{p}} > \alpha > 0$ , if we further set

$$n = \tilde{O}(4^{p(p+1)} \log(1/\beta) D_p^2 p^5 \epsilon^{-2} \alpha^{-4}),$$

then there is an  $\epsilon$ -LDP algorithm, with  $O(1)$  running time and 1-bit communication per player, and  $\text{Poly}(\frac{1}{\alpha}, \log \frac{1}{\beta})$  running time for the server. Furthermore, with probability at least  $1 - 5\beta$ , the output  $\tilde{w}_{\text{priv}}$  satisfies  $\text{Err}_D(\tilde{w}_{\text{priv}}) \leq O(\alpha)$ .

Note that comparing with the sample complexity in Theorem 5.1.6 and Corollary 5.1.2, we have an additional factor of  $O(p^4)$ ; however, the  $\alpha$  terms are the same.

---

<sup>5</sup>Note that since here we assume  $n$  is at least exponential in  $p$ , thus the algorithm is not fully polynomial.

### 5.1.4 LDP-ERM with Convex Generalized Linear Loss Functions

In Section 5.1.3, we have seen that under the condition of  $(\infty, T)$ -smoothness for the loss function, the sample complexity can actually have polynomial dependence on  $p$  and  $\alpha$ . However, as shown in (5.5), there is still another exponential term  $c^{p^2}$  in the sample complexity that needs to be removed.

In this section, we show that if the loss function is generalized linear, the sample complexity for achieving error  $\alpha$  is only linear in the dimensionality  $p$ . We first give the assumptions that will be used throughout this section.

**Assumption 2:** We assume that  $\|x_i\|_2 \leq 1$  and  $|y_i| \leq 1$  for each  $i \in [n]$  and the constraint set  $\|\mathcal{C}\|_2 \leq 1$ . Unless specified otherwise, the loss function is assumed to be generalized linear, that is, the loss function  $\ell(w; x_i, y_i) \equiv f(y_i \langle x_i, w \rangle)$  for some 1-Lipschitz convex function  $f$ .

The generalized linear assumption holds for a large class of functions such as Generalized Linear Model and SVM. We also note that there is another definition for general linear functions,  $\ell(w; x, y) = f(\langle w, x \rangle, y)$ , which is more general than our definition. This class of functions has been studied in [177]; we leave as future research to extend our work to this class of loss functions.

#### Sample Complexity for Hinge Loss Function

We first consider LDP-ERM with hinge loss function and then extend the obtained result to general convex linear functions.

The hinge loss function is defined as  $\ell(w; x_i, y_i) = f(y_i \langle x_i, w \rangle) = [\frac{1}{2} - y_i \langle w, x_i \rangle]_+$ , where the plus function  $[x]_+ = \max\{0, x\}$ , i.e.,  $f(x) = \max\{0, \frac{1}{2} - x\}$  for  $x \in [-1, 1]$ .<sup>6</sup> Note that to avoid the scenario that  $1 - y_i \langle w, x_i \rangle$  is always greater than or equal to 0, we use  $\frac{1}{2}$ , instead of 1 as in the classical setting.

Before showing our idea, we first smooth the function  $f(x)$ . The following lemma shows

---

<sup>6</sup>The reader should think about about particular function  $f$ , not just a general  $f$ .

one of the smooth functions that is close to  $f$  in the domain of  $[-1, 1]$  (note that there are other ways to smooth  $f$ ; see [71] for details).

**Lemma 5.1.8.** Let  $f_\beta(x) = \frac{\frac{1}{2}-x+\sqrt{(\frac{1}{2}-x)^2+\beta^2}}{2}$  be a function with parameter  $\beta > 0$ . Then, we have

1.  $|f_\beta(x) - f(x)|_\infty \leq \frac{\beta}{2}, \forall x \in \mathbb{R}$ .
2.  $f_\beta(x)$  is 1-Lipschitz, that is,  $f'(x)$  is bounded by 1 for  $x \in \mathbb{R}$ .
3.  $f_\beta$  is  $\frac{1}{\beta}$ -smooth and convex.
4.  $f'_\beta(x)$  is  $(2, O(\frac{1}{\beta^2}))$ -smooth if  $\beta \leq 1$ .

The above lemma indicates that  $f_\beta(x)$  is a smooth and convex function which well approximates  $f(x)$ . This suggests that we can focus on  $f_\beta(y_i \langle w, x_i \rangle)$ , instead of  $f$ . Our idea is to construct a locally private  $(\gamma, \beta, \sigma)$  stochastic oracle for some  $\gamma, \beta, \sigma$  to approximate  $f'_\beta(y_i \langle w, x_i \rangle)$  in each iteration, and then run the SIGM step of [102]. By Lemma 5.1.8, we know that  $f'_\beta$  is  $(2, O(\frac{1}{\beta^2}))$ -smooth; thus, we can use Lemma 5.1.2 to approximate  $f'_\beta(x)$  via Bernstein polynomials.

Let  $P_d(x) = \sum_{i=0}^d c_i \binom{d}{i} x^i (1-x)^{d-i}$  be the  $d$ -th order Bernstein polynomial ( $c_i = f'_\beta(\frac{i}{d})$ , where  $\max_{x \in [-1, 1]} |P_d(x) - f'_\beta(x)| \leq \frac{\alpha}{4}$  (i.e.,  $d = c \frac{1}{\beta^2 \alpha}$  for some constant  $c > 0$ ). Then, we have  $\nabla_w \ell(w; x, y) = f'(y \langle w, x \rangle) y x^T$ , which can be approximated by  $[\sum_{i=0}^d c_i \binom{d}{i} (y \langle w, x \rangle)^i (1 - y \langle w, x \rangle)^{d-i}] y x^T$ . The idea is that if  $(y \langle w, x \rangle)^i, (1 - y \langle w, x \rangle)^{d-i}$  and  $y x^T$  can be approximated locally differentially privately by directly adding  $d+1$  numbers of independent Gaussian noises, which means it is possible to form an unbiased estimator of the term  $[\sum_{i=0}^d c_i \binom{d}{i} (y \langle w, x \rangle)^i (1 - y \langle w, x \rangle)^{d-i}] y x^T$ . The error of this procedure can be estimated by Lemma 5.1.5. Details of the algorithm are given in Algorithm 5.1.33.

**Theorem 5.1.7.** For each  $i \in [n]$ , the term  $G(w_t, i)$  generated by Algorithm 5.1.33 will be an  $(\frac{\alpha}{2}, \frac{1}{\beta}, O(\frac{d^{3d} C_4^d \sqrt{p}}{\epsilon^{2d+2}} + \alpha + 1))$  stochastic oracle (see Definition 5.1.9) for function

---

**Algorithm 5.1.33** Hinge Loss-LDP

---

```

1: Input: Player  $i \in [n]$  holds data  $(x_i, y_i) \in \mathcal{D}$ , where  $\|x_i\|_2 \leq 1, \|y_i\|_2 \leq 1$ ; privacy
   parameters  $\epsilon, \delta$ ;  $P_d(x) = \sum_{j=0}^d c_i \binom{d}{j} x^j (1-x)^{d-j}$  be the  $d$ -th order Bernstein polynomial
   for the function of  $f'_\beta$ , where  $c_i = f'_\beta(\frac{i}{d})$  and  $f_\beta(x)$  is the function in Lemma 5.1.8.
2: for Each Player  $i \in [n]$  do
3:   Calculate  $x_{i,0} = x_i + \sigma_{i,0}$  and  $y_{i,0} = y_i + z_{i,0}$ , where  $\sigma_{i,0} \sim \mathcal{N}(0, \frac{32 \log(1.25/\delta)}{\epsilon^2} I_p)$  and
       $z_{i,0} \sim \mathcal{N}(0, \frac{32 \log(1.25/\delta)}{\epsilon^2})$ .
4:   for  $j = 1, \dots, d(d+1)$  do
5:      $x_{i,j} = x_i + \sigma_{i,j}$ , where  $\sigma_{i,j} \sim \mathcal{N}(0, \frac{8 \log(1.25/\delta) d^2 (d+1)^2}{\epsilon^2} I_p)$ 
6:      $y_{i,j} = y_i + z_{i,j}$ , where  $z_{i,j} \sim \mathcal{N}(0, \frac{8 \log(1.25/\delta) d^2 (d+1)^2}{\epsilon^2})$ 
7:   end for
8:   Send  $\{x_{i,j}\}_{j=0}^{d(d+1)}$  and  $\{y_{i,j}\}_{j=0}^{d(d+1)}$  to the server.
9: end for
10: for the Server side do
11:   for  $t = 1, 2, \dots, n$  do
12:     Randomly sample  $i \in [n]$  uniformly.
13:     Set  $t_{i,0} = 1$ 
14:     for  $j = 0, \dots, d$  do
15:        $t_{i,j} = \prod_{k=jd+1}^{jd+j} y_{i,k} \langle w_t, x_{i,k} \rangle$  and  $t_{i,0} = 1$ 
16:        $s_{i,j} = \prod_{k=jd+j+1}^{jd+d} (1 - y_{i,k} \langle w_t, x_{i,k} \rangle)$  and  $s_{i,d} = 1$ 
17:     end for
18:     Denote  $G(w_t, i) = (\sum_{j=0}^d c_j \binom{d}{j} t_{i,j} s_{i,j}) y_{i,0} x_{i,0}^T$ .
19:     Update SIGM in [102] by  $G(w_t, i)$ 
20:   end for
21: end for
22: return  $w_n$ 

```

---

$L_\beta(w; D) = \frac{1}{n} \sum_{i=1}^n f_\beta(y_i \langle x_i, w \rangle)$ , where  $f_\beta$  is the function in Lemma 5.1.8, where  $C_4$  is some constant.

From Lemmas 5.1.8, 5.1.5 and Theorem 5.1.7, we have the following sample complexity bound for the hinge loss function under the non-interactive local model.

**Theorem 5.1.8.** For any  $\epsilon > 0$  and  $0 < \delta < 1$ , Algorithm 5.1.33 is  $(\epsilon, \delta)$  non-interactively locally differentially private.<sup>7</sup> Furthermore, for the target error  $\alpha$ , if we take  $\beta = \frac{\alpha}{4}$  and  $d = \frac{2}{\beta^2 \alpha} = O(\frac{1}{\alpha^3})$ . Then with the sample size  $n = \tilde{O}(\frac{d^{6d} C^d p}{\epsilon^{4d+4} \alpha^2})$ , the output  $w_n$  satisfies the

---

<sup>7</sup>Note that in the non-interactive local model,  $(\epsilon, \delta)$ -LDP is equivalent to  $\epsilon$ -LDP by using the protocol given in [49]; this allows us to omit the term of  $\delta$ .

following inequality

$$\mathbb{E}L(w_n, D) - \min_{w \in \mathcal{C}} L(w, D) \leq \alpha,$$

where  $C$  is some constant.

**Remark 5.1.1.** Note that the sample complexity bound in Theorem 5.1.8 is quite loose for parameters other than  $p$ . This is mainly due to the fact that we use only the basic composition theorem to ensure local differential privacy.<sup>8</sup> It is possible to obtain a tighter bound by using Advanced Composition Theorem [105] (this is the same for other algorithms in this section). Details of the improvement are omitted from this version. We can also extend to the population risk by the same algorithm, the main difference is that now  $G(w, i)$  is a  $(\frac{\alpha}{2}, \frac{1}{\beta}, O(\frac{d^{3d} C_4^d \sqrt{p}}{\epsilon^{2d+2}} + \alpha + 1))$  stochastic oracle, where  $\sigma^2 = \mathbb{E}_{(x,y) \sim \mathcal{P}} \|\ell(w; x, y) - \mathbb{E}_{(x,y) \sim \mathcal{P}} \ell(w; x, y)\|_2^2$ . For simplicity of presentation, we omitt the details here.

### Extension to Generalized Linear Convex Loss Functions

In this section, we extend our results for the hinge loss function to generalized linear convex loss functions  $L(w, D) = \frac{1}{n} \sum_{i=1}^n f(y_i \langle x_i, w \rangle)$  for any 1-Lipschitz convex function  $f$ .

One possible way (for the extension) is to follow the same approach used in previous section. That is, we first smooth the function  $f$  by some function  $f_\beta$ . Then, we use Bernstein polynomials to approximate the derivative function  $f'_\beta$ , and apply an algorithm similar to Algorithm 5.1.33. One of the main issues of this approach is that we do not know whether Bernstein polynomials can be directly used for every smooth convex function. Instead, we will use some ideas in approximation theory, which says that every 1-Lipschitz convex function can be expressed by a linear combination of the absolute value functions and linear functions.

To implement this approach, we first note that for the plus function  $f(x) \equiv \max\{0, x\}$ , by using Algorithm 5.1.33 we can get the same result as in Theorem 5.1.8. Since the

---

<sup>8</sup>There could be some improvement on the term of  $\frac{1}{\alpha}$  if we use advanced composition theorem. However, since the dependency of  $\frac{1}{\alpha}$  is already exponential, and it will be still exponential after the improvement. So here the improvement will be very incremental.

absolute value function  $|x| = 2 \max\{0, x\} - x$ , Theorem 5.1.8 clearly also holds for the absolute function. The following key lemma shows that every 1-dimensional 1-Lipschitz convex function  $f : [-1, 1] \mapsto [-1, 1]$  is contained in the convex hull of the set of absolute value and identity functions. We need to point out that [257] gave a similar lemma. Their proof is, however, somewhat incomplete and thus we give a complete one in this paper.

**Lemma 5.1.9.** Let  $f : [-1, 1] \mapsto [-1, 1]$  be a 1-Lipschitz convex function. If we define the distribution  $\mathcal{Q}$  which is supported on  $[-1, 1]$  as the output of the following algorithm:

1. first sample  $u \in [f'(-1), f'(1)]$  uniformly,
2. then output  $s$  such that  $u \in \partial f(s)$  (note that such an  $s$  always exists due to the fact that  $f$  is convex and thus  $f'$  is non-decreasing); if multiple number of such as  $s$  exist, return the maximal one,

then, there exists a constant  $c$  such that

$$\forall \theta \in [-1, 1], f(\theta) = \frac{f'(1) - f'(-1)}{2} \mathbb{E}_{s \sim \mathcal{Q}} |\theta - s| + \frac{f'(1) + f'(-1)}{2} \theta + c.$$

Using Lemma 5.1.9 and the ideas discussed in the previous section, we can now show that the sample complexity in Theorem 5.1.8 also holds for any general linear convex function. See Algorithm 5.1.34 for the details.

**Theorem 5.1.9.** Under Assumption 2, where the loss function  $\ell$  is  $\ell(w; x, y) = f(y \langle w, x \rangle)$  for any 1-Lipschitz convex function  $f$ , for any  $\epsilon, \delta \in (0, 1]$ , Algorithm 5.1.34 is  $(\epsilon, \delta)$  non-interactively differentially private. Moreover, given the target error  $\alpha$ , if we take  $\beta = \frac{\alpha}{4}$  and  $d = \frac{2}{\beta^2 \alpha} = O(\frac{1}{\alpha^3})$ . Then with the sample size  $n = \tilde{O}(\frac{d^{6d} C^d p}{\epsilon^{4d+4} \alpha^2})$ , the output  $w_n$  satisfies the following inequality

$$\mathbb{E} L(w_n, D) - \min_{w \in \mathcal{C}} L(w, D) \leq \alpha,$$

where  $C$  is some universal constant independent of  $f$ .

---

**Algorithm 5.1.34** General Linear-LDP

---

```

1: Input: Player  $i \in [n]$  holds raw data record  $(x_i, y_i) \in \mathcal{D}$ , where  $\|x_i\|_2 \leq 1$  and  $\|y_i\|_2 \leq 1$ ; privacy parameters  $\epsilon, \delta$ ;  $h_\beta(x) = \frac{x + \sqrt{x^2 + \beta^2}}{2}$  and  $P_d(x) = \sum_{j=0}^d c_j \binom{d}{j} x^j (1-x)^{d-j}$  is the  $d$ -th order Bernstein polynomial approximation of  $h'_\beta(x)$ . Loss function  $\ell$  can be represented by  $\ell(w; x, y) = f(y\langle w, x \rangle)$ .
2: for Each Player  $i \in [n]$  do
3:   Calculate  $x_{i,0} = x_i + \sigma_{i,0}$  and  $y_{i,0} = y_i + z_{i,0}$ , where  $\sigma_{i,0} \sim \mathcal{N}(0, \frac{32 \log(1.25/\delta)}{\epsilon^2} I_p)$  and  $z_{i,0} \sim \mathcal{N}(0, \frac{32 \log(1.25/\delta)}{\epsilon^2})$ 
4:   for  $j = 1, \dots, d(d+1)$  do
5:      $x_{i,j} = x_i + \sigma_{i,j}$ , where  $\sigma_{i,j} \sim \mathcal{N}(0, \frac{8 \log(1.25/\delta) d^2 (d+1)^2}{\epsilon^2} I_p)$ 
6:      $y_{i,j} = y_i + z_{i,j}$ , where  $z_{i,j} \sim \mathcal{N}(0, \frac{8 \log(1.25/\delta) d^2 (d+1)^2}{\epsilon^2})$ 
7:   end for
8:   Send  $\{x_{i,j}\}_{j=0}^{d(d+1)}$  and  $\{y_{i,j}\}_{j=0}^{d(d+1)}$  to the server.
9: end for
10: for the Server side do
11:   for  $t = 1, 2, \dots, n$  do
12:     Randomly sample  $i \in [n]$  uniformly.
13:     Randomly sample  $d(d+1)$  numbers of i.i.d  $s = \{s_k\}_{k=1}^{d(d+1)} \in [-1, 1]$  based on the distribution  $\mathcal{Q}$  in Lemma 5.1.9.
14:     Set  $t_{i,0} = 1$ 
15:     for  $j = 0, \dots, d$  do
16:        $t_{i,j} = \prod_{k=jd+1}^{jd+i} \left( \frac{y_{i,k} \langle w_t, x_{i,k} \rangle - s_k}{2} \right)$  and  $t_{i,0} = 1$ 
17:        $r_{i,j} = \prod_{k=jd+i+1}^{jd+d} \left( 1 - \frac{y_{i,k} \langle w_t, x_{i,k} \rangle - s_k}{2} \right)$  and  $r_{i,d} = 1$ 
18:     end for
19:     Denote  $G(w_t, i, s) = (f'(1) - f'(-1)) (\sum_{j=0}^d c_j \binom{d}{j} t_{i,j} r_{i,j}) y_{i,0} x_{i,0}^T + f'(-1)$ .
20:     Update SIGM in [102] by  $G(w_t, i, s)$ 
21:   end for
22: end for
23: return  $w_n$ 

```

---

**Remark 5.1.2.** The above theorem suggests that the sample complexity for any generalized linear loss function depends only linearly on  $p$ . However, there are still some not so desirable issues. Firstly, the dependence on  $\alpha$  is exponential, while we have already shown in the Section 5.1.3 that it is only polynomial (*i.e.*,  $\alpha^{-4}$ ) for sufficiently smooth loss functions. Secondly, the term of  $\epsilon$  is not optimal in the sample complexity, since it is  $\epsilon^{-\Omega(\frac{1}{\alpha^3})}$ , while the optimal one is  $\epsilon^{-2}$  [257]. We leave it as an open problem to remove the exponential dependency. Thirdly, the assumption on the loss function is that  $\ell(w; x, y) = f(y\langle w, x \rangle)$ ,

which includes the generalized linear models and SVM. However, as mentioned earlier, there is another slightly more general function class  $\ell(w; x, y) = f(\langle w, x \rangle, y)$  which does not always satisfy our assumption, *e.g.*, linear regression and  $\ell_1$  regression. For linear regression, we have already known its optimal bound  $\Theta(p\alpha^{-2}\epsilon^{-2})$ ; for  $\ell_1$  regression, we can use a method similar to Algorithm 5.1.33 to achieve a sample complexity which is linear in  $p$ . Thus, a natural question is whether the sample complexity is still linear in  $p$  for all loss functions  $\ell(w; x, y)$  that can be written as  $f(\langle w, x \rangle, y)$ .

We can see from Algorithm 5.1.33 and 5.1.34 that, both of the computation and communication cost of each user will be  $O(d^2) = O(\frac{1}{\alpha^6})$ . So, our question is, can we reduce these costs just as in the Section 5.1.3? We will leave it as future research.

There are still many open problems left. Firstly, as we showed in this paper, the  $\alpha$  term can be polynomial in the sample complexity when the loss function is smooth enough while the  $p$  term can be polynomial when the loss function is generalized linear. Thus, a natural question is to determine whether it is possible to get an algorithm whose sample complexity is fully polynomial in all the terms when the loss function is generalized linear and smooth enough, such as logistic regression. Secondly, although we have shown the advantages of these two methods, we do not know the practical performance of these methods.

Additional to the aforementioned improvements, another advantage of our method is that it can be extended to other LDP problems. Below we show how it can be used to answer the class of k-way marginals and smooth queries under LDP.

### 5.1.5 LDP Algorithms for Learning k-way Marginals Queries and Smooth Queries

In this section, we show further applications of our idea by giving LDP algorithms for answering sets of queries. All the queries considered in this section are linear, that is, of the form  $q_f(D) = \frac{1}{|D|} \sum_{x \in D} f(x)$  for some function  $f$ . It will be convenient to have a notion of

accuracy for the algorithm to be presented with respect to a set of queries. This is defined as follow:

**Definition 5.1.11.** Let  $\mathcal{Q}$  denote a set of queries. An algorithm  $\mathcal{A}$  is said to have  $(\alpha, \beta)$ -accuracy for size  $n$  databases with respect to  $\mathcal{Q}$ , if for every  $n$ -size dataset  $D$ , the following holds:  $\Pr[\exists q \in \mathcal{Q}, |\mathcal{A}(D, q) - q(D)| \geq \alpha] \leq \beta$ .

### k-way Marginals Queries

Now we consider a database  $D = (\{0, 1\}^p)^n$ , where each row corresponds to an individuals record. A marginal query is specified by a set  $S \subseteq [p]$  and a pattern  $t \in \{0, 1\}^{|S|}$ . Each such query asks: ‘What fraction of the individuals in  $D$  has each of the attributes set to  $t_j$ ?’. We will consider here k-way marginals which are the subset of marginal queries specified by a set  $S \subseteq [p]$  with  $|S| \leq k$ . K-way marginals could represent several statistics over datasets, including contingency tables, and the problem is to release them under differential privacy has been studied extensively in the literature [143, 138, 278, 119]. All these previous works have considered the central model of differential privacy, and only the recent work [185] studies this problem in the local model, while their methods are based on Fourier Transform. We now use the LDP version of Chebyshev polynomial approximation to give an efficient way of constructing a sanitizer for releasing k-way marginals.

Since learning the class of  $k$ -way marginals is equivalent to learning the class of monotone k-way disjunctions [143], we will only focus on the latter. The reason of why we can locally privately learning them is that they form a  $\mathcal{Q}$ -Function Family.

**Definition 5.1.12** ( $\mathcal{Q}$ -Function Family). Let  $\mathcal{Q} = \{q_y\}_{y \in Y_{\mathcal{Q}} \subseteq \{0, 1\}^m}$  be a set of counting queries on a data universe  $\mathcal{D}$ , where each query is indexed by an  $m$ -bit string. We define the index set of  $\mathcal{Q}$  to be the set  $Y_{\mathcal{Q}} = \{y \in \{0, 1\}^m | q_y \in \mathcal{Q}\}$ . We define a  $\mathcal{Q}$ -Function Family  $\mathcal{F}_{\mathcal{Q}} = \{f_{\mathcal{Q},x} : \{0, 1\}^m \mapsto \{0, 1\}\}_{x \in \mathcal{D}}$  as follows: for every data record  $x \in D$ , the function  $f_{\mathcal{Q},x} : \{0, 1\}^m \mapsto \{0, 1\}$  is defined as  $f_{\mathcal{Q},x}(y) = q_y(x)$ . Given a database  $D \in \mathcal{D}^n$ ,

we define  $f_{\mathcal{Q},D}(y) = \frac{1}{n} \sum_{i=1}^n f_{\mathcal{Q},x^i}(y) = \frac{1}{n} \sum_{i=1}^n q_y(x^i) = q_y(D)$ , where  $x^i$  is the  $i$ -th row of  $D$ .

This definition guarantees that  $\mathcal{Q}$ -function queries can be computed from their values on the individual's data  $x^i$ . We can now formally define the class of monotone k-way disjunctions.

**Definition 5.1.13.** Let  $\mathcal{D} = \{0, 1\}^p$ . The query set  $\mathcal{Q}_{disj,k} = \{q_y\}_{y \in Y_k \subseteq \{0,1\}^p}$  of monotone  $k$ -way disjunctions over  $\{0, 1\}^p$  contains a query  $q_y$  for every  $y \in Y_k = \{y \in \{0, 1\}^p \mid |y| \leq k\}$ . Each query is defined as  $q_y(x) = \vee_{j=1}^p y_j x_j$ . The  $\mathcal{Q}_{disj,k}$ -function family  $\mathcal{F}_{\mathcal{Q}_{disj,k}} = \{f_x\}_{x \in \{0,1\}^p}$  contains a function  $f_x(y_1, y_2, \dots, y_p) = \vee_{j=1}^p y_j x_j$  for each  $x \in \{0, 1\}^p$ .

Definition 5.1.13 guarantees that if we can uniformly approximate the function  $f_{\mathcal{Q},x}$  by polynomials  $p_x$ , then we can also have an approximation of  $f_{\mathcal{Q},D}$ , i.e., we can approximate  $q_y(D)$  for every  $y$  or all the queries in the class  $\mathcal{Q}$ . Thus, if we can locally privately estimate the sum of coefficients of the monomials for the  $m$ -multivariate functions  $\{p_x\}_{x \in D}$ , then we can uniformly approximate  $f_{\mathcal{Q},D}$ . Clearly, this can be done by Lemma 5.1.6, if the coefficients of the approximated polynomial are bounded.

In order to uniformly approximate the  $\mathcal{Q}_{disj,k}$ -function, we use Chebyshev polynomials.

**Definition 5.1.14** (Chebyshev Polynomials). For every  $k \in \mathbb{N}$  and  $\gamma > 0$ , there exists a univariate real polynomial  $p_k(x) = \sum_{i=0}^{t_k} c_i x^i$  of degree  $t_k$  such that  $t_k = O(\sqrt{k} \log(\frac{1}{\gamma}))$ ; for every  $i \in [t_k]$ ,  $|c_i| \leq 2^{O(\sqrt{k} \log(\frac{1}{\gamma}))}$ ; and  $p(0) = 0$ ,  $|p_k(x) - 1| \leq \gamma, \forall x \in [k]$ .

**Lemma 5.1.10** ([278]). For every  $k, p \in \mathbb{N}$ , such that  $k \leq p$ , and every  $\gamma > 0$ , there is a family of  $p$ -multivariate polynomials of degree  $t = O(\sqrt{k} \log(\frac{1}{\gamma}))$  with coefficients bounded by  $T = p^{O(\sqrt{k} \log(\frac{1}{\gamma}))}$ , which uniformly approximate the family  $\mathcal{F}_{\mathcal{Q}_{disj,k}}$  over the set  $Y_k$  (Definition 5.1.13) with error bound  $\gamma$ . That is, there is a family of polynomials  $\mathcal{P}$  such that for every  $f_x \in \mathcal{F}_{\mathcal{Q}_{disj,k}}$ , there is  $p_x \in \mathcal{P}$  which satisfies  $\sup_{y \in Y_k} |p_x(y) - f_x(y)| \leq \gamma$ .

By combining the ideas discussed above and Lemma 5.1.10, we have Algorithm 5.1.35 and the following theorem.

---

**Algorithm 5.1.35** Local Chebyshev Mechanism for  $\mathcal{Q}_{\text{disj},k}$ 


---

- 1: **Input:** Player  $i \in [n]$  holds a data record  $x_i \in \{0, 1\}^p$ , privacy parameter  $\epsilon > 0$ , error bound  $\alpha$ , and  $k \in \mathbb{N}$ .
  - 2: **for** Each Player  $i \in [n]$  **do**
  - 3:     Consider the  $p$ -multivariate polynomial  $q_{x_i}(y_1, \dots, y_p) = p_k(\sum_{j=1}^p y_j[x_i]_j)$ , where  $p_k$  is defined as in Lemma 5.1.10 with  $\gamma = \frac{\alpha}{2}$ .
  - 4:     Denote the coefficients of  $q_{x_i}$  as a vector  $\tilde{q}_i \in \mathbb{R}^{(p+t_k)}$  (since there are  $(p+t_k)$  coefficients in a  $p$ -variate polynomial with degree  $t_k$ ), note that each  $\tilde{q}_i$  can be seen as a  $p$ -multivariate polynomial  $q_{x_i}(y)$ .
  - 5: **end for**
  - 6: **for** The Server **do**
  - 7:     Run LDP-AVG from Lemma 5.1.1 on  $\{\tilde{q}_i\}_{i=1}^n \in \mathbb{R}^{(p+t_k)}$  with parameter  $\epsilon$ ,  $b = p^{O(\sqrt{k} \log(\frac{1}{\gamma}))}$ , denote the output as  $\tilde{q}_D \in \mathbb{R}^{(p+t_k)}$ , note that  $\tilde{q}_D$  also corresponds to a  $p$ -multivariate polynomial.
  - 8:     For each query  $y$  in  $\mathcal{Q}_{\text{disj},k}$  (seen as a  $d$  dimension vector), compute the  $p$ -multivariate polynomial  $\tilde{q}_D(y_1, \dots, y_p)$ .
  - 9: **end for**
- 

**Theorem 5.1.10.** For  $\epsilon > 0$  Algorithm 5.1.35 is  $\epsilon$ -LDP. Also, for  $0 < \beta < 1$ , there are constants  $C, C_1$  such that for every  $k, p, n \in \mathbb{N}$  with  $k \leq p$ , if

$$n = O\left(\max\left\{\frac{p^{C\sqrt{k} \log \frac{1}{\alpha}} \log \frac{1}{\beta}}{\epsilon^2 \alpha^2}, \frac{\log \frac{1}{\beta}}{\epsilon^2}, p^{C_1 \sqrt{k} \log \frac{1}{\alpha}} \log \frac{1}{\beta}\right\}\right),$$

this algorithm is  $(\alpha, \beta)$ -accurate with respect to  $\mathcal{Q}_{\text{disj},k}$ . The running time for each player is  $\text{Poly}(p^{O(\sqrt{k} \log \frac{1}{\alpha})})$ , and the running time for the server is at most  $O(n)$  and the time for answering a query is  $O(p^{C_2 \sqrt{k} \log \frac{1}{\alpha}})$  for some constant  $C_2$ . Moreover, as in Section 5.1.3, the communication complexity can be improved to 1-bit per player.

## Smooth Queries

We now consider the case where each player  $i \in [n]$  holds a data record in the continuous interval  $x_i \in [-1, 1]^p$  and we want to estimate the kernel density for a given point  $x_0 \in \mathbb{R}^p$ . A natural question is: If we want to estimate Gaussian kernel density of a given point  $x_0$  with many different bandwidths, can we do it simultaneously under  $\epsilon$  local differential privacy?

We can view this kind of queries as a subclass of the smooth queries. So, like in the case

---

**Algorithm 5.1.36** Local Trigonometry Mechanism for  $\mathcal{Q}_{C_T^h}$ 


---

- 1: **Input:** Player  $i \in [n]$  holds a data record  $x_i \in [-1, 1]^p$ , privacy parameter  $\epsilon > 0$ , error bound  $\alpha$ , and  $t \in \mathbb{N}$ .  $\mathcal{T}_t^p = \{0, 1, \dots, t-1\}^p$ . For a vector  $x = (x_1, \dots, x_p) \in [-1, 1]^p$ , denote operators  $\theta_i(x) = \arccos(x_i)$ ,  $i \in [p]$ .
  - 2: **for** Each Player  $i \in [n]$  **do**
  - 3:   **for** Each  $v = (v_1, v_2, \dots, v_p) \in \mathcal{T}_t^p$  **do**
  - 4:     Compute  $p_{i;v} = \cos(v_1 \theta_1(x_i)) \cdots \cos(v_p \theta_p(x_i))$
  - 5:   **end for**
  - 6:   Let  $p_i = (p_{i;v})_{v \in \mathcal{T}_t^p}$ .
  - 7: **end for**
  - 8: **for** The Server **do**
  - 9:   Run LDP-AVG from Lemma 5.1.1 on  $\{p_i\}_{i=1}^n \in \mathbb{R}^{t^p}$  with parameter  $\epsilon, b = 1$ , denote the output as  $\tilde{p}_D$ .
  - 10:   For each query  $q_f \in \mathcal{Q}_{C_T^h}$ . Let  $g_f(\theta) = f(\cos(\theta_1), \cos(\theta_2), \dots, \cos(\theta_p))$ .
  - 11:   Compute the trigonometric polynomial approximation  $p_t(\theta)$  of  $g_f(\theta)$ , where  $p_t(\theta) = \sum_{r=(r_1, r_2, \dots, r_p), \|r\|_\infty \leq t-1} c_r \cos(r_1 \theta_1) \cdots \cos(r_p \theta_p)$  as in (5.6). Denote the vector of the coefficients  $c \in \mathbb{R}^{t^p}$ .
  - 12:   Compute  $\tilde{p}_D \cdot c$ .
  - 13: **end for**
- 

of k-way marginals queries, we will give an  $\epsilon$ -LDP sanitizer for smooth queries. Now we consider the data universe  $\mathcal{D} = [-1, 1]^p$ , and dataset  $D \in \mathcal{D}^n$ . For a positive integer  $h$  and constant  $T > 0$ , we denote the set of all  $p$ -dimensional  $(h, T)$ -smooth function (Definition 5.1.4) as  $C_T^h$ , and  $\mathcal{Q}_{C_T^h} = \{q_f(D) = \frac{1}{n} \sum_{x \in D} f(D), f \in C_T^h\}$  the corresponding set of queries. The idea of the algorithm is similar to the one used for the k-way marginals; but instead of using Chebyshev polynomials, we will use trigonometric polynomials. We now assume that the dimensionality  $p$ ,  $h$  and  $T$  are constants so all the result in big  $O$  notation will be omitted. The idea of Algorithm 5.1.36 is based on the following Lemma.

**Lemma 5.1.11** ([340]). Assume  $\gamma > 0$ . For every  $f \in C_T^h$ , defined on  $[-1, 1]^p$ , let  $g_f(\theta_1, \dots, \theta_p) = f(\cos(\theta_1), \dots, \cos(\theta_p))$ , for  $\theta_i \in [-\pi, \pi]$ . Then there is an even trigonometric polynomial  $p$  whose degree for each variable is  $t(\gamma) = (\frac{1}{\gamma})^{\frac{1}{h}}$ :

$$p(\theta_1, \dots, \theta_p) = \sum_{0 \leq r_1, \dots, r_p < t(\gamma)} c_{r_1, \dots, r_p} \prod_{i=1}^p \cos(r_i \theta_i), \quad (5.6)$$

such that 1)  $p$   $\gamma$ -uniformly approximates  $g_f$ , i.e.  $\sup_{x \in [-\pi, \pi]^p} |p(x) - g_f(x)| \leq \gamma$ , 2) the coefficients are uniformly bounded by a constant  $M$  which only depends on  $h, T$  and  $p$ , 3) moreover, the entire set of the coefficients can be computed in time  $O\left(\left(\frac{1}{\gamma}\right)^{\frac{p+2}{h} + \frac{2p}{h^2}} \text{poly log } \frac{1}{\gamma}\right)$ .

By (5.6), we can see that all the  $p(x)$  which corresponds to  $g_f(x)$ , representing functions  $f \in C_T^h$ , have the same basis  $\prod_{i=1}^p \cos(r_i \theta_i)$ . So we can use Lemma 5.1.1 and 5.1.6 to estimate the average of the basis. Then, for each query  $f$  the server can only compute the corresponding coefficients  $\{c_{r_1, r_2, \dots, r_p}\}$ . This idea is implemented in Algorithm 5.1.36 for which we have the following result.

**Theorem 5.1.11.** For any  $\epsilon > 0$ , Algorithm 5.1.36 is  $\epsilon$ -LDP. Also for  $\alpha > 0$ ,  $0 < \beta < 1$ , if

$$n = O\left(\max\left\{\log^{\frac{5p+2h}{2h}}\left(\frac{1}{\beta}\right)\epsilon^{-2}\alpha^{-\frac{5p+2h}{h}}, \frac{1}{\epsilon^2} \log\left(\frac{1}{\beta}\right)\right\}\right)$$

and  $t = O((\sqrt{n}\epsilon)^{\frac{2}{5p+2h}})$ , then Algorithm 5.1.36 is  $(\alpha, \beta)$ -accurate with respect to  $\mathcal{Q}_{C_T^h}$ . Moreover, the time for answering each query is  $\tilde{O}((\sqrt{n}\epsilon)^{\frac{4p+4}{5p+2h} + \frac{4p}{5ph+2h^2}})$ , where  $O$  omits  $h, T, p$  and some log terms. For each player, the computation and communication cost could be improved to  $O(1)$  and 1 bit, respectively, as in Section 5.1.3.

## 5.1.6 Omitted Proofs

In this section, we provide the details of the omitted proofs for the theorems, lemmas, and corollaries stated in previous sections.

### Proof of Lemma 5.1.1

We first provide the following lemma:

**Lemma 5.1.12** ([234]). Suppose that  $x_1, \dots, x_n$  are i.i.d sampled from  $\text{Lap}(\frac{1}{\epsilon})$ . Then for every  $0 \leq t < \frac{2n}{\epsilon}$ , we have

$$\Pr\left(\left|\sum_{i=1}^n x_i\right| \geq t\right) \leq 2 \exp\left(-\frac{\epsilon^2 t^2}{4n}\right).$$

Consider Algorithm 5.1.30. We have  $|a - \frac{1}{n} \sum_{i=1}^n v_i| = |\frac{\sum_{i=1}^n x_i}{n}|$ , where  $x_i \sim \text{Lap}(\frac{b}{\epsilon})$ . Taking  $t = \frac{2\sqrt{n}\sqrt{\log \frac{2}{\beta}}}{\epsilon}$  and applying Lemma 5.1.12, we prove the lemma.

### Proof of Theorem 5.1.2

The proof of the  $\epsilon$ -LDP comes from Lemma 5.1.1 and the basic composition theorem of differential privacy. Without loss of generality, we assume that  $T=1$ .

To prove the theorem, it is sufficient to estimate  $\sup_{w \in \mathcal{C}} |\tilde{L}(w; D) - L(w; D)| \leq \alpha$  for some  $\alpha$ . Since if it is true, denoting  $w^* = \arg \min_{w \in \mathcal{C}} L(w; D)$ , we have  $L(w_{\text{priv}}; D) - L(w^*; D) \leq L(w_{\text{priv}}; D) - \tilde{L}(w_{\text{priv}}; D) + \tilde{L}(w_{\text{priv}}; D) - \tilde{L}(w^*; D) + \tilde{L}(w^*; D) - L(w^*; D) \leq L(w_{\text{priv}}; D) - \tilde{L}(w_{\text{priv}}; D) + \tilde{L}(w^*; D) - L(w^*; D) \leq 2\alpha$ .

Since we have

$$\sup_{w \in \mathcal{C}} |\tilde{L}(w; D) - L(w; D)| \leq \sup_{w \in \mathcal{C}} |\tilde{L}(w; D) - B_k^{(h)}(\hat{L}, w)| + \sup_{w \in \mathcal{C}} |B_k^{(h)}(\hat{L}, w) - L(w; D)|.$$

The second term is bounded by  $O(D_h p \frac{1}{k^h})$  by Lemma 5.1.3.

For the first term, by (5.2) and Algorithm 5.1.31, we have

$$\sup_{w \in \mathcal{C}} |\tilde{L}(w; D) - B_k^{(h)}(\hat{L}, w)| \leq \max_{v \in \mathcal{T}} |\tilde{L}(v; D) - \hat{L}(v; D)| \sup_{w \in \mathcal{C}} \sum_{j=1}^p \sum_{v_j=0}^k |\prod_{i=1}^p b_{v_i, k}^{(h)}(w_i)|. \quad (5.7)$$

By Proposition 4 in [7], we have

$$\sum_{j=1}^p \sum_{v_j=0}^k |\prod_{i=1}^p b_{v_i, k}^{(h)}(w_i)| \leq (2^h - 1)^p.$$

The following lemma bounds the term of  $\max_{v \in \mathcal{T}} |\tilde{L}(v; D) - L(v; D)|$ , which is obtained by Lemma 5.1.1.

**Lemma 5.1.13.** If  $0 < \beta < 1$ ,  $k$  and  $n$  satisfy the condition of  $n \geq p \log(2/\beta) \log(k+1)$ ,

then with probability at least  $1 - \beta$ , for each  $v \in \mathcal{T}$ , the following holds

$$|\tilde{L}(v; D) - L(v; D)| \leq O\left(\frac{\sqrt{\log \frac{1}{\beta}} \sqrt{p} \sqrt{\log(k)} (k+1)^p}{\sqrt{n}\epsilon}\right).$$

*Proof of Lemma 5.1.13.* By Lemma 5.1.1, for a fixed  $v \in \mathcal{T}$ , if  $n \geq \log \frac{2}{\beta}$ , we have, with probability  $1 - \beta$ ,  $|\tilde{L}(v; D) - L(v; D)| \leq \frac{2\sqrt{\log \frac{2}{\beta}}}{\sqrt{n}\epsilon}$ . Taking the union of all  $v \in \mathcal{T}$  and then taking  $\beta = \frac{\beta}{(k+1)^p}$  (since there are  $(k+1)^p$  elements in  $\mathcal{T}$ ) and  $\epsilon = \frac{\epsilon}{(k+1)^p}$ , we get the proof.  $\square$

By the fact that  $(k+1) < 2k$ , we have in total

$$\sup_{w \in \mathcal{C}} |\tilde{L}(w; D) - L(w; D)| \leq O\left(\frac{D_h p}{k^h} + \frac{2^{(h+1)p} \sqrt{\log \frac{1}{\beta}} \sqrt{p \log k} k^p}{\sqrt{n}\epsilon}\right). \quad (5.8)$$

Now, we take  $k = O\left(\frac{D_h \sqrt{pn\epsilon}}{2^{(h+1)p} \sqrt{\log \frac{1}{\beta}}}\right)^{\frac{1}{h+p}}$ . Since  $n = O\left(\frac{4^{p(h+1)}}{\epsilon^2 p D_h^2}\right)$ , we have  $\log k > 1$ . Plugging it into (5.8), we get

$$\begin{aligned} \sup_{w \in \mathcal{C}} |\tilde{L}(w; D) - L(w; D)| &\leq \tilde{O}\left(\frac{\log^{\frac{h}{2(h+p)}}\left(\frac{1}{\beta}\right) D_h^{\frac{p}{p+h}} p^{\frac{1}{2} + \frac{p}{2(h+p)}} 2^{(h+1)p} \frac{h}{h+p}}{\sqrt{h+pn}^{\frac{h}{2(h+p)}} \epsilon^{\frac{h}{h+p}}}\right) \\ &= \tilde{O}\left(\frac{\log^{\frac{h}{2(h+p)}}\left(\frac{1}{\beta}\right) D_h^{\frac{p}{p+h}} p^{\frac{p}{2(h+p)}} 2^{(h+1)p}}{n^{\frac{h}{2(h+p)}} \epsilon^{\frac{h}{h+p}}}\right). \end{aligned}$$

Also, we can see that  $n \geq p \log(2/\beta) \log(k+1)$  is true for  $n = O\left(\frac{4^{p(h+1)}}{\epsilon^2 p D_h^2}\right)$ . Thus, the theorem follows.

### Proof of Corollaries 5.1.1 and 5.1.2

Since the loss function is  $(\infty, T)$ -smooth, it is  $(2p, T)$ -smooth for all  $p$ . Thus, taking  $h = p$  in Theorem 5.1.2, we get the proof.

### Proof of Theorem 5.1.3

**Lemma 5.1.14.** [[252]] If the loss function  $\ell$  is L-Lipschitz and  $\mu$ -strongly convex, then with probability at least  $1 - \beta$  over the randomness of sampling the data set  $\mathcal{D}$ , the following is true,

$$\text{Err}_{\mathcal{P}}(\theta) \leq \sqrt{\frac{2L^2}{\mu}} \sqrt{\text{Err}_{\mathcal{D}}(\theta)} + \frac{4L^2}{\beta\mu n}.$$

For the general convex loss function  $\ell$ , we let  $\hat{\ell}(\theta; x) = \ell(\theta; x) + \frac{\mu}{2}\|\theta\|^2$  for some  $\mu > 0$ . Note that in this case the new empirical risk becomes  $\bar{L}(\theta; D) = \hat{L}(\theta; D) + \frac{\mu}{2}\|\theta\|^2$ . Since  $\frac{\mu}{2}\|\theta\|^2$  does not depend on the dataset, we can still use the Bernstein polynomial approximation for the original empirical risk  $\hat{L}(\theta; D)$  as in Algorithm 5.1.31, and the error bound for  $\bar{L}(\theta; D)$  is the same. Thus, we can get the population excess risk of the loss function  $\hat{\ell}$ ,  $\text{Err}_{\mathcal{P}, \hat{\ell}}(\theta_{\text{priv}})$  by Corollary 5.1.2 and have the following relation,

$$\text{Err}_{\mathcal{P}, \ell}(\theta_{\text{priv}}) \leq \text{Err}_{\mathcal{P}, \hat{\ell}}(\theta_{\text{priv}}) + \frac{\mu}{2}.$$

By Lemma 5.1.14 for  $\text{Err}_{\mathcal{P}, \hat{\ell}}(\theta_{\text{priv}})$ , where  $\hat{\ell}(\theta; x)$  is  $1 + \|\mathcal{C}\|_2 = O(1)$ -Lipschitz, we have the following,

$$\text{Err}_{\mathcal{P}, \ell}(\theta_{\text{priv}}) \leq \tilde{O}\left(\sqrt{\frac{2 \log^{\frac{1}{8}} \frac{1}{\beta} D_p^{\frac{1}{4}} p^{\frac{1}{8}} \sqrt[4]{2}^{(p+1)p}}{\mu n^{\frac{1}{8}} \epsilon^{\frac{1}{4}}}} + \frac{4}{\beta\mu n} + \frac{\mu}{2}\right).$$

Taking  $\mu = O(\frac{1}{\sqrt[12]{n}})$ , we get

$$\text{Err}_{\mathcal{P}, \ell}(\theta_{\text{priv}}) \leq \tilde{O}\left(\frac{\log^{\frac{1}{8}} \frac{1}{\beta} D_p^{\frac{1}{4}} p^{\frac{1}{8}} \sqrt[4]{2}^{(p+1)p}}{\beta n^{\frac{1}{12}} \epsilon^{\frac{1}{4}}}\right).$$

Thus, we have the theorem.

### Proof of Theorem 5.1.4

By [27] it is  $\epsilon$ -LDP. The time complexity and communication complexity is obvious. As in [27], it is sufficient to show that the LDP-AVG is sampling resilient.

The STAT in [27] corresponds to the average in our problem, and  $\phi(x, y)$  corresponds to  $\max_{j \in [p]} |[x]_j - [y]_j|$ . By Lemma 5.1.6, we can see that with probability at least  $1 - \beta$ ,

$$\phi(\text{Avg}(v_1, v_2, \dots, v_n); a) = O\left(\frac{bp}{\sqrt{n}\epsilon} \sqrt{\log \frac{p}{\beta}}\right).$$

Now let  $\mathcal{S}$  be the set obtained by sampling each point  $v_i, i \in [n]$  independently with probability  $\frac{1}{2}$ . Note that by Lemma 5.1.6, we have the subset  $\mathcal{S}$ . If  $|\mathcal{S}| \geq \Omega(\max\{p \log(\frac{p}{\beta}), \frac{1}{\epsilon^2} \log \frac{1}{\beta}\})$  with probability  $1 - \beta$ ,

$$\phi(\text{Avg}(\mathcal{S}); \text{LDP-AVG}(\mathcal{S})) = O\left(\frac{b\sqrt{p}}{\sqrt{|\mathcal{S}|}\epsilon} \sqrt{\log \frac{p}{\beta}}\right).$$

Now by Hoeffdings inequality, we can get  $|n/2 - |\mathcal{S}|| \leq \sqrt{n \log \frac{4}{\beta}}$  with probability  $1 - \beta$ .

Also since  $n = \Omega(\log \frac{1}{\beta})$ , we know that  $|\mathcal{S}| \geq O(n) \geq \Omega(p \log(\frac{p}{\beta}))$  is true. Thus, with probability at least  $1 - 2\beta$ ,  $\phi(\text{Avg}(\mathcal{S}); \text{LDP-AVG}(\mathcal{S})) = O\left(\frac{bp}{\sqrt{n}\epsilon} \sqrt{\log \frac{p}{\beta}}\right)$ .

Actually, we can also get  $\phi(\text{Avg}(\mathcal{S}); \text{Avg}(v_1, v_2, \dots, v_n)) \leq O\left(\frac{bp}{\sqrt{n}\epsilon} \sqrt{\log \frac{p}{\beta}}\right)$ . We now assume that  $v_i \in \mathbb{R}$ . Note that  $\text{Avg}(\mathcal{S}) = \frac{v_1x_1 + \dots + v_nx_n}{x_1 + \dots + x_n}$ , where each  $x_i \sim \text{Bernoulli}(\frac{1}{2})$ . Denote  $M = x_1 + x_2 + \dots + x_n$ . By Hoeffdings Inequality, we have with probability at least  $1 - \frac{\beta}{2}$ ,  $|M - \frac{n}{2}| \leq \sqrt{n \log \frac{4}{\beta}}$ . We further denote  $N = v_1x_1 + \dots + v_nx_n$ . Also, by Hoeffdings inequality, with probability at least  $1 - \beta$ , we get  $|N - \frac{v_1 + \dots + v_n}{2}| \leq b\sqrt{n \log \frac{2}{\beta}}$ . Thus, with probability at least  $1 - \beta$ , we have:

$$\begin{aligned} \left| \frac{N}{M} - \frac{v_1 + \dots + v_n}{n} \right| &\leq \frac{|N - \sum_{i=1}^n v_i/2|}{M} + \left| \sum_{i=1}^n v_i/2 \right| \left| \frac{1}{M} - \frac{2}{n} \right| \\ &\leq \frac{|N - \sum_{i=1}^n v_i/2|}{M} + \frac{nb}{2} \left| \frac{1}{M} - \frac{2}{n} \right|. \end{aligned} \tag{5.9}$$

For the second term of (5.9),  $|\frac{1}{M} - \frac{2}{n}| = \frac{|n/2 - M|}{M^{\frac{n}{2}}}$ . We know from the above  $|n/2 - M| \leq \sqrt{n \log \frac{4}{\beta}}$ . Also since  $n = \Omega(\log \frac{1}{\beta})$ , we get  $M \geq O(n)$ . Thus,  $|\frac{1}{M} - \frac{2}{n}| \leq O(\frac{\sqrt{\log \frac{1}{\beta}}}{\sqrt{nn}})$ . The upper bound of the second term is  $O(\frac{b\sqrt{\log \frac{1}{\beta}}}{\sqrt{n}})$ , and the same for the first term. For  $p$  dimensions, we just choose  $\beta = \frac{\beta}{p}$  and take the union. Thus in total we have  $\phi(\text{Avg}(\mathcal{S}); \text{Avg}(v_1, v_2, \dots, v_n)) \leq O(\frac{b}{\sqrt{n}\epsilon} \sqrt{\log \frac{p}{\beta}}) \leq O(\frac{bp}{\sqrt{n}\epsilon} \sqrt{\log \frac{p}{\beta}})$ .

In summary, we have shown that

$$\phi(\text{AVG-LDP}(\mathcal{S}); \text{Avg}(v_1, v_2, \dots, v_n)) \leq O(\frac{bp}{\sqrt{n}\epsilon} \sqrt{\log \frac{p}{\beta}})$$

with probability at least  $1 - 4\beta$ .

### Proof of Theorem 5.1.5

Let  $\theta^* = \arg \min_{\theta \in \mathcal{C}} L(\theta; D)$ ,  $\theta_{\text{priv}} = \arg \min_{\theta \in \mathcal{C}} \tilde{L}(\theta; D)$ . Under the assumptions of  $\alpha, n, k, \epsilon, \beta$ , we know from the proof of Theorem 5.1.2 and Corollary 5.1.2 that  $\sup_{\theta \in \mathcal{C}} |\tilde{L}(\theta; D) - L(\theta; D)| \leq \alpha$ . Also by setting  $\epsilon = 16348p\alpha$  and  $\alpha \leq \frac{1}{16348} \frac{\mu}{p\sqrt{p}}$ , we can see that the condition in Lemma 5.1.7 holds for  $\Delta = \alpha$ . So there is an algorithm whose output  $\tilde{\theta}_{\text{priv}}$  satisfies

$$\tilde{L}(\tilde{\theta}_{\text{priv}}; D) \leq \min_{\theta \in \mathcal{C}} \tilde{L}(\theta; D) + O(p\alpha).$$

Thus, we have

$$L(\tilde{\theta}_{\text{priv}}; D) - L(\theta^*; D) \leq L(\tilde{\theta}_{\text{priv}}; D) - \tilde{L}(\theta_{\text{priv}}; D) + \tilde{L}(\theta_{\text{priv}}; D) - L(\theta^*; D),$$

where

$$\begin{aligned} L(\tilde{\theta}_{\text{priv}}; D) - \tilde{L}(\theta_{\text{priv}}; D) &\leq L(\tilde{\theta}_{\text{priv}}; D) - \tilde{L}(\tilde{\theta}_{\text{priv}}; D) + \tilde{L}(\tilde{\theta}_{\text{priv}}; D) - \tilde{L}(\theta_{\text{priv}}; D) \\ &\leq O(p\alpha). \end{aligned}$$

Also  $\tilde{L}(\theta_{\text{priv}}; D) - \hat{L}(\theta^*; D) \leq \tilde{L}(\theta^*; D) - \hat{L}(\theta^*; D) \leq \alpha$ . Thus, the theorem follows. The running time is determined by  $n$ . This is because when we use the algorithm in Lemma 5.1.7, we have to use the first order optimization. That is, we have to evaluate some points at  $\tilde{L}(\theta; D)$ , which will cost at most  $O(\text{Poly}(n, \frac{1}{\alpha}))$  time (note that  $\tilde{L}$  is a polynomial with  $(k+1)^p \leq n$  coefficients).

### Proof of Lemma 5.1.8

It is easy to see that items 1 is true. Item 2 is due to the following  $|f'_\beta(x)| = \left| \frac{-1 + \frac{x-\frac{1}{2}}{\sqrt{(x-\frac{1}{2})^2 + \beta^2}}}{2} \right| \leq$

1. Item 3 is because of the following  $0 \leq f''_\beta(x) = \frac{\beta^2}{((x-\frac{1}{2})^2 + \beta^2)^{\frac{3}{2}}} \leq \frac{1}{\beta}$ . For item 4 we have  $|f_\beta^{(3)}(x)| = \frac{3\beta^2 x}{(x^2 + \beta^2)^{\frac{5}{2}}} \leq \frac{3}{\beta^2}$ .

### Proof of Theorem 5.1.7

For simplicity, we omit the term of  $\delta$ , which will not affect the linear dependency. Let

$$\hat{G}(w, i) = [\sum_{j=0}^d c_j \binom{d}{j} (y_i \langle w, x_i \rangle)^j (1 - y_i \langle w, x_i \rangle)^{d-j}] y_i x_i^T,$$

where  $c_j = f'_\beta(\frac{j}{d})$  and

$$\mathbb{E}_i \hat{G}(w, i) = \frac{1}{n} \sum_{i=1}^n \hat{G}(w, i) = \hat{G}(w).$$

For the term of  $G(w, i)$ , the randomness comes from sampling the index  $i$  and the Gaussian noises added for preserving local privacy.

Note that in total  $\mathbb{E}_{\sigma, z, i} G(w, i) = \hat{G}(w)$ , where  $\sigma = \{\sigma_{i,j}\}_{j=0}^{\frac{d(d+1)}{2}}$  and  $z = \{z_{i,j}\}_{j=0}^{\frac{d(d+1)}{2}}$ .

It is easy to see that  $\mathbb{E}_{\sigma, z} G(w, i) = \mathbb{E}[(\sum_{j=0}^d c_j \binom{d}{j} t_{i,j} s_{i,j}) y_{i,0} x_{i,0}^T \mid i] = \hat{G}(w, i)$ , which is due to the fact that  $\mathbb{E} t_{i,j} = (y_i \langle w, x_i \rangle)^j$ ,  $\mathbb{E} s_{i,j} = (1 - y_i \langle w, x_i \rangle)^{d-j}$  and each  $t_{i,j}, s_{i,j}$  is independent. We now calculate the variance for this term with fixed  $i$ . Firstly, we have

$\text{Var}(y_{i,0}x_{i,0}^T) = O(\frac{p}{\epsilon^4})$ . For each  $t_{i,j}$ , we get

$$\text{Var}(t_{i,j}) \leq \prod_{k=jd+1}^{jd+j} \text{Var}(y_{i,k})(\text{Var}(\langle w_i, x_{i,k} \rangle) + (\mathbb{E}(w_i^T x_{i,k}))^2) \leq \tilde{O}\left((C_1 \frac{d(d+1)}{\epsilon^2})^{2j}\right).$$

and similarly we have

$$\text{Var}(s_{i,j}) \leq \tilde{O}\left((C_2 \frac{d(d+1)}{\epsilon^2})^{2(d-j)}\right).$$

Thus we have

$$\text{Var}(t_{i,j}s_{i,j}) \leq \tilde{O}\left((C_3 \frac{d(d+1)}{\epsilon^2})^{2d}\right).$$

Since function  $f'_\beta$  is bounded by 1 and  $\binom{d}{j} \leq d^d$  for each  $j$ . In total, we have

$$\text{Var}(G(w_t, i)|i) \leq O(d \cdot d^d \cdot (C_3 \frac{d(d+1)}{\epsilon^2})^{2d} \cdot \frac{p}{\epsilon^4}) = \tilde{O}\left(\frac{d^{6d} C^d p}{\epsilon^{4d+4}}\right).$$

Next we consider  $\text{Var}(\hat{G}(w, i))$ . Since

$$\begin{aligned} \|\hat{G}(w, i) - f'_\beta(y_i x_i^T w) y_i x_i^T\|_2^2 &= \|[\sum_{j=0}^d c_j \binom{d}{j} (y_i \langle w, x_i \rangle)^j (1 - y_i \langle w, x_i \rangle)^{d-j} - f'_\beta(w)] y_i x_i^T\|_2^2 \\ &\leq (\frac{1}{\beta^2 d})^2 \leq \frac{\alpha^2}{4}, \end{aligned}$$

we get

$$\begin{aligned} \text{Var}(\hat{G}(w, i)) &\leq O(\mathbb{E}[\|\hat{G}(w, i) - f'_\beta(y_i x_i^T w) y_i x_i^T\|_2^2] + \mathbb{E}[\|\hat{G}(w) - \nabla L_\beta(w; D)\|_2^2] \\ &\quad + \mathbb{E}[\|f'_\beta(y_i x_i^T w) y_i x_i^T - \nabla L_\beta(w; D)\|_2^2]) \leq O((\alpha + 1)^2). \end{aligned}$$

In total, we have  $\mathbb{E}[\|G(w, i) - \hat{G}(w)\|_2^2] \leq \mathbb{E}[\|G(w, i) - \hat{G}(w, i)\|_2^2] + \mathbb{E}[\|\hat{G}(w, i) - \hat{G}(w)\|_2^2] \leq \tilde{O}\left((\frac{d^{3d} C_4^d \sqrt{p}}{\epsilon^{2d+2}} + \alpha + 1)^2\right)$ .

Also, we know that

$$\begin{aligned}
& L_\beta(v; D) - L_\beta(w; D) - \langle \hat{G}(w), v - w \rangle = \\
& L_\beta(v; D) - L_\beta(w; D) - \langle \nabla L_\beta(w; D), v - w \rangle + \langle \nabla L_\beta(w; D) - G(w), v - w \rangle \\
& \leq \frac{1}{2\beta} \|v - w\|_2^2 + \frac{\alpha}{2},
\end{aligned}$$

since  $L_\beta$  is  $\frac{1}{\beta}$ -smooth and  $|\langle \nabla L_\beta(w) - G(w), v - w \rangle| \leq \frac{\alpha}{2}$ .

Thus,  $G(w, i)$  is an  $(\frac{\alpha}{2}, \frac{1}{\beta}, O(\frac{d^{3d}C_4^d\sqrt{p}}{\epsilon^{2d+2}} + \alpha + 1))$  stochastic oracle of  $L_\beta$ .

### Proof of Theorem 5.1.8

The guarantee of differential privacy is by Gaussian mechanism and composition theorem.

By Theorem 5.1.7, Lemma 5.1.8 and 5.1.5, we have

$$\mathbb{E}L_\beta(w_n, D) - \min_{w \in \mathcal{C}} L_\beta(w, D) \leq O\left(\frac{(\frac{d^{3d}C_4^d\sqrt{p}}{\epsilon^{2d+2}} + \alpha + 1)}{\beta\sqrt{n}} + \frac{1}{\beta^2d}\right) = O\left(\frac{d^{3d}C_4^d\sqrt{p}}{\epsilon^{2d+2}\beta\sqrt{n}} + \frac{\alpha}{2}\right).$$

By Lemma 5.1.8, we know that

$$\mathbb{E}L(w_n, D) - \min_{w \in \mathcal{C}} L(w, D) \leq O\left(\beta + \frac{d^{3d}C_4^d\sqrt{p}}{\epsilon^{2d+2}\beta\sqrt{n}} + \frac{\alpha}{2}\right).$$

Thus, if we take  $\beta = \frac{\alpha}{4}$ ,  $d = \frac{2}{\beta^2\alpha} = O(\frac{1}{\alpha^3})$  and  $n = \Omega(\frac{d^{6d}C_5^dp}{\epsilon^{4d+4}\alpha^2})$ , we have

$$\mathbb{E}L(w_n, D) - \min_{w \in \mathcal{C}} L(w, D) \leq \alpha.$$

### Proof of Lemma 5.1.9

Let  $g(\theta) = \mathbb{E}_{s \sim Q}|s - \theta|$ . Then, we have the following for every  $\theta$ , where  $f'(\theta)$  is well defined,

$$\begin{aligned} g'(\theta) &= \mathbb{E}_{s \sim Q}[1_{s \leq \theta}] - \mathbb{E}_{s \sim Q}[1_{s > \theta}] \\ &= \frac{[f'(\theta) - f'(-1)] - [f'(1) - f'(\theta)]}{f'(1) - f(-1)} \\ &= \frac{2f'(\theta) - (f'(1) + f'(-1))}{f'(1) - f'(-1)}. \end{aligned}$$

Thus, we get

$$F'(\theta) = \frac{f'(1) - f'(-1)}{2}g'(\theta) + \frac{f'(1) + f'(-1)}{2} = f'(\theta).$$

Next, we show that if  $F'(\theta) = f'(\theta)$  for every  $\theta \in [0, 1]$ , where  $f'(\theta)$  is well defined, there is a constant  $c$  which satisfies the condition of  $F(\theta) = f(\theta) + c$  for all  $\theta \in [0, 1]$ .

**Lemma 5.1.15.** If  $f$  is convex and 1-Lipschitz, then  $f$  is differentiable at all but countably many points. That is,  $f'$  has only countable many discontinuous points.

*Proof of Lemma 5.1.15.* Since  $f$  is convex, we have the following for  $0 \leq s < u \leq v < t \leq 1$

$$\frac{f(u) - f(s)}{u - s} \leq \frac{f(t) - f(v)}{t - v},$$

This is due to the property of 3-point convexity, where

$$\frac{f(u) - f(s)}{u - s} \leq \frac{f(t) - f(u)}{t - u} \leq \frac{f(t) - f(v)}{t - v}.$$

Thus, we can obtain the following inequality of one-sided derivation, that is,

$$f'_-(x) \leq f'_+(x) \leq f'_-(y) \leq f'_+(y)$$

for every  $x < y$ . For each point where  $f'_-(x) < f'_+(x)$ , we pick a rational number  $q(x)$  which satisfies the condition of  $f'_-(x) < q(x) < f'_+(x)$ . From the above discussion, we can see that all these  $q(x)$  are different. Thus, there are at most countable many points where  $f$  is non-differentiable.  $\square$

From the above lemma, we can see that the Lebesgue measure of these dis-continuous points is 0. Thus,  $f'$  is Riemann Integrable on  $[-1, 1]$ . By Newton-Leibniz formula, we have the following for any  $\theta \in [0, 1]$ ,

$$\int_{-1}^{\theta} f'(x)dx = f(\theta) - f(-1) = \int_{-1}^{\theta} F'(x)dx = F(\theta) - F(-1).$$

Therefore, we get  $F(\theta) = f(\theta) + c$  and complete the proof.

### Proof of Theorem 5.1.9

Let  $h_\beta$  denote the function  $h_\beta(x) = \frac{x + \sqrt{x^2 + \beta^2}}{2}$ . By Lemma 5.1.9 we have

$$f(\theta) = (f'(1) - f'(-1))\mathbb{E}_{s \sim Q} \frac{|s - \theta|}{2} + \frac{f'(1) + f'(-1)}{2}\theta + c.$$

Now, we consider function  $F_\beta(\theta)$ , which is

$$F_\beta(\theta) = (f'(1) - f'(-1))\mathbb{E}_{s \sim Q} [2h_\beta(\frac{\theta - s}{2}) - \frac{\theta - s}{2}] + \frac{f'(1) + f'(-1)}{2}\theta + c.$$

From this, we have

$$\nabla F_\beta(\theta) = (f'(1) - f'(-1))\mathbb{E}_{s \sim Q} [\nabla h_\beta(\frac{\theta - s}{2})] + \frac{f'(1) + f'(-1)}{2} - \frac{f'(1) - f'(-1)}{2}.$$

Note that since  $|x| = 2 \max\{x, 0\} - x$ , we can get 1)  $|F_\beta(\theta) - f(\theta)| \leq O(\beta)$  for any  $\theta \in \mathbb{R}$ , 2)  $F_\beta(x)$  is  $O(\frac{1}{\beta})$ -smooth and convex since  $h_\beta(\theta - s)$  is  $\frac{1}{\beta}$ -smooth and convex, and 3)  $F_\beta(\theta)$  is  $O(1)$ -Lipschitz. Now, we optimize the following problem in the non-interactive local

model:

$$F_\beta(w; D) = \frac{1}{n} \sum_{i=1}^n F_\beta(y_i \langle x_i, w \rangle).$$

For each fixed  $i$  and  $s$ , we let

$$\hat{G}(w, i, s) = (f'(1) - f'(-1)) \left[ \sum_{j=1}^d c_j \binom{d}{j} t_{i,j} r_{i,j} \right] y_i x_i^T + f'(-1).$$

Then, we have  $\mathbb{E}_{\sigma,z} G(w, i, s) = \hat{G}(w, i, s)$ . By using a similar argument given in the proof of Theorem 5.1.7, we get

$$\text{Var}(\hat{G}(w, i, s) | i, s) \leq \tilde{O}\left(\frac{d^{6d} C^d p}{\epsilon^{4d+4}}\right).$$

Thus, for each fixed  $i$  we have

$$\begin{aligned} \mathbb{E}_s \hat{G}(w, i, s) &= \bar{G}(w, i) = (f'(1) - f'(-1)) \left[ \mathbb{E}_{s \sim \mathcal{Q}} \sum_{j=1}^d c_j \binom{d}{j} \left( \frac{y_i \langle w, x_i \rangle - s}{2} \right)^j \right. \\ &\quad \left. \left( 1 - \frac{y_i \langle w, x_i \rangle - s}{2} \right)^{d-j} \right] y_i x_i^T + f'(-1). \end{aligned}$$

Next, we bound the term of  $\text{Var}(\hat{G}(w, i, s) | i) \leq O(d^{2d+2})$ .

Let  $t_{i,j} = \prod_{k=j+1}^{jd+j} \left( \frac{y_i \langle w_t, x_i \rangle - s_k}{2} \right)$ . Then, we have

$$\text{Var}(t_{i,j}) \leq \prod_{k=j+1}^{jd+j} |y_i|^2 \text{Var}(\langle w_t, x_i \rangle - s_k) \leq O(1).$$

And similarly for  $\text{Var}(r_{i,j})$ . Thus, we get

$$\text{Var}(\hat{G}(w, i, s) | i) \leq O\left(\sum_{j=1}^d c_j^2 \binom{d}{j}^2 \text{Var}(t_{i,j} r_{i,j})\right) = O(d^{2d+2}).$$

Since  $\mathbb{E}_i \bar{G}(w, i) = \hat{G} = \frac{1}{n} \sum_{i=1}^n \bar{G}(w, i)$ , we have  $\text{Var}(\bar{G}(w, i)) \leq O((\alpha+1)^2)$  by a similar

argument given in the proof of Theorem 5.1.7. Thus, in total we have

$$\mathbb{E}\|G(w, i, s) - \hat{G}\|_2^2 \leq \tilde{O}\left(\frac{d^{6d}C^d p}{\epsilon^{4d+4}}\right)$$

The other part of the proof is the same as that of Theorem 5.1.7.

### Proof of Theorem 5.1.10

It is sufficient to prove that

$$\sup_{y \in Y_k} |\tilde{q}_D(y) - q_y(D)| \leq \gamma + \frac{T \binom{p+t_k}{t_k}^2 \sqrt{\log \frac{\binom{p+t_k}{t_k}}{\beta}}}{\sqrt{n}\epsilon},$$

where  $T = p^{O(\sqrt{k} \log(\frac{1}{\gamma}))}$ . Now we denote  $p_D \in \mathbb{R}^{\binom{p+t_k}{t_k}}$  as the average of  $\tilde{q}_i$ . That is, it is the unperturbed version of  $\tilde{p}_D$ . By Lemma 5.1.10, we have  $\sup_{y \in Y_k} |p_D(y) - q_y(D)| \leq \gamma$ . Thus it is sufficient to prove that

$$\sup_{y \in Y_k} |\tilde{q}_D(y) - p_D(y)| \leq \frac{T \binom{p+t_k}{t_k}^2 \sqrt{\log \frac{\binom{p+t_k}{t_k}}{\beta}}}{\sqrt{n}\epsilon}.$$

Since both  $\tilde{q}_D$  and  $p_D$  can be viewed as  $\binom{p+t_k}{t_k}$ -dimensional vectors, we then have

$$\sup_{y \in Y_k} |\tilde{p}_D(y) - p_D(y)| \leq \|\tilde{p}_D - p_D\|_1.$$

Also, since each coordinate of  $p_D(y)$  is bounded by  $T$  by Lemma 5.1.10, we can see that if  $n = \Omega(\max\{\frac{1}{\epsilon^2} \log \frac{1}{\beta}, \binom{p+t_k}{t_k} \log \binom{p+t_k}{t_k} \log 1/\beta\})$ , then by Lemma 5.1.1, with probability at least  $1 - \beta$ , the following is true

$$\|\tilde{p}_D - p_D\|_1 \leq \frac{T \binom{p+t_k}{t_k}^2 \sqrt{\log \frac{\binom{p+t_k}{t_k}}{\beta}}}{\sqrt{n}\epsilon}.$$

Thus, if taking  $\gamma = \frac{\alpha}{2}$  and by the fact that  $\binom{p+t_k}{t_k} = p^{O(t_k)}$ , we get the proof.

### Proof of Theorem 5.1.11

Let  $t = (\frac{1}{\gamma})^{\frac{1}{h}}$ . It is sufficient to prove that  $\sup_{q_f \in \mathcal{Q}_{C_T^h}} |\tilde{p}_D \cdot c_f - q_f(D)| \leq \alpha$ . Let  $p_D$  denote the average of  $\{p_i\}_{i=1}^n$ , i.e. the unperturbed version of  $\tilde{p}_D$ . Then by Lemma 5.1.11, we have  $\sup_{q_f \in \mathcal{Q}_{C_T^h}} |p_D \cdot c_f - q_f(D)| \leq \gamma$ . Also since  $\|c_f\|_\infty \leq M$ , we have  $\sup_{q_f \in \mathcal{Q}_{C_T^h}} |\tilde{p}_D \cdot c_f - p_D \cdot c_f| \leq O(\|\tilde{p}_D - p_D\|_1)$ . By Lemma 5.1.1, we know that if  $n = \Omega(\max\{\frac{1}{\epsilon^2} \log \frac{1}{\beta}, t^{2p} \log \frac{1}{\beta}\})$ , then  $\|\tilde{p}_D - p_D\|_1 \leq O(\frac{t^{\frac{5p}{2}} \sqrt{\log(\frac{1}{\beta})}}{\sqrt{n}\epsilon})$  with probability at least  $1 - \beta$ . Thus, we have  $\sup_{q_f \in \mathcal{Q}_{C_T^h}} |\tilde{p}_D \cdot c_f - q_f(D)| \leq O(\gamma + \frac{(\frac{1}{\gamma})^{\frac{5p}{2h}} \sqrt{\log(\frac{1}{\beta})}}{\sqrt{n}\epsilon})$ . Taking  $\gamma = O((1/\sqrt{n}\epsilon)^{\frac{2h}{5p+2h}})$ , we get  $\sup_{q_f \in \mathcal{Q}_{C_T^h}} |\tilde{p}_D \cdot c_f - q_f(D)| \leq O(\sqrt{\log(\frac{1}{\beta})(\frac{1}{\sqrt{n}\epsilon})^{\frac{2h}{5p+2h}}}) \leq \alpha$ . The computational cost for answering a query follows from Lemma 5.1.11 and  $b \cdot c = O(t^p)$ .

### 5.1.7 Omitted Details in Section 5.1.3

Recently, [49] proposed a generic transformation, GenProt, which could transform any  $(\epsilon, \delta)$  (so as for  $\epsilon$ ) non-interactive LDP protocol to an  $O(\epsilon)$ -LDP protocol with the communication complexity for each player being  $O(\log \log n)$  (at the expense of increasing the shared randomness in the protocol), which removes the condition of 'sample resilient' in [27]. The detail is in Algorithm 5.1.37. The transformation uses  $O(n \log \frac{n}{\beta})$  independent public string. The reader is referred to [49] for details. Actually, by Algorithm 5.1.37, we can easily get an  $O(\epsilon)$ -LDP algorithm with the same error bound.

**Theorem 5.1.12.** For any given  $\epsilon \leq \frac{1}{4}$ , under the condition of Corollary 5.1.2, Algorithm 5.1.37 is  $10\epsilon$ -LDP. If  $T = O(\log \frac{n}{\beta})$ , then with probability at least  $1 - 2\beta$ , Corollary 5.1.2 holds. Moreover, the communication complexity of each layer is  $O(\log \log n)$  bits, and the computational complexity for each player is  $O(\log \frac{n}{\beta})$ .

---

**Algorithm 5.1.37** Player-Efficient Local Bernstein Mechanism with  $O(\log \log n)$  bits communication complexity.

---

- 1: **Input:** Each user  $i \in [n]$  has data  $x_i \in \mathcal{D}$ , privacy parameter  $\epsilon$ , public loss function  $\ell : [0, 1]^p \times \mathcal{D} \mapsto [0, 1]$ , and parameter  $k, T$ .
  - 2: **Preprocessing:**
  - 3: For every  $(i, T) \in [n] \times [T]$ , generate independent public string  $y_{i,t} = \text{Lap}(\perp)$ .
  - 4: Construct the grid  $\mathcal{T} = \{\frac{v_1}{k}, \frac{v_2}{k}, \dots, \frac{v_p}{k}\}_{v_1, v_2, \dots, v_p}$ , where  $\{v_1, v_2, \dots, v_p\} = \{0, 1, \dots, k\}^p$ .
  - 5: Randomly partition  $[n]$  into  $d = (k+1)^p$  subsets  $I_1, I_2, \dots, I_d$ , with each subset  $I_j$  corresponding to a grid in  $\mathcal{T}$  denoted as  $\mathcal{T}(j)$ .
  - 6: **for** Each Player  $i \in [n]$  **do**
  - 7:     Find the subset  $I_\ell$  such that  $i \in I_\ell$ . Calculate  $v_i = \ell(\mathcal{T}(l); x_i)$ .
  - 8:     For each  $t \in [T]$ , compute  $p_{i,t} = \frac{1}{2} \frac{\Pr[v_i + \text{Lap}(\frac{1}{\epsilon}) = y_{i,t}]}{\Pr[\text{Lap}(\perp) = y_{i,t}]}$
  - 9:     For every  $t \in [T]$ , if  $p_{i,t} \notin [\frac{e^{-2\epsilon}}{2}, \frac{e^{2\epsilon}}{2}]$ , then set  $p_{i,t} = \frac{1}{2}$ .
  - 10:     For every  $t \in [T]$ , sample a bit  $b_{i,t}$  from  $\text{Bernoulli}(p_{i,t})$ .
  - 11:     Denote  $H_i = \{t \in [T] : b_{i,t} = 1\}$
  - 12:     If  $H_i = \emptyset$ , set  $H_i = [T]$
  - 13:     Sample  $g_i \in H_i$  uniformly, and send  $g_i$  to the server.
  - 14: **end for**
  - 15: **for** The Server **do**
  - 16:     **for** Each  $l \in [d]$  **do**
  - 17:         Compute  $v_\ell = \frac{n}{|I_\ell|} \sum_{i \in I_\ell} g_i$ .
  - 18:         Denote the corresponding grid point  $(\frac{v_1}{k}, \frac{v_2}{k}, \dots, \frac{v_p}{k}) \in \mathcal{T}$  as  $\ell$ ; then let  $\hat{L}((\frac{v_1}{k}, \frac{v_2}{k}, \dots, \frac{v_p}{k}); D) = v_\ell$ .
  - 19:     **end for**
  - 20:     Construct perturbed Bernstein polynomial of the empirical loss  $\tilde{L}$  as in Algorithm 2. Denote the function as  $\tilde{L}(\cdot, D)$ .
  - 21:     Compute  $w_{\text{priv}} = \arg \min_{w \in \mathcal{C}} \tilde{L}(w; D)$ .
  - 22: **end for**
- 

### 5.1.8 Detailed Algorithm of SIGM in Lemma 5.1.5

Let  $a \geq 1, b \geq 0, p \geq 1$  be some parameters. Let us assume that we know a number  $R$  such that  $\|w^*\|_2 \leq R$ . We choose

$$\alpha_i = \frac{1}{a} \left( \frac{i+p}{p} \right)^{p-1} \quad (5.10)$$

$$\beta_i = \beta + \frac{b\sigma}{R} (i+p+1)^{\frac{2p-1}{2}} \quad (5.11)$$

$$B_i = a\alpha_i^2 = \frac{1}{a} \left( \frac{i+p}{p} \right)^{2p-2}. \quad (5.12)$$

We also define  $A_k = \sum_{i=0}^n \alpha_i$  and  $\eta_i = \frac{\alpha_{i+1}}{B_{i+1}}$  and  $\alpha_0 = A_0 = B_0$

**Lemma 5.1.16** (Theorem 3.4 in [102]). Assume that  $f(w)$  is endowed with a  $(\gamma, \beta, \sigma)$  stochastic oracle  $(F_{\gamma, \beta, \sigma}(w; \xi), G_{\gamma, \beta, \sigma}(w; \xi))$  with  $\beta \geq O(1)$ . By choosing the parameters above with  $a = 2^{\frac{p-1}{2}}$  and  $b = 2^{\frac{5-2p}{4}} p^{\frac{1-2p}{2}}$ , then the sequence  $y_k$  generated by Algorithm 5.1.38

$$\mathbb{E}_{x_0, x_1, \dots, x_k} [f(y_k)] - \min_{y \in \mathcal{C}} f(y) \leq \Theta\left(\frac{\beta R^2}{k^p} + \frac{\sigma R}{\sqrt{k}} + k^{p-1} \gamma\right).$$

Taking  $p = 1$ , this is just Lemma 5.1.5.

---

#### Algorithm 5.1.38 Stochastic Intermediate Gradient Method

---

- 1: **Input:** The sequences  $\{\alpha_i\}_{i \geq 0}$ ,  $\{\beta_i\}_{i \geq 0}$ ,  $\{B_i\}_{i \geq 0}$ , functions  $d(x) = \frac{1}{2}\|x\|^2$ , Bregman distance  $V(x, z) = d(X) - d(Z) - \langle \nabla d(z), x - z \rangle$ .
- 2: Compute  $x_0 = \arg \min_{x \in \mathcal{C}} \{d(x)\}$ .
- 3: Let  $\xi_0$  be a realization of the random variable  $\xi$ .
- 4: Computer  $G_{\gamma, \beta, \sigma}(x_0; \xi_0)$ .
- 5: Compute

$$y_0 = \arg \min_{x \in \mathcal{C}} \{\beta_0 d(x) + \alpha_0 \langle G_{\gamma, \beta, \sigma}(x_0; \xi_0), x - x_0 \rangle\}. \quad (5.13)$$

- 6: **for**  $k = 0, \dots, T - 1$  **do**
- 7:     Compute

$$z_k = \arg \min_{x \in \mathcal{C}} \beta_k d(x) + \sum_{i=0}^k \alpha_i \langle G_{\gamma, \beta, \sigma}(x_i; \xi_i), x - x_i \rangle \quad (5.14)$$

- 8:     Let  $x_{k+1} = \eta_k z_k + (1 - \eta_k) y_k$ .
- 9:     Let  $\xi_{k+1}$  be a realization of the random variable  $\xi$ .
- 10:    Compute  $G_{\gamma, \beta, \sigma}(x_{k+1}; \xi_{k+1})$
- 11:    Compute

$$\hat{x}_{k+1} = \arg \min_{x \in \mathcal{C}} \beta_k V(x, z_k) + \alpha_{k+1} \langle G_{\gamma, \beta, \sigma}(x_{k+1}; \xi_{k+1}), x - z_k \rangle. \quad (5.15)$$

- 12:    Let  $w_{k+1} = \eta_k \hat{x}_{k+1} + (1 - \eta_k) y_k$ .
  - 13:    Let  $y_{k+1} = \frac{A_{k+1} - B_{k+1}}{A_{k+1}} y_k + \frac{B_{k+1}}{A_{k+1}} w_{k+1}$ .
  - 14: **end for**
  - 15: **return**  $y_T$ .
-

## 5.2 ERM in a Relaxed Non-interactive LDP model

From the previous chapter, we can see that, although we have already improved the sample complexity given by [257], here our improved sample complexity for achieving an error of  $\alpha$  still need to be exponential in  $\alpha$  [307, 363] or exponential in the dimensionality  $p$ . Due to these negative results, there is no study on the practical performance of these algorithms.

To address high sample complexity and practical issues of NLDP, a possible way is to make use of some recent developments on the central DP model. Quite a few results [26, 139, 239, 238, 30] have suggested that by allowing the server to access some public but unlabeled data in addition to the private data, it is possible to reduce the sample complexity in the central DP model, under the assumption that these public data have the same marginal distribution as the private ones. It has also shown that such a relaxed setting is likely to enable better practical performance for problems like Empirical Risk Minimization (ERM) [139, 239]. Thus, it would be interesting to know whether the relaxed setting can also help reduce sample complexity in the NLDP model.

In this section, we will focus on a subclass of ERM, Generalized Linear Model (GLM), in a relaxed version of the NLDP model. GLM is one of the most fundamental models in statistics and machine learning. It generalizes ordinary linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value. GLM was introduced as a way of unifying various statistical models, including linear, logistic and Poisson regressions. It has a wide range of applications in various domains, such as social sciences [341], genomics research [268], finance [217] and medical research [199]. The model can be formulated as follows.

**GLM:** Let  $y \in [0, 1]$  be the response variable that belongs to an exponential family with natural parameter  $\eta$ . That is, its probability density function can be written as  $p(y|\eta) =$

$\exp(\eta y - \Phi(\eta))h(y)$ , where  $\Phi$  is the *cumulative generating function*. Given observations  $y_1, \dots, y_n$  such that  $y_i \sim p(y_i|\eta_i)$  for  $\eta = (\eta_1, \dots, \eta_n)$ , the maximum likelihood estimator (MLE) can be written as  $p(y_1, y_2, \dots | \eta) = \exp(\sum_{i=1}^n y_i \eta_i - \Phi(\eta_i)) \prod_{i=1}^n h(y_i)$ . In GLM, we assume that  $\eta$  is modeled by linear relations, *i.e.*,  $\eta_i = \langle x_i, w^* \rangle$  for some  $w^* \in \mathbb{R}^p$  and feature vector  $x_i$ . Thus, maximizing MLE is equivalent to minimizing  $\frac{1}{n} \sum_{i=1}^n [\Phi(\langle x_i, w \rangle) - y_i \langle x_i, w \rangle]$ . The goal is to find  $w^*$ , which is equivalent to minimizing its population version

$$w^* = \arg \min_{w \in \mathbb{R}^p} \mathbb{E}_{(x,y)}[\Phi(\langle x, w \rangle) - y \langle x, w \rangle]. \quad (5.16)$$

Thus, in this chapter, our main questions now become the follows. **Can we further reduce the sample complexity of GLM in the NLDP model if the server has additional public but unlabeled data? Moreover, is there any efficient algorithm for this problem in the relaxed setting?**

In this paper, we provide positive answers to the above two questions. Our contributions can be summarized as follows:

1. We first show that when the feature vector  $x$  of GLM is sub-Gaussian with bounded  $\ell_1$ -norm, there is an  $(\epsilon, \delta)$ -NLDP algorithm for GLM (under some mild assumptions) whose sample complexities of the private and public data, for achieving an error of  $\alpha$  (in  $\ell_\infty$ -norm), are  $O(p^2 \epsilon^{-2} \alpha^{-2})$  and  $O(p^2 \alpha^{-2})$  (with other terms omitted), respectively, if  $\alpha$  is not too small (*i.e.*,  $\alpha \geq \Omega(\frac{1}{\sqrt{p}})$ ). We note that this is the first result that achieves a **fully polynomial** sample complexity for a general class of loss functions in the NLDP model with public unlabeled data. Another nice feature of this algorithm is that, instead of just answering one GLM query, it can answer, with constant probability, multiple (at most  $\exp(O(p))$ ) GLM queries and achieve an error of  $\alpha \geq \Omega(\frac{1}{\sqrt{p}})$  with the same sample complexities as in the single query case.
2. We then extend our idea to the non-linear regression problem. By using the zero-bias transformation [130], we show that when  $x$  is sub-Gaussian with bounded  $\ell_1$ -norm, it

exhibits the same phenomenon as GLM.

3. Finally, we provide an experimental study of our algorithms on both synthetic and real world datasets. The experimental results suggest that our methods are efficient and effective, which is consistent with our theoretical analysis. To our best knowledge, these are the **first** effective algorithms in the NLDP model with public unlabeled data for both the GLM and non-linear regression problems. Moreover, these experimental results also provide a clear message as to which aspects need further theoretical investigation.

### 5.2.1 Related Work

Private learning with public unlabeled data has been studied previously in [139, 239, 238, 30]. These results differ from ours in quite a few ways. Firstly, all of them consider either the multiparty setting or the centralized model. Consequently, none of them can be used to solve our problems. Specifically, [139] considered the multiparty setting where each party possesses several data records, while each party in our NLDP model has only one data record. [239, 238] investigated the DP model, used sub-sample and aggregate to train some deep learning models, but provided no provable sample complexity. [30] also studied the DP model by combining the distance to instability and the sparse vector techniques, and showed some theoretical guarantees. However, both the sub-sample/aggregate and the sparse vector methods cannot be used in the NLDP model. Moreover, public data in their methods are also used quite differently from ours. Secondly, all of the above results use the private data to label the public data and conduct the learning process on the public data, while we use the public data to approximate some crucial constants. Finally, all of the previous methods rely on the known model or loss functions, while in our algorithms the loss functions could be unknown to the users; also the server could use multiple loss functions with the same sample complexity.

## 5.2.2 Our Model

**Our Model:** Different from the classical NLDP model where only one private dataset  $\{(x_i, y_i)\}_{i=1}^n$  exists, the NLDP model in our setting allows the server to have an additional public but unlabeled dataset  $D' = \{x_j\}_{j=n+1}^{n+m} \subset \mathcal{X}^m$ , where each  $x_j$  is sampled from  $\mathcal{P}_x$ , which is the marginal distribution of  $\mathcal{P}$  (*i.e.*, they have the same distribution as  $\{x_i\}_{i=1}^n$ ).

## 5.2.3 Privately Learning Generalized Linear Models

In this section, we study GLM in our model and privately estimate  $w^*$  in (5.16) by using both the private data  $\{(x_i, y_i)\}_{i=1}^n$  and the public unlabeled data  $\{x_j\}_{j=n+1}^{n+m}$ . Our goal is to achieve a fully polynomial sample complexity for  $n$  and  $m$ , *i.e.*,  $n, m = \text{Poly}(p, \frac{1}{\epsilon}, \frac{1}{\alpha}, \log \frac{1}{\delta})$ , such that there is an  $(\epsilon, \delta)$ -NLDP algorithm with estimation error less than  $\alpha$  (with high probability). Before presenting our ideas, we first consider the following lemma for  $x \sim \mathcal{N}(0, \Sigma)$ , which is from Stein's lemma [45].

---

### Algorithm 5.2.39 Non-interactive LDP for smooth GLM with public data

---

**Input:** Private data  $\{(x_i, y_i)\}_{i=1}^n \subset (\mathbb{R}^p \times \{0, 1\})^n$ , where  $\|x_i\|_1 \leq r$  and  $|y_i| \leq 1$ , public unlabeled data  $\{x_j\}_{j=n+1}^{n+m}$ , loss function  $\Phi : \mathbb{R} \mapsto \mathbb{R}$ , privacy parameters  $\epsilon, \delta$ , and initial value  $c \in \mathbb{R}$ .

- 1: **for** Each user  $i \in [n]$  **do**
  - 2:     Release  $\widehat{x_i x_i^T} = x_i x_i^T + E_{1,i}$ , where  $E_{1,i} \in \mathbb{R}^{p \times p}$  is a symmetric matrix and each entry of the upper triangle matrix is sampled from  $\mathcal{N}(0, \frac{32r^4 \log \frac{2.5}{\delta}}{\epsilon^2})$ .
  - 3:     Release  $\widehat{x_i y_i} = x_i y_i + E_{2,i}$ , where  $E_{2,i} \in \mathbb{R}^p$  is sampled from  $\mathcal{N}(0, \frac{32r^2 \log \frac{2.5}{\delta}}{\epsilon^2} I_p)$ .
  - 4: **end for**
  - 5: **for** The server **do**
  - 6:     Let  $\widehat{X^T X} = \sum_{i=1}^n \widehat{x_i x_i^T}$  and  $\widehat{X^T y} = \sum_{i=1}^n \widehat{x_i y_i}$ . Calculate  $\widehat{w}^{ols} = (\widehat{X^T X})^{-1} \widehat{X^T y}$ .
  - 7:     Calculate  $\tilde{y}_j = x_j^T \widehat{w}^{ols}$  for each  $j = n+1, \dots, n+m$ . Find the root  $\hat{c}_\Phi$  such that  $1 = \frac{\hat{c}_\Phi}{m} \sum_{j=n+1}^{n+m} \Phi^{(2)}(\hat{c}_\Phi \tilde{y}_j)$  by using Newton's root-finding method (or other methods):
  - 8:         **for**  $t = 1, 2, \dots$  until convergence **do**
  - 9:              $c = c - \frac{c \frac{1}{m} \sum_{j=n+1}^{n+m+1} \Phi^{(2)}(c \tilde{y}_j) - 1}{\frac{1}{m} \sum_{j=n+1}^{n+m+1} \{\Phi^{(2)}(c \tilde{y}_j) + c \tilde{y}_j \Phi^{(3)}(c \tilde{y}_j)\}}$ .
  - 10:       **end for**
  - 11: **end for**
  - 12: Return  $\hat{w}^{glm} = \hat{c}_\Phi \cdot \widehat{w}^{ols}$ .
-

**Lemma 5.2.1** ([45]). If  $x \sim \mathcal{N}(0, \Sigma)$ , then  $w^*$  in (5.16) can be written as

$$w^* = c_\Phi \times w^{ols},$$

where  $c_\Phi$  is the fixed point of  $z \mapsto (\mathbb{E}[\Phi^{(2)}(\langle x, w^{ols} \rangle z)])^{-1}$  (if we assume taht  $\mathbb{E}[\Phi^{(2)}(\langle x, w^{ols} \rangle z)] \neq 0$ ) and  $w^{ols} = \Sigma^{-1}\mathbb{E}[xy]$  is the Ordinary Least Squares (OLS) vector.

From Lemma 5.2.1, we can see that to obtain  $w^*$ , it is sufficient to estimate  $w^{ols}$  and the underlying constant  $c_\Phi$ . Specifically, to estimate  $w^{ols}$  in a non-interactive local differentially private manner, a direct way is to let each player perturb her sufficient statistics, *i.e.*,  $x_i x_i^T$  and  $y_i x_i$ . After receiving the private OLS estimator  $\hat{w}^{ols}$ , the server can then estimate the constant  $c_\Phi$  by using the public unlabeled data and  $\hat{w}^{ols}$ . From the definition, it is easy to see that  $c_\Phi$  is independent of the label  $y$ . Thus,  $c_\Phi$  can be estimated by using the empirical version of  $\mathbb{E}[\Phi^{(2)}(\langle x, w^{ols} \rangle z)]$ . That is, find the root of the function  $1 - \frac{c}{m} \sum_{j=n+1}^{n+m} \Phi^{(2)}(c\langle x_j, \hat{w}^{ols} \rangle)$ . Several methods are available for finding roots, such as the Newton's method which has a quadratic convergence rate.

One problem with the above approach is that Lemma 5.2.1 needs  $x$  to be Gaussian, which implies that the sensitivity of the term  $x_i x_i^T$  could be unbounded. We also note that Lemma 5.2.1 is only for Gaussian distribution. The following lemma extends Lemma 5.2.1 to bounded sub-Gaussian with an additional additive error of  $O(\frac{\|w^*\|_\infty^2}{\sqrt{p}})$ .

**Lemma 5.2.2** ([108]). Let  $x_1, \dots, x_n \in \mathbb{R}^p$  be i.i.d realizations of a random vector  $x$  that is sub-Gaussian with zero mean, whose covariance matrix  $\Sigma$  has its corresponding  $\Sigma^{\frac{1}{2}}$  being diagonally dominant<sup>9</sup>, and whose distribution is supported on a  $\ell_2$ -norm ball of radius  $r$ . Let  $v = \Sigma^{-\frac{1}{2}}x$  be the whitened random vector of  $x$  with sub-Gaussian norm  $\|v\|_{\psi_2} = \kappa_x$ . If each  $v_i$  has constant first and second conditional moments (*i.e.*,  $\forall j \in [p]$  and  $\tilde{w} = \Sigma^{\frac{1}{2}}w^*$ ,  $\mathbb{E}[v_{ij} | \sum_{k \neq j} \tilde{w} v_{ik}]$  and  $\mathbb{E}[v_{ij}^2 | \sum_{k \neq j} \tilde{w} v_{ik}]$  are deterministic) and the function  $\Phi^{(2)}$  is Lipschitz

---

<sup>9</sup>A square matrix is said to be diagonally dominant if, for every row of the matrix, the magnitude of the diagonal entry in a row is larger than or equal to the sum of the magnitudes of all the other (non-diagonal) entries in that row.

continuous with constant  $G$ , then for  $c_\Phi = \frac{1}{\mathbb{E}[\Phi^{(2)}(\langle x_i, w^* \rangle)]}$  (assuming  $\mathbb{E}[\Phi^{(2)}(\langle x_i, w^* \rangle)] \neq 0$ ), the following holds for GLM in (5.16)

$$\left\| \frac{1}{c_\Phi} \cdot w^* - w^{ols} \right\|_\infty \leq 16Gr\kappa_x^3 \sqrt{\rho_2} \rho_\infty \frac{\|w^*\|_\infty^2}{\sqrt{p}}, \quad (5.17)$$

where  $\rho_q$  for  $q = \{2, \infty\}$  is the conditional number of  $\Sigma$  in  $\ell_q$  norm and  $w^{ols} = (\mathbb{E}[xx^T])^{-1}\mathbb{E}[xy]$  is the OLS vector.

Lemma 5.2.2 indicates that we can use the same idea as above to estimate  $w^*$ . Note that the forms of  $c_\Phi$  in Lemmas 5.2.1 and 5.2.2 are different. However, due to the closeness of  $w^*$  and  $w^{ols}$  in (5.17), we can still use  $\frac{1}{\mathbb{E}[\Phi^{(2)}(\langle x_i, w^{ols} \rangle) \bar{c}_\Phi]}$  to approximate  $c_\Phi$ , where  $\bar{c}_\Phi$  is the root of  $c\mathbb{E}[\Phi^{(2)}(\langle x_i, w^{ols} \rangle)c] - 1$ . Combining these ideas, we have Algorithm 5.2.39.

**Theorem 5.2.1.** For any  $0 < \epsilon, \delta < 1$ , Algorithm 5.2.39 is  $(\epsilon, \delta)$  non-interactive LDP.

The following theorem shows the sample complexity of the bounded sub-Gaussian case.

**Theorem 5.2.2.** Under the assumptions of Lemma 5.2.2, if further assume that the distribution of  $x$  is supported on the  $\ell_1$ -norm ball with radius  $r$ ,  $|\Phi^{(2)}(\cdot)| \leq L$ , and for some constant  $\bar{c}$  and  $\tau > 0$ , the function  $f(c) = c\mathbb{E}[\Phi^{(2)}(\langle x, w^{ols} \rangle)c]$  satisfies the condition of  $f(\bar{c}) \geq 1 + \tau$ , and the derivative of  $f$  in the interval  $[0, \max\{\bar{c}, c_\Phi\}]$  does not change the sign (*i.e.*, its absolute value is lower bounded by some constant  $M > 0$ ), then for sufficiently large  $m, n$  such that

$$m \geq \Omega\left(\|\Sigma\|_2 \|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\} \rho_2 \rho_\infty^2 p^2 \max\{1, \frac{1}{c_\Phi}\}^2\right) \quad (5.18)$$

$$n \geq \Omega\left(\frac{\rho_2 \rho_\infty^2 \|\Sigma\|_2^2 p^2 \|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\} \log \frac{1}{\delta} \log \frac{p}{\xi}}{\epsilon^2 \lambda_{\min}(\Sigma) \min\{\lambda_{\min}(\Sigma), 1\}} \max\{1, \frac{1}{c_\Phi}\}^2\right), \quad (5.19)$$

with probability at least  $1 - \exp(-\Omega(p)) - \xi$ , the output  $\hat{w}^{glm}$  in Algorithm 5.2.39 satisfies

$$\begin{aligned} \|\hat{w}^{glm} - w^*\|_\infty &\leq O\left(\frac{\rho_2 \rho_\infty^2 \|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\} \|\Sigma\|_2^{\frac{1}{2}} p \max\{\frac{1}{c_\Phi}, 1\}^2}{\sqrt{m}}\right. \\ &\quad + \frac{\rho_2 \rho_\infty^2 \|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\} \|\Sigma\|_2^{\frac{1}{2}} p \sqrt{\log \frac{1}{\delta} \log \frac{p}{\xi^2}}}{\epsilon \lambda_{\min}^{\frac{1}{2}}(\Sigma) \min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma), 1\} \sqrt{n}} \times \max\{\frac{1}{c_\Phi}, 1\}^2 \\ &\quad \left. + \rho_2 \rho_\infty^2 \|\Sigma^{\frac{1}{2}}\|_\infty \frac{\|w^*\|_\infty^3 \max\{1, \|w^*\|_\infty\}}{\sqrt{p}} \times \max\{1, \frac{1}{c_\Phi}\}\right), \quad (5.20) \end{aligned}$$

where  $G, L, \tau, M, \bar{c}, r, \kappa_x$  are assumed to be  $O(1)$  and thus omitted in the Big- $O$  notations.

Theorem 5.2.2 suggests that if we omit all the other terms and assume that  $\|w^*\|_\infty = O(1)$ , then for any given error  $\alpha \geq \Omega(\frac{1}{\sqrt{p}})$ , there is an  $(\epsilon, \delta)$ -LDP algorithm whose sample complexity of private ( $n$ ) and public unlabeled ( $m$ ) data, to achieve an estimation error of  $\alpha$  (in  $\ell_\infty$ -norm), is  $O(p^2 \epsilon^{-2} \alpha^{-2})$  and  $O(p^2 \alpha^{-2})$ , respectively. We note that  $m \leq n$ , which means that the sample complexity of the public data is less than that of the private data. We also note that the sample complexity of the public data is independent of the privacy parameters  $\epsilon, \delta$ . All these are quite reasonable in practice. We will also see that in practice we do not need large amount of public data (see Experiments section for details).

There are also some previous work on LDP linear regression. [257] proposed an algorithm with a sample complexity of  $\tilde{O}(p\alpha^{-2}\epsilon^{-2})$  and [363] achieved a sample complexity of  $O(\log p\alpha^{-4}\epsilon^{-2})$ . It seems that our sample complexity for the more general GLM is worse than theirs. However, these results are not really comparable due to their different settings. Firstly, [257, 363] considered the optimization error and [318] measured the  $\ell_2$ -norm statistical error, while we estimate the  $\ell_\infty$ -norm statistical error. Secondly,  $w^*$  is assumed to be bounded in  $\ell_2$ -norm in [257],  $\ell_1$ -norm in [363], and  $\ell_\infty$ -norm in ours. There is also a result on NLDP linear regression [318]. It relies on assumptions that  $\|x\|_2 = O(\sqrt{p})$  and  $w^*$  is 1-sparse, which are not needed in ours.

Also note that in Theorem 5.2.2,  $\Phi^{(2)}$  is assumed to be bounded. This is a quite common assumption in related works such as [301, 297]. Actually, this condition can be relaxed by

only assuming that  $\Phi^{(2)}(\langle x, w \rangle)$  is sub-Gaussian in some range of  $w$ .

**Theorem 5.2.3.** Under the assumptions of Lemma 5.2.2, if further assume that the distribution of  $x$  is supported on the  $\ell_1$ -norm ball with radius  $r$ ,  $\sup_{w: \|w - \Sigma^{\frac{1}{2}} w^{ols}\|_2 \leq 1} \|\Phi^{(2)}(\langle x, w \rangle)\|_{\psi_2} \leq \kappa_g$ , the function  $f(c) = c\mathbb{E}[\Phi^{(2)}(\langle x, w^{ols} \rangle c)]$  satisfies the inequality of  $f(\bar{c}) \geq 1 + \tau$  for some constant  $\bar{c}$  and  $\tau > 0$ , and the derivative of  $f$  in the interval of  $[0, \max\{\bar{c}, c_\Phi\}]$  does not change the sign (*i.e.*, its absolute value is lower bounded by some constant  $M > 0$ ), then for sufficiently large  $m, n$  such that

$$m \geq \tilde{\Omega}\left(\frac{1}{\tilde{\mu}^2} \epsilon^2 n\right), \quad (5.21)$$

$$n \geq \Omega\left(\|\Sigma\|_2^2 \frac{p^2 \rho_2 \rho_\infty^2 \|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\} \log \frac{1}{\delta} \log \frac{p^2}{\xi}}{\epsilon^2 \lambda_{\min}(\Sigma) \min\{\lambda_{\min}(\Sigma), 1\}} \max\{1, \frac{1}{c_\Phi}\}^2\right), \quad (5.22)$$

the following holds with probability at least  $1 - \exp(-\Omega(p)) - \xi$ ,

$$\begin{aligned} \|\hat{w}^{glm} - w^*\|_\infty &\leq O\left(\rho_2 \rho_\infty^2 \|\Sigma^{\frac{1}{2}}\|_2 \times \frac{p \|w^*\|_\infty \max\{1, \|w^*\|_\infty^3\} \sqrt{\log \frac{1}{\delta} \log \frac{p^2}{\xi}}}{\epsilon \sqrt{n} \lambda_{\min}^{1/2}(\Sigma) \min\{\lambda_{\min}^{1/2}(\Sigma), 1\}} \max\{1, \frac{1}{c_\Phi}\}^2\right. \\ &\quad + \rho_2 \rho_\infty^2 \frac{\|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\}}{\sqrt{p}} \max\{1, \frac{1}{c_\Phi}\} \\ &\quad \left. + \sqrt{\rho_2} \rho_\infty \|w^*\|_\infty \max\{1, \|w^*\|_\infty\} \frac{1}{\tilde{\mu}} \sqrt{\frac{p^2 \log m}{m}} \max\{1, \frac{1}{c_\Phi}\}\right), \end{aligned} \quad (5.23)$$

where  $\tilde{\mu} = \frac{\mathbb{E}[\|x\|_2]}{\sqrt{p}}$ , the terms of  $r, \kappa_x, \kappa_g, G, M, \tau, \bar{c}$  are assumed to be constants, and thus omitted in the Big- $O$  notations.

From the above theorem, we can see that with more relaxed assumptions, the sample complexity in Theorem 5.2.3 increases by a factor of  $O(\log m)$  to achieve an upper bound on the statistical error (in  $\ell_\infty$ -norm) that is asymptotically the same as the one in Theorem 5.2.2.

Algorithm 5.2.39 has several advantages over existing techniques. Firstly, different from the approach of using Gradient Descent methods to solve DP-ERM (*e.g.*, [328]), our algorithm is parameter-free. That is, we do not need to choose a specific step size, an iteration number or initial vectors. Secondly, comparing with some previous work such

as [363, 257, 307], all of our above results do not need to assume that the loss function is convex. Thirdly, since the private data contributes only to obtaining the OLS estimator, and only the constant  $\hat{c}$  depends on the loss function  $\Phi$ , this means that with probability at least  $1 - T \exp(-\Omega(p)) - \xi$ , our algorithm can simultaneously use  $T$  different loss functions to achieve the same errors and with the same sample complexity. This implies that we can answer at most  $O(\exp(O(p)))$  number of GLM queries with constant probability to achieve error  $\alpha$  for each query with the same sample complexity as in Theorem 5.2.2. To our best knowledge, this is the first result which can answer multiple non-linear queries in the NLDP model with polynomial sample complexity. Previous results are either for linear queries [40, 25], or in the central DP model [285].

A not so desirable issue of Theorems 5.2.2 and 5.2.3 is that they need quite a few assumptions/conditions. Although almost all of them commonly appear in some related work, the assumptions on function  $f$  seem to be a little weird. They are introduced to ensure that the function  $f - 1$  has a root and  $\hat{c}_\Phi$  is close to  $c_\Phi$  for large enough  $m$ . Fortunately, this is a not big issue in practice. As shown in [108], these conditions actually hold for many loss functions, such as logistic and boosting loss. Also, as we will see later, our experiments show that the algorithm actually performs quite well for many loss functions that may not satisfy these assumptions. Also, we note that the error bounds in Theorems 5.2.2 and 5.2.3 are dependent on the  $\ell_1$ -norm of the upper bound of  $x_i$ , while such a dependency is on the  $\ell_2$ -norm in previous work such as [257, 363]. We leave the problem of relaxing/lifting these assumptions to future research.

### 5.2.4 Privately Learning Non-linear Regressions

In this section, we extend our ideas in the previous section to the problem of estimating non-linear regression in the NLDP model with public unlabeled data. We assume that there

is an underlying vector  $w^* \in \mathbb{R}^p$  with  $\|w^*\|_2 \leq 1$  such that

$$y = f(\langle x, w^* \rangle) + \sigma, \quad (5.24)$$

where  $x$  is the feature vector sampled from some distribution (for simplicity, we assume that the mean is zero) and  $y$  is the response.  $\sigma$  is the zero-mean noise which is independent of  $x$  and bounded by some constant  $C = O(1)$  (*i.e.*,  $\sigma \in [-C, C]$ ).  $f$  is some known differentiable link function with  $f(0) \neq \infty$ <sup>10</sup>. We note that these assumptions are quite common in related work such as [318, 99]. In our model, the goal is to obtain some estimator  $w^{\text{priv}}$  of  $w^*$ , based on the private dataset  $\{(x_i, y_i)\}_{i=1}^n$  and the public unlabeled dataset  $\{x_j\}_{j=n+1}^{n+m+1}$  via some NLDP algorithms.

To solve this problem, we first use the zero-bias transformation [130] and the techniques in [108] to get a lemma similar to Lemma 5.2.2.

**Definition 5.2.1** (Zero-bias Transformation). Let  $z$  be a random variable with mean 0 and variance  $\sigma^2$ . Then, there exists a random variable  $z^*$  that satisfies  $\mathbb{E}[zf(z)] = \sigma^2 \mathbb{E}[f'(z^*)]$  for all differentiable functions  $f$ . The distribution of  $z^*$  is called the  $z$ -zero-bias distribution.

Normal distribution is a unique distribution whose zero-bias transformation is itself. This is the basic Stein's lemma.

**Theorem 5.2.4.** Let  $x_1, \dots, x_n \in \mathbb{R}^p$  be  $n$  i.i.d realizations of a random vector  $x$  which is sub-Gaussian with zero mean, whose covariance matrix  $\Sigma$  has its  $\Sigma^{\frac{1}{2}}$  being diagonally dominant, and whose distribution is supported on an  $\ell_2$ -norm ball of radius  $r$ . Let  $v = \Sigma^{-\frac{1}{2}}x$  be the whitened random vector of  $x$  with sub-Gaussian norm  $\|v\|_{\psi_2} = \kappa_x$ . If each  $v_i$  has constant first and second conditional moments and function  $f'$  is Lipschitz continuous with constant  $G$ , then for  $c_f = \frac{1}{\mathbb{E}[f'(\langle x_i, w^* \rangle)]}$ , the following holds

$$\left\| \frac{1}{c_f} \cdot w^* - w^{ols} \right\|_\infty \leq O(Gr\kappa_x^3 \sqrt{\rho_2} \rho_\infty \frac{\|w^*\|_\infty^2}{\sqrt{p}}),$$

---

<sup>10</sup>This assumption can be relaxed to "there is a point  $x$  such that  $f(x) \neq 0$ ".

where  $w^{ols}$  is the OLS vector.

---

**Algorithm 5.2.40** Non-interactive LDP for smooth Non-linear Regression with public data

---

**Input:** Private data  $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^p \times \{0, 1\}$  with  $\|x_i\|_1 \leq r$ , public unlabeled data  $\{x_j\}_{j=n+1}^{n+m}$ . Link function  $f : \mathbb{R} \mapsto \mathbb{R}$ , privacy parameters  $\epsilon, \delta$ , and initial value  $c \in \mathbb{R}$ .

- 1: **for** Each user  $i \in [n]$  **do**
  - 2:     Release  $\widehat{x_i x_i^T} = x_i x_i^T + E_{1,i}$ , where  $E_{1,i} \in \mathbb{R}^{p \times p}$  is a symmetric matrix and each entry of the upper triangle matrix is sampled from  $\mathcal{N}(0, \frac{32r^4 \log \frac{2.5}{\delta}}{\epsilon^2})$ . Release  $\widehat{x_i y_i} = x_i y_i + E_{2,i}$ , where the vector  $E_{2,i} \in \mathbb{R}^p$  is sampled from  $\mathcal{N}(0, \frac{32r^2(Lr + |f(0)| + C)^2 \log \frac{2.5}{\delta}}{\epsilon^2} I_p)$ .
  - 3: **end for**
  - 4: **for** The server **do**
  - 5:     Denote  $\widehat{X^T X} = \sum_{i=1}^n \widehat{x_i x_i^T}$  and  $\widehat{X^T y} = \sum_{i=1}^n \widehat{x_i y_i}$ . Calculate  $\hat{w}^{ols} = (\widehat{X^T X})^{-1} \widehat{X^T y}$ .
  - 6:     Calculate  $\tilde{y}_j = x_j^T \hat{w}^{ols}$  for each  $j = n+1, \dots, n+m$ . Find the root  $\hat{c}_\Phi$  such that  $1 = \frac{\hat{c}_\Phi}{m} \sum_{j=n+1}^{n+m} f'(\hat{c}_\Phi \tilde{y}_j)$  using Newton's root finding method:
  - 7:     **for**  $t = 1, 2, \dots$  until convergence **do**
  - 8:          $c = c - \frac{c \frac{1}{m} \sum_{j=n+1}^{n+m+1} f'(c \tilde{y}_j) - 1}{\frac{1}{m} \sum_{j=n+1}^{n+m+1} \{f'(c \tilde{y}_j) + c \tilde{y}_j f''(c \tilde{y}_j)\}}$ .
  - 9:     **end for**
  - 10: **end for**
  - 11: Return  $\hat{w}^{nlr} = \hat{c}_\Phi \cdot \hat{w}^{ols}$ .
- 

From Theorem 5.2.4, we can see that it shares the same phenomenon as Lemma 5.2.2 (*i.e.*, the OLS vector with some constant could approximate  $w^*$  well). Thus, a similar idea to Algorithm 5.2.39 can be used to solve this problem for the bounded sub-Gaussian case, which gives us Algorithm 5.2.40 and the following theorem.

**Theorem 5.2.5.** Under the assumptions of Theorem 5.2.4, if further assume that the assumptions in Theorem 5.2.2 hold for function  $f'(\cdot)$  instead of  $\Phi^{(2)}(\cdot)$ , then for sufficiently large  $m, n$  such that

$$m \geq \Omega\left(\|\Sigma\|_2 \|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\} \rho_2 \rho_\infty^2 p^2 \max\{1, \frac{1}{c_f}\}^2\right) \quad (5.25)$$

$$n \geq \Omega\left(\frac{\rho_2 \rho_\infty^2 \|\Sigma\|_2^2 p^2 \|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\} \log \frac{1}{\delta} \log \frac{p}{\xi}}{\epsilon^2 \lambda_{\min}(\Sigma) \min\{\lambda_{\min}(\Sigma), 1\}} \max\{1, \frac{1}{c_f}\}^2\right), \quad (5.26)$$

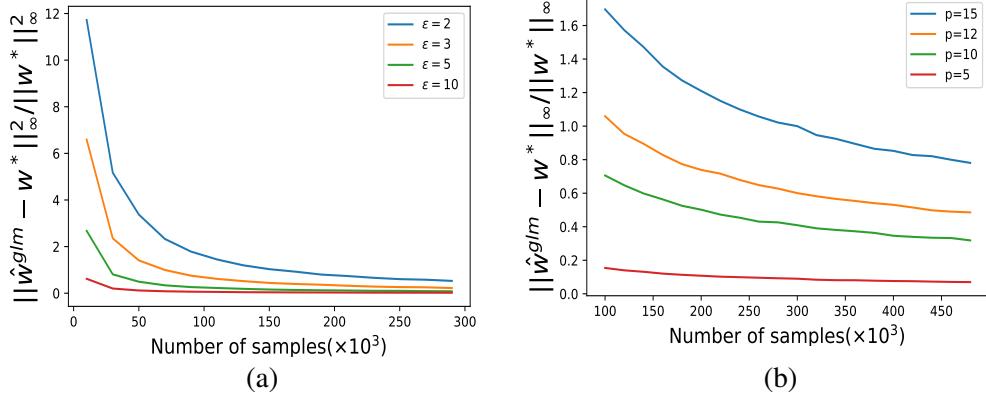


Figure 5.1: GLM with logistic loss under i.i.d Bernoulli design. The left plot shows the squared relative error under different levels of privacy. The right one shows relative error under different dimensionality.

with probability at least  $1 - \exp(-\Omega(p)) - \xi$ , the output of Algorithm 5.2.40 satisfies

$$\begin{aligned} \|\hat{w}^{nlr} - w^*\|_\infty &\leq O\left(\frac{\rho_2 \rho_\infty^2 \|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\} \|\Sigma\|_2^{\frac{1}{2}} p}{\sqrt{m}} \max\left\{\frac{1}{c_f}, 1\right\}^2 \right. \\ &\quad + \frac{\rho_2 \rho_\infty^2 \|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\} \|\Sigma\|_2^{\frac{1}{2}} p \sqrt{\log \frac{1}{\delta} \log \frac{p}{\xi^2}}}{\epsilon \lambda_{\min}^{\frac{1}{2}}(\Sigma) \min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma), 1\} \sqrt{n}} \max\left\{\frac{1}{c_f}, 1\right\}^2 \\ &\quad \left. + \rho_2 \rho_\infty^2 \|\Sigma^{\frac{1}{2}}\|_\infty \frac{\|w^*\|_\infty^3 \max\{1, \|w^*\|_\infty\}}{\sqrt{p}} \times \max\{1, \frac{1}{c_f}\}\right), \quad (5.27) \end{aligned}$$

where the terms of  $G, L, \tau, M, \bar{c}, r, \kappa_x, C$  are assumed to be  $O(1)$  and thus omitted in the Big- $O$  notations.

## 5.2.5 Experiments

### Evaluation on synthetic data

**Experimental Setting** For GLM, we consider the problem of binary logistic loss *i.e.*,  $\Phi(\langle x, w \rangle) = \ln(1 + \exp(\langle x, w \rangle))$  in (5.16) while for non-linear regression we set  $f(x) = \frac{1}{3}x^3$  in (5.24). For each problem we first compare the squared relative error  $\frac{\|\hat{w} - w^*\|_\infty^2}{\|w^*\|_\infty^2}$  with respect to different privacy parameters  $\epsilon \in \{10, 5, 3, 2\}$  with  $\delta = \frac{1}{n}$ . In these ex-

periments, we estimate the squared relative error with the fixed dimensionality  $p = 10$  and the population parameter  $w^* = (1, 1, \dots, 1)/\sqrt{p}$ . The sample size  $n$  is chosen from the set  $10^4 \cdot \{1, 3, 5, \dots, 29\}$ . We assume that the same amount of public unlabeled data is available. The features are generated independently from a Bernoulli distribution  $\Pr(x_{i,j} = \pm \frac{1}{p}) = 0.5$  and the label is generated according to the logistic model or the model (5.24). In non-linear regression model,  $\sigma$  is bounded by  $C = 0.001$ . The results are shown in Figure 5.1a and 5.2a. For each problem we then evaluate the impact of the dimensionality. In these experiments, we fix the privacy parameters <sup>11</sup>  $\epsilon = 10$ ,  $\delta = \frac{1}{n}$ , and tune the dimensionality  $p \in \{5, 10, 12, 15\}$ .  $w^*$ 's are the same as above. The sample size takes values from  $n \in 10^4 \cdot \{10, 12, 14, \dots, 48\}$  and the same amount of public unlabeled data is assumed. The responses are generated as the same as above. We measure the performance directly by the relative error. For each experiments above, we run 1000 times and take the average of the errors. The results are shown in Figure 5.1b and 5.2b.

From Figure 5.1a and 5.2a, we can see that the square of relative error is inversely proportional to the number of samples  $n$ . In other words, in order to achieve relative error  $\alpha$ , we only need the number of private samples  $n \sim \frac{1}{\alpha^2}$  if we omit the dependency on the other parameters. Besides, we also observe that the square of relative error is proportional to  $\frac{1}{\epsilon^2}$ , which matches our theoretical result.

From Figure 5.1b and 5.2b, we can see that the relative error increases as the dimensionality increases. It may seem a little weird that it is not linear in the dimensionality. We note that as the dimensionality  $p$  changes, some other parameters, for example, the  $l_2$  norm of the covariance matrix and  $w_\infty^*$  also change, which bring other effects to the relative error.

## Evaluation on real data

We first conduct experiment for GLM with logistic loss on the Covertype dataset [94]. Before running our algorithm, we first normalize the data and remove some co-related

---

<sup>11</sup>Note that in the studies on LDP ERM,  $\epsilon$  is always chosen as a large value such as [37].

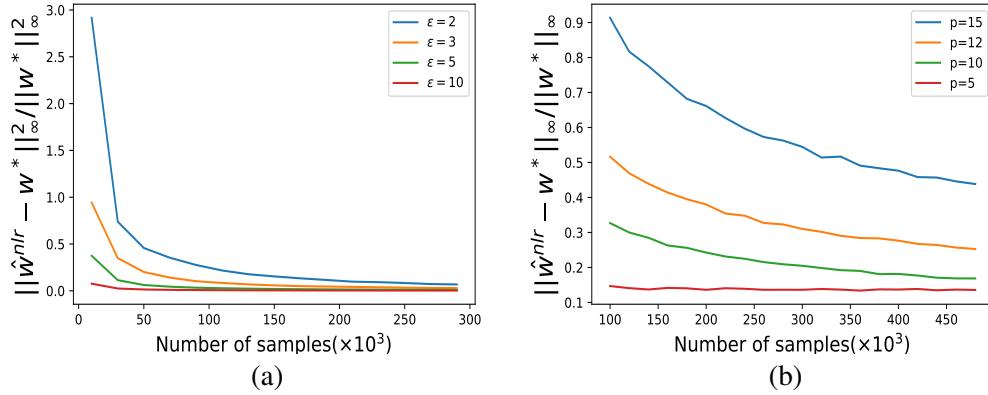


Figure 5.2: Cubic regression with i.i.d Bernoulli design. The left plot shows the squared relative error under different level of privacy. The right one shows relative error under different dimensionality.

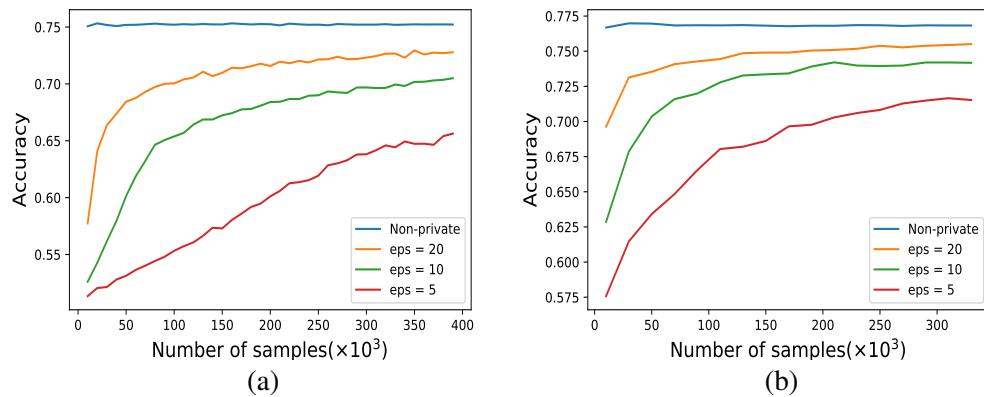


Figure 5.3: GLM with logistic loss on real dataset. The dataset we use is Covertype (left) and SUSY (right).

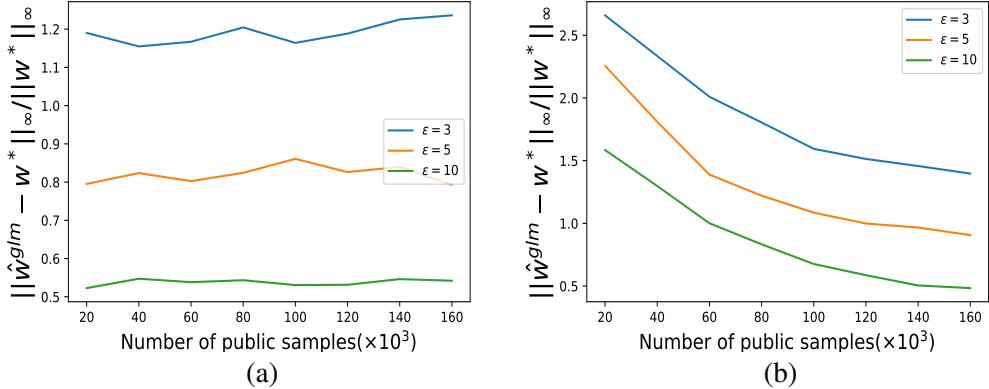


Figure 5.4: The effect of the number of public unlabeled samples. The left plot shows the relative error of GLM with logistic loss. The right one shows the relative error of cubic regression.

features. After the pre-processing, the dataset contains 581012 samples and 44 features. There are seven possible values for the label. Since multinomial logistic regression can not be regarded as a Generalized Linear Model, we consider a weaker test, which is to classify whether the label is Lodgepole Pine (type 2) or not. The chosen algorithm is still binary logistic regression. We divide the data into training and testing, where  $n_{\text{training}} = 406708$  and  $n_{\text{testing}} = 174304$  and randomly choose the sample size  $n \in 10^4 \cdot \{1, 2, 3, \dots, 39\}$  from the training data and use the same amount of public data. Regarding the privacy parameter, we take  $\delta = \frac{1}{n}$  and let  $\epsilon$  take value from  $\{20, 10, 5\}$ . We measure the performance by the prediction accuracy. For each combination of  $\epsilon$  and  $n$ , the experiment is repeated 1000 times. Through Figure 5.3a we observe that when  $\epsilon$  takes a reasonable value, the performance is approaching to the non-private case, provided that the size of private dataset is large enough. Thus, our algorithm is practical and is comparable to the non-private one.

We also conduct experiment for GLM with logistic loss on the SUSY dataset [22]. The task is to classify whether the class label is signal or background. After the pre-processing and sampling, the dataset contains 500000 samples and 18 features. Then we divide the data into training and testing, where  $n_{\text{training}} = 350000$  and  $n_{\text{testing}} = 150000$  and randomly choose the sample size  $n \in 10^4 \cdot \{1, 3, \dots, 33\}$  from the training data and use the same

amount of public data. Regarding the privacy parameter, we take  $\delta = \frac{1}{n}$  and let  $\epsilon$  take value from  $\{20, 10, 5\}$ . We measure the performance by the prediction accuracy. For each combination of  $\epsilon$  and  $n$ , the experiment is repeated 1000 times. As shown in Figure 5.3b, we have almost the same conclusion as in the Covertype case.

### The effect of public unlabeled data

We use similar setting as our synthetic experiments in Section 5.2.5. For GLM we consider the problem of binary logistic loss while for non-linear regression we will set  $f(x) = \frac{1}{3}x^3$  in (5.24). We compare relative error  $\frac{\|\hat{w}-w^*\|_\infty}{\|w^*\|_\infty}$  with respect to different privacy parameters  $\epsilon \in \{10, 5, 3\}$  with  $\delta = \frac{1}{n}$ . In these experiments, we fix dimensionality  $p = 10$  and the population parameter  $w^* = (1, 1, \dots, 1)/\sqrt{p}$ . We also fix the private sample size  $n = 200000$  and the public data size is chosen from the set  $10^4 \cdot \{2, 4, \dots, 16\}$ . We assume that the same amount of public unlabeled data is available. The features are generated independently from a Bernoulli distribution  $\Pr(x_{i,j} = \pm \frac{1}{p}) = 0.5$  and the label is generated according to the logistic model or the model (5.24). In non-linear regression model,  $\sigma$  is bounded by  $C = 0.001$ . The results are shown in Figure 5.4a and 5.4b.

### Further theoretical investigation motivated by experiments

Through the previous experimental results, we can also get some further theoretical investigation:

- Firstly, in all the previous experiments, we use the logistic loss for GLM and cubic function for non-linear regression. Actually, here these loss functions may not satisfy all the assumptions in Theorem 5.2.2 and 5.2.4. This indicate that theoretically we may relax these assumptions to get the same estimation error. Second, since our algorithm has good performance for real data, as shown in Figure 3. However, these real data may do not satisfy the assumptions in Theorem 5.2.2 and 5.2.4. Thus, we conjecture that it is possible to further relax the assumptions on the distribution of

samples.

- Theoretically, in Theorem 5.2.2, 5.2.3 and 5.2.4 we show that to achieve good performance we need sufficient large number of public unlabeled data, and this can be supported via Figure 5.4b. However, as shown in Figure 5.4a, sometimes there is no need to use as large amount of public data. Thus, this motivate us to further improve the sample complexity of public unlabeled data as future research.

## 5.2.6 Omitted Proofs

### Background and Auxiliary Lemmas

**Notations** For a positive semi-definite matrix  $M \in \mathbb{R}^{p \times p}$ , we define the  $M$ -norm for a vector  $w$  as  $\|w\|_M^2 = w^T M w$ .  $\lambda_{\min}(A)$  is the minimal singular value of the matrix  $A$ . For a semi positive definite matrix  $M \in \mathbb{R}^{p \times p}$ , let its SVD composition be  $\Sigma = U^T \Sigma U$ , where  $\Sigma = \text{diag}(\lambda_1, \dots, \lambda_p)$ , then  $M^{\frac{1}{2}}$  is defined as  $M^{\frac{1}{2}} = U^T \Sigma^{\frac{1}{2}} U$ , where  $\Sigma^{\frac{1}{2}} = \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_p})$ .

**Definition 5.2.2** (Sub-Gaussian). For a given constant  $\kappa$ , a random variable  $x \in \mathbb{R}$  is said to be sub-Gaussian if it satisfies  $\sup_{m \geq 1} \frac{1}{\sqrt{m}} \mathbb{E}[|x|^m]^{\frac{1}{m}} \leq \kappa$ . The smallest such  $\kappa$  is the **sub-Gaussian norm** of  $x$  and it is denoted by  $\|x\|_{\psi_2}$ . A random vector  $x \in \mathbb{R}^p$  is called a sub-Gaussian vector if there exists a constant  $\kappa$  such that for any unit vector  $v$ , we have  $\|\langle x, v \rangle\|_{\psi_2} \leq \kappa$ .

**Lemma 5.2.3** (Weyl's Inequality [264]). Let  $X, Y \in \mathbb{R}^{p \times p}$  be two symmetric matrices, and  $E = X - Y$ . Then, for all  $i = 1, \dots, p$ , we have

$$|\sigma_i(X) - \sigma_i(Y)| \leq \|E\|_2.$$

**Lemma 5.2.4.** Let  $w \in \mathbb{R}^p$  be a fixed vector and  $E$  be a symmetric Gaussian random matrix where the upper triangle entries are i.i.d Gaussian distribution  $\mathcal{N}(0, \sigma^2)$ . Then, with

probability at least  $1 - \xi$ , the following holds for a fixed positive semi-definite matrix

$$M \in \mathbb{R}^{p \times p}$$

$$\|Ew\|_M^2 \leq \sigma^2 \text{Tr}(M) \|w\|^2 \log \frac{2p^2}{\xi}.$$

*Proof of Lemma 5.2.4.* Let  $M = U^T \Sigma U$  denote the eigenvalue decomposition of  $M$ . Then, we have

$$\|Ew\|_M^2 = w^T E^T U^T \Sigma U E w = \sum_{i=1}^p \sigma_i \sum_{j=1}^p [UE]_{ij}^2 w_i^2.$$

Note that  $[UE]_{i,j} = \sum_{k=1}^p U_{i,k} E_{j,k}$  where  $E_{i,j}$  is Gaussian. Since  $U$  is orthogonal, we know that  $[UE]_{i,j} \sim \mathcal{N}(0, \sigma^2)$ . Using the Gaussian tail bound for all  $i, j \in [d]^2$ , we have

$$\mathbb{P}\left(\max_{i,j \in [p]^2} |[UE]_{i,j}| \geq \sqrt{\sigma^2 \log \frac{2p^2}{\xi}}\right) \leq \xi.$$

□

**Lemma 5.2.5** (Theorem 4.7.1 in [288]). Let  $x$  be a random vector in  $\mathbb{R}^p$  that is sub-Gaussian with covariance matrix  $\Sigma$  and  $\|\Sigma^{-\frac{1}{2}}x\|_{\psi_2} \leq \kappa_x$ . Then, with probability at least  $1 - \exp(-p)$ , the empirical covariance matrix  $\frac{1}{n}X^T X = \frac{1}{n} \sum_{i=1}^n x_i x_i^T$  satisfies

$$\left\| \frac{1}{n}X^T X - \Sigma \right\|_2 \leq C \kappa_x^2 \sqrt{\frac{p}{n}} \|\Sigma\|_2.$$

**Lemma 5.2.6** (Corollary 2.3.6 in [274]). Let  $M \in \mathbb{R}^{p \times p}$  be a symmetric matrix whose entries  $m_{ij}$  are independent for  $j > i$ , have mean zero, and are uniformly bounded in magnitude by 1. Then, there exists absolute constants  $C_2, c_1 > 0$  such that with probability at least  $1 - \exp(-C_2 c_1 p)$ , the following inequality holds  $\|M\|_2 \leq C \sqrt{p}$ .

Below we introduce some concentration lemmas given in [108].

**Lemma 5.2.7.** Let  $\mathbb{B}^\delta(\tilde{w})$  denote the ball centered at  $\tilde{w}$  and with radius  $\delta$  (*i.e.*,  $\mathbb{B}^\delta(\tilde{w}) = \{w : \|w - \tilde{w}\|_2 \leq \delta\}$ ). For  $i = 1, 2, \dots, n$ , let  $x_i \in \mathbb{R}^p$  be i.i.d isotropic sub-Gaussian random vectors with  $\|x_i\|_{\psi_2} \leq k_x$ , and  $\tilde{\mu} = \frac{\mathbb{E}[x]}{\sqrt{p}}$ . For any given function  $g : \mathbb{R} \mapsto \mathbb{R}$  that is

Lipschitz continuous with  $G$  and satisfies  $\sup_{w \in \mathbb{B}^\delta(\tilde{w})} \|g(\langle x, w \rangle)\|_{\psi_2} \leq \kappa_g$ , with probability at least  $1 - 2 \exp(-p)$ , the following holds for  $np > 51 \max\{\chi, \chi^2\}$

$$\sup_{w \in \mathbb{B}^\delta(\tilde{w})} \left| \frac{1}{m} \sum_{i=1}^m g(\langle x_i, w \rangle) - \mathbb{E}[g(\langle x, w \rangle)] \right| \leq c(\kappa_g + \frac{\kappa_x}{\tilde{u}}) \sqrt{\frac{p \log m}{m}},$$

where  $\chi = \frac{(\kappa_g + \frac{\kappa_x}{\tilde{u}})^2}{c \delta^2 G^2 \tilde{u}^2}$ .  $c$  is some absolute constant.

**Lemma 5.2.8.** Let  $\mathbb{B}^\delta(\tilde{w})$  be the ball centered at  $\tilde{w}$  and with radius  $\delta$  (i.e.,  $\mathbb{B}^\delta(\tilde{w}) = \{w : \|w - \tilde{w}\|_2 \leq \delta\}$ ). For  $i = 1, 2, \dots, n$ , let  $x_i \in \mathbb{R}^p$  be i.i.d sub-Gaussian random vectors with covariance matrix  $\Sigma$ . For any given function  $g : \mathbb{R} \mapsto \mathbb{R}$  that is uniformly bounded by  $L$  and Lipschitz continuous with  $G$ , the following holds with probability at least  $1 - \exp(-p)$

$$\sup_{w \in \mathbb{B}^\delta(\tilde{w})} \left| \frac{1}{m} \sum_{i=1}^m g(\langle x_i, w \rangle) - \mathbb{E}[g(\langle x, w \rangle)] \right| \leq 2\{G(\|\tilde{w}\|_2 + \delta)\|\Sigma\|_2 + L\} \sqrt{\frac{p}{m}}.$$

The following lemma shows that the private estimator  $\hat{w}^{ols}$  is close to the unperturbed one.

**Lemma 5.2.9.** Let  $X = [x_1^T; x_2^T; \dots; x_n^T] \in \mathbb{R}^{n \times d}$  be a matrix such that  $X^T X$  is invertible, and  $x_1, \dots, x_n$  are realizations of a sub-Gaussian random variable  $x$  which satisfies the condition of  $\|\Sigma^{-\frac{1}{2}}x\|_{\psi_2} \leq \kappa_x = O(1)$  and  $\Sigma = \mathbb{E}[xx^T]$  is the the population covariance matrix. Let  $\tilde{w}^{ols} = (X^T X)^{-1} X^T y$  denote the empirical linear regression estimator. Then, for sufficiently large  $n \geq \Omega(\frac{\kappa_x^4 \|\Sigma\|_2^2 p r^4 \log \frac{1}{\delta}}{\epsilon^2 \lambda_{\min}^2(\Sigma)})$ , the following holds with probability at least  $1 - \exp(-\Omega(p)) - \xi$ ,

$$\|\hat{w}^{ols} - \tilde{w}^{ols}\|_2^2 = O\left(\frac{pr^2(1 + r^2 \|\tilde{w}^{ols}\|_2^2) \log \frac{1}{\delta} \log \frac{p^2}{\xi}}{\epsilon^2 n \lambda_{\min}^2(\Sigma)}\right), \quad (5.28)$$

where  $r = r$  if  $x_i$  is sampled from some bounded distribution.

*Proof of Lemma 5.2.9.* It is obvious that  $\widehat{X^T X} = X^T X + E_1$ , where  $E_1$  is a symmetric Gaussian matrix with each entry sampled from  $\mathcal{N}(0, \sigma_1^2)$  and  $\sigma_1^2 = O(\frac{nr^4 \log \frac{1}{\delta}}{\epsilon^2})$ .  $\widehat{X^T y} =$

$X^T y + E_2$ , where  $E_2$  is a Gaussian vector sampled from  $\mathcal{N}(0, \sigma_2^2 I_p)$  and  $\sigma_2^2 = O\left(\frac{nr^2 \log \frac{1}{\delta}}{\epsilon^2}\right)$ .

We first show that  $\widehat{X^T X}$  is invertible with high probability under our assumption.

It is sufficient to show that  $X^T X + E_1 \succ \frac{X^T X}{2}$ , i.e.,  $\|E_1\|_2 \leq \frac{\lambda_{\min}(X^T X)}{2}$ . By Lemma 5.2.6, we can see that with probability  $1 - \exp(-\Omega(p))$ ,

$$\|E_1\|_2 \leq O\left(\frac{r^2 \sqrt{pn \log \frac{1}{\delta}}}{\epsilon}\right).$$

Also, by Lemma 5.2.5 and Lemma 5.2.3 we know that with probability at least  $1 - \exp(-\Omega(p))$ ,

$$\lambda_{\min}(X^T X) \geq n \lambda_{\min}(\Sigma) - O(\kappa_x^2 \|\Sigma\|_2 \sqrt{pn}).$$

Thus, it is sufficient to show that  $n \lambda_{\min}(\Sigma) \geq O\left(\frac{\kappa_x^2 \|\Sigma\|_2 r^2 \sqrt{pn \log \frac{1}{\delta}}}{\epsilon}\right)$ , which is true under the assumption of  $n \geq \Omega\left(\frac{\kappa_x^4 \|\Sigma\|_2^2 p r^4 \log \frac{1}{\delta}}{\epsilon^2 \lambda_{\min}^2(\Sigma)}\right)$ . Thus, with probability at least  $1 - \exp(-\Omega(p))$ , it is invertible. In the following we will always assume that this event holds.

By direct calculation we have

$$\|\hat{w}^{ols} - \tilde{w}^{ols}\|_2 = -(X^T X + E_1)^{-1} E_1 \tilde{w}^{ols} + (X^T X + E_1)^{-1} E_2.$$

Thus, by Cauchy-Schwartz inequality we get

$$\|\hat{w}^{ols} - \tilde{w}^{ols}\|_2^2 = O\left(\|E_1 \tilde{w}^{ols}\|_{(X^T X + E_1)^{-2}}^2 + \|E_2\|_{(X^T X + E_1)^{-2}}^2\right).$$

Since we already assume that  $X^T X + E_1 \succ \frac{X^T X}{2}$ , by Lemma 5.2.4 we can obtain the following with probability at least  $1 - \xi$

$$\begin{aligned} \|E_1 \tilde{w}^{ols}\|_{(X^T X + E_1)^{-2}}^2 &\leq O\left(\frac{nr^4 \log \frac{1}{\delta}}{\epsilon^2} \|\tilde{w}^{ols}\|_2^2 \text{Tr}((X^T X)^{-2}) \log \frac{4p^2}{\xi}\right) \\ \|E_2\|_{(X^T X + E_1)^{-2}}^2 &\leq O\left(\frac{nr^2 \log \frac{1}{\delta}}{\epsilon^2} \text{Tr}((X^T X)^{-2}) \frac{4p}{\xi}\right). \end{aligned}$$

Thus, we have

$$\|\hat{w}^{ols} - \tilde{w}^{ols}\|_2^2 \leq C_1 n \cdot \frac{r^2(1 + r^2\|\tilde{w}^{ols}\|_2^2) \log \frac{1}{\delta} \log \frac{p^2}{\xi}}{\epsilon^2} \text{Tr}((X^T X)^{-2}).$$

For the term of  $\text{Tr}((X^T X)^{-2})$ , we get

$$\text{Tr}((X^T X)^{-2}) \leq (\text{Tr}((X^T X)^{-1}))^2 \leq p\|(X^T X)^{-2}\|_2^2 = \frac{p}{\lambda_{\min}^2(X^T X)} \leq O\left(\frac{p}{n^2 \lambda_{\min}^2(\Sigma)}\right),$$

where the last inequality is due to the fact that  $\lambda_{\min}(X^T X) \geq n\lambda_{\min}(\Sigma) - O(\kappa_x^2 \|\Sigma\|_2 \sqrt{pn}) \geq \frac{1}{2}n\lambda_{\min}(\Sigma)$  (by the assumption on  $n$ ). This completes the proof.  $\square$

Let  $w^{ols} = (\mathbb{E}[xx^T])^{-1}\mathbb{E}[xy]$  denote the population linear regression estimator. The following lemma bounds the estimation error between  $\tilde{w}^{ols}$  and  $w^{ols}$ . The proof could be found in [108] or [91].

**Lemma 5.2.10** (Prop. 7 in [108]). Assume that  $\mathbb{E}[x_i] = 0$ ,  $\mathbb{E}[x_i x_i^T] = \Sigma$ , and  $\Sigma^{-\frac{1}{2}}x_i$  and  $y_i$  are sub-Gaussian with norms  $\kappa_x$  and  $\gamma$ , respectively. If  $n \geq \Omega(\kappa_x \gamma p)$ , the following holds

$$\|\tilde{w}^{ols} - w^{ols}\|_2 \leq O\left(\gamma \kappa_x \sqrt{\frac{p}{n \lambda_{\min}(\Sigma)}}\right),$$

with probability at least  $1 - 3 \exp(-p)$ .

## Proofs of LDP

The LDP proof of Algorithm 5.2.39 follows from Gaussian mechanism and the composition property of DP.

For Algorithm 5.2.40, it is  $(\epsilon, \delta)$ -LDP due to the  $\ell_2$ -norm bound on  $\|x_i y_i\|_2 = \|x_i\|_2 \|f(\langle x, w^* \rangle) + \sigma_i\|_2 \leq \|x_i\|_2 (L\|x\|_2 + |f(0)| + C)$ , where the last inequality is due to the fact that  $f'$  is  $L$ -bounded and  $\|w^*\|_2 \leq 1$ . That is,  $|f(\langle x, w^* \rangle) - f(0)| \leq L|\langle x, w^* \rangle - 0| \leq L\|x\|_2 \|w^*\|_2$ .

### 5.2.2.

### Proof of Theorem 5.2.3

Since Theorem 5.2.3 is the most complicated one, we will first prove it and then Theorem

Since  $r = O(1)$  (by assumption), combining this with Lemmas 5.2.9 and 5.2.10, we have that with probability at least  $1 - \exp(-\Omega(p)) - \xi$  and under the assumption on  $n$ , there is a constant  $C_3 > 0$  such that

$$\|\hat{w}^{ols} - w^{ols}\|_2 \leq C_3 \frac{\kappa_x \sqrt{pr^2} \|w^{ols}\|_2 \sqrt{\log \frac{1}{\delta} \log \frac{p^2}{\xi}}}{\epsilon \sqrt{n} \lambda_{\min}^{1/2}(\Sigma) \min\{\lambda_{\min}^{1/2}(\Sigma), 1\}}. \quad (5.29)$$

**Lemma 5.2.11.** Let  $\Phi^{(2)}$  be a function that is Lipschitz continuous with constant  $G$ , and  $f : \mathbb{R} \times \mathbb{R}^p \mapsto \mathbb{R}$  be another function such that  $f(c, w) = c\mathbb{E}[\Phi^{(2)}(\langle x, w \rangle c)]$  and its empirical one is

$$\hat{f}(c, w) = \frac{c}{m} \sum_{j=1}^m \Phi^{(2)}(\langle x_j, w \rangle c).$$

Let  $\mathbb{B}^\delta(\bar{w}^{ols}) = \{w : \|w - \bar{w}^{ols}\|_2 \leq \delta\}$ , where  $\bar{w}^{ols} = \Sigma^{\frac{1}{2}} w^{ols}$ . Under the assumptions in Lemma 5.2.9 and Eq. (5.29), if further assume that  $\|\Sigma^{-\frac{1}{2}} x\|_{\psi_2} \leq \kappa_x$ , and  $\sup_{w \in \mathbb{B}^\delta(\bar{w}^{ols})} \|\Phi^{(2)}(\langle x, w \rangle)\|_{\psi_2} \leq \kappa_g$ , and there exist  $\bar{c} > 0$  and  $\tau > 0$  such that  $f(\bar{c}, w^{ols}) \geq 1 + \tau$ , then there is  $\bar{c}_\Phi \in (0, \bar{c})$  such that  $1 = f(\bar{c}_\Phi, w^{ols})$ . Also, for sufficiently large  $n$  and  $m$  such that

$$m \geq \Omega\left((\kappa_g + \frac{\kappa_x}{\tilde{\mu}})^2 \max\{p \log m \tau^{-2}, \frac{1}{G^2 \tilde{\mu}^2} \frac{\epsilon^2 n}{pr^4 \|w^{ols}\|_2^2 \log \frac{1}{\delta} \log \frac{p^2}{\xi} \|\Sigma\|_2}\}\right), \quad (5.30)$$

$$n \geq \Omega\left(\kappa_x^4 G^2 \bar{c}^4 \|\Sigma\|_2 \frac{pr^4 \|w^{ols}\|_2^2 \log \frac{1}{\delta} \log \frac{p^2}{\xi}}{\tau^2 \epsilon^2 \lambda_{\min}(\Sigma) \min\{\lambda_{\min}(\Sigma), 1\}}\right), \quad (5.31)$$

with probability at least  $1 - 2 \exp(-p)$ , there exists a  $\hat{c}_\Phi \in [0, \bar{c}]$  such that  $\hat{f}(\hat{c}_\Phi, \hat{w}^{ols}) = 1$ . Furthermore, if the derivative of  $c \mapsto f(c, w^{ols})$  is bounded below in the absolute value (*i.e.*,

does not change sign) by  $M > 0$  in the interval  $c \in [0, \bar{c}]$ , then the following holds

$$|\hat{c}_\Phi - \bar{c}_\Phi| \leq O\left(M^{-1}\bar{c}(\kappa_g + \frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p \log m}{m}} + M^{-1}G\kappa_x^2\bar{c}^2\|\Sigma\|_2^{\frac{1}{2}}\frac{\sqrt{pr^2}\|w^{ols}\|_2\sqrt{\log \frac{1}{\delta} \log \frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma), 1\}}\right). \quad (5.32)$$

**Proof of Lemma 5.2.11.** We divide the proof into three parts.

**Part 1: Existence of  $\bar{c}_\Phi$ :** From the definition, we know that  $f(0, w^{ols}) = 0$  and  $f(\bar{c}, w^{ols}) > 1$ . Since  $f$  is continuous, we known that there exists a constant  $\bar{c}_\Phi \in (0, \bar{c})$  which satisfying  $f(\bar{c}_\Phi, w^{ols}) = 0$ .

**Part 2: Existence of  $\hat{c}_\Phi$ :** For simplicity, we use the following notations.

$$\delta = C_3 \frac{\kappa_x \sqrt{pr^2} \|w^{ols}\|_2 \sqrt{\log \frac{1}{\delta} \log \frac{p^2}{\xi}}}{\epsilon \sqrt{n} \min\{\lambda_{\min}^{1/2}(\Sigma), 1\}}, \quad \delta' = \frac{\|\Sigma\|_2^{\frac{1}{2}} \delta}{\lambda_{\min}^{\frac{1}{2}}(\Sigma)}, \quad (5.33)$$

where  $C_3$  is the one in (5.29). Thus,  $\|\Sigma^{\frac{1}{2}}\hat{w}^{ols} - \Sigma^{\frac{1}{2}}w^{ols}\|_2 \leq \delta'$ .

Now consider the term of  $|\hat{f}(c, \hat{w}^{ols}) - f(c, \hat{w}^{ols})|$  for  $c \in [0, \bar{c}]$ . We have

$$\sup_{c \in [0, \bar{c}]} |\hat{f}(c, \hat{w}^{ols}) - f(c, \hat{w}^{ols})| \leq \sup_{c \in [0, \bar{c}]} \sup_{w \in \mathbb{B}_\Sigma^{\delta'}(w^{ols})} |\hat{f}(c, w) - f(c, w)|, \quad (5.34)$$

where  $\mathbb{B}_\Sigma^{\delta'}(w^{ols}) = \{w : \|\Sigma^{\frac{1}{2}}w - \Sigma^{\frac{1}{2}}w^{ols}\|_2 \leq \delta'\}$ .

Note that for any  $x$ , we have  $\langle x, w \rangle = \langle v, \Sigma^{\frac{1}{2}}w \rangle$ , where  $v = \Sigma^{-\frac{1}{2}}x$  follows an isotropic sub-Gaussian distribution. Also, by definition we know that  $w \in \mathbb{B}_\Sigma^{\delta'}(w^{ols})$  is equivalent to

$\Sigma^{\frac{1}{2}}w \in \mathbb{B}^{\delta'}(\bar{w}^{ols})$ . Thus, we have

$$\begin{aligned}
& \sup_{c \in [0, \bar{c}]} \sup_{w \in \mathbb{B}_{\Sigma}^{\delta'}(w^{ols})} |\hat{f}(c, \hat{w}^{ols}) - f(c, \hat{w}^{ols})| \\
& \leq \bar{c} \sup_{c \in [0, \bar{c}]} \sup_{w \in \mathbb{B}_{\Sigma}^{\delta'}(w^{ols})} \left| \frac{1}{m} \sum_{j=1}^m \Phi^{(2)}(\langle v_i, \Sigma^{\frac{1}{2}}w \rangle c) - \mathbb{E}\Phi^{(2)}(\langle v, \Sigma^{\frac{1}{2}}w \rangle c) \right| \\
& = \bar{c} \sup_{c \in [0, \bar{c}]} \sup_{\Sigma^{\frac{1}{2}}w \in \mathbb{B}^{\delta'}(\bar{w}^{ols})} \left| \frac{1}{m} \sum_{j=1}^m \Phi^{(2)}(\langle v_i, \Sigma^{\frac{1}{2}}w \rangle c) - \mathbb{E}\Phi^{(2)}(\langle v, \Sigma^{\frac{1}{2}}w \rangle c) \right| \\
& = \bar{c} \sup_{w' \in \mathbb{B}^{\bar{c}\delta'}(\bar{w}^{ols})} \left| \frac{1}{m} \sum_{j=1}^m \Phi^{(2)}(\langle v_i, w' \rangle) - \mathbb{E}\Phi^{(2)}(\langle v, w' \rangle) \right|. \tag{5.35}
\end{aligned}$$

By Lemma 5.2.7, we know that when  $mp \geq 51 \max\{\chi, \chi^{-1}\}$ , where

$$\chi = \frac{(\kappa_g + \frac{\kappa_x}{\tilde{\mu}})^2}{c\delta'^2 G^2 \tilde{\mu}^2} = \Theta\left(\frac{(\kappa_g + \frac{\kappa_x}{\tilde{\mu}})^2 \epsilon^2 n \lambda_{\min}(\Sigma) \min\{\lambda_{\min}(\Sigma), 1\}}{G^2 \tilde{\mu}^2 pr^4 \|w^{ols}\|_2^2 \log \frac{1}{\delta} \log \frac{p^2}{\xi} \|\Sigma\|_2}\right),$$

the following holds with probability at least  $1 - 2 \exp(-p)$

$$\sup_{w' \in \mathbb{B}^{\bar{c}\delta'}(\bar{w}^{ols})} \left| \frac{1}{m} \sum_{j=1}^m \Phi^{(2)}(\langle v_i, w' \rangle) - \mathbb{E}\Phi^{(2)}(\langle v, w' \rangle) \right| \leq O((\kappa_g + \frac{\kappa_x}{\tilde{\mu}}) \sqrt{\frac{p \log m}{m}}). \tag{5.36}$$

By the Lipschitz property of  $\Phi^{(2)}$ , we have that for any  $w_1$  and  $w_2$ ,

$$\begin{aligned}
& \sup_{c \in [0, \bar{c}]} |f(c, w_1) - f(c, w_2)| \leq G\bar{c}^2 \mathbb{E}[\langle v, \Sigma^{\frac{1}{2}}(w_1 - w_2) \rangle] \\
& \leq \kappa_x G\bar{c}^2 \|\Sigma^{\frac{1}{2}}(w_1 - w_2)\|_2. \tag{5.37}
\end{aligned}$$

Taking  $w_1 = \hat{w}^{ols}$  and  $w_2 = w^{ols}$ , we have

$$\sup_{c \in [0, \bar{c}]} |f(c, \hat{w}^{ols}) - f(c, w^{ols})| \leq O\left(\kappa_x G\bar{c}^2 \|\Sigma\|_2^{\frac{1}{2}} \frac{\delta}{\lambda_{\min}^{\frac{1}{2}}(\Sigma)}\right).$$

Combining this with (5.35), (5.36), (5.37), and taking  $\delta$  as in (5.33), we get

$$\sup_{c \in [0, \bar{c}]} |\hat{f}(c, \hat{w}^{ols}) - f(c, w^{ols})| \leq O\left(\bar{c}(\kappa_g + \frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p \log m}{m}} + G\bar{c}^2\|\Sigma\|_2^{\frac{1}{2}} \frac{\kappa_x^2 \sqrt{pr^2}\|w^{ols}\|_2 \sqrt{\log \frac{1}{\delta} \log \frac{p^2}{\xi}}}{\epsilon \sqrt{n} \lambda_{\min}^{1/2}(\Sigma) \min\{\lambda_{\min}^{1/2}(\Sigma), 1\}}\right). \quad (5.38)$$

Let  $B$  denote the RHS of (5.38). If  $c = \bar{c}$ , we have  $\hat{f}(\bar{c}, \hat{w}^{ols}) \geq 1 + \tau - B$ . Thus, if  $B \leq \tau$ , there must exist a  $\hat{c}_\Phi \in [0, \bar{c}]$  such that  $\hat{f}(\hat{c}_\Phi, \hat{w}^{ols}) = 1$ .

To ensure that  $B \leq \tau$  holds, it is sufficient to have

$$O\left(\bar{c}(\kappa_g + \frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p \log m}{m}}\right) \leq \frac{\tau}{2}$$

and

$$O\left(G\bar{c}^2\|\Sigma\|_2^{\frac{1}{2}} \frac{\kappa_x^2 \sqrt{pr^2}\|w^{ols}\|_2 \sqrt{\log \frac{1}{\delta} \log \frac{p^2}{\xi}}}{\epsilon \sqrt{n} \lambda_{\min}^{1/2}(\Sigma) \min\{\lambda_{\min}^{1/2}(\Sigma), 1\}}\right) \leq \frac{\tau}{2}.$$

This means that

$$\begin{aligned} m &\geq \Omega\left(\bar{c}^2(\kappa_g + \frac{\kappa_x}{\tilde{\mu}})^2 p \log m \tau^{-2}\right), \\ n &\geq \Omega\left(\kappa_x^4 G^2 \bar{c}^4 \|\Sigma\|_2 \frac{pr^4 \|w^{ols}\|_2^2 \log \frac{1}{\delta} \log \frac{p^2}{\xi}}{\tau^2 \epsilon^2 \lambda_{\min}(\Sigma) \min\{\lambda_{\min}(\Sigma), 1\}}\right), \end{aligned}$$

which are assumed in the lemma.

**Part 3: Estimation Error:** So far, we know that  $\hat{f}(\hat{c}_\Phi, \hat{w}^{ols}) = f(\bar{c}_\Phi, w^{ols}) = 1$  with high probability. By (5.34), (5.35) and (5.36), we have

$$|1 - f(\hat{c}_\Phi, \hat{w}^{ols})| = |\hat{f}(\hat{c}_\Phi, \hat{w}^{ols}) - f(\hat{c}_\Phi, \hat{w}^{ols})| \leq O\left(\bar{c}(\kappa_g + \frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p \log m}{m}}\right).$$

By the same argument for (5.38), we have

$$|f(\hat{c}_\Phi, \hat{w}^{ols}) - f(\hat{c}_\Phi, w^{ols})| \leq G\kappa_x \bar{c}^2 \|\Sigma\|_2^{\frac{1}{2}} \frac{\delta}{\lambda_{\min}^{\frac{1}{2}}(\Sigma)}.$$

Thus, using Taylor expansion on  $f(c, w^{ols})$  around  $c_\Phi$  and by the assumption of the bounded derivative of  $f$ , we have

$$\begin{aligned} M|\hat{c}_\Phi - \bar{c}_\Phi| &\leq |f(\hat{c}_\Phi, w^{ols}) - f(\bar{c}_\Phi, w^{ols})| \\ &\leq |f(\hat{c}_\Phi, w^{ols}) - f(\hat{c}_\Phi, \hat{w}^{ols})| + |f(\hat{c}_\Phi, \hat{w}^{ols}) - 1| \\ &\leq O\left(\bar{c}(\kappa_g + \frac{\kappa_x}{\tilde{\mu}})\sqrt{\frac{p \log m}{m}} + G\kappa_x^2 \bar{c}^2 \|\Sigma\|_2^{\frac{1}{2}} \frac{\sqrt{pr^2} \|w^{ols}\|_2 \sqrt{\log \frac{1}{\delta} \log \frac{p^2}{\xi}}}{\epsilon \sqrt{n} \lambda_{\min}^{1/2}(\Sigma) \min\{\lambda_{\min}^{1/2}(\Sigma), 1\}}\right). \end{aligned}$$

□

Next, we prove our main theorem.

**Proof of Theorem 5.2.3.** By definition, we have

$$\begin{aligned} \|\hat{w}^{glm} - w^*\|_\infty &\leq \|\hat{c}_\Phi \hat{w}^{ols} - \bar{c}_\Phi w^{ols}\|_\infty + \|\bar{c}_\Phi w^{ols} - w^*\|_\infty \\ &\leq \|\hat{c}_\Phi \hat{w}^{ols} - \bar{c}_\Phi w^{ols}\|_\infty + \|\bar{c}_\Phi w^{ols} - c_\Phi w^{ols}\|_\infty + \|c_\Phi w^{ols} - w^*\|_\infty. \end{aligned} \tag{5.39}$$

We first bound the term of  $|\bar{c}_\Phi - c_\Phi|$ . Since  $\bar{c}_\Phi \mathbb{E}[\Phi^{(2)}(\langle x, w^{ols} \rangle \bar{c}_\Phi)] = 1$  and  $c_\Phi \mathbb{E}[\Phi^{(2)}(\langle x, w^* \rangle)] = 1$  (by definition), we get

$$\begin{aligned} |f(\bar{c}_\Phi, w^{ols}) - f(c_\Phi, w^{ols})| &= |c_\Phi \mathbb{E}[\Phi^{(2)}(\langle x, w^* \rangle)] - f(c_\Phi, w^{ols})| \\ &\leq c_\Phi |\mathbb{E}[\Phi^{(2)}(\langle x, w^* \rangle)] - \Phi^{(2)}(\langle x, w^{ols} \rangle c_\Phi)| \\ &\leq c_\Phi G |\mathbb{E}[\langle x, (w^* - c_\Phi w^{ols}) \rangle]| \\ &\leq c_\Phi G \|(w^* - c_\Phi w^{ols})\|_\infty \mathbb{E}\|x\|_1 \\ &\leq c_\Phi G r \|c_\Phi w^{ols} - w^*\|_\infty, \end{aligned}$$

where the last inequality is due to the assumption that  $\|x\|_1 \leq r$ .

Thus, by the assumption of the bounded deviation of  $f(c, w^{ols})$  on  $[0, \max\{\bar{c}, c_\Phi\}]$ , we

have

$$M|\bar{c}_\Phi - c_\Phi| \leq |f(\bar{c}_\Phi, w^{ols}) - f(c_\Phi, w^{ols})| \leq c_\Phi Gr \|c_\Phi w^{ols} - w^*\|_\infty.$$

By Lemma 5.2.2 in the context, we have

$$|\bar{c}_\Phi - c_\Phi| \leq 16M^{-1}c_\Phi G^2 r^2 \kappa_x^3 \sqrt{\rho_2} \rho_\infty \frac{\|w^*\|_\infty^2}{\sqrt{p}}. \quad (5.40)$$

Thus, the second term of (5.39) is bounded by

$$\begin{aligned} \|\bar{c}_\Phi w^{ols} - c_\Phi w^{ols}\|_\infty &\leq 16M^{-1}c_\Phi G^2 r^2 \kappa_x^3 \sqrt{\rho_2} \rho_\infty \frac{\|w^*\|_\infty^2}{\sqrt{p}} \|w^{ols}\|_\infty \\ &\leq 16M^{-1}c_\Phi G^2 r^2 \kappa_x^3 \sqrt{\rho_2} \rho_\infty \frac{\|w^*\|_\infty^3}{\sqrt{p}} \left( \frac{1}{c_\Phi} + 16Gr\kappa_x^3 \sqrt{\rho_2} \rho_\infty \frac{\|w^*\|_\infty}{\sqrt{p}} \right) \\ &= O\left(M^{-1}r^3 \kappa_x^6 G^3 \rho_2 \rho_\infty^2 \frac{\|w^*\|_\infty^3 \max\{1, \|w^*\|_\infty\}}{\sqrt{p}} \max\{1, c_\Phi\}\right), \end{aligned} \quad (5.41)$$

where the last inequality is due to Lemma 5.2.2 in the context.

By Lemma 5.2.2 in the context, the third term of (5.39) is bounded by  $16c_\Phi Gr\kappa_x^3 \sqrt{\rho_2} \rho_\infty \frac{\|w^*\|_\infty^2}{\sqrt{p}}$ .

For the first term of (5.39), by (5.29) and Lemma 5.2.11 we have

$$\begin{aligned} \|\hat{c}_\Phi \hat{w}^{ols} - \bar{c}_\Phi w^{ols}\|_\infty &\leq |\hat{c}_\Phi| \cdot \|\hat{w}^{ols} - w^{ols}\|_\infty + |\hat{c}_\Phi - \bar{c}_\Phi| \cdot \|w^{ols}\|_\infty \\ &\leq O\left(\bar{c} \frac{\kappa_x \sqrt{pr^2} \|w^{ols}\|_2 \sqrt{\log \frac{1}{\delta} \log \frac{p^2}{\xi}}}{\epsilon \sqrt{n} \lambda_{\min}^{1/2}(\Sigma) \min\{\lambda_{\min}^{1/2}(\Sigma), 1\}}\right. \\ &\quad \left. + \|w^{ols}\|_\infty (M^{-1} \bar{c} (\kappa_g + \frac{\kappa_x}{\tilde{\mu}}) \sqrt{\frac{p \log m}{m}} + M^{-1} G \kappa_x^2 \bar{c}^2 \|\Sigma\|_2^{\frac{1}{2}} \frac{\sqrt{pr^2} \|w^{ols}\|_2 \sqrt{\log \frac{1}{\delta} \log \frac{p^2}{\xi}}}{\epsilon \sqrt{n} \lambda_{\min}^{1/2}(\Sigma) \min\{\lambda_{\min}^{1/2}(\Sigma), 1\}})\right). \end{aligned} \quad (5.42)$$

For the first term of (5.42), we have

$$\begin{aligned}
& \bar{c} \frac{\kappa_x \sqrt{pr^2} \|w^{ols}\|_2 \sqrt{\log \frac{1}{\delta} \log \frac{p^2}{\xi}}}{\epsilon \sqrt{n} \lambda_{\min}^{1/2}(\Sigma) \min\{\lambda_{\min}^{1/2}(\Sigma), 1\}} \leq \bar{c} \frac{\kappa_x pr^2 \|w^{ols}\|_\infty \sqrt{\log \frac{1}{\delta} \log \frac{p^2}{\xi}}}{\epsilon \sqrt{n} \lambda_{\min}^{1/2}(\Sigma) \min\{\lambda_{\min}^{1/2}(\Sigma), 1\}} \\
& \leq \bar{c} \frac{\kappa_x pr^2 \|w^*\|_\infty \sqrt{\log \frac{1}{\delta} \log \frac{p^2}{\xi}}}{\epsilon \sqrt{n} \lambda_{\min}^{1/2}(\Sigma) \min\{\lambda_{\min}^{1/2}(\Sigma), 1\}} \left( \frac{1}{c_\Phi} + 16Gr\kappa_x^3 \sqrt{\rho_2} \rho_\infty \frac{\|w^*\|_\infty}{\sqrt{p}} \right) \\
& = O\left(\bar{c} \frac{p\kappa_x^4 \sqrt{\rho_2} \rho_\infty Gr^3 \|w^*\|_\infty \max\{1, \|w^*\|_\infty\} \sqrt{\log \frac{1}{\delta} \log \frac{p^2}{\xi}}}{\epsilon \sqrt{n} \lambda_{\min}^{1/2}(\Sigma) \min\{\lambda_{\min}^{1/2}(\Sigma), 1\}} \max\{1, \frac{1}{c_\Phi}\}\right). \quad (5.43)
\end{aligned}$$

For the second term of (5.42), we have

$$\begin{aligned}
& \|w^{ols}\|_\infty M^{-1} \bar{c} \left( \kappa_g + \frac{\kappa_x}{\tilde{\mu}} \right) \sqrt{\frac{p \log m}{m}} \\
& \leq \bar{c} \|w^*\|_\infty \left( \kappa_g + \frac{\kappa_x}{\tilde{\mu}} \right) \sqrt{\frac{p \log m}{m}} \left( \frac{1}{c_\Phi} + 16Gr\kappa_x^3 \sqrt{\rho_2} \rho_\infty \frac{\|w^*\|_\infty}{\sqrt{p}} \right) \\
& \leq O\left( Gr\kappa_x^3 \sqrt{\rho_2} \rho_\infty \bar{c} \|w^*\|_\infty \max\{1, \|w^*\|_\infty\} \left( \kappa_g + \frac{\kappa_x}{\tilde{\mu}} \right) \sqrt{\frac{p \log m}{m}} \max\{1, \frac{1}{c_\Phi}\} \right). \quad (5.44)
\end{aligned}$$

For the third term of (5.42), we have

$$\begin{aligned}
& \|w^{ols}\|_\infty M^{-1} G \kappa_x^2 \bar{c}^2 \|\Sigma\|_2^{\frac{1}{2}} \frac{\sqrt{pr^2} \|w^{ols}\|_2 \sqrt{\log \frac{1}{\delta} \log \frac{p^2}{\xi}}}{\epsilon \sqrt{n} \lambda_{\min}^{1/2}(\Sigma) \min\{\lambda_{\min}^{1/2}(\Sigma), 1\}} \\
& \leq M^{-1} G \kappa_x^2 \bar{c}^2 \|\Sigma\|_2^{\frac{1}{2}} \frac{pr^2 \|w^*\|_\infty^2 \sqrt{\log \frac{1}{\delta} \log \frac{p^2}{\xi}}}{\epsilon \sqrt{n} \lambda_{\min}^{1/2}(\Sigma) \min\{\lambda_{\min}^{1/2}(\Sigma), 1\}} \left( \frac{1}{c_\Phi} + 16Gr\kappa_x^3 \sqrt{\rho_2} \rho_\infty \frac{\|w^*\|_\infty}{\sqrt{p}} \right)^2 \\
& \leq O\left(M^{-1} G^3 \kappa_x^8 \bar{c}^2 \rho_2 \rho_\infty^2 \|\Sigma\|_2^{\frac{1}{2}} \frac{pr^4 \|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\} \sqrt{\log \frac{1}{\delta} \log \frac{p^2}{\xi}}}{\epsilon \sqrt{n} \lambda_{\min}^{1/2}(\Sigma) \min\{\lambda_{\min}^{1/2}(\Sigma), 1\}} \max\{1, \frac{1}{c_\Phi}\}^2\right). \quad (5.45)
\end{aligned}$$

Thus, the first term of (5.39) is bounded by (since  $m \geq \Omega(n)$ )

$$\begin{aligned}
\|\hat{c}_\Phi \hat{w}^{ols} - \bar{c}_\Phi w^{ols}\|_\infty &\leq O\left(\bar{c} \frac{p\kappa_x^4 \sqrt{\rho_2} \rho_\infty G r^3 \|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty\} \sqrt{\log \frac{1}{\delta} \log \frac{p^2}{\xi}}}{\epsilon \sqrt{n} \lambda_{\min}^{1/2}(\Sigma) \min\{\lambda_{\min}^{1/2}(\Sigma), 1\}} \max\{1, \frac{1}{c_\Phi}\}\right. \\
&\quad \left. + Gr\kappa_x^3 \sqrt{\rho_2} \rho_\infty \bar{c} \|w^*\|_\infty \max\{1, \|w^*\|_\infty\} (\kappa_g + \frac{\kappa_x}{\tilde{\mu}}) \sqrt{\frac{p \log m}{m}} \max\{1, \frac{1}{c_\Phi}\} + \right. \\
&\quad \left. M^{-1} G^3 \kappa_x^8 \bar{c}^2 \rho_2 \rho_\infty^2 \|\Sigma^{\frac{1}{2}}\|_2 \frac{pr^4 \|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\} \sqrt{\log \frac{1}{\delta} \log \frac{p^2}{\xi}}}{\epsilon \sqrt{n} \lambda_{\min}^{1/2}(\Sigma) \min\{\lambda_{\min}^{1/2}(\Sigma), 1\}} \max\{1, \frac{1}{c_\Phi}\}^2 \right. \\
&= O\left(M^{-1} (\kappa_g + \frac{\kappa_x}{\tilde{\mu}}) G^3 \kappa_x^8 \bar{c}^2 \rho_2 \rho_\infty^2 \|\Sigma^{\frac{1}{2}}\|_2 \right. \\
&\quad \left. \times \frac{pr^4 \|w^*\|_\infty \max\{1, \|w^*\|_\infty^3\} \sqrt{\log m \log \frac{1}{\delta} \log \frac{p^2}{\xi}}}{\epsilon \sqrt{n} \lambda_{\min}^{1/2}(\Sigma) \min\{\lambda_{\min}^{1/2}(\Sigma), 1\}} \max\{1, \frac{1}{c_\Phi}\}^2\right).
\end{aligned}$$

Putting all the bounds together, we have

$$\begin{aligned}
\|\hat{w}^{glm} - w^*\|_\infty &\leq \tilde{O}\left(M^{-1} G^3 \kappa_x^8 \bar{c}^2 \rho_2 \rho_\infty^2 \|\Sigma^{\frac{1}{2}}\|_2 \right. \\
&\quad \times \frac{pr^4 \|w^*\|_\infty \max\{1, \|w^*\|_\infty^3\} \sqrt{\log \frac{1}{\delta} \log \frac{p^2}{\xi}}}{\epsilon \sqrt{n} \lambda_{\min}^{1/2}(\Sigma) \min\{\lambda_{\min}^{1/2}(\Sigma), 1\}} \max\{1, \frac{1}{c_\Phi}\}^2 \\
&\quad + M^{-1} r^3 \kappa_x^6 c_\Phi G^3 \rho_2 \rho_\infty^2 \frac{\|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\}}{\sqrt{p}} \max\{1, \frac{1}{c_\Phi}\} + \\
&\quad \left. Gr\kappa_x^3 \sqrt{\rho_2} \rho_\infty \bar{c} \|w^*\|_\infty \max\{1, \|w^*\|_\infty\} (\kappa_g + \frac{\kappa_x}{\tilde{\mu}}) \sqrt{\frac{p \log m}{m}} \max\{1, \frac{1}{c_\Phi}\}\right). \quad (5.46)
\end{aligned}$$

Next, we bound the probability. We assume that Lemma 5.2.9, 5.2.10 and 5.2.11 hold with probability at least  $1 - \exp(-\Omega(p)) - \rho$ . They hold when

$$m \geq \Omega\left((\kappa_g + \frac{\kappa_x}{\tilde{\mu}})^2 \max\{p \log m \tau^{-2}, \frac{1}{G^2 \tilde{\mu}^2} \frac{\epsilon^2 n}{pr^4 \|w^{ols}\|_2^2 \log \frac{1}{\delta} \log \frac{p^2}{\xi}}\}\right), \quad (5.47)$$

$$n \geq \Omega\left(\max\{\kappa_x^4 G^2 \bar{c}^4 \|\Sigma\|_2 \frac{pr^4 \|w^{ols}\|_2^2 \log \frac{1}{\delta} \log \frac{p^2}{\xi}}{\tau^2 \epsilon^2 \lambda_{\min}(\Sigma) \min\{\lambda_{\min}(\Sigma), 1\}}, \frac{\kappa_x^4 \|\Sigma\|_2^2 pr^4 \log \frac{1}{\delta}}{\epsilon^2 \lambda_{\min}^2(\Sigma)}\}\right). \quad (5.48)$$

Since  $\|w^{ols}\|_2 \leq \sqrt{p}\|w^*\|_\infty(\frac{1}{c_\Phi} + 16Gr\kappa_x^3\sqrt{\rho_2}\rho_\infty\frac{\|w^*\|_\infty}{\sqrt{p}})$ , it suffices for  $n$

$$n \geq \Omega\left(G^4\bar{c}^4\|\Sigma\|_2^2 \frac{p^2r^6\kappa_x^{10}\rho_2\rho_\infty^2\|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\} \log\frac{1}{\delta} \log\frac{p^2}{\xi}}{\tau^2\epsilon^2\lambda_{\min}(\Sigma) \min\{\lambda_{\min}(\Sigma), 1\}} \max\{1, \frac{1}{c_\Phi}\}^2\right). \quad (5.49)$$

□

### Proof of Theorem 5.2.2

**Lemma 5.2.12.** Let  $\bar{c}_\Phi, \bar{c}, \tau, f, \hat{f}$  be defined the same as in Lemma 5.2.11. If further assume that  $|\Phi^{(2)}(\cdot)| \leq L$  for some constant  $L > 0$  and is Lipschitz continuous with constant  $G$ , then, under the assumptions in Lemma 5.2.9 and (5.29), with probability at least  $1 - 4\exp(-p)$  there exists a constant  $\hat{c}_\Phi \in [0, \bar{c}]$  such that  $\hat{f}(\hat{c}_\Phi, \hat{w}^{ols}) = 1$ . Furthermore, if the derivative of  $c \mapsto f(c, w^{ols})$  is bounded below in absolute value (*i.e.*, does not change the sign) by  $M > 0$  in the interval  $c \in [0, \bar{c}]$ , then with probability at least  $1 - 4\exp(-p)$ , the following holds

$$|\hat{c}_\Phi - \bar{c}_\Phi| \leq O\left(\frac{M^{-1}GL\bar{c}^2\kappa_x^2r^2\|\Sigma\|_2^{\frac{1}{2}}\sqrt{p}\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta} \log\frac{p}{\xi^2}}}{\epsilon\lambda_{\min}^{\frac{1}{2}}(\Sigma) \min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma), 1\}\sqrt{n}} + M^{-1}LG\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{\frac{p}{m}}\right) \quad (5.50)$$

for sufficiently large  $m, n$  such that

$$n \geq \Omega\left(\frac{LG^2\tau^{-2}\bar{c}^4\|\Sigma\|_2\kappa_x^4pr^4\|w^{ols}\|_2^2 \log\frac{1}{\delta} \log\frac{p^2}{\xi}}{\epsilon^2\lambda_{\min}(\Sigma) \min\{\lambda_{\min}(\Sigma), 1\}}\right) \quad (5.51)$$

$$m \geq \Omega\left(G^2L^2\|\Sigma\|_2\|w^{ols}\|_2^2 p\tau^{-2}\right). \quad (5.52)$$

*Proof of Lemma 5.2.12.* The main idea of this proof is almost the same as the one for Lemma 5.2.11. The only difference is that instead of using Lemma 5.2.7 to get (5.36), we

use here Lemma 5.2.8 to obtain the following with probability at least  $1 - \exp(-p)$

$$\begin{aligned}
& \sup_{w' \in \mathbb{B}^{\bar{c}\delta'}(\bar{w}^{ols})} \left| \frac{1}{m} \sum_{j=1}^m \Phi^{(2)}(\langle v_i, w' \rangle) - \mathbb{E} \Phi^{(2)}(\langle v, w' \rangle) \right| \\
& \leq O((G(\|\bar{w}^{ols}\|_2 + \bar{c}\delta')\|I\|_2 + L)\sqrt{\frac{p}{m}} \\
& \leq O((G\|\Sigma\|_2^{\frac{1}{2}}(\|w^{ols}\|_2 + \bar{c}\frac{\delta}{\lambda_{\min}^{\frac{1}{2}}(\Sigma)}) + L)\sqrt{\frac{p}{m}}). \tag{5.53}
\end{aligned}$$

Thus, by (5.35), (5.37) and (5.53), we have

$$\begin{aligned}
\sup_{c \in [0, \bar{c}]} |\hat{f}(c, \hat{w}^{ols}) - f(c, w^{ols})| & \leq O(G\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{\frac{p}{m}} + \\
& \frac{G\kappa_x \bar{c}\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{pr^2\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma), 1\}}\sqrt{\frac{p}{mn}} + L\sqrt{\frac{p}{m}}). \tag{5.54}
\end{aligned}$$

Let D denote the RHS of (5.54), we have

$$\hat{f}(\bar{c}, \hat{w}^{ols}) \geq 1 + \tau - D.$$

It is sufficient to show that  $\tau > D$ , which holds when

$$O(G\bar{c}^2\|\Sigma\|_2^{\frac{1}{2}}\frac{\kappa_x^2\sqrt{pr^2}\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\sqrt{n}\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma), 1\}}) \leq \frac{\tau}{2}$$

and

$$O(\frac{G\kappa_x \bar{c}\|\Sigma\|_2^{\frac{1}{2}}L\|w^{ols}\|_2\sqrt{pr^2\log\frac{1}{\delta}\log\frac{p^2}{\xi}}}{\epsilon\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma), 1\}}\sqrt{\frac{p}{mn}}) \leq \frac{\tau}{2}.$$

That is,

$$n \geq \Omega\left(\frac{G^2\tau^{-2}\bar{c}^4\|\Sigma\|_2\kappa_x^4pr^4\|w^{ols}\|_2^2\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma), 1\}}\right) \tag{5.55}$$

$$m \geq \Omega(G^2L^2\|\Sigma\|_2\|w^{ols}\|_2^2p\tau^{-2}). \tag{5.56}$$

Then, there exists  $\hat{c}_\Phi \in [0, \bar{c}]$  such that  $\hat{f}(\hat{c}_\Phi, \hat{w}^{ols}) = 1$ . We can easily get

$$\begin{aligned} M|\hat{c}_\Phi - \bar{c}_\Phi| &\leq |f(\hat{c}_\Phi, \hat{w}^{ols}) - f(\bar{c}_\Phi, \hat{w}^{ols})| \\ &\leq O\left(\frac{G\bar{c}^2\kappa_x^2r^2\|\Sigma\|_2^{\frac{1}{2}}\sqrt{p}\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p}{\xi^2}}}{\epsilon\lambda_{\min}^{\frac{1}{2}}(\Sigma)\min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma), 1\}\sqrt{n}}\right. \\ &\quad \left. + \frac{G\kappa_x\bar{c}\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{pr^2\log\frac{1}{\delta}\log\frac{p^2}{\xi^2}}}{\epsilon\lambda_{\min}^{1/2}(\Sigma)\min\{\lambda_{\min}^{1/2}(\Sigma), 1\}}\sqrt{\frac{p}{mn}} + LG\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{\frac{p}{m}}\right) \quad (5.57) \\ &\leq O\left(\frac{GL\bar{c}^2\kappa_x^2r^2\|\Sigma\|_2^{\frac{1}{2}}\sqrt{p}\|w^{ols}\|_2\sqrt{\log\frac{1}{\delta}\log\frac{p}{\xi^2}}}{\epsilon\lambda_{\min}^{\frac{1}{2}}(\Sigma)\min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma), 1\}\sqrt{n}} + LG\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{\frac{p}{m}}\right). \quad (5.58) \end{aligned}$$

□

**Proof of Theorem 5.2.2 .** The proof is almost the same as the one for Theorem 5.2.3. By definition, we have

$$\begin{aligned} \|\hat{w}^{glm} - w^*\|_\infty &\leq \|\hat{c}_\Phi\hat{w}^{ols} - \bar{c}_\Phi w^{ols}\|_\infty + \|\bar{c}_\Phi w^{ols} - w^*\|_\infty \\ &\leq \|\hat{c}_\Phi\hat{w}^{ols} - \bar{c}_\Phi w^{ols}\|_\infty + \|\bar{c}_\Phi w^{ols} - c_\Phi w^{ols}\|_\infty + \|c_\Phi w^{ols} - w^*\|_\infty. \end{aligned} \quad (5.59)$$

The second term of (5.59) is bounded by

$$\|\bar{c}_\Phi w^{ols} - c_\Phi w^{ols}\|_\infty \leq O\left(M^{-1}r^2\kappa_x^7c_\Phi G^3\rho_2\rho_\infty^2 \frac{\|w^*\|_\infty^3 \max\{1, \|w^*\|_\infty\}}{\sqrt{p}} \max\{1, \frac{1}{c_\Phi}\}\right). \quad (5.60)$$

By Lemma 5.2.2 in the context, the third term of (5.59) is bounded by  $16c_\Phi Gr\kappa_x^3\sqrt{\rho_2}\rho_\infty \frac{\|w^*\|_\infty}{\sqrt{p}}$ .

The first term is bounded by

$$\begin{aligned} \|\hat{c}_\Phi \hat{w}^{ols} - \bar{c}_\Phi w^{ols}\|_\infty &\leq \\ O\Big( &\frac{M^{-1}G^3L\bar{c}^2\kappa_x^8r^4\rho_2\rho_\infty^2\|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\} \|\Sigma\|_2^{\frac{1}{2}} p \sqrt{\log \frac{1}{\delta} \log \frac{p}{\xi^2}}}{\epsilon \lambda_{\min}^{\frac{1}{2}}(\Sigma) \min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma), 1\} \sqrt{n}} \times \max\left\{\frac{1}{c_\Phi}, 1\right\}^2 \\ &+ \frac{M^{-1}G^3L\bar{c}^2\kappa_x^6r^2\rho_2\rho_\infty^2\|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\} \|\Sigma\|_2^{\frac{1}{2}} p}{\sqrt{m}} \times \max\left\{\frac{1}{c_\Phi}, 1\right\}^2 \Big). \end{aligned} \quad (5.61)$$

Thus, in total we have

$$\begin{aligned} \|\hat{w}^{glm} - w^*\|_\infty &\leq O\Big( \frac{M^{-1}G^3L\bar{c}^2\kappa_x^6r^2\rho_2\rho_\infty^2\|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\} \|\Sigma\|_2^{\frac{1}{2}} p}{\sqrt{m}} \times \max\left\{\frac{1}{c_\Phi}, 1\right\}^2 \\ &+ \frac{G^3L\bar{c}^2\kappa_x^6r^4\rho_2\rho_\infty^2\|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\} \|\Sigma\|_2^{\frac{1}{2}} p \sqrt{\log \frac{1}{\delta} \log \frac{p}{\xi^2}}}{\epsilon \lambda_{\min}^{\frac{1}{2}}(\Sigma) \min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma), 1\} \sqrt{n}} \max\left\{\frac{1}{c_\Phi}, 1\right\}^2 \\ &+ M^{-1}r^2\kappa_x^7c_\Phi G^3\rho_2\rho_\infty^2\|\Sigma^{\frac{1}{2}}\|_\infty \frac{\|w^*\|_\infty^3 \max\{1, \|w^*\|_\infty\}}{\sqrt{p}} \max\left\{1, \frac{1}{c_\Phi}\right\} \Big). \end{aligned} \quad (5.62)$$

The probability of success is at least  $1 - \exp(-\Omega(p)) - \xi$ . The sample complexity should satisfy

$$m \geq \Omega\left(G^2L^2\|\Sigma\|_2\|w^*\|_\infty^2 \max\{1, \|w^*\|_\infty^2\} G^2r^2\kappa_x^6\rho_2\rho_\infty^2p^2\tau^{-2} \max\left\{1, \frac{1}{c_\Phi}\right\}^2\right) \quad (5.63)$$

$$n \geq \Omega\left(\frac{\rho_2\rho_\infty^2G^4\tau^{-2}\bar{c}^4\|\Sigma\|_2^2\kappa_x^{10}p^2\|w^*\|_\infty^2r^6 \max\{1, \|w^*\|_\infty^2\} \log \frac{1}{\delta} \log \frac{p^3}{\xi}}{\epsilon^2\lambda_{\min}(\Sigma) \min\{\lambda_{\min}(\Sigma), 1\}} \max\left\{1, \frac{1}{c_\Phi}\right\}^2\right). \quad (5.64)$$

□

### Proof of Theorem 5.2.4

The idea of the proof follows the one in [108].

By assumption, we have

$$\mathbb{E}[xy] = \mathbb{E}[xf(\langle x, w^* \rangle)] = \Sigma^{\frac{1}{2}} \mathbb{E}[vf(\langle v, \hat{w}^* \rangle)],$$

where  $\hat{w}^* = \Sigma^{\frac{1}{2}} w^*$ . Now, consider each coordinate  $j \in [p]$  for the term  $\mathbb{E}[vf(\langle v, \hat{w}^* \rangle)]$ . Let  $v_j^*$  denote the zero-bias transformation of  $v_j$  conditioned on  $V_j = \langle v, \hat{w}^* \rangle - v_j \hat{w}_j^*$ . Then, we have

$$\begin{aligned}\mathbb{E}[v_j f(\langle v, \hat{w}^* \rangle)] &= \mathbb{E}\mathbb{E}[v_j f(v_j \hat{w}_j^* + V_j) | V_j] \\ &= \hat{w}_j^* \mathbb{E}\mathbb{E}[f'(v_j^* \hat{w}_j^* + V_j) | V_j] \\ &= \hat{w}_j^* \mathbb{E}\mathbb{E}[f'((v_j^* - v_j) \hat{w}_j^* + \langle v, \hat{w}^* \rangle) | V_j] \\ &= \hat{w}_j^* \mathbb{E}[f'((v_j^* - v_j) \hat{w}_j^* + \langle v, \hat{w}^* \rangle)].\end{aligned}$$

Thus, we have  $w^{ols} = \Sigma^{-\frac{1}{2}} D \Sigma^{\frac{1}{2}} w^*$ , where  $D$  is a diagonal matrix whose  $i$ -th entry is  $\mathbb{E}[f'((v_j^* - v_j) \hat{w}_j^* + \langle v, \hat{w}^* \rangle)]$ .

By the Lipschitz condition, we have

$$|\mathbb{E}[f'((v_j^* - v_j) \hat{w}_j^* + \langle v, \hat{w}^* \rangle)] - \mathbb{E}[f'(\langle v, \hat{w}^* \rangle)]| \leq G |\hat{w}_j^*| \mathbb{E}|(v_j^* - v_j)|.$$

By the same argument given in [108], we have

$$\mathbb{E}|(v_j^* - v_j)| \leq 1.5 \mathbb{E}[|v_j|^3].$$

Using the bound of the third moment induced by the sub-Gaussian norm, we have

$$L |\hat{w}_j^*| \mathbb{E}|(v_j^* - v_j)| \leq 8G \kappa_x^3 \max_{j \in [p]} |\hat{w}_j^*| \leq 8G \kappa_x^3 \|\Sigma^{\frac{1}{2}} w^*\|_\infty.$$

Thus, we get

$$\max_{j \in [d]} |D_{jj} - \frac{1}{c_f}| \leq 8G\kappa_x^3 \|\Sigma^{\frac{1}{2}} w^*\|_\infty.$$

This means that

$$\begin{aligned} \|w^{ols} - \frac{1}{c_f} w^*\|_\infty &= \|\Sigma^{-\frac{1}{2}}(D - \frac{1}{c_f}I)\Sigma^{\frac{1}{2}} w^*\|_\infty \\ &\leq \max_{j \in [p]} |D_{jj} - \frac{1}{c_f}| \|\Sigma^{-\frac{1}{2}}\|_\infty \|\Sigma^{\frac{1}{2}}\|_\infty \|w^*\|_\infty \\ &\leq 8L\kappa_x^3 \rho_\infty L \|\Sigma^{\frac{1}{2}}\|_\infty \|w^*\|_\infty^2. \end{aligned}$$

Due to the diagonal dominance property we have

$$\|\Sigma^{\frac{1}{2}}\|_\infty = \max_i \sum_{j=1}^p |\Sigma_{ij}^{\frac{1}{2}}| \leq 2 \max_i \Sigma_{ii}^{\frac{1}{2}} \leq 2 \|\Sigma\|_2^{\frac{1}{2}}.$$

Since we have  $\|x\|_2 \leq r$ , we write

$$r^2 \geq \mathbb{E}[\|x\|_2^2] = \text{Trace}(\Sigma) \geq p\|\Sigma\|_2 \geq \frac{p\|\Sigma\|_2}{\rho_2}.$$

Thus we have  $\|\Sigma^{\frac{1}{2}}\|_\infty \leq 2r \sqrt{\frac{\rho_2}{p}}$ .

### Proof of Theorem 5.2.5

By the same argument in the proof of Lemma 5.2.9, we can show that when  $n \geq \Omega(\frac{\kappa_x^4 \|\Sigma\|_2^2 p r^4 \log \frac{1}{\delta}}{\epsilon^2 \lambda_{\min}^2(\Sigma)})$ , with probability at least  $1 - \exp(-\Omega(p)) - \xi$ , the following holds

$$\|\hat{w}^{ols} - \tilde{w}^{ols}\|_2^2 = O\left(\frac{pC^2 r^2 (L^2 r^2 + C^2 + r^2 \|\tilde{w}^{ols}\|_2^2) \log \frac{1}{\delta} \log \frac{p^2}{\xi}}{\epsilon^2 n \lambda_{\min}^2(\Sigma)}\right). \quad (5.65)$$

Thus, by Lemma 5.2.10 we have

$$\|\hat{w}^{ols} - w^{ols}\|_2 \leq O\left(\frac{CL\kappa_x \sqrt{pr^2} \|w^{ols}\|_2 \sqrt{\log \frac{1}{\delta} \log \frac{p^2}{\xi}}}{\epsilon \sqrt{n} \lambda_{\min}^{1/2}(\Sigma) \min\{\lambda_{\min}^{1/2}(\Sigma), 1\}}\right). \quad (5.66)$$

In the following, we will always assume that (5.66) holds. By the same argument given in Lemma 5.2.12, we have the following Lemma, which can be proved in the same way as Lemma 5.2.12.

**Lemma 5.2.13.** Let  $f'$  be a function that is Lipschitz continuous with constant  $G$  and  $|f'(\cdot)| \leq L$ , and  $g : \mathbb{R} \times \mathbb{R}^p \mapsto \mathbb{R}$  be another function such that  $g(c, w) = c\mathbb{E}[f'(\langle x, w \rangle c)]$  and its empirical one is

$$\hat{g}(c, w) = \frac{c}{m} \sum_{j=1}^m f'(\langle x_j, w \rangle c).$$

Let  $\mathbb{B}^\delta(\bar{w}^{ols}) = \{w : \|w - \bar{w}^{ols}\|_2 \leq \delta\}$ , where  $\bar{w}^{ols} = \Sigma^{\frac{1}{2}}w^{ols}$ . Then, under the assumptions in Lemma 5.2.9 and Eq. (5.66), with probability at least  $1 - 4\exp(-p)$ , there exists a constant  $\hat{c}_\Phi \in [0, \bar{c}]$  such that  $\hat{g}(\hat{c}_\Phi, \hat{w}^{ols}) = 1$ . Furthermore, if the derivative of  $c \mapsto g(c, w^{ols})$  is bounded below in absolute value (*i.e.*, does not change the sign) by  $M > 0$  in the interval of  $c \in [0, \bar{c}]$ , then with probability at least  $1 - 4\exp(-p)$ , the following holds

$$|\hat{c}_\Phi - \bar{c}_\Phi| \leq O\left(\frac{M^{-1}CGL\bar{c}^2r^2\|\Sigma\|_2^{\frac{1}{2}}\sqrt{p}\|w^{ols}\|_2\log\frac{1}{\delta}\log\frac{p}{\xi^2}}{\epsilon\lambda_{\min}^{\frac{1}{2}}(\Sigma)\min\{\lambda_{\min}^{\frac{1}{2}}(\Sigma), 1\}\sqrt{n}} + M^{-1}LG\|\Sigma\|_2^{\frac{1}{2}}\|w^{ols}\|_2\sqrt{\frac{p}{m}}\right) \quad (5.67)$$

for sufficiently large  $m, n$  such that

$$n \geq \Omega\left(\frac{LG^2\tau^{-2}\bar{c}^4\|\Sigma\|_2\kappa_x^4pr^4\|w^{ols}\|_2^2\log\frac{1}{\delta}\log\frac{p^2}{\xi}}{\epsilon^2\lambda_{\min}(\Sigma)\min\{\lambda_{\min}(\Sigma), 1\}}\right) \quad (5.68)$$

$$m \geq \Omega(G^2L^2\|\Sigma\|_2\|w^{ols}\|_2^2p\tau^{-2}). \quad (5.69)$$

where  $r = \max_{i \in [n]} \|x_i\|_2$ .

### 5.3 Sparse Linear Regression in LDP model

In the previous two sections, we studied ERM in the NLDP model. in this section, we will study ERM in the general LDP model. Specifically, we wish to understand the high

dimensional (sparse) ERM in LDP model. To advance our understanding on the local model, we study, in this paper, the LDP version of the most simplest problem in ERM, *i.e.*, the sparse linear regression problem. Linear regression is a fundamental and classical tool for data analysis, and finds numerous applications in social sciences [212], genomics research [55] and signal recovery [48]. One frequently encountered challenge for such a technique is how to deal with the high dimensionality of the dataset, such as those in genomics, educational and psychological research. A commonly adopted strategy for dealing with such an issue is to assume that the unknown regression vector is sparse.

There are two commonly used ways for measuring the performance of this problem, which correspond to two different settings, the statistical learning and the statistical estimation settings. For the first setting, the measurement is based on the optimization error, *i.e.*  $F(\theta^{\text{priv}}) - \min_{\theta \in \mathcal{C}} F(\theta)$ , where  $F(\theta) = \mathbb{E}_{(x,y) \sim \mathcal{P}} (\langle x, \theta \rangle - y)^2$ , and  $\mathcal{P}$  is an unknown distribution. For the second setting,  $y$  is assumed to be  $y = \langle x, \theta^* \rangle + \sigma$ , where  $x \sim \mathcal{D}$ ,  $\mathcal{D}$  is a known distribution,  $\sigma$  is a random noise, and  $\theta^* \in \mathbb{R}^p$  is the to-be-estimated vector that satisfies the condition of  $\|\theta^*\|_0 \leq s$ . The estimation error for this setting is represented by the loss of the squared  $\ell_2$  norm, *i.e.*,  $\|\theta^{\text{priv}} - \theta^*\|_2^2$ . In this paper, we will focus on the latter setting, and assume that  $x \sim \text{Uniform}\{+1, -1\}^p$ .

Our contributions can be summarized as follows:

- We first present a negative result which suggests that the  $\epsilon$  non-interactive private minimax risk of  $\|\theta^{\text{priv}} - \theta^*\|_2^2$  is lower bounded by  $\Omega(\frac{p \log p}{n \epsilon^2})$  if the privacy of the whole dataset  $\{(x_i, y_i)\}_{i=1}^n$  needs to be preserved. This indicates that it is impossible to obtain any non-trivial error bound in high dimensional space (*i.e.*  $p \gg n$ ). The private minimax risk is still lower bounded by  $\Omega(\frac{p}{n \epsilon^2})$ , even in the sequentially interactive local model. Our proofs are based on a locally differentially private version of the Fano and Le Cam method [97, 98, 100]. We further reveal that this polynomial dependency on  $p$  cannot be avoided even if the measurement of the loss function or definitions of differential privacy is relaxed.

- With the understanding of this limitation, we then propose an  $\epsilon$ -sequential interactive LDP algorithm for the low dimensional sparse case, called Locally Differentially Private Iterative Hard Thresholding (LDP-IHT), which achieves a near optimal upper bound. Furthermore, we show that the idea of DP-IHT is actually rather general and can be used to achieve differential privacy for quite a few other problems. Specifically, it can be applied to the (Locally) Differentially Private Empirical Risk Minimization (DP-ERM) problem with sparsity constraints, and achieves an upper bound that depends only logarithmically on  $p$  (i.e.,  $\log p$ ) and the sparsity parameter of the optimal estimator, making it suitable for applications in high dimensions. To our best knowledge, this is the first paper studying DP-ERM with non-convex constraint set. Another application of LDP-IHT is the sparse regression problem with non-linear measurements [357, 350].
- We also give a positive result for high dimensions. Particularly, we consider the restricted case where only the responses (labels) are required to be private, *i.e.*, the dataset  $\{x_i\}_{i=1}^n$  is assumed to be public and  $\{y_i\}_{i=1}^n$  is private (note that this is a valid assumption as shown in [65, 33]). For this case, we propose a general algorithm which achieves an upper bound of  $O(\frac{s \log p}{n \epsilon^2})$  for the estimation error. We show that this bound is actually optimal, as the  $\epsilon$  non-interactive private minimax risk can also be lower bounded by  $\Omega(\frac{s \log p}{n \epsilon^2})$ .
- Finally, we perform our algorithms on both synthetic and real world datasets. Experimental results also support our theoretical analysis.

### 5.3.1 Related Work

There is a vast number of existing results studying the differentially private linear regression problem (or more generally, DP-ERM) from different perspectives, such as [73, 24, 334, 253, 181, 257, 277]. Below, we focus only on those with theoretical guarantees on the error.

For the central model, [334] recently conducted a comprehensive study, from both theoretical and practical points of views, on the differentially private linear regression problem. The author gave upper bounds of the optimization error in the statistical learning setting and the estimation error in the statistical estimation setting, as well as a general lower bound of the optimization error. There are also other works on this problem (we refer the reader to the Related Work section in [334] for more details). But all these results are only for the low dimensional case (*i.e.* the dimensionality  $p$  is a small constant number). Contrarily, we study mainly, in this paper, the high dimensional sparse case under the statistical estimation setting and provide both upper and lower bounds of the estimation error for the non-interactive and sequentially interactive models. A couple of results also exist for the high dimensional sparse linear regression problem in the central model [181, 270]; but all of them consider only the optimization error. [34] studied the problem of Bayesian linear regression, which is incomparable to our problem. [253] focused the confidence interval of Ordinary Linear Regression while we mainly focus on the estimation error. It is notable that recently [59] studied the optimal rates of the estimation error of linear regression in both low dimension and high dimensional sparse settings. Specifically, for  $(\epsilon, \delta)$ -DP, they showed that in the low dimension setting, the near optimal rate of estimation error is  $\tilde{O}(\sqrt{\frac{p}{n}} + \frac{p\sqrt{\log 1/\delta}}{n\epsilon})$ , while in the high dimensional setting it is  $\tilde{O}(\sqrt{\frac{s\log p}{n}} + \frac{s\log p\sqrt{\log 1/\delta}}{n\epsilon})$ , here  $\tilde{O}$ -term omits  $\log n$  factor. We will show more details in Remark 5.3.2 for the comparison between sparse linear regression in the central model and the local model.

Unlike the central model where tremendous progresses have been made, linear regression in the local model is still not well understood. The only known results are [257, 363, 98, 97]. [97] studied the low dimensional, non-interactive private minimax risk of the estimation error for the restricted case of keeping the responses private, while we consider the high dimensional case of the problem in the interactive local model. [257] gave the optimal lower bound of the optimization error,  $\Theta(\sqrt{\frac{p}{n\epsilon^2}})$ , for the low dimensional case which was later improved to  $O((\frac{\log p}{n\epsilon^2})^{\frac{1}{4}})$  by [363, 299] in the case where the constraint set is a unit  $\ell_1$  norm

ball. However, their settings are different from ours since they all assume that the norm of  $x_i$  is bounded by 1, *i.e.*  $\|x_i\|_2 \leq 1$ , while in our statistical setting,  $\|x_i\|_2 = \sqrt{p}$ . Thus, our results are incomparable with theirs.

DP-ERM has been studied in [154, 305, 310, 325, 299, 97, 161] under different settings. However, none of these considered the non-convex constraint case.

To proof the low bounds in this paper, we mainly use private version of the Fano and Le Cam method, which are initially given by [97, 98, 100]. Based on different settings or problems, there are different versions of private Fano and Le Cam method. For example, [316] proposed a generalized private Assouad method to deal with the lower bounds of some matrix estimation problems in the local differential privacy model. [3] proposed private Fano, Le Cam and Assouad method under central differential privacy. [2] proved lower bounds for various testing and estimation problems under local differential privacy using a notion of chi-squared contractions based on Le Cam's method and Fano's inequality.

### 5.3.2 Problem Set-up

The focus of this paper is the sparse linear regression problem. In this problem, we have  $n$  pair of observations  $\{(x_i, y_i)\}_{i=1}^n$ , where each  $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$ . Moreover, there is some unknown parameter vector  $\theta^* \in \mathbb{R}^p$  that links each pair  $(x_i, y_i)$  by the standard linear model

$$y_i = \langle x_i, \theta^* \rangle + \sigma_i,$$

where  $|\sigma_i| \leq C$  is observation noise and  $C > 0$  is some constant. Here  $\theta^*$  satisfies the sparsity constraint, meaning that  $\theta^*$  has no more than  $s \ll p$  non-zero entries. The goal is to estimate the unknown vector  $\theta^*$  based on these  $n$  observations while also under the local differential privacy constraint. Specifically, we want to find an estimator  $\theta^{priv}$  via some locally differentially private algorithm to make its estimation error  $\|\theta^{priv} - \theta^*\|_2^2$  be as small as possible. Specifically, in this paper we will focus on the following collection of samples

$(x, y) \in \{+1, -1\}^p \times \mathbb{R}$ :

$$\begin{aligned}\mathcal{P}_{s,p,C} = \{P_{\theta,\sigma} \mid x \sim \text{Uniform}\{+1, -1\}^p, y = \langle \theta, x \rangle + \sigma, \text{ where } \sigma \text{ is the random noise s.t} \\ \mathbb{E}[\sigma|x] = 0, |\sigma| \leq C \text{ for some constant } C > 0, \|\theta\|_2 \leq 1, \|\theta\|_0 \leq s\}.\end{aligned}\quad (5.70)$$

In the above definition,  $\sigma$  is sampled from a bounded stochastic noise domain such as uniform distribution and could depend on  $x$ .

It is notable that in the non-private setting, [244] showed the following optimal minimax rate  $\mathcal{M}_n(\theta(\mathcal{P}_{s,p,C}), \|\cdot\|_2^2) = \Theta\left(\frac{C^2 s \log \frac{p}{s}}{n}\right)$ .

It is worth noting that there is some difference between our model (5.70) and the sub-Gaussian linear model, which is a classic model in statistics [244]. That is, here  $x$  is assumed to follow a uniform distribution (which is an often adopted assumption in estimating lower bounds in differential privacy [53]) in our model, while it is often sampled from general sub-Gaussian distribution in a sub-Gaussian model. Even though the uniform distribution can be viewed as a sub-Gaussian distribution, the way of using it in our paper is different.

### 5.3.3 Keeping the Whole Dataset Private

#### Lower Bounds of Private Minimax Risk

In this section, we investigate the private minimax risk in the case where the whole dataset  $\{(x_i, y_i)\}_{i=1}^n$  needs to be locally private, and show that even if the parameter vector  $\theta^*$  is 1-sparse, the polynomial dependence on the dimensionality  $p$  in the estimation error cannot be avoided. This implies that achieving  $\epsilon$ -LDP for the high dimensional sparse linear regression problem is unlikely.

To show the limitations of the problem with respect to the private minimax risk, we first give some intuition. Consider a raw data record  $(x_i, y_i)$  which is sampled from some  $P_{\theta,\sigma} \in \mathcal{P}_{1,p,C}$ , where  $\mathcal{P}_{1,p,C}$  has the form as in (5.70). Suppose that we want to use a Gaussian or Laplacian mechanism on  $(x_i, y_i)$  in order to make the algorithm locally differentially

private. Then, due to sensitivity, the  $\ell_1$  or  $\ell_2$  norm of  $(x_i, y_i)$  is a polynomial of  $p$ . The scale of the added random noise will also be a polynomial of  $p$ , which makes the final estimation error large.

The following theorem indicates that for some fixed privacy parameter  $\epsilon \in (0, 1)$ , the optimal rate of the  $\epsilon$  non-interactive private minimax risk is lower bounded by  $\Omega(\min\{1, \frac{p \log p}{n\epsilon^2}\})$ .

**Theorem 5.3.1.** For a given fixed privacy parameter  $\epsilon \in (0, \frac{1}{2}]$ , the  $\epsilon$  non-interactive private minimax risk (measured by the  $\|\cdot\|_2^2$  metric) of the 1-sparse high dimensional sparse linear regression problem  $\mathcal{P}_{1,p,2}$  needs to satisfy the following inequality,

$$\mathcal{M}_n^{\text{Nint}}(\theta(\mathcal{P}_{1,p,2}), \|\cdot\|_2^2, \epsilon) \geq \Omega(\min\{1, \frac{p \log p}{n\epsilon^2}\}). \quad (5.71)$$

With the above theorems, our question now is to determine whether there are other factors in the local model that might allow us to avoid the polynomial dependency on  $p$  in the estimation error.

We first consider the necessity of interaction in the model, since for some problems, such as convex Empirical Risk Minimization (ERM), there exists a large gap in the estimation error between the interactive and non-interactive local models [257]. The following theorem suggests that even if sequential interaction is allowed in the local model, the polynomial dependence on  $p$  is still unavoidable. Note that sequential interaction is a commonly used model in LDP [97, 257].

**Theorem 5.3.2.** For a given fixed privacy parameter  $\epsilon \in (0, \frac{1}{2}]$ , the  $\epsilon$  sequential private minimax risk (measured by the  $\|\cdot\|_2^2$  metric) of the 1-sparse high dimensional sparse linear regression problem  $\mathcal{P}_{1,p,2}$  needs to satisfy the following inequality,

$$\mathcal{M}_n^{\text{int}}(\theta(\mathcal{P}_{1,p,2}), \|\cdot\|_2^2, \epsilon) \geq \Omega(\min\{1, \frac{p}{n\epsilon^2}\}). \quad (5.72)$$

**Remark 5.3.1.** Since the lower bound of the non-private minimax risk is  $O(\frac{\log p}{n})$  [244], we

conjecture that the lower bound in Theorem 5.3.2 is not tight and the tightest bound should be  $O(\frac{p \log p}{n\epsilon^2})$ , which is the same as Theorem 5.3.1. Later, we will propose a near optimal algorithm (compared with (5.72)) in Section 5.3.3 and leave the problem of finding a tighter lower bound as future research.

**Corollary 5.3.1.** Recently, [170] proposed a general framework which could transfer any  $k$ -compositional *fully* interactive LDP algorithm to sequentially interactive LDP algorithm with an  $O(k)$  blowup in the same complexity. Combining with Theorem 5.3.2, we can claim that even in the  $O(p)$ -compositional fully interactive LDP model, the dependence on the polynomial of the dimensionality  $p$  still cannot be avoided.

**Remark 5.3.2.** Recently [59] studied the lower bound of linear regression with statistical error in both low and high dimensional case under central  $(\epsilon, \delta)$ -DP model. Specifically, they show that for  $s$ -sparse high dimensional case, the private minimax risk under the  $\ell_2$  norm measurement is lower bound by  $\Omega(\sqrt{\frac{s \log p}{n}} + \frac{s \log p \sqrt{\log 1/\delta}}{n\epsilon})$  while for the low dimensional case it is lower bounded by  $\Omega(\sqrt{\frac{p}{n}} + \frac{p \sqrt{\log 1/\delta}}{n\epsilon})$ , all of these bounds are optimal up to factors of  $\text{Poly}(\log n)$ . From Theorem 5.3.1 and 5.3.2, we can see that for sparse linear regression problem, LDP and DP are quite different.

Then, we investigate whether the loss function in the estimation error is too strong. For example, if let  $\theta^* = e_j$  and the private estimator  $\theta^{\text{priv}} = e_i$  for some  $i \neq j$ , then by the squared  $\ell_2$  norm loss, we have  $\|\theta^{\text{priv}} - \theta^*\|_2^2 = 2$ . Since it is possible to get  $|\langle 1, \theta^{\text{priv}} - \theta^* \rangle| = 0$ , this seems to suggest that relaxing the loss function could possibly lower the dependency on  $p$ . However, our next theorem gives a negative answer.

**Theorem 5.3.3.** Consider the loss function  $L : \Theta \times \Theta \mapsto \mathbb{R}_+$ , where  $L(\theta, \theta') = |1^T(\theta - \theta')|$ . Then, for any fixed  $\epsilon \in (0, \frac{1}{2}]$ , the  $\epsilon$  sequential private minimax risk of the loss function  $L$  in the 1-sparse high dimensional sparse linear regression problem  $\mathcal{P}_{1,p,2}$  needs to satisfy the

following inequality,

$$\mathcal{M}_n^{\text{int}}(\theta(\mathcal{P}_{1,p,2}), L, \epsilon) \geq \Omega(\min\{1, \sqrt{\frac{p}{n\epsilon^2}}\}). \quad (5.73)$$

Finally, we consider the possibility of lowering the dependence of  $p$  by relaxing the definition of  $\epsilon$  local differential privacy. This is motivated by the following fact in the central model, where there is a big difference between  $\epsilon$  and  $(\epsilon, \delta)$ -differential privacy for a number of problems, such as the Empirical Risk Minimization [28] and the 1-way marginal [53]. However, as shown in a recent study [50], any non-interactive  $(\epsilon, \delta)$ -LDP protocol can be transformed to an  $\epsilon$ -LDP protocol. This implies that relaxing to  $(\epsilon, \delta)$  LDP cannot avoid the polynomial dependence.

To further investigate the problem, we consider other types of relaxation for LDP, such as Local Rényi Differential Privacy (LRDP) [221] and Local Zero-Concentrated Differential Privacy (LzCDP) [52]. The following theorem shows that the lower bounds on the minimax risk of the  $(2, \log(1 + \epsilon^2))$  sequential LRDP and  $(\kappa, \rho)$  sequential LzCDP still have polynomial dependence on  $p$ .

We first recall the definitions of Rényi Differential Privacy and Zero-Concentrated Differential Privacy and then extend them to the sequentially interactive model. For any  $\alpha \geq 1$ , we denote the Rényi divergence of distribution  $P$  and  $Q$  as

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log \int \left(\frac{dP}{dQ}\right)^\alpha dQ.$$

For  $\alpha = 1$ , it is just the KL-divergence.

**Definition 5.3.1.** Similar to the Definition of local differential privacy, a random variable  $Z_i$  is a  $(\kappa, \rho)$  locally zero-concentrated differentially private view of  $X_i$  if for all  $\alpha > 1$ ,  $z_1, z_2, \dots, z_{i-1}$  and  $x, x' \in \mathcal{X}$ ,

$$D_\alpha(Q_i(Z_i \in S | x_i, z_{1:i-1}) \| Q_i(Z_i \in S | x'_i, z_{1:i-1})) \leq \kappa + \rho\alpha$$

holds for all events  $S$ . Similar to the locally differentially private case, we have  $(\kappa, \rho)$  local zero-concentrated differential privacy (LzCDP) and  $(\kappa, \rho)$  sequential zero-concentrated differential private minimax risk (sequential zCDP minimax risk).

**Definition 5.3.2.** Similarly, we have  $(\alpha, \epsilon)$  local Rényi differential privacy and  $(\alpha, \epsilon)$  (sequential) Renyi differential private minimax risk (called sequential RDP minimax risk) if

$$D_\alpha(Q_i(Z_i \in S \mid x_i, z_{1:i-1}) \| Q_i(Z_i \in S \mid x'_i, z_{1:i-1})) \leq \epsilon.$$

**Theorem 5.3.4.** For given fixed privacy parameters  $0 < \epsilon \leq 1, \kappa, \rho > 0$ , the  $(\kappa, \rho)$  sequential zCDP minimax risk (under the  $\|\cdot\|_2^2$  metric) of the 1-sparse high dimensional sparse linear regression problem  $\mathcal{P}_{1,p,2}$  needs to satisfy the following inequality,

$$\mathcal{M}_n^{\text{int}}(\theta(\mathcal{P}_{1,p,2}), \|\cdot\|_2^2, (\kappa, \rho)) \geq \Omega(\min\{1, \frac{p}{n(e^{\kappa+2\rho} - 1)}\}).$$

The  $(2, \log(1 + \epsilon^2))$  sequential RDP minimax risk (under the  $\|\cdot\|_2^2$  metric) of the 1-sparse high dimensional sparse linear regression problem  $\mathcal{P}_{1,p,2}$  needs to satisfy :

$$\mathcal{M}_n^{\text{int}}(\theta(\mathcal{P}_{1,p,2}), \|\cdot\|_2^2, (2, \log(1 + \epsilon^2))) \geq \Omega(\min\{1, \frac{p}{n\epsilon^2}\}).$$

### Near Optimal Upper Bound for Sequential Interactive Local Model

With the understanding of the limitation in high dimensions, we focus, in this section, on the low dimensional sparse case (i.e.,  $n \geq \Omega(\frac{p}{\epsilon^2})$ ) and propose an  $\epsilon$  sequential interactive LDP algorithm that achieves a near optimal upper bound on the estimation error (compared with (5.72)). Instead of considering the 1-sparse case as in Theorem 5.3.2, we study here the general case, that is,  $\{(x_i, y_i)\}_{i=1}^n \sim P_{\theta^*, \sigma}$ , where  $P_{\theta^*, \sigma} \in \mathcal{P}_{s^*, p, C}$ , and assume that some upper bound of  $s^*$  is already known.

Our method is called Locally Differentially Private Iterative Hard Thresholding (LDP-IHT), which is a locally differentially private version of the traditional Iterative Hard

Thresholding method [43]. We consider the following more general optimization problem, with the intention to extend it to other problems (see Section 5.3.5),

$$\begin{aligned} \min L(\theta; D) &= \frac{1}{2n} \sum_{i=1}^n (\langle x_i, \theta \rangle - y_i)^2 \\ \text{s.t. } &\|\theta\|_2 \leq 1, \|\theta\|_0 \leq s. \end{aligned} \quad (5.74)$$

The key ideas for solving (5.74) in our Algorithm 5.3.41 are the follows. First, we partition the users into  $T$  groups  $\{S_t\}_{t=1}^T$  (where the value of  $T$  will be specified later). Then, in the  $i$ -th iteration, each user receives the current estimator  $\theta_{i-1}$ , and all users in group  $S_i$  conduct the  $\epsilon$ -LDP randomizer procedure [98] on their current gradients  $x_i^T(\langle x_i, \theta_{i-1} \rangle - y_i)$  (see below for the definition of the Randomizer). After receiving the noisy version of the gradient from each user, the server runs the iterative hard thresholding algorithm and produces a new estimator. That is, it executes first a gradient descent step, and then a truncation step  $\theta'_{t+1} = \text{Trunc}(\tilde{\theta}_{t+1}, s)$ , where the truncation function simply keeps the largest  $s$  entries of  $\tilde{\theta}_{t+1}$  (in terms of the magnitude) and converts the rest of the entries to zero. This can be done by first sorting  $\{|\tilde{\theta}_{t+1,j}|\}_{j=1}^p$ , where  $\tilde{\theta}_{t+1,j}$  is the  $j$ -th coordinate of the vector, then keeping the  $s$ -largest ones, and making the entries of all other coordinates 0. Finally, the algorithm projects  $\theta'_{t+1}$  onto the unit  $\ell_2$  norm ball  $\mathbb{B}_1$ .

**Randomizer  $\mathcal{R}_\epsilon^r(\cdot)$  [98]** On input  $x \in \mathbb{R}^p$ , where  $\|x\|_2 \leq r$ , the randomizer  $\mathcal{R}_\epsilon(x)$  does the following. It first sets  $\tilde{x} = \frac{bx}{\|x\|_2}$  where  $b \in \{-1, +1\}$  a Bernoulli random variable  $\text{Ber}(\frac{1}{2} + \frac{\|x\|_2}{2r})$ . We then sample  $T \sim \text{Ber}(\frac{e^\epsilon}{e^\epsilon + 1})$  and outputs  $O(r\sqrt{p})\mathcal{R}_\epsilon(x)$ , where

$$\mathcal{R}_\epsilon(x) = \begin{cases} \text{Uni}(u \in \mathbb{S}^{p-1} : \langle u, \tilde{x} \rangle > 0) & \text{if } T = 1 \\ \text{Uni}(u \in \mathbb{S}^{p-1} : \langle u, \tilde{x} \rangle \leq 0) & \text{if } T = 0 \end{cases} \quad (5.75)$$

Using the same proof as in [257] we can show that each coordinate of the the randomizer  $\mathcal{R}_\epsilon^r(x)$  is sub-Gaussian.

**Lemma 5.3.1** ([257]). Given any vector  $x \in \mathbb{R}^p$ , where  $\|x\|_2 \leq r$ , each coordinate of the randomizer  $\mathcal{R}_\epsilon^r(x)$  defined above is a sub-Gaussian random vector with variance  $\sigma^2 = O(\frac{r^2}{\epsilon^2})$  and  $\mathbb{E}[\mathcal{R}_\epsilon(x)] = x$ .

---

**Algorithm 5.3.41 LDP-IHT**

---

**Input:** Private data records  $\{(x_i, y_i)\}_{i=1}^n \sim P_{\theta^*, \sigma}$ , where  $P_{\theta^*, \sigma} \in \mathcal{P}_{s^*, p, C}$ , iteration number  $T$ , privacy parameter  $\epsilon$ , step size  $\eta$ . Set  $\theta_0 = 0$ .  $s = 8s^*$ .

- 1: For  $t = 1, \dots, T$ , define the index set  $S_t = \{(t-1)\lfloor \frac{n}{T} \rfloor, \dots, t\lfloor \frac{n}{T} \rfloor - 1\}$ ; if  $t = T$ , then  $S_t = S_t \cup \{t\lfloor \frac{n}{T} \rfloor, \dots, n\}$ .
  - 2: **for**  $t = 1, 2, \dots, T$  **do**
  - 3:     The server sends  $\theta_{t-1}$  to all the users. Every user  $i$ ,  $i \in S_t$ , conducts the following operation: let  $\nabla_i = x_i^T (\langle \theta_{t-1}, x_i \rangle - y_i)$ , compute  $z_i = \mathcal{R}_\epsilon^r(\nabla_i)$ , where  $\mathcal{R}_\epsilon^r$  is the randomizer defined above with  $r = O(C\sqrt{p})$  and send back to the server.
  - 4:     The server compute  $\tilde{\nabla}_{t-1} = \frac{1}{|S_t|} \sum_{i \in S_t} z_i$ .
  - 5:     Perform the gradient descent updating  $\tilde{\theta}_t = \theta_{t-1} - \eta \tilde{\nabla}_{t-1}$ .
  - 6:      $\theta'_t = \text{Trunc}(\tilde{\theta}_t, s)$ .
  - 7:      $\theta_t = \arg_{\theta \in \mathbb{B}_1} \|\theta - \theta'_t\|_2^2$ .
  - 8: **end for**
  - 9: Return  $\theta_T$
- 

Before giving the theoretical analysis of Algorithm 5.3.41, we first show the assumption of the partitioned datasets  $\{X_{S_t}\}_{t=1}^T$ .

**Assumption 5.3.1.**  $\{X_{S_t}\}_{t=1}^T$  satisfies the Restricted Isometry Property (RIP) with parameter  $2s + s^*$ , where  $s = 8s^*$ . That is, for any  $v \in \mathbb{R}^p$  with  $\|v\|_0 \leq 2s + s^*$ , there exists a constant  $\Delta$  which satisfies  $(1 - \Delta)\|v\|^2 \leq \frac{1}{|S_t|}\|X_{S_t}v\|_2^2 \leq (1 + \Delta)\|v\|_2^2$  for any  $t \in [T]$ .

Note that for an  $m \times p$  matrix  $X = (x_1^T, \dots, x_m^T)^T \sim \text{Uniform}\{+1, -1\}^{m \times p}$ , it satisfies the RIP condition (with parameter  $s^*$ ) with probability at least  $1 - \epsilon$  if  $m \geq c\Delta^{-2}(s^* \log p + \ln(1/\epsilon))$  for some universal constant  $c$  (see Theorem 2.12 in [245]). Thus, with probability at least  $1 - \xi$ ,  $\{X_{S_t}\}_{t=1}^T$  satisfies Assumption 5.3.1 if  $n \geq \Omega(\Delta^{-2}(Ts^* \log p \log \frac{T}{\xi}))$ . Later, we will see that  $T = O(\log n)$ . Thus, in order to ensure that Assumption 5.3.1 and  $n \geq \Omega(\frac{p}{\epsilon^2})$  hold, we need to assume that  $\frac{n}{\log n} \geq \Omega(\frac{ps^* \log p}{\epsilon^2})$ .

**Theorem 5.3.5.** For any  $\epsilon > 0$ , Algorithm 5.3.41 is  $\epsilon$  sequentially interactive LDP. Moreover, under Assumption 5.3.1 with  $\Delta = O(1)$  and  $\frac{n}{\log n} \geq \Omega(\frac{ps^* \log p}{\epsilon^2})$ , if  $\{(x_i, y_i)\}_{i=1}^n \sim P_{\theta^*, \sigma}$ ,

where  $P_{\theta^*, \sigma} \in \mathcal{P}_{s^*, p, C}$ , then by taking  $s = 8s^*$  and  $\eta = O(1)$ , the output  $\theta_T$  of the algorithm satisfies

$$\|\theta_T - \theta^*\|_2 \leq \left(\frac{1}{2}\right)^T \|\theta^*\|_2 + O\left(\frac{C\sqrt{p \log p} \sqrt{T} \sqrt{s^*}}{\sqrt{n}\epsilon}\right), \quad (5.76)$$

with probability at least  $1 - \frac{2T}{p^c}$  for some constant  $c > 0$ .

Note that Theorem 5.3.5 shows that if  $s^* = 1$ ,  $T = O(\log \frac{n\epsilon^2}{p \log p})$ , then  $\|\theta_T - \theta^*\|_2^2 = O(\frac{p \log p \log n}{n\epsilon^2})$ . Compared with the lower bound in Theorem 5.3.2, it is an optimal upper bound up to a factor of  $\sqrt{\log p}$ .

We notice that recently [122] also used IHT to distributed DP-sparse PCA. However, compared with theirs, our method is  $\epsilon$ -sequentially LDP while theirs is  $(\epsilon, \delta)$ -fully interactive LDP. Thus, the algorithms are quite different.

### 5.3.4 Keeping the Responses Private

In this section, we consider a restricted case where only the responses or labels (*i.e.*,  $\{y_i\}_{i=1}^n$ ) are required to be locally differentially private and all the observations  $\{x_i\}_{i=1}^n$  are assumed to be public. Preserving the privacy of the labels has been studied in [65, 33] for private PAC learning. We also note that keeping the responses private is related to some issues of physical sensory data and the sparse recovery problem, which has been studied in [216]. In this case, we can actually assume that  $\{x_i\}_{i=1}^n \sim \text{Uniform}(\{+1, -1\}^p)^n$  are public, and the collection of probability  $\mathcal{P}_{s, p, C}$  in (5.70) is now reduced to the following model:

$$\mathcal{P}'_{s, p, C} = \{P_{\theta, \sigma}(y_1, \dots, y_n) \mid y_i = \langle \theta^*, x_i \rangle + \sigma_i, \text{ where } \|\theta\|_0 \leq s, \|\theta\|_2 \leq 1 \text{ and the random noise } |\sigma_i| \leq C\} \quad (5.77)$$

The following theorem shows that, for every set of data  $\{(x_i, y_i)\}_{i=1}^n$ , if only  $\{y_i\}_{i=1}^n$  needs to be private, then there is an  $(\epsilon, \delta)$  non-interactively locally differentially private algorithm DP-IHT, which yields a non-trivial upper bound on the squared  $\ell_2$  norm of the estimation error (see Algorithm 5.3.42). More specifically, the algorithm first perturbs each

$y_i$  by Gaussian noise to ensure that it is  $(\epsilon, \delta)$ -LDP. Then, it performs the classical IHT procedure on the server side. Note that we can combine our algorithm with the protocol in [50] to obtain an  $\epsilon$  non-interactive LDP algorithm.

---

**Algorithm 5.3.42** Label-LDP-IHT

---

**Input:** Public dataset  $\{x_i\}_{i=1}^n$ , private  $\{y_i\}_{i=1}^n \in P_{\theta^*, \sigma}$ , where  $P_{\theta^*, \sigma} \in \mathcal{P}'_{s^*, p, C}$ ,  $\epsilon, \delta$  are privacy parameters,  $T$  is the number of iteration,  $\eta$  is the step size, and  $s = 8s^*$ . Set  $\theta_0 = 0$ .

```

1: for Each  $i \in [n]$  do
2:   Denote  $\tilde{y}_i = y_i + z_i$ , where  $z_i \sim \mathcal{N}(0, \tau^2)$ ,  $\tau^2 = \frac{32C^2 \ln(1.25/\delta)}{\epsilon^2}$ .
3: end for
4: for  $t = 0, 1, \dots, T - 1$  do
5:    $\tilde{\theta}_{t+1} = \theta_t - \eta(\frac{1}{n} \sum_{i=1}^n (\tilde{y}_i - \langle x_i, \theta_t \rangle) x_i^T)$ .
6:    $\theta'_{t+1} = \text{Trunc}(\tilde{\theta}_{t+1}, s)$ .
7:    $\theta_{t+1} = \arg_{\theta \in \mathbb{B}_1} \|\theta - \theta'_{t+1}\|_2^2$ .
8: end for
9: Return  $\theta_T$ .
```

---

**Assumption 5.3.2.**  $X = (x_1^T, \dots, x_n^T)^T \in \{-1, +1\}^{n \times p}$  satisfies the Restricted Isometry Property (RIP) with parameter  $2s + s^*$ , where  $s = 8s^*$ . That is, for any  $v \in \mathbb{R}^p$  with  $\|v\|_0 \leq 2s + s^*$ , there exists a constant  $\Delta$  which satisfies  $(1 - \Delta)\|v\|^2 \leq \frac{1}{n}\|Xv\|_2^2 \leq (1 + \Delta)\|v\|_2^2$ .

**Theorem 5.3.6.** For any  $0 < \epsilon \leq 1$  and  $0 < \delta < 1$ , Algorithm 5.3.42 is  $(\epsilon, \delta)$  (non-interactively) locally differentially private for  $\{y_i\}_{i=1}^n$ . Moreover, if  $X$  satisfies Assumption 5.3.2 with  $0 < \Delta \leq \frac{2}{7}$ , then by setting  $s = 8s^*$  in Algorithm 5.3.42, there is an  $\eta = \eta(\Delta)$  which ensures that the output  $\theta_T$  satisfies the following inequality with probability at least  $1 - \exp(-n) - \frac{2}{p^c}$

$$\|\theta_T - \theta^*\|_2 \leq \left(\frac{1}{2}\right)^T \|\theta^*\|_2 + O\left(\frac{C \log(1/\delta) \sqrt{s^* \log p}}{\sqrt{n} \epsilon}\right). \quad (5.78)$$

Note that if  $T = O(\log \frac{\sqrt{n}\epsilon}{C\sqrt{s^* \log p}})$  in (5.78), we have  $\|\theta_T - \theta^*\|_2^2 \leq O(C^2 \frac{s \log p}{n \epsilon^2})$ . Compared with the bounds in Theorem 5.3.1 and 5.3.2, the dependency on  $p$  is reduced from polynomial to logarithmic, which makes it suitable for handling high dimensional data. We note that the term  $O(\frac{s \log p}{n})$  also appears in the optimal minimax rate of the high dimensional sparse sub-Gaussian linear model [244].

Also note that after obtaining  $\{(x_i, \tilde{y}_i)\}_{i=1}^n$ , we can get another private estimator, which has the same upper bound of  $O(\frac{s \log p}{n\epsilon^2})$ , by performing Lasso  $\theta^{\text{priv}} \in \arg_{\theta \in \mathbb{R}^p} \{ \frac{1}{2n} \sum_{i=1}^n (\tilde{y}_i - \langle \theta, x_i \rangle)^2 + \lambda \|\theta\|_1 \}$ , for some  $\lambda = O(\sqrt{\frac{\log p}{n\epsilon^2}})$  [227]. However, we would like to point out that our algorithm is more practical and can be extended to the case of non-linear measurements.

With the above theorem, a natural question is to determine whether the upper bound in Theorem 5.3.6 can be further improved. The following theorem (adopted from [244]) suggests that it is actually tight as the  $\epsilon$  non-interactive local private minimax risk (under the  $\|\cdot\|^2$  metric) is lower bounded by  $\Omega(\frac{C^2 s^* \log p}{n\epsilon^2})$ .

**Theorem 5.3.7.** Under Assumption 5.3.2 and for a given fixed privacy parameter  $\epsilon \in (0, \frac{1}{2}]$ , the  $\epsilon$  non-interactive local private minimax risk (under the  $\|\cdot\|^2$  metric) satisfies the following inequality if only  $\{y_i\}_{i=1}^n$  needs to be kept locally private

$$\mathcal{M}_n^{\text{Nint}}(\theta(\mathcal{P}'_{s,p,C}), \|\cdot\|_2^2, \epsilon) \geq \Omega\left(\min\left\{1, \frac{C^2 s \log \frac{p}{s}}{n\epsilon^2(1+\Delta)}\right\}\right).$$

### 5.3.5 Extension to Other Problems

As mentioned earlier, the (Local) DP-IHT method is actually quite general for achieving differential privacy. In this section, we extend it to other problems. Specifically, we use it to the DP-ERM problem <sup>12</sup> under some sparsity constraint and the sparse regression problem with non-linear monotone measurements.

#### ERM with sparsity constraint

In this section, we consider the sparsity-constrained  $(\epsilon, \delta)$  DP-ERM problem. That is, the constraint set  $\mathcal{C}$  in ERM problem is defined as  $\mathcal{C} = \{x : \|x\|_0 \leq k\}$ , where  $\|x\|_0$  denotes the number of non-zero entries in vector  $x$ . We note that such a formulation encapsulates several important problems such as the  $\ell_0$ -constrained linear/logistic regression [17].

---

<sup>12</sup>It is easy to extend to LDP model

We first introduce some assumptions to the loss function, which are commonly used in the research of ERM under the sparsity-constrained optimization.

**Definition 5.3.3** (Restricted Strong Convexity, RSC). A differentiable function  $f(x)$  is restricted  $\rho_s$ -strongly convex with parameter  $s$  if there exists a constant  $\rho_s > 0$  such that for any  $x, x'$  with  $\|x - x'\|_0 \leq s$ , we have

$$f(x) - f(x') - \langle \nabla f(x'), x - x' \rangle \geq \frac{\rho_s}{2} \|x - x'\|_2^2.$$

**Definition 5.3.4** (Restricted Strong Smoothness, RSS). A differentiable function  $f(x)$  is restricted  $\ell_s$ -strong smooth with parameter  $s$  if there exists a constant  $\ell_s > 0$  such that for any  $x, x'$  with  $\|x - x'\|_0 \leq s$ , we have

$$f(x) - f(x') - \langle \nabla f(x'), x - x' \rangle \leq \frac{\ell_s}{2} \|x - x'\|_2^2.$$

**Assumption 5.3.3.** Denote  $x^* = \arg \min_{x \in \mathcal{C}} L(x; D)$  and  $\|x^*\|_0 = k^*$ . We assume that the objective function  $L(x; D)$  is  $\rho_s$ -RSC and  $\ell(x, z)$  is  $\ell_s$ -RSS for all  $z \in \mathcal{X}$  with parameter  $s = 2k + k^*$ . We also assume that  $\ell(x, z)$  is  $G$ -Lipshitz w.r.t  $\ell_2$  norm for all  $z \in \mathcal{X}$ .

For the sparsity-constrained DP-ERM problem, we follow the idea in Algorithm 5.3.41 to solve the optimization problem (5.74). That is, we first execute a DP-Gradient Descent step and then perform a hard thresholding operation (see Algorithm 5.3.43 for details).

**Theorem 5.3.8.** Under Assumption 5.3.3, for any  $1 \geq \epsilon, \delta > 0$ , there exists a constant  $c > 0$  which makes Algorithm 5.3.43  $(\epsilon, \delta)$ -DP. Moreover, if the sparsity level  $k \geq (1 + 64\kappa_s^2)k^*$ , where  $\kappa_s = \frac{\ell_s}{\rho_s}$ , then by setting  $\eta = \frac{1}{2\ell_s}$  and  $T = O(\kappa_s \log \frac{n^2 \epsilon^2}{k^*})$ , we have

$$\mathbb{E}L(x_T; D) - L(x^*; D) \leq O\left(\frac{\log n \log p k^* \log \frac{1}{\delta}}{n^2 \epsilon^2}\right), \quad (5.79)$$

where the big  $O$ -notation omits the terms of  $G, \rho_s$  and  $\ell_s$ .

---

**Algorithm 5.3.43** DP-IHT

---

**Input:** Initial point  $x_0$ , learning rate  $\eta$ , empirical risk  $L(x; D)$ , privacy parameters  $1 > \epsilon, \delta > 0$ , and iteration number  $T$ .

- 1: **for**  $t = 0, 1, \dots, T - 1$  **do**
  - 2:     Let  $\tilde{x}_{t+1} = x_t - \eta(\nabla L(x_t; D) + z_t)$ , where  $z_t \sim \mathcal{N}(0, \sigma^2 I_p)$ ,  $\sigma^2 = \frac{cT \log \frac{1}{\delta} G^2}{n^2 \epsilon^2}$  for some constant  $c$ .
  - 3:     Let  $x_{t+1} = \text{Trun}(\tilde{x}_{t+1}, k)$ .
  - 4: **end for**
  - 5: Return  $x_T$ .
- 

**Remark 5.3.3.** We note that the upper bound in (5.79) depends only logarithmically on  $p$  (i.e.,  $\log p$ ), rather than polynomially (i.e.,  $\text{Poly}(p)$ ) as in general DP-ERM with (strongly) convex loss functions [325, 29]. This means that we have obtained a non-trivial upper bound for the high dimensional case ( $p \gg n$ ) of the problem. Recently, [270, 269] also studied the case of high dimensional DP-ERM with specified constraint set. However, there are considerable differences. Firstly, the [270] paper considers only linear regression and  $\ell_1$ -norm Lipschitz with the constraint set restricted to an  $\ell_1$ -norm ball. Secondly, the [269] paper shows that its upper bound depends only on the Gaussian width of the underlying constraint set, instead of  $p$ .<sup>13</sup> However, their algorithm is based on the mirror descent method, which needs the constraint set to be convex. But it is non-convex in our problem. Thus, these previous results are not comparable with ours.

It would be interesting to find a general condition on the constraint set such that the upper bound of the problem can be independent of  $\text{Poly}(p)$ . Also, we note that to achieve the bound in (5.79), the gradient complexity of Algorithm 5.3.43 needs to be  $\tilde{O}(n\kappa_s)$ , which is quite large. We leave it as an open problem to make it more practical.

## Non-linear Regression

We now study a model with non-linear non-convex measurement:  $y_i = f(\langle \theta^*, x_i \rangle) + \sigma$ , where  $f$  is some known function and  $\theta^*$  is sparse. This model has recently been studied in

---

<sup>13</sup>For a constraint set  $\mathcal{C} \subset \mathbb{R}^P$ , its Gaussian width can depend on  $p$  in general.

[357, 350]. Note that when  $f$  is the identity function, it reduces to the sparse linear regression model. In this paper, we focus on a special class of functions called  $(a, b)$  monotone:

**Definition 5.3.5.** A function  $f : \mathbb{R} \mapsto \mathbb{R}$  is  $(a, b)$  monotone for some  $0 < a \leq b$  if  $f$  is differentiable and  $f'(x) \in [a, b]$  for all  $x \in \mathbb{R}$ .

Like in the linear model, we also consider the cases of keeping the whole dataset and only the responses  $\{y_i\}_{i=1}^n$  locally differentially private.

### Keeping the Whole Dataset Private

Same as in the linear model case, we consider the following distribution collection of samples  $(x, y) \in \{+1, -1\}^p \times \mathbb{R}$ :

$$\begin{aligned} \mathcal{P}_{s,p,C,f,a,b} = \{P_{\theta,\sigma} \mid x \sim \text{Uniform}\{+1, -1\}^p, y = f(\langle \theta, x \rangle) + \sigma, \text{ where } \sigma \text{ is the random noise} \\ |\sigma| \leq C, C > 0 \text{ is some constant } \|\theta\|_2 \leq 1, \|\theta\|_0 \leq s, f \text{ is } (a, b) \text{ monotone} \}. \end{aligned} \quad (5.80)$$

We note that when  $f(x) = x$ , it reduces to (1).

To obtain an upper bound of the empirical risk, we can easily extend Algorithm 5.3.42 to the non-linear measurement case (see Algorithm 5.3.44) to solve the following problem

$$\begin{aligned} \min L(\theta; D) &= \frac{1}{n} \sum_{i=1}^n (f(\langle x_i, \theta \rangle) - y_i)^2 \\ \text{s.t. } &\|\theta\|_2 \leq 1, \|\theta\|_0 \leq s. \end{aligned} \quad (5.81)$$

**Theorem 5.3.9.** For any  $\epsilon > 0$ , Algorithm 5.3.44 is  $\epsilon$  sequential interactive LDP. Moreover, if  $\{X_{S_t}\}$  satisfies Assumption 1 with  $0 \leq \delta' \leq \frac{9a^2 - 5b^2}{14}$  in Section 4.2 and  $\frac{n}{\log n} \geq \Omega\left(\frac{ps^* \log p}{\epsilon^2}\right)$ , and  $\{(x_i, y_i)\}_{i=1}^n \sim P_{\theta^*, \sigma}$ , where  $P_{\theta^*, \sigma} \in \mathcal{P}_{s^*, p, C, f, a, b}$  (we assume  $\frac{a^2}{b^2} \geq \frac{5}{9}$ ), then after taking

---

**Algorithm 5.3.44** LDP-IHT

---

**Input:** Private data records  $\{(x_i, y_i)\}_{i=1}^n \sim P_{\theta^*, \sigma}$ , where  $P_{\theta^*, \sigma} \in \mathcal{P}_{s, p, C, f, a, b}$ ,  $T$  is the Iteration number,  $\epsilon$  is the privacy parameter, and  $\eta$  is the step size. Set  $\theta_0 = 0$ .  $s$  is a parameter to be specified later.

- 1: For  $t = 1, \dots, T$ , define the index set  $S_t = \{(t-1)\lfloor \frac{n}{T} \rfloor, \dots, t\lfloor \frac{n}{T} \rfloor - 1\}$ , if  $t = T$ , then  $S_t = S_t \cup \{t\lfloor \frac{n}{T} \rfloor, \dots, n\}$ .
  - 2: **for**  $t = 1, 2, \dots, T$  **do**
  - 3:     The server sends  $\theta_{t-1}$  to all the users. Every user  $i$  which  $i \in S_t$  does the following operation: let  $\nabla_i = x_i^T f'(\langle \theta_{t-1}, x_i \rangle)(f(\langle \theta_{t-1}, x_i \rangle) - y_i)$ , compute  $z_i = \mathcal{R}_\epsilon^r(\nabla_i)$ , where  $\mathcal{R}_\epsilon^r$  is the randomizer defined in the previous section with  $r = O(bC\sqrt{p})$  and send back to the server.
  - 4:     The server compute  $\tilde{\nabla}_{t-1} = \frac{1}{|S_t|} \sum_{i \in S_t} z_i$ .
  - 5:     Do the gradient descent updating  $\tilde{\theta}_t = \theta_{t-1} - \eta \tilde{\nabla}_{t-1}$ .
  - 6:      $\theta'_t = \text{Trunc}(\tilde{\theta}_t, s)$ .
  - 7:      $\theta_t = \arg_{\theta \in \mathbb{B}_1} \|\theta - \theta'_t\|_2^2$ .
  - 8: **end for**
  - 9: Return  $\theta_T$
- 

$s = 8s^*$  and  $\eta = \eta(a, b)$ , the output  $\theta_T$  satisfies

$$\|\theta_T - \theta^*\|_2 \leq (\frac{1}{2})^T \|\theta^*\|_2 + O\left(\frac{\sqrt{p \log p} \sqrt{T} \sqrt{s}}{\sqrt{n} \epsilon}\right), \quad (5.82)$$

with probability at least  $1 - \frac{2T}{p^c}$  for some constant  $c > 0$ .

### Keeping the Labels Private

For a fixed  $X = (x_1^T, \dots, x_n^T)^T \in \{+1, -1\}^{n \times p}$ , we consider the following collection of distributions:

$$\mathcal{P}'_{s, p, C, f, a, b} = \{P_{\theta, \sigma}(\{y_i\}_{i=1}^n) \mid y_i = f(\langle \theta^*, x_i \rangle) + \sigma_i, \text{ where } \|\theta\|_0 \leq s, \|\theta\|_2 \leq 1,$$

the random noise  $|\sigma_i| \leq C$  for some constant  $C > 0$ , and  $f$  is  $(a, b)$  monotone}.

The following theorem shows the lower bound of the private minimax risk (under the  $\|\cdot\|_2^2$  metric) with respect to the above collection of distributions, which is similar to the one in Theorem 5.3.6.

**Theorem 5.3.10.** Under Assumption 5.3.2 and for a given fixed privacy parameter  $\epsilon \in (0, \frac{1}{2}]$ , the  $\epsilon$  non-interactive local private minimax risk (under the  $\|\cdot\|_2^2$  metric) in the case of keeping  $\{y_i\}_{i=1}^n$  locally private satisfies the following inequality

$$\mathcal{M}_n^{\text{Nint}}(\theta(\mathcal{P}'_{s,p,C,f,a,b}), \|\cdot\|_2^2, \epsilon) \geq \Omega(\min\{1, C^2 \frac{s \log \frac{p}{s}}{nb^2 \epsilon^2 (1 + \Delta)}\}).$$

Comparing to the lower bound in Theorem 5.3.6 in the previous section, we can see that there is an additional factor of  $b^2$  in Theorem 5.3.10, which is due to the fact that the model is more complicated.

For the upper bound, we adopt a similar approach as in DP-IHT for linear regression. Particularly, we let  $L(\theta) = \frac{1}{2n} \sum_{i=1}^n (\tilde{y}_i - \langle x_i, \theta \rangle)^2$  and then apply the ideas of IHT.

---

**Algorithm 5.3.45** General DP-Iterative Hard Thresholding

---

**Input:** Public dataset  $\{x_i\}_{i=1}^n$ , private  $\{y_i\}_{i=1}^n \in P_{\theta^*, \sigma}$ , where  $P_{\theta^*, \sigma} \in \mathcal{P}_{s^*, p, C, f, a, b}$ ,  $\epsilon, \delta$  are privacy parameters,  $T$  is the number of iteration,  $\eta$  is the step size, and  $s$  is a parameter to be specified. Set  $\theta_0 = 0$ .

```

1: for Each  $i \in [n]$  do
2:   Denote  $\tilde{y}_i = y_i + z_i$ , where  $z_i \sim \mathcal{N}(0, \tau^2)$ ,  $\tau^2 = \frac{32C^2 \ln(1.25/\delta)}{\epsilon^2}$ .
3: end for
4: for  $t = 0, 1, \dots, T-1$  do
5:    $\tilde{\theta}_{t+1} = \theta_t - \eta \nabla L(\theta_t)$ .
6:    $\theta'_{t+1} = \text{Trunc}(\tilde{\theta}_{t+1}, s)$ .
7:    $\theta_{t+1} = \arg_{\theta \in \mathbb{B}_1} \|\theta - \theta'_{t+1}\|_2^2$ .
8: end for
9: Return  $\theta_T$ .
```

---

**Theorem 5.3.11.** For any  $0 < \epsilon \leq 1$  and  $0 < \delta < 1$ , Algorithm 5.3.45 is  $(\epsilon, \delta)$  (non-interactively) locally differentially private for  $\{y_i\}_{i=1}^n$ . Moreover, if  $\{y_i\}_{i=1}^n \in P_{\theta^*, \sigma}$  (where  $P_{\theta^*, \sigma} \in \mathcal{P}'_{s^*, p, C, f, a, b}$  with  $1 \geq \frac{a}{b} > \frac{\sqrt{5}}{3}$ ) and  $X$  satisfies Assumption 1 with  $0 < \Delta \leq \frac{9a^2 - 5b^2}{14}$ , then by setting  $s = 8s^*$  in Algorithm 5.3.45, there is an  $\eta = \eta(\Delta)$  which ensures that the output  $\theta_T$  satisfies the following inequality

$$\|\theta_T - \theta^*\|_2 \leq \left(\frac{1}{2}\right)^T \|\theta^*\|_2 + O\left(\frac{bC \log(1/\delta) \sqrt{s^* \log p}}{\sqrt{n} \epsilon}\right),$$

with probability at least  $1 - T \exp(-n) - \frac{2T}{p^c}$ .

### 5.3.6 Experiments

#### Experiments on Sparse Linear Regression

**Data Generation** Our data generation process is similar to the one in [161]. We first fix a parameter vector  $\theta^*$  by randomly choosing  $s^*$  coordinates, with each of them sampled independently from a uniform distribution in interval  $[0, 1]$ , and setting the remaining coordinates/entries to zero. Then, we generate the data samples using equation  $y_i = \langle x_i, \theta^* \rangle + \sigma_i$ , where  $x_i \in \text{Uniform}\{-1, +1\}^p$  and  $\sigma_i \in \text{Uniform}[-C, C]$ . We assume  $C = 0.05$  in our experiment.

**Experiment Results** We compare the relative error, *i.e.*  $\frac{\|\theta_T - \theta^*\|_2}{\|\theta^*\|_2}$ , with the sample size  $n$  in three different settings, *i.e.*, under varying dimensionality, sparsity and privacy level, respectively. We run algorithms Label-LDP-IHT with  $\eta = 0.2$  or  $\eta = 0.1$ ,  $s = s^*$ ,  $T = \lceil \log \frac{n}{p} \rceil$ ,  $\delta = 10^{-3}$  and a random normal Gaussian vector as the initial point to obtain  $\theta_T$ . For each experiment, we run the algorithm 10 times and take the one with the lowest relative error as the final value.

Figure 5.5 and 5.6 depict the results of Algorithm 5.3.41 and 5.3.42, respectively. From Figure 5.5, we can see that when the dimensionality and the sparsity level increase or the privacy parameter  $\epsilon$  decreases, the relative error increases, especially when the sample size  $n$  is small. When the sample size increases, the relative error will decrease. From Figure 5.6, we can learn that when the dimensionality  $p$  increases, unlike Figure 5.5, it does not cause the relative error to change significantly. This can be explained by the fact that the error bound is only logarithmically depending on  $p$ . Moreover, when the privacy parameter increases, the relative error decreases. These results confirm our theoretical claims.

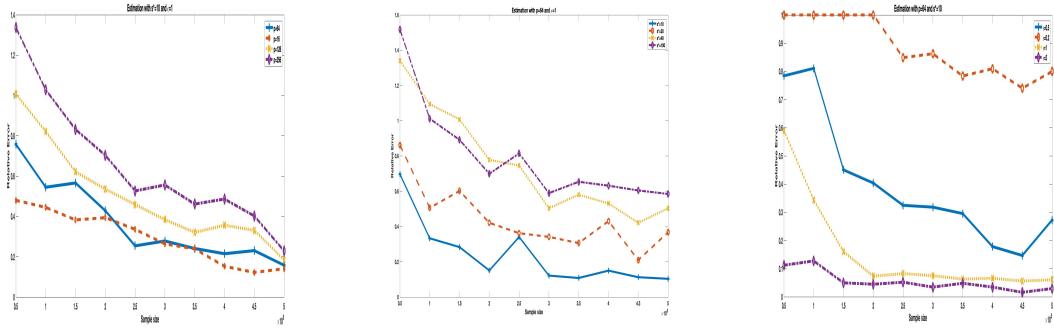
## Experiments on Sparsity-constrained DP-ERM

In this section, we test Algorithm 5.3.43 on real world datasets Covertype and rcv1 [64]. Particularly, we study the sparsity-constrained logistic regression problem with  $\ell(w, z) = \log(1 + \exp(-y_i \langle w, x_i \rangle)) + \frac{\lambda}{2} \|w\|_2^2$ , where  $y_i$  is the label of  $x_i$ . As pre-processing, the data is first normalized. Since there is no ground truth on real data, we run the algorithm in [161] sufficiently long until  $\|w_t - w_{t+1}\|_2 / \|w_t\|_2 \leq 10^{-4}$  and then use the output  $L(w_t; D)$  as the approximate optimal value. With this, we can calculate the optimality gap of our estimator. In the experiments, we set  $\lambda = 10^{-3}$ ,  $\eta = 0.1$  and  $\delta = 10^{-3}$ , and use zCDP [52] to achieve the  $(\epsilon, \delta)$ -DP.

From Figure 5.7 and 5.8, we can see that when the dimensionality  $p$  increases, the optimality gap does not change too much, which is due to the fact that the error bound is only logarithmically depending on  $p$ . Also, when the sparsity level increases or  $\epsilon$  decreases, the optimality gap increases. Clearly, all these experimental results are consistent with Theorem 5.3.8.

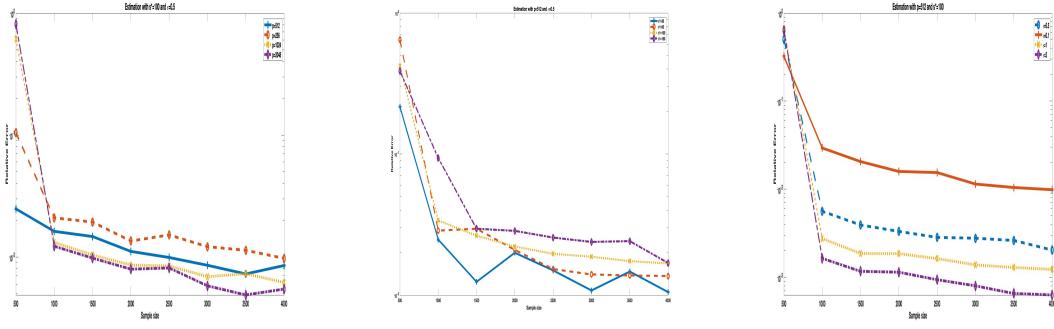
## Tests on Synthetic Datasets For Linear Regression with Non-linear Measurements

Our data generation process is similar to the one in [161]. We first fix a parameter vector  $\theta^*$  by randomly choosing  $s^*$  coordinates, with each of them sampled independently from a uniform distribution in interval  $[0, 1]$ , and setting the remaining coordinates/entries to zero. For the case of non-linear measurements, we assume that  $y_i = f(\langle x_i, \theta^* \rangle) + \sigma_i$ , where  $f(x) := 8x + \cos x$  where  $x_i \in \text{Uniform}\{-1, +1\}^p$  and  $\sigma_i \in \text{Uniform}[-C, C]$  so that it satisfies the assumptions in Theorem 5.3.9 . The results are shown in Figure 5.10 and 5.9. We can see that these results are almost the same as in Figure 5.5 and 5.6, respectively.



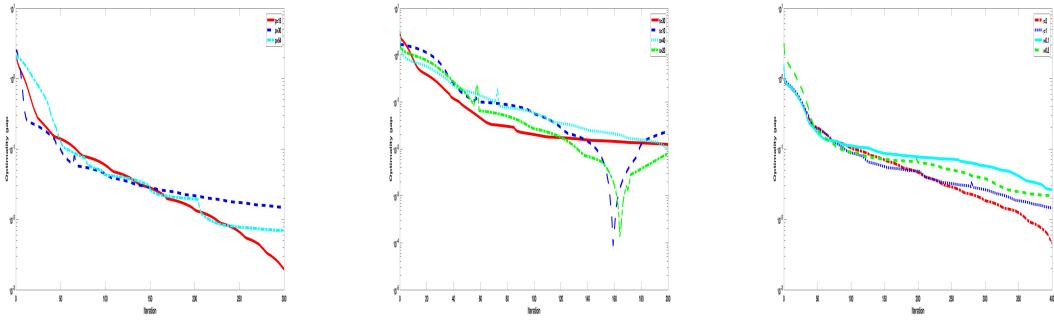
(a) Relative error w.r.t dimensionality      (b) Relative error w.r.t sparsity level      (c) Relative error w.r.t privacy level

Figure 5.5: Experimental results on sparse linear regression under LDP while keeping the whole dataset private (Algorithm 5.3.41).



(a) Relative error w.r.t dimensionality      (b) Relative error w.r.t sparsity level      (c) Relative error w.r.t privacy level

Figure 5.6: Experimental results on sparse linear regression under LDP while keeping the labels private (Algorithm 5.3.42).



(a) Optimality gap w.r.t dimensionality with fixed  $s = 10$  and  $\epsilon = 2$ .      (b) Optimality gap w.r.t sparsity level with fixed  $p = 54$  and  $\epsilon = 2$ .      (c) Optimality gap w.r.t privacy level with fixed  $p = 54$  and  $s = 10$

Figure 5.7: Experimental results on Covertype dataset [90] for  $\ell_0$ -constrained logistic regression under  $(\epsilon, \delta)$ -DP (Algorithm 5.3.43).

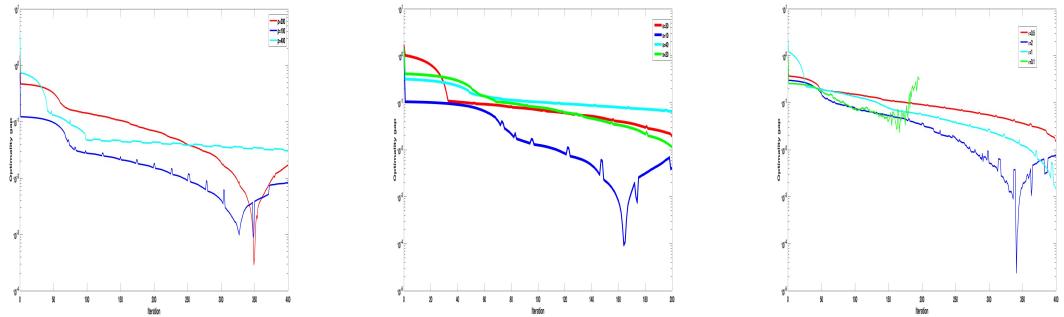


Figure 5.8: Experimental results on rcv1 dataset [64] for  $\ell_0$ -constrained logistic regression under  $(\epsilon, \delta)$ -DP (Algorithm 5.3.43).

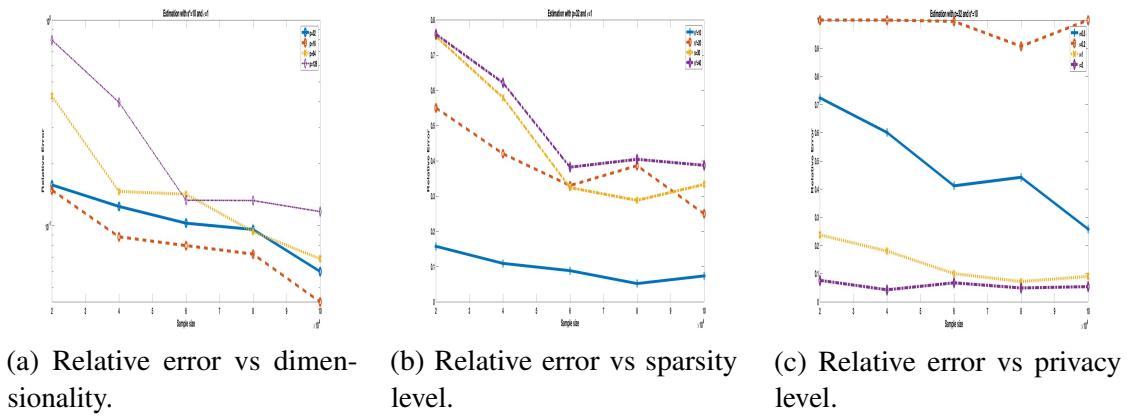


Figure 5.9: Experimental results for sparse regression with non-linear measurement under LDP when keeping the whole dataset private (Algorithm 5.3.44).

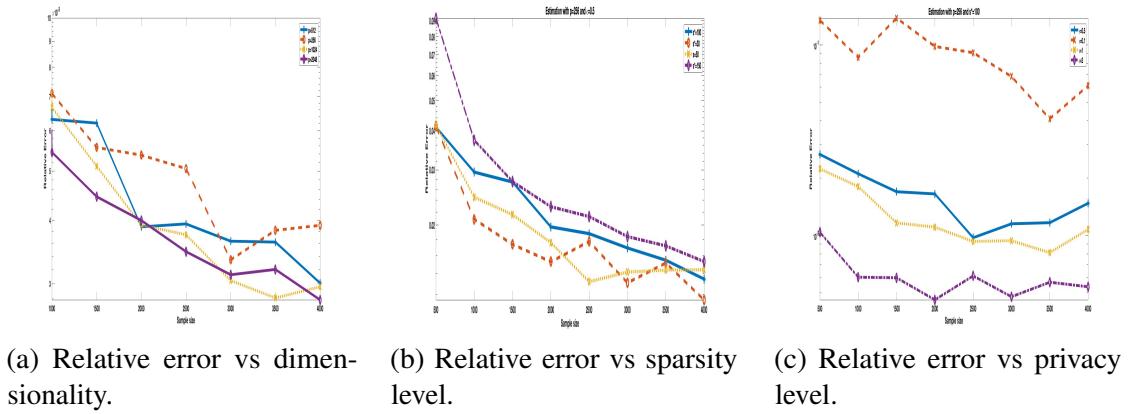


Figure 5.10: Experimental results for sparse regression with non-linear measurement under LDP when keeping the label private (Algorithm 5.3.45).

### 5.3.7 Omitted Proofs

#### Technical Lemmas

For the estimation error, we first give some definitions and lemmas.

**Definition 5.3.6.** A random variable  $X$  is said to be sub-Gaussian with  $\sigma^2$  if  $\mathbb{E}(X) = 0$  and

$$\mathbb{E}[\exp(sX)] \leq \exp\left(\frac{\sigma^2 s^2}{2}\right), \forall s \in \mathbb{R}.$$

For the case that  $X$  is a  $d$ -dimensional random vector, it is sub-Gaussian with  $\sigma^2$  if for any unit vector  $u \in \mathbb{S}^{d-1}$ ,  $u^T X$  is sub-Gaussian with  $\sigma^2$ .

It is well known that if  $X_1, X_2, \dots, X_n$  are all sub-Gaussian with  $\sigma^2$ , then  $a_1 X_1 + \dots + a_n X_n$  is sub-Gaussian with  $(\sum_{i=1}^n a_i^2) \sigma^2$ .

We can easily see that if  $x \sim \text{Uniform}\{+1, -1\}^d$ ,  $x$  is sub-Gaussian with  $\sigma^2 = 1$ .

**Lemma 5.3.2** ([289]). Let  $X_1, X_2, \dots, X_n$  be  $n$  random variables such that each  $X_i$  is sub-Gaussian with  $\sigma^2$ . Then the following holds

$$\begin{aligned} \Pr\left[\max_{i \in n} X_i \geq t\right] &\leq ne^{-\frac{t^2}{2\sigma^2}}, \\ \Pr\left[\max_{i \in n} |X_i| \geq t\right] &\leq 2ne^{-\frac{t^2}{2\sigma^2}}. \end{aligned}$$

**Lemma 5.3.3** ([161]). For any  $\theta \in \mathbb{R}^k$  and an integer  $s \leq k$ , if  $\theta_t = \text{Trunc}(\theta, s)$  then for any  $\theta^* \in \mathbb{R}^k$  with  $\|\theta^*\|_0 \leq s$ , we have  $\|\theta_t - \theta\|_2 \leq \frac{k-s}{k-s^*} \|\theta^* - \theta\|_2^2$ .

**Lemma 5.3.4.** Let  $\mathcal{K}$  be a convex body in  $\mathbb{R}^p$ , and  $v \in \mathbb{R}^p$ . Then for every  $u \in \mathcal{K}$ , we have

$$\|\mathcal{P}_{\mathcal{K}}(v) - u\|_2 \leq \|v - u\|_2,$$

where  $\mathcal{P}_{\mathcal{K}}$  is the operator of projection onto  $\mathcal{K}$ .

The following theorem says that when  $X \in \text{Uniform}\{+1, -1\}^{n \times p}$ , with high probability it satisfies the Restricted Isometry Property if  $n$  is sufficiently large.

**Lemma 5.3.5** (Theorem 2.12 in [245]). Let  $X \in \{+1, -1\}^{n \times p}$  be a Bernoulli Random Matrix and  $\xi, \Delta \in (0, 1)$ . Assume that

$$n \geq C\Delta^{-2}(s \log(p/s) + \log(1/\xi)).$$

Then with probability at least  $1 - \xi$ ,  $X$  satisfies the Restricted Isometry Property (RIP) with sparsity level  $s$  and parameter  $\Delta$ , that is, for every  $\|v\|_0 \leq s$ ,

$$(1 - \Delta)\|v\|^2 \leq \frac{1}{n}\|Xv\|_2^2 \leq (1 + \Delta)\|v\|_2^2.$$

Note that if  $X$  satisfies the Restricted Isometry Property (RIP) with sparsity level  $s$  and parameter  $\Delta$ , it means that

$$\Delta = \max_{\|x\|_2=1, \|x\|_0 \leq s} \|(\frac{1}{n}X^T X - I_{p \times p})x\|_2.$$

**Lemma 5.3.6** ([189]). If  $z \sim \chi_n^2$ , where  $\chi_n^2$  is the Chi-square distribution with parameter  $n$ , then

$$\Pr[z - n \geq 2\sqrt{nx} + 2x] \leq \exp(-x).$$

### Private Fano and Le Cam Method

Our lower bounds are basic on the locally private version Fano and Le Cam method [98, 100]. Given a finite set  $\mathcal{V}$ , a family of distributions  $\{P_v, v \in \mathcal{V}\}$  with  $P_v \in \mathcal{P}$  is  $2\delta$ -separated in a metric  $\rho$  if  $\rho(\theta(P_v), \theta(P_{v'})) \geq 2\delta$  for all distinct pairs  $v, v' \in \mathcal{V}$ . Given any  $2\delta$ -separated set, the private Fano's method for the  $\epsilon$  non-interactive private minimax risk can be summarized by the following lemma.

**Lemma 5.3.7** (Prop. 2 in [98]). Given any  $2\delta$ -separated set  $\{P_v, v \in \mathcal{V}\}$ , and  $\alpha \in (0, \frac{1}{2}]$ , the  $\epsilon$  non-interactive private minimax risk satisfies the following inequality

$$\mathcal{M}_n^{\text{Nint}}(\theta(\mathcal{P}), \Phi \circ \rho, \epsilon) \geq \frac{\Phi(\delta)}{2} \left( 1 - \frac{n\alpha^2 \mathcal{C}_{\infty}^{\text{Nint}}(\{P_v\}_{v \in \mathcal{V}}) + \log 2}{\log |\mathcal{V}|} \right),$$

where  $\mathcal{C}_{\infty}^{\text{Nint}}(\{P_v\}_{v \in \mathcal{V}}) = \frac{1}{|\mathcal{V}|} \sup_{\gamma \in \mathbb{B}_{\infty}} \sum_{v \in \mathcal{V}} (\psi_v(\gamma))^2$ ,  $\mathbb{B}_{\infty}$  is the 1-ball of the supremum norm  $\mathbb{B}_{\infty} = \{\gamma \in L^{\infty}(\mathcal{X}) \mid \|\gamma\|_{\infty} \leq 1\}$ , and  $L^{\infty}(\mathcal{X}) = \{f : \mathcal{X} \mapsto \mathbb{R} \mid \|f\|_{\infty} < \infty\}$  is the space of uniformly bounded functions with the supremum norm  $\|f\|_{\infty} = \sup_x |f(x)|$ . Also, for each  $v \in \mathcal{V}$ ,  $\psi_v : L^{\infty}(\mathcal{X}) \mapsto \mathbb{R}$  is a linear function defined by

$$\psi_v(\gamma) = \int_{\mathcal{X}} \gamma(x) dP_v(x) - d\bar{P}(x),$$

where  $\bar{P}$  is the mixture distribution  $\bar{P} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} P_v^n$ .

A useful corollary is the following:

**Lemma 5.3.8** (Corollaries 2 and 4 in [97]). Let  $V$  be randomly and uniformly distributed in  $\mathcal{V}$ . Assume that given  $V = v$ ,  $X_i$  is sampled independently according to the distribution of  $P_{v,i}$  for  $i = 1, \dots, n$ . Then, there is a universal constant  $c < 19$  such that for  $\alpha \in (0, \frac{1}{2}]$ ,

$$I(Z_1, Z_2, \dots, Z_n; V) \leq c\epsilon^2 \sum_{i=1}^n \frac{1}{|\mathcal{V}|^2} \sum_{v, v' \in \mathcal{V}} \|P_{v,i} - P_{v',i}\|_{TV}^2.$$

The  $\epsilon$  non-interactive private minimax risk satisfies

$$\mathcal{M}_n^{\text{Nint}}(\theta(\mathcal{P}), \Phi \circ \rho, \epsilon) \geq \frac{\Phi(\delta)}{2} \left( 1 - \frac{I(Z_1, \dots, Z_n; V) + \log 2}{\log |\mathcal{V}|} \right).$$

Now we introduce the generalized private Le Cam method. Let  $\mathcal{P}_0$  and  $\mathcal{P}_1$  be two collections of distributions in  $\mathcal{P}$ . We say that  $\mathcal{P}_0$  and  $\mathcal{P}_1$  are  $\delta$ -separated for loss function  $L$  if  $d_L(P_0, P_1) \geq \delta$  for all  $P_0 \in \mathcal{P}_0$  and  $P_1 \in \mathcal{P}_1$ , where  $d_L(P_0, P_1) = \inf_{\theta \in \Theta} \{L(\theta, \theta(P_0)) + L(\theta, \theta(P_1))\}$ . Then we have the following lemma.

**Lemma 5.3.9** (Theorem 2 in [100]). Consider a set of distributions  $\mathcal{P}$ , a collection of distributions on  $\mathcal{X}$ ,  $\{P_v\}_{v \in \mathcal{V}} \subset \mathcal{P}$ , indexed by  $v \in \mathcal{V}$ , as well as a distribution  $P_0 \in \mathcal{P}$ . For each of these distributions, we have i.i.d. observations  $X_i$ , that is, samples from the product with density

$$dP_v^n = \prod_{i=1}^n dP_v(x_i).$$

We also define the marginal distributions  $M_v^n(\cdot) = \int Q(\cdot|x_{1:n})dP_v^n(x_{1:n})$  and  $\bar{M}^n = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} M_v^n$ , where  $Q$  is a private channel. For any  $\epsilon \in (0, \frac{1}{2}]$ , the  $\epsilon$  sequential private minimax risk in the loss function  $L$  satisfies the following inequality

$$\mathcal{M}_n^{\text{Int}}(\theta(\mathcal{P}), L, \epsilon) \geq \frac{1}{2} \min_{v \in \mathcal{V}} d_L(P_0, P_v) \left(1 - \frac{1}{2} \sqrt{D_{kl}(M_0^n \parallel \bar{M}^n)}\right),$$

where

$$D_{kl}(M_0^n \parallel \bar{M}^n) \leq \frac{n\epsilon^2}{4} \mathcal{C}_\infty(\{P_v\}_{v \in \mathcal{V}}) \min\{e^\epsilon, \max_{v \in \mathcal{V}} \left\| \frac{dP}{dP_v} \right\|_\infty\}$$

for any distribution  $P$  supported on  $\mathcal{X}$ . Here

$$\mathcal{C}_\infty(\{P_v\}_{v \in \mathcal{V}}) = \inf_{\text{supp } P^* \in \mathcal{X}} \sup_{\gamma} \left\{ \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \phi_v(\gamma)^2 \|\gamma\|_{L^\infty(P^*)} \right\}.$$

Where the linear functional  $\phi_v(f)$  is defined as

$$\phi_v(f) := \int f(x) (dP_0(x) - dP_v(x)).$$

### Proof of Theorem 5.3.1

The main idea of the proof is :

- Find an index set  $\mathcal{V}$  which corresponds to a  $2\delta$ -separated set  $\{P_v, v \in \mathcal{V}\}$ .
- Obtain an upper bound on  $\mathcal{C}_\infty(\{P_v\}_{v \in \mathcal{V}})$ , use Lemma 5.3.7 to specify  $\delta$ , and then get an lower bound.

We consider  $\mathcal{V}$  as the set of  $\{\pm e_j, j \in [p]\}$ , where  $\{e_j\}_{j=1}^n$  is the standard basis of  $\mathbb{R}^p$ .

Let  $\theta_v = \delta v$  for some  $\delta < 1$  and every  $v \in \mathcal{V}$ . Then for each  $\theta_v$ , we define the distribution

$P_{\theta_v}$  as

$$P_{\theta_v} = \left\{ x \in \text{Uniform}\{+1, -1\}^p; p_{\theta_v}(y | x, \sigma) = \langle x, \theta_v \rangle + \sigma; \text{ where } \sigma = \begin{cases} 1 - \langle x, \theta_v \rangle \text{ w.p. } \frac{1+\langle x, \theta_v \rangle}{2} \\ -1 - \langle x, \theta_v \rangle \text{ w.p. } \frac{1-\langle x, \theta_v \rangle}{2} \end{cases} \right\}. \quad (5.83)$$

It is easy to see that  $P_{\theta_v} \in \mathcal{P}_{1,p,2}$  since the noise  $|\sigma| \leq 1 + |\langle x, \theta_v \rangle| \leq 2$ . Note that the distribution in (5.83) is equivalent to

$$p_{\theta_v}((x, y)) = \frac{1 + y\langle x, \theta_v \rangle}{2^{p+1}} \text{ for } (x, y) \in \{+1, -1\}^{p+1}. \quad (5.84)$$

Also for every fixed  $(x, y) \in \{+1, -1\}^{p+1}$ , we have  $\bar{p}((x, y)) := \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} p_{\theta_v}((x, y)) = \frac{1}{2^{p+1}}$ .

Now we show our main lemma used in the proof.

**Lemma 5.3.10.** The term  $\mathcal{C}_\infty^{Nint}(\{P_v\}_{v \in \mathcal{V}})$  satisfies the following inequality

$$\mathcal{C}_\infty^{Nint}(\{P_v\}_{v \in \mathcal{V}}) \leq \frac{\delta^2}{p}. \quad (5.85)$$

*Proof of Lemma 5.3.10.* By definition, for each  $v \in \mathcal{V}$  we have

$$\begin{aligned} \psi_v(\gamma) &= \sum_{(x,y) \in \{+1, -1\}^{p+1}} \gamma(x, y)[p_v((x, y)) - \bar{p}((x, y))] \\ &= \frac{\delta}{2^{p+1}} \sum_{(x,y) \in \{+1, -1\}^{p+1}} \gamma(x, y)y\langle x, v \rangle \\ &= \frac{\delta}{2^{p+1}} \sum_{x \in \{+1, -1\}^p} [\gamma(x, 1)\langle x, v \rangle - \gamma(x, -1)\langle x, v \rangle] \end{aligned}$$

Thus, we can get

$$\begin{aligned}
\frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \psi_v^2(\gamma) &\leq 2 \times \frac{1}{2p} \sum_{v \in \mathcal{V}} \left[ \left( \frac{\delta}{2^{p+1}} \sum_{x \in \{+1, -1\}^p} \gamma(x, 1) \langle x, v \rangle \right)^2 + \right. \\
&\quad \left. \left( \frac{\delta}{2^{p+1}} \sum_{x \in \{+1, -1\}^p} \gamma(x, -1) \langle x, v \rangle \right)^2 \right] \\
&= \frac{\delta^2}{p4^{p+1}} \sum_{v \in \mathcal{V}} \sum_{x_1, x_2 \in \{+1, -1\}^p} [(\gamma(x_1, 1)\gamma(x_2, 1) + \gamma(x_1, -1)\gamma(x_2, -1))x_1^T v v^T x_2] \\
&= \frac{2\delta^2}{p4^{p+1}} \sum_{x_1, x_2 \in \{+1, -1\}^p} (\gamma(x_1, 1)\gamma(x_2, 1)x_1^T x_2 + \gamma(x_1, -1)\gamma(x_2, -1)x_1^T x_2),
\end{aligned}$$

where the last equation is due to  $\sum_{v \in \mathcal{V}} v v^T = 2I_{p \times p}$ . Thus by the definition of  $\mathcal{C}_\infty^{Nint}(\{P_v\}_{v \in \mathcal{V}})$  we have

$$\begin{aligned}
\mathcal{C}_\infty^{Nint}(\{P_v\}_{v \in \mathcal{V}}) &\leq \frac{1}{2} \frac{\delta^2}{p4^p} \left[ \sup_{\gamma \in \mathbb{B}_\infty} \sum_{x_1, x_2 \in \mathcal{X}} \gamma(x_1, 1)\gamma(x_2, 1)x_1^T x_2 \right. \\
&\quad \left. + \sup_{\gamma \in \mathbb{B}_\infty} \sum_{x_1, x_2 \in \mathcal{X}} \gamma(x_1, -1)\gamma(x_2, -1)x_1^T x_2 \right] \\
&= \frac{\delta^2}{2p} \left[ \sup_{\gamma \in \mathbb{B}_\infty} \|\mathbb{E}_{P_0}[\gamma(X, 1)X]\|^2 + \sup_{\gamma \in \mathbb{B}_\infty} \|\mathbb{E}_{P_0}[\gamma(X, -1)X]\|^2 \right],
\end{aligned}$$

where  $P_0$  is the uniform distribution on  $\{+1, -1\}^p$ . Note that since  $\|a\|_2^2 = \sup_{\|v\| \leq 1} \langle v, a \rangle^2$  for any vector  $a$ , by Cauchy-Schwartz inequality we have

$$\begin{aligned}
&\sup_{\gamma \in \mathbb{B}_\infty} \|\mathbb{E}_{P_0}[\gamma(X, 1)X]\|^2 \\
&= \sup_{\gamma \in \mathbb{B}_\infty, \|v\|_2 \leq 1} (\mathbb{E}_{P_0}[\gamma(X, 1)v^T X])^2 \\
&\leq \sup_{\gamma \in \mathbb{B}_\infty} \mathbb{E}_{P_0}[\gamma(X, 1)^2] \times \sup_{\|v\|_2 \leq 1} \mathbb{E}_{P_0}[(v^T X)^2] \\
&\leq \sup_{\|v\|_2 \leq 1} v^T \sum_{x \in \{-1, 1\}^p} \frac{xx^T}{2^p} v \leq 1,
\end{aligned}$$

where the second inequality is due to the definition of  $X$  and  $\gamma$ . Similarly, we can bound the term  $\sup_{\gamma \in \mathbb{B}_\infty} \|\mathbb{E}_{P_0}[\gamma(X, -1)X]\|^2 \leq 1$ . This completes the proof.  $\square$

By Lemma 5.3.7 and Lemma 5.3.10, we can get

$$\mathcal{M}_n^{\text{Nint}}(\theta(\mathcal{P}_{1,p,2}), \Phi \circ \rho, \alpha) \geq \frac{\delta^2}{2} \left( 1 - \frac{n\epsilon^2 \frac{\delta^2}{p} + \log 2}{\log 2p} \right).$$

If we take  $\delta^2 = \Omega(\min\{1, \frac{p \log 2p}{n\epsilon^2}\})$ , we can get the proof of the lower bound in Theorem 5.3.1.

### Proof of Theorem 5.3.2

Now we use the squared loss as the loss function  $L(\theta, \theta') = \|\theta - \theta'\|_2^2$ . Then,  $d_L(P_0, P_1) = \frac{1}{2}\|\theta(P_0) - \theta(P_1)\|_2^2$ . Define  $P_0 \in \mathcal{P}_{1,p,C}$  as the uniform distribution on  $\{+1, -1\}^p \times \{+1, -1\}$ , that is,

$$P_0 = \{x \in \text{Uniform}\{+1, -1\}^p; p_{\theta_v}(y | x, \sigma) = \langle x, 0 \rangle + \sigma; \\ \text{where } \sigma = \begin{cases} 1 - \langle x, 0 \rangle \text{ w.p. } \frac{1+\langle x, 0 \rangle}{2} \\ -1 - \langle x, 0 \rangle \text{ w.p. } \frac{1-\langle x, 0 \rangle}{2} \end{cases}\}.$$

Thus,  $\theta(P_0) = 0$ .

Define the set of distributions  $\{P_v, v \in \mathcal{V}\}$  in the same way as in the proof of Theorem 5.3.1. Then, we have  $d_L(P_0, P_1) = \frac{1}{2}\delta^2$ . As in Lemma 5.3.9, we have  $M_0^n$  and  $\bar{M}^n$ . For the KL-divergence  $D_{kl}$  between  $M_0^n$  and  $\bar{M}^n$ , by Lemma 5.3.9 we have

$$D_{kl}(M_0^n \| \bar{M}^n) \leq \frac{n\epsilon^2}{4} \mathcal{C}_\infty(\{P_v\}_{v \in \mathcal{V}}) \min\{e^\epsilon, \max_{v \in \mathcal{V}} \|\frac{dP}{dP_v}\|_\infty\}.$$

We can easily see that for each  $\gamma \in \mathbb{B}_\infty$  and  $v \in \mathcal{V}$ , we have that  $\psi_v(\gamma)$  in the proof of Lemma 5.3.10 is equivalent to  $\phi_v(\gamma)$  in Lemma 5.3.9 for our construction. Thus, by Lemma

5.3.10 we have  $\mathcal{C}_\infty(\{P_v\}_{v \in \mathcal{V}}) \leq \frac{\delta^2}{p}$ . Taking  $P = P_0$ , we get  $\max_{v \in \mathcal{V}} \|\frac{dP}{dP_v}\|_\infty = \frac{1}{1-\delta}$ . Thus, if choosing  $\delta^2 = \Omega(\min\{1, \frac{p}{n\epsilon^2}\})$ , we have

$$D_{kl}(M_0^n \parallel \bar{M}^n) \leq \frac{n\epsilon^2\delta^2(1+\delta)}{8p}.$$

By Lemma 5.3.9, we can get

$$\mathcal{M}_n^{\text{Int}}(\theta(\mathcal{P}_{1,p,2}), \Phi \circ \rho, \alpha) \geq \frac{\delta^2}{4} \left(1 - \sqrt{\frac{n\epsilon^2\delta^2(1+\delta)}{8p}}\right).$$

Thus, if taking  $\delta^2 = \Omega(\min\{1, \frac{p}{n\epsilon^2}\})$ , we have the proof.

### Proof of Theorem 5.3.3

Now consider the case of  $L(\theta, \theta') = |\mathbf{1}^T(\theta - \theta')|$ . We can easily obtain  $d_L(P_1, P_2) \geq |\mathbf{1}^T(\theta(P_2) - \theta(P_1))|$ . Consider the same distributions  $P_0, \{P_v, v \in \mathcal{V}\}$  as in the proof of Theorem 5.3.2, we have  $\min_{v \in \mathcal{V}} d_L(P_0, P_v) \geq \delta$ . Since  $D_{kl}(M_0^n \parallel \bar{M}^n) \leq \frac{n\epsilon^2\delta^2(1+\delta)}{8p}$  for  $\delta^2 = \Omega(\min\{1, \frac{p}{n\epsilon^2}\})$ , we have

$$\mathcal{M}_n^{\text{int}}(\theta(\mathcal{P}_{1,p,2}), L, \alpha) \geq \frac{\delta}{2} \left(1 - \sqrt{\frac{n\epsilon^2\delta^2(1+\delta)}{8p}}\right).$$

Thus, we have the proof if set  $\delta^2 = \Omega(\min\{1, \frac{p}{n\epsilon^2}\})$ .

### Proof of Theorem 5.3.4

Before the proof, let us recall the definition of  $\chi^2$ -local differential privacy [100]:

For any convex function  $f$  on  $\mathbb{R}_+$  with  $f(1) = 0$ , the  $f$ -divergence of distributions  $P$  and  $Q$  is

$$D_f(P \parallel Q) := \int f\left(\frac{dP}{dQ}\right) dQ.$$

**Definition 5.3.7.** Let  $f(x) = (x - 1)^2$ . Following the above definitions, we have  $\epsilon^2$ - $\chi^2$ -

divergence local differential privacy and  $\epsilon$ - $\chi^2$ -divergence (sequentially) private minimax risk if

$$D_f(Q_i(Z_i \in S | x_i, z_{1:i-1}) \| Q_i(Z_i \in S | x'_i, z_{1:i-1})) \leq \epsilon^2.$$

From the above definitions, it is easy to see that if a channel  $Q$  is  $(\kappa, \rho)$  sequentially locally zero-concentrated differentially private, it is  $(\epsilon^2 = e^{\kappa+2\rho} - 1)$ - $\chi^2$ -divergence sequentially locally differentially private. Also, since  $(2, \log(1 + \epsilon^2))$  local Renyi differential privacy is equivalent to  $\epsilon^2$ - $\chi$ -divergence local differential privacy, to prove Theorem 5.3.4, we only need to show the lower bound of  $\epsilon^2$ - $\chi^2$ -divergence sequential local private minimax risk, which is denoted as  $\mathcal{M}_{n,\chi^2}^{\text{Int}}(\theta(\mathcal{P}), L, \epsilon^2)$ . To do that, we need the following lemma.

**Lemma 5.3.11.** [Theorem 2 in [100]] For any  $\epsilon \in (0, 1]$ , the  $\epsilon^2$ - $\chi^2$ -divergence sequential private minimax risk in the loss function  $L$  satisfies the following inequality

$$\mathcal{M}_{n,\chi^2}^{\text{Int}}(\theta(\mathcal{P}), L, \epsilon^2) \geq \frac{1}{2} \min_{v \in \mathcal{V}} d_L(P_0, P_v) \times \left(1 - \frac{1}{2} \sqrt{D_{kl}(M_0^n \| \bar{M}^n)}\right),$$

where

$$D_{kl}(M_0^n \| \bar{M}^n) \leq n\epsilon^2 \mathcal{C}_2(\{P_v\}_{v \in \mathcal{V}}) \min\{e^\epsilon, \max_{v \in \mathcal{V}} \|\frac{dP_v}{dP}\|_\infty\}$$

for any distribution  $P$  supported on  $\mathcal{X}$ , and  $\mathcal{C}_2(\{P_v\}_{v \in \mathcal{V}}) = \frac{1}{|\mathcal{V}|} \inf_{\text{supp } P \subset \mathcal{X}} \sup_{\gamma} \{\sum_{v \in \mathcal{V}} (\phi_v(\gamma))^2 | \|\gamma\|_{L^2(P)} \leq 1\}$ , where  $\phi(\gamma)$  is defined in Lemma 5.3.9.

Now, we will proof Theorem 5.3.4.

The construction of  $P_0$  and  $\{P_v, v \in \mathcal{V}\}$  is the same as in the proof of Theorem 5.3.3. Thus, by Lemma 5.3.11, we only need to bound  $\mathcal{C}_2(\{P_v\}_{v \in \mathcal{V}})$ , instead of  $\mathcal{C}_\infty(\{P_v\}_{v \in \mathcal{V}})$ . From the proof of Lemma 5.3.10, we can see that if taking  $P$  as a uniform distribution, then for any  $\gamma$  with  $\|\gamma\|_{L^2(P_0)} \leq 1$ , we always have  $\mathbb{E}_{P_0}[\gamma(X, 1)^2] \leq 1$ . This means that  $\frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} (\psi_v(\gamma))^2 \leq \frac{\delta^2}{p}$ . Thus, we have  $\mathcal{C}_2(\{P_v\}_{v \in \mathcal{V}}) \leq \frac{\delta^2}{p}$ . The remaining part of the proof is the same as the one in the proof of Theorem 5.3.2.

### Proof of Theorem 5.3.5

Follow from the fact that the linear model is a special case of the non-linear measurement.

See the proof of Theorem 5.3.9 in Section 5.3.5 for the case  $f(x) = x$  and  $a = b = 1$ .

### Proof of Theorem 5.3.6

Follow from the fact that the linear model is a special case of the non-linear measurement.

See the proof of Theorem 5.3.11 in Section 5.3.5 for the case  $f(x) = x$  and  $a = b = 1$ .

### Proof of Theorem 5.3.7

Follow from the fact that the linear model is a special case of the non-linear measurement.

See the proof of Theorem 5.3.10 in Section 5.3.5 for the case  $f(x) = x$  and  $a = b = 1$ .

### Proof of Theorem 5.3.8

For the guarantee of  $(\epsilon, \delta)$ -DP, it follows from the Moment accountant and composition theorem, see [1, 325] for details.

Let  $\mathcal{I} = \mathcal{I}^{t+1} \cup \mathcal{I}^t \cup \mathcal{I}^*$ , where  $\mathcal{I}^* = \text{supp}(x^*)$ ,  $\mathcal{I}^t = \text{supp}(x_t)$  and  $\mathcal{I}^{t+1} = \text{supp}(x_{t+1})$ , and  $g_t = \nabla L(x_t) + z_t$ . Since  $\|x_{t+1} - x_t\|_0 \leq 2k$ . By the assumption of RSS, we have

$$\begin{aligned}
L(x_{t+1}) &\leq L(x_t) + \langle \nabla L(x_t), x_{t+1} - x_t \rangle + \frac{\ell_s}{2} \|x_{t+1} - x_t\|^2 \\
&\leq L(x_t) + \langle (g_t)_\mathcal{I}, (x_{t+1} - x_t)_\mathcal{I} \rangle + \frac{\ell_s}{2} \|x_{t+1} - x_t\|^2 + \|z_{t,\mathcal{I}}\| \| (x_{t+1} - x_t)_\mathcal{I} \|_2 \\
&= L(x_t) + \frac{1}{2\eta} \|x_{t+1,\mathcal{I}} - x_{t,\mathcal{I}} + \eta g_{t,\mathcal{I}}\|^2 - \frac{\eta \|g_{t,\mathcal{I}}\|^2}{2} \\
&\quad - \frac{1 - \eta \ell_s}{2\eta} \|x_{t+1} - x_t\|^2 + \|z_{t,\mathcal{I}}\| \| (x_{t+1} - x_t)_\mathcal{I} \|_2 \\
&= L(x_t) + \frac{1}{2\eta} (\|x_{t+1,\mathcal{I}} - x_{t,\mathcal{I}} + \eta g_{t,\mathcal{I}}\|^2 - \eta^2 \|g_{t,\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}\|^2) - \frac{\eta \|g_{t,\mathcal{I}^t \cup \mathcal{I}^*}\|^2}{2} \\
&\quad - \frac{1 - \eta \ell_s}{2\eta} \|x_{t+1} - x_t\|^2 + \|z_{t,\mathcal{I}}\| \| (x_{t+1} - x_t)_\mathcal{I} \|_2,
\end{aligned} \tag{5.86}$$

where the second inequality is due to  $x_{t+1} - x_t = x_{t+1,\mathcal{I}} - x_{t,\mathcal{I}}$ .

We now bound the term of  $\|x_{t+1,\mathcal{I}} - x_{t,\mathcal{I}} + \eta g_{t,\mathcal{I}}\|^2 - \eta^2 \|g_{t,\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}\|^2$  by the idea in [161]. Since  $\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*) = \mathcal{I}^{t+1} \setminus (\mathcal{I}^t \cup \mathcal{I}^*) \subseteq \mathcal{I}^{t+1}$ , we have

$$x_{t+1,\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)} = x_{t,\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)} - \eta g_{t,\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}.$$

Also, since  $x_{t,\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)} = 0$ , this means that  $x_{t+1,\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)} = -\eta g_{t,\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}$ . Next, we choose a set  $\mathcal{R} \subseteq \mathcal{I}^t \setminus \mathcal{I}^{t+1}$  such that  $|\mathcal{R}| = |\mathcal{I}^{t+1} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)|$ . Note that such  $\mathcal{R}$  can be found since  $|\mathcal{I}^{t+1} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)| = |\mathcal{I}^t \setminus \mathcal{I}^{t+1}| - |(\mathcal{I}^{t+1} \cap \mathcal{I}^*) \setminus \mathcal{I}^t|$  (which is a consequence of  $|\mathcal{I}^t| = |\mathcal{I}^{t+1}|$ ). Thus, we have

$$\eta^2 \|g_{t,\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}\|^2 = \|x_{t+1,\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}\|^2 \geq \|x_{t,\mathcal{R}} - \eta g_{t,\mathcal{R}}\|^2. \quad (5.87)$$

With (5.87) and the fact that  $x_{t+1,\mathcal{R}} = 0$ , we have

$$\begin{aligned} & \|x_{t+1,\mathcal{I}} - x_{t,\mathcal{I}} + \eta g_{t,\mathcal{I}}\|^2 - \eta^2 \|g_{t,\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}\|^2 \\ & \leq \|x_{t+1,\mathcal{I}} - x_{t,\mathcal{I}} + \eta g_{t,\mathcal{I}}\|^2 - \|x_{t+1,\mathcal{R}} - x_{t,\mathcal{R}} + \eta g_{t,\mathcal{R}}\|^2 \\ & = \|x_{t+1,\mathcal{I} \setminus \mathcal{R}} - x_{t,\mathcal{I} \setminus \mathcal{R}} + \eta g_{t,\mathcal{I} \setminus \mathcal{R}}\|^2. \end{aligned} \quad (5.88)$$

We then bound the size of  $|\mathcal{I} \setminus \mathcal{R}|$  as  $|\mathcal{I} \setminus \mathcal{R}| \leq |\mathcal{I}^{t+1}| + |(\mathcal{I}^t \setminus \mathcal{I}^{t+1}) \setminus \mathcal{R}| + |\mathcal{I}^*| \leq k + |(\mathcal{I}^{t+1} \cap \mathcal{I}^*) \setminus \mathcal{I}^t| + k^* \leq k + 2k^*$ . Also, since  $\mathcal{I}^{t+1} \subseteq (\mathcal{I} \setminus \mathcal{R})$ , we have  $x_{t+1,\mathcal{I} \setminus \mathcal{R}} =$

$\text{Trun}(x_{t,\mathcal{I}\setminus\mathcal{R}} - \eta g_{t,\mathcal{I}\setminus\mathcal{R}}, k)$ . Thus, by (5.87) and Lemma 5.3.3 we have

$$\begin{aligned}
& \|x_{t+1,\mathcal{I}} - x_{t,\mathcal{I}} + \eta g_{t,\mathcal{I}}\|^2 - \eta^2 \|g_{t,\mathcal{I}\setminus(\mathcal{I}^t \cup \mathcal{I}^*)}\|^2 \\
& \leq \|x_{t+1,\mathcal{I}\setminus\mathcal{R}} - x_{t,\mathcal{I}\setminus\mathcal{R}} + \eta g_{t,\mathcal{I}\setminus\mathcal{R}}\|^2 \\
& \leq \frac{2k^*}{k+k^*} \|x_{\mathcal{I}\setminus\mathcal{R}}^* - x_{t,\mathcal{I}\setminus\mathcal{R}} + \eta g_{t,\mathcal{I}\setminus\mathcal{R}}\|^2 \\
& \leq \frac{2k^*}{k+k^*} \|x_{\mathcal{I}}^* - x_{t,\mathcal{I}} + \eta g_{t,\mathcal{I}}\|^2 \\
& = \frac{2k^*}{k+k^*} (\|x^* - x^t\|^2 + 2\eta \langle g_{t,\mathcal{I}}, (x^* - x_t)_{\mathcal{I}} \rangle + \eta^2 \|g_{t,\mathcal{I}}\|^2) \\
& = \frac{2k^*}{k+k^*} (\|x^* - x^t\|^2 + 2\eta \langle \nabla L(x_t), (x^* - x_t) \rangle + \eta^2 \|g_{t,\mathcal{I}}\|^2) + \frac{4k^*}{k+k^*} \langle z_{t,\mathcal{I}}, (x^* - x_t)_{\mathcal{I}} \rangle \\
& \leq \frac{2k^*}{k+k^*} [\|x^* - x^t\|^2 + 2\eta (L(x^*) - L(x_t) - \frac{\rho_s}{2} \|x^* - x_t\|^2) \\
& \quad + \eta^2 \|g_{t,\mathcal{I}}\|^2] + \frac{4k^*}{k+k^*} \langle z_{t,\mathcal{I}}, (x^* - x_t)_{\mathcal{I}} \rangle \\
& = \frac{4\eta k^*}{k+k^*} (L(x^*) - L(x_t)) + \frac{2(1-\eta\rho_s)k^*}{k+k^*} \|x^* - x_t\|^2 + \frac{2\eta^2 k^*}{k+k^*} \|g_{t,\mathcal{I}\setminus(\mathcal{I}^t \cup \mathcal{I}^*)}\|^2 \\
& \quad + \frac{2\eta^2 k^*}{k+k^*} \|g_{t,(\mathcal{I}^t \cup \mathcal{I}^*)}\|^2 + \frac{4k^*}{k+k^*} \langle z_{t,\mathcal{I}}, (x^* - x_t)_{\mathcal{I}} \rangle.
\end{aligned}$$

Plugging this into (5.86), we get

$$\begin{aligned}
L(x_{t+1}) & \leq L(x_t) + \frac{2k^*}{k+k^*} (L(x^*) - L(x_t)) + \frac{(1-\eta\rho_s)k^*}{\eta(k+k^*)} \|x^* - x_t\|^2 \\
& \quad + \frac{\eta k^*}{k+k^*} \|g_{t,\mathcal{I}\setminus(\mathcal{I}^t \cup \mathcal{I}^*)}\|^2 + \left(\frac{\eta k^*}{k+k^*} - \frac{\eta}{2}\right) \|g_{t,\mathcal{I}^t \cup \mathcal{I}^*}\|^2 + \frac{2k^*}{\eta(k+k^*)} \langle z_{t,\mathcal{I}}, (x^* - x_t)_{\mathcal{I}} \rangle \\
& \quad + \|z_{t,\mathcal{I}}\| \|(x_{t+1} - x_t)_{\mathcal{I}}\|_2 - \frac{1-\eta\ell_s}{2\eta} \|x_{t+1} - x_t\|^2
\end{aligned} \tag{5.90}$$

$$\begin{aligned}
& \leq L(x_t) + \frac{2k^*}{k+k^*} (L(x^*) - L(x_t)) + \frac{(1-\eta\rho_s)k^*}{\eta(k+k^*)} \|x^* - x_t\|^2 \\
& \quad - \left(\frac{1-\eta\ell_s}{2\eta} - \frac{k^*}{\eta(k+k^*)}\right) \|x_{t+1} - x_t\|^2 + \\
& \quad \left(\frac{\eta k^*}{k+k^*} - \frac{\eta}{2}\right) \|g_{t,\mathcal{I}^t \cup \mathcal{I}^*}\|^2 + \frac{2k^*}{\eta(k+k^*)} \langle z_{t,\mathcal{I}}, (x^* - x_t)_{\mathcal{I}} \rangle + \|z_{t,\mathcal{I}}\| \|(x_{t+1} - x_t)_{\mathcal{I}}\|_2
\end{aligned} \tag{5.91}$$

$$\begin{aligned}
& \leq L(x_t) + \frac{2k^*}{k+k^*} (L(x^*) - L(x_t)) + \frac{(1-\eta\rho_s)k^*}{\eta(k+k^*)} \|x^* - x_t\|^2 + \left(\frac{\eta k^*}{k+k^*} - \frac{\eta}{2}\right) \|g_{t,\mathcal{I}^t \cup \mathcal{I}^*}\|^2 \\
& \quad + \frac{2k^*}{\eta(k+k^*)} \langle z_{t,\mathcal{I}}, (x^* - x_t)_{\mathcal{I}} \rangle + \frac{\eta(k+k^*)}{2((1-\eta\ell_s)k - (1+\eta\ell_s)k^*)} \|z_{t,\mathcal{I}}\|^2,
\end{aligned} \tag{5.92}$$

where the second inequality is due to the fact that  $\|x_{t+1} - x_t\| \geq \eta \|g_{t,\mathcal{I} \setminus (\mathcal{I}^t \cup \mathcal{I}^*)}\|$  and the third inequality is due to the fact that  $ab \leq \frac{a^2}{4c} + cb^2$  for any  $c > 0$ .

For the term  $\|x_t - x^*\|^2$ , we have the following lemma:

**Lemma 5.3.12.**

$$\|x_t - x^*\|^2 \leq \frac{4}{\rho} [L(x^*) - L(x_t)] + \frac{8}{\rho_s^2} \|g_{t,\mathcal{I}^t \cup \mathcal{I}^*}\|^2 + \frac{8}{\rho_s^2} \|z_{t,\mathcal{I}}\|^2. \quad (5.93)$$

*Proof.* From RSC, we have

$$\begin{aligned} L(x^*) &\geq L(x_t) + \langle \nabla L(x_t), x^* - x_t \rangle + \frac{\rho_s}{2} \|x^* - x_t\|^2 \\ &= L(x_t) + \langle \nabla_{\mathcal{I}^t \cup \mathcal{I}^*} L(x_t) - g_{t,\mathcal{I}^t \cup \mathcal{I}^*} + g_{t,\mathcal{I}^t \cup \mathcal{I}^*}, x^* - x_t \rangle + \frac{\rho_s}{2} \|x^* - x_t\|^2 \\ &\geq L(x_t) - \frac{2}{\rho_s} \|z_{t,\mathcal{I}}\|^2 - \frac{2}{\rho_s} \|g_{t,\mathcal{I}^t \cup \mathcal{I}^*}\|^2 + \frac{\rho_s}{4} \|x^* - x_t\|^2, \end{aligned}$$

where the last inequality is due to  $ab \leq \frac{a^2}{4c} + cb^2$ .  $\square$

With this lemma, we get

$$\begin{aligned} L(x_{t+1}) &\leq L(x_t) + \frac{2k^*}{k + k^*} \left(1 + \frac{2(1 - \eta\rho_s)}{\eta\rho_s}\right) (L(x^*) - L(x_t)) \\ &\quad - \left(\frac{\eta}{2} - \frac{(\eta^2\rho_s^2 + 8(1 - \eta\rho_s))k^*}{\eta\rho_s^2(k + k^*)}\right) \|g_{t,\mathcal{I}^t \cup \mathcal{I}^*}\|^2 + \frac{2k^*}{\eta(k + k^*)} \langle z_{t,\mathcal{I}}, (x^* - x_t)_{\mathcal{I}} \rangle \\ &\quad + \left(\frac{\eta(k + k^*)}{2((1 - \eta\rho_s)k - (1 + \eta\rho_s)k^*)} + \frac{8(1 - \eta\rho_s)k^*}{\eta\rho_s^2(k + k^*)}\right) \|z_{t,\mathcal{I}}\|^2. \end{aligned} \quad (5.94)$$

Taking  $\eta = \frac{1}{2\ell_s}$  and  $k \geq (1 + \frac{64\ell_s^2}{\rho_s^2})k^*$ , we further get

$$L(x_{t+1}) \leq L(x_t) + \frac{\rho_s}{8\ell_s} (L(x^*) - L(x_t)) + \frac{4k^*\ell_s}{(k + k^*)} \langle z_{t,\mathcal{I}}, (x^* - x_t)_{\mathcal{I}} \rangle + \frac{37\ell_s}{\rho_s^2} \|z_{t,\mathcal{I}}\|^2. \quad (5.95)$$

**Lemma 5.3.13.** For  $x \sim \mathcal{N}(0, \sigma^2 I_p)$

$$\mathbb{E}|x|_\infty^2 \leq O(\sigma^2 \log p)$$

*Proof.* By definition of expectation, we have

$$\begin{aligned} \mathbb{E}|x|_\infty^2 &= \int_0^\infty \Pr[|x|_\infty^2 \geq t] dt = \int_0^{O(\sigma^2 \log p)} \Pr[|x|_\infty^2 \geq t] dt + \int_{O(\sigma^2 \log p)}^\infty \Pr[|x|_\infty^2 \geq t] dt \\ &\leq O(\sigma^2 \log p) + \int_{O(\sigma^2 \log p)}^\infty 2p \exp\left(-\frac{t}{2\sigma^2}\right) dt \\ &\leq O(\sigma^2 \log p) + 2\sqrt{2}p\sigma^2 \exp(-O(\log p)) = O(\sigma^2 \log p). \end{aligned}$$

□

Note that  $\mathbb{E}\langle z_{t,\mathcal{I}}, (x^* - x_t)_\mathcal{I} \rangle = \mathbb{E}\langle z_t, x^* - x_t \rangle = 0$ . Taking the expectation w.r.t  $z_t$  and by the fact that  $\|z_{t,\mathcal{I}}\|^2 \leq |I| |z_t|_\infty^2$  (from the above lemma), we have

$$\mathbb{E}L(x_{t+1}) \leq L(x_t) + \frac{\rho_s}{8\ell_s}(L(x^*) - L(x_t)) + O\left(\frac{\kappa_s k^* G^2 \log \frac{1}{\delta} \log p T}{\rho_s n^2 \epsilon^2}\right). \quad (5.96)$$

That is

$$\mathbb{E}[L(x_{t+1}) - L(x^*)] \leq (1 - \frac{\rho_s}{8\ell_s})\mathbb{E}[L(x_t) - L(x^*)] + O\left(\frac{\kappa_s k^* G^2 \log p \log \frac{1}{\delta} T}{\rho_s n^2 \epsilon^2}\right). \quad (5.97)$$

Thus, taking  $T = O(\kappa_s \log(\frac{n^2}{k^*}))$ , we get the theorem.

### Proof of Theorem 5.3.9

We first show that each stochastic gradient

$$\|x_i^T f'(\langle x_i, \theta_{t-1} \rangle)(f(\langle x_i, \theta_{t-1} \rangle) - y_i)\|_2 \leq O(bC\sqrt{p}),$$

this is due to that

$$\begin{aligned} \|x_i^T f'(\langle x_i, \theta_{t-1} \rangle)(f(\langle x_i, \theta_{t-1} \rangle) - y_i)\|_2 &\leq b\|x_i^T\|_2(f(\langle x_i, \theta_{t-1} \rangle) - y_i) \\ &\leq b\sqrt{p}(f(1) - y_i) \leq O(bC\sqrt{p}), \end{aligned}$$

where the second inequality is due to that  $\langle x_i, \theta_{t-1} \rangle \leq \|x_i\|_\infty \|\theta_{t-1}\|_2 \leq 1$ ,  $f$  is monotone and  $|y_i| = |f(\langle \theta^*, x_i \rangle) + \sigma_i| \leq O(C)$ .

W.o.l.g we assume that each  $|S_t| = \frac{n}{T}$ . From the randomizer  $\mathcal{R}_\epsilon(\cdot)$  and Lemma 5.3.1, we can see that  $\tilde{\nabla}_t = \frac{T}{n} \sum_{i \in S_t} x_i^T f'(\langle x_i, \theta_{t-1} \rangle)(f(\langle x_i, \theta_{t-1} \rangle) - y_i) + \zeta_t$ , where each coordinate of  $\zeta_t$  is a sub-Gaussian vector with  $\sigma^2 = O(\frac{bCpT}{n\epsilon^2})$ .

Let  $\mathcal{S}^* = \text{supp}(\theta^*)$  denote the support of  $\theta^*$ , and  $s^* = |\mathcal{S}^*|$ . Similarly, we define  $\mathcal{S}^t = \text{supp}(\theta_t)$ , and  $\mathcal{F}^{t-1} = \mathcal{S}^{t-1} \cup \mathcal{S}^t \cup \mathcal{S}^*$ . Thus, we have  $|\mathcal{F}^{t-1}| \leq 2s + s^*$ .

We let  $\tilde{\theta}_{t-\frac{1}{2}}$  denote the following

$$\tilde{\theta}_{t-\frac{1}{2}} = \theta_{t-1} - \eta \tilde{\nabla}_{t-1, \mathcal{F}^{t-1}},$$

where  $v_{\mathcal{F}^{t-1}}$  means keeping  $v_i$  for  $i \in \mathcal{F}^{t-1}$  and converting all other terms to 0. By the definition of  $\mathcal{F}^{t-1}$ , we have  $\theta'_t = \text{Trunc}(\tilde{\theta}_{t-\frac{1}{2}}, s)$ . Denote by  $\Delta_t$  the difference of  $\theta_t - \theta^*$ .

We have the following

$$\|\tilde{\theta}_{t-\frac{1}{2}} - \theta^*\|_2 = \|\Delta_{t-1} - \eta([\nabla L_t(\theta_{t-1}) + \zeta_t]_{\mathcal{F}^{t-1}})\|_2,$$

where  $\nabla L_t(\theta_{t-1}) = \frac{T}{n} \sum_{i \in S_t} (f(\langle x_i, \theta_{t-1} \rangle) - y_i) f'(\langle x_i, \theta_{t-1} \rangle) x_i^T$ . Taking  $y_i = \langle x_i, \theta^* \rangle + \sigma_i$  and by the triangle inequality we can get

$$\begin{aligned} \|\tilde{\theta}_{t-\frac{1}{2}} - \theta^*\|_2 &\leq \|\Delta_{t-1} - \eta[\frac{T}{n} \sum_{i \in S_t} (f(\langle x_i, \theta_{t-1} \rangle) - f(\langle x_i, \theta^* \rangle)) f'(\langle x_i, \theta_{t-1} \rangle) x_i^T]_{\mathcal{F}^{t-1}}\|_2 + \\ &\quad \eta \sqrt{|\mathcal{F}^{t-1}|} [\|\frac{T}{n} \sum_{i \in S_t} f'(\langle x_i, \theta_{t-1} \rangle) \sigma_i x_i^T\|_\infty + |\zeta_t|_\infty]. \end{aligned}$$

We denote the followings:

$$A^{t-1} = \|\Delta_{t-1} - \eta \left[ \frac{T}{n} \sum_{i \in S_t} (f(\langle x_i, \theta_{t-1} \rangle) - f(\langle x_i, \theta^* \rangle)) f'(\langle x_i, \theta_{t-1} \rangle) x_i^T \right]_{\mathcal{F}^{t-1}} \|_2 \quad (5.98)$$

$$B^{t-1} = \eta \sqrt{|\mathcal{F}^{t-1}|} \left| \frac{T}{n} \sum_{i \in S_t} f'(\langle x_i, \theta_{t-1} \rangle) \sigma_i x_i^T \right|_\infty \quad (5.99)$$

$$C^{t-1} = \eta \sqrt{|\mathcal{F}^{t-1}|} |\zeta_t|_\infty \quad (5.100)$$

We first bound  $B^{t-1}$ . Since each  $x_i \in \text{Uniform}\{+1, -1\}^p$ , which is sub-Gaussian with 1, we know that for each coordinate  $j \in [p]$ ,  $\frac{T}{n} \sum_{i \in S_t} f'(\langle x_i, \theta_{t-1} \rangle) \sigma_i x_{i,j}$  is sub-Gaussian with  $\sigma^2 = \frac{T^2}{n^2} \sum_{i \in S_t} f'^2(\langle x_i, \theta_{t-1} \rangle) \sigma_i^2 \leq \frac{Tb^2C^2}{n}$ . Thus, by Lemma 5.3.2 we have

$$\Pr \left[ \left| \frac{1}{n} \sum_{i=1}^n f'(\langle x_i, \theta_t \rangle) \sigma_i x_i^T \right|_\infty \leq O \left( \frac{\sqrt{T \log pbC}}{\sqrt{n}} \right) \right] \geq 1 - \frac{1}{p^c}.$$

This means that with probability at least  $1 - \frac{1}{p^c}$ , we have

$$B^t \leq \eta \sqrt{2s + s^*} O \left( \frac{\sqrt{T \log pbC}}{\sqrt{n}} \right). \quad (5.101)$$

For the term  $C^{t-1}$ , by Lemma 5.3.1 and 5.3.2 and since each coordinate  $\zeta_{t,i}$  is sub-Gaussian, we have  $C^{t-1} \leq \eta \sqrt{2s + s^*} O \left( \frac{\sqrt{TpbC \log p}}{\sqrt{n\epsilon^2}} \right)$  with probability at least  $1 - \frac{1}{p^c}$  for some constant  $c > 0$ .

Finally, we bound the term  $A^{t-1}$ . By the mean value theorem, we know that there exists a  $\theta_{t-1,i}$  line between  $\theta_{t-1}$  and  $\theta^*$  which satisfies the equation  $f(\langle x_i, \theta_{t-1} \rangle) - f(\langle x_i, \theta^* \rangle) = f'(\langle x_i, \theta_{t-1,i} \rangle) \langle x_i, \theta_{t-1} - \theta^* \rangle$ . Hence, we have

$$\frac{T}{n} \sum_{i \in S_t} (f(\langle x_i, \theta_{t-1} \rangle) - f(\langle x_i, \theta^* \rangle)) f'(\langle x_i, \theta_{t-1} \rangle) x_i^T = D^{t-1} \Delta_{t-1},$$

where  $D^{t-1} = \frac{T}{n} \sum_{i \in S_t} f'(\langle x_i, \theta_{t-1,i} \rangle) f'(\langle x_i, \theta_{t-1} \rangle) x_i x_i^T \in \mathbb{R}^{p \times p}$ .

Since  $\text{Supp}(D^{t-1}\Delta_{t-1}) \subset \mathcal{F}^{t-1}$  (by assumption), we have

$$A^{t-1} = \|\Delta_{t-1} - \eta D_{\mathcal{F}^{t-1}, \mathcal{F}^{t-1}}^{t-1} \Delta_{t-1}\|_2 \leq \|(I - \eta D_{\mathcal{F}^{t-1}, \mathcal{F}^{t-1}}^{t-1})\|_2 \|\Delta_{t-1}\|_2.$$

Now we bound the term  $\|(I - \eta D_{\mathcal{F}^{t-1}, \mathcal{F}^{t-1}}^{t-1})\|_2$ , where  $I$  is the  $|\mathcal{F}^{t-1}|$ -dimensional identity matrix.

By the RIP property of  $X$  and  $|\mathcal{F}^{t-1}| \leq 2s + s^*$ , we can easily get the following for any  $|\mathcal{F}^{t-1}|$ -dimensional vector  $v$

$$a^2[1 - \Delta(2s + s^*)] \|v\|_2^2 \leq v^T D_{\mathcal{F}^{t-1}, \mathcal{F}^{t-1}}^{t-1} v \leq b^2[1 + \Delta(2s + s^*)].$$

Thus,  $\|(I - \eta D_{\mathcal{F}^{t-1}, \mathcal{F}^{t-1}}^{t-1})\|_2 \leq \max\{1 - \eta a^2[1 - \Delta(2s + s^*)], \eta b^2[1 + \Delta(2s + s^*)] - 1\}$ .

This means that if we can find an  $\eta$  satisfying the condition of

$$\frac{5}{7} \frac{1}{a[1 - \Delta(2s + s^*)]} \leq \eta \leq \frac{9}{7} \frac{1}{b^2[1 + \Delta(2s + s^*)]},$$

then we have  $\|(I - \eta D_{\mathcal{F}^{t-1}, \mathcal{F}^{t-1}}^{t-1})\|_2 \leq \frac{2}{7}$ . Note that such an  $\eta$  can indeed be found if  $\Delta(2s + s^*) \leq \frac{5a^2 - 9b^2}{14}$ . This means that  $\frac{a}{b} > \frac{\sqrt{5}}{3}$ .

Thus, in total we have the following with probability at least  $1 - \frac{2}{p^c}$

$$\|\tilde{\theta}_{t-\frac{1}{2}} - \theta^*\|_2 \leq \frac{2}{7} \|\Delta_{t-1}\|_2 + O\left(\frac{\sqrt{Tp(2s + s^*) \log p} b C}{\sqrt{n} \epsilon}\right).$$

Our next task is to bound  $\|\theta'_t - \theta^*\|_2$  by  $\|\tilde{\theta}_{t-\frac{1}{2}} - \theta^*\|_2$  by Lemma 5.3.3.

Thus, we have  $\|\theta'_t - \tilde{\theta}_{t-\frac{1}{2}}\|_2^2 \leq \frac{|\mathcal{F}^{t-1}| - s}{|\mathcal{F}^{t-1}| - s^*} \|\tilde{\theta}_{t-\frac{1}{2}} - \theta^*\|_2^2 \leq \frac{s + s^*}{2s} \|\tilde{\theta}_{t-\frac{1}{2}} - \theta^*\|_2^2$ .

Taking  $s = 8s^*$ , we get

$$\|\theta'_t - \tilde{\theta}_{t-\frac{1}{2}}\|_2 \leq \frac{3}{4} \|\tilde{\theta}_{t-\frac{1}{2}} - \theta^*\|_2$$

and

$$\|\theta'_t - \theta^*\|_2 \leq \frac{7}{4} \|\tilde{\theta}_{t-\frac{1}{2}} - \theta^*\|_2 \leq \frac{1}{2} \|\Delta_{t-1}\|_2 + O\left(\frac{\sqrt{Tps^* \log pbC}}{\sqrt{n}\epsilon}\right).$$

Finally, we need to show that  $\|\Delta_t\|_2 = \|\theta_t - \theta^*\|_2 \leq \|\theta'_t - \theta^*\|_2$ , which is due to the Lemma 5.3.4.

Putting all together, we have the following with probability at least  $1 - \frac{2}{p^c}$ ,

$$\|\Delta_t\| \leq \frac{1}{2} \|\Delta_{t-1}\|_2 + O\left(\frac{\sqrt{Tps^* \log pbC}}{\sqrt{n}\epsilon}\right).$$

Thus, we get with probability at least  $1 - \frac{2T}{p^c}$ ,

$$\|\Delta_T\|_2 \leq \left(\frac{1}{2}\right)^T \|\theta^*\|_2 + O\left(\frac{\sqrt{Tps^* \log pbC}}{\sqrt{n}\epsilon}\right).$$

### Proof of Theorem 5.3.10

Our proof is inspired by the ones in [97, 350] and [244]. Since it is reduced to the linear model when  $f(x) \equiv x$ , we only need to consider the general case. Similar to the proof of Theorem 1, we first construct a packing set  $\{P_v : v \in \mathcal{V}\}$  and then bound  $\mathcal{C}_\infty(\{P_v\})$ . To do so, we need the following lemma.

**Lemma 5.3.14.** [[244]] For any  $s \in [p]$ , define the set

$$\mathcal{H}(s) := \{z \in \{-1, 0, +1\}^d \mid \|z\|_0 = s\}$$

with Hamming distance  $\rho_H(z, z') = \sum_{i=1}^d 1[z_j \neq z'_j]$  between the vectors  $z$  and  $z'$ . Then, there exists a subset  $\tilde{\mathcal{H}} \subset \mathcal{H}$  with cardinality  $|\tilde{\mathcal{H}}| \geq \exp\left(\frac{s}{2} \log \frac{p-s}{s/2}\right)$  such that  $\rho_H(z, z') \geq \frac{s}{2}$  for all  $z, z' \in \tilde{\mathcal{H}}$ .

Now consider the rescaled version of  $\tilde{\mathcal{H}}$ ,  $\sqrt{\frac{2}{\delta}} \tilde{\mathcal{H}}$ , for some  $\delta \leq \frac{1}{\sqrt{2}}$ . For any two  $\theta, \theta' \in \tilde{\mathcal{H}}$ ,

we have

$$8\delta^2 \geq \|\theta - \theta'\|_2^2 \geq \delta^2. \quad (5.102)$$

Then,  $\sqrt{\frac{2}{\delta}}\tilde{\mathcal{H}}$  is a  $\delta$  packing in  $\ell_2$  norm with  $M = |\tilde{\mathcal{H}}|$  elements, denoted as  $\{\theta_1, \theta_2, \dots, \theta_M\}$ .

For each  $\theta_i$ , let  $\sigma_i$  denote the uniform distribution on the interval  $[-C, C]$ . Thus, we have  $P_{\theta_i}$ , which can be easily verified that  $P_{\theta_i} \in \mathcal{P}'_{s,p,C,f,a,b}$ .

Our idea is to use Lemma 5.3.8. Thus, our goal is to bound the sum of the Total Variance  $\sum_{v,v' \in \mathcal{V}} \|P_{v,i} - P_{v',i}\|_{TV}^2$ . Now consider the case of  $P_{\theta,i}$  and  $P_{\theta',i}$ , where (due to our construction)  $P_{\theta,i}$  is the uniform distribution on the interval of  $[f(\langle x_i, \theta \rangle) - C, f(\langle x_i, \theta \rangle) + C]$ . Thus, we have

$$\begin{aligned} \|P_{\theta,i} - P_{\theta',i}\|_{TV} &= \frac{1}{2} \int |p_{\theta,i}(y) - p_{\theta',i}(y)| dy \\ &\leq \frac{1}{2C} |f(\langle \theta, x_i \rangle) - f(\langle \theta', x_i \rangle)| \leq \frac{b}{2C} |\langle \theta - \theta', x_i \rangle|, \end{aligned}$$

where the last inequality is due to the assumption on  $f$ . Hence, we have

$$\begin{aligned} \sum_{i=1}^n \frac{1}{|\mathcal{V}|^2} \sum_{v,v' \in \mathcal{V}} \|P_{v,i} - P_{v',i}\|_{TV}^2 &\leq \sum_{i=1}^n \frac{b^2}{4C^2} \sum_{v,v \in \mathcal{V}} (\theta_v - \theta_{v'})^T x_i x_i^T (\theta_v - \theta_{v'}) \\ &= \frac{b^2}{4C^2} \frac{1}{|\mathcal{V}|^2} \sum_{v,v \in \mathcal{V}} (\theta_v - \theta_{v'}) X^T X (\theta_v - \theta_{v'}) \\ &\leq 8 \frac{b^2(1 + \Delta)}{4C^2} \delta^2 = \frac{2b^2(1 + \Delta)\delta^2}{C^2}, \end{aligned}$$

where the last inequality is due to the fact that for every pair  $(v, v')$  with  $\|\theta_v - \theta_{v'}\|_0 \leq 2s$ ,  $(\theta_v - \theta_{v'}) X^T X (\theta_v - \theta_{v'}) \leq n(1 + \Delta)$  holds (by Assumption 1).

Thus by Lemmas 5.3.14 and 5.3.8, we have

$$\frac{\Phi(\delta)}{2} \geq \frac{\delta^2}{8} \left( 1 - \frac{2cn\epsilon^2 \delta^2 \frac{b^2(1+\Delta)}{C^2} + \log 2}{\frac{s}{2} \log \frac{p-s}{s/2}} \right).$$

Taking  $\delta^2 = \Omega(\min\{1, \frac{s \log p / sC^2}{(1+\Delta)b^2 n \epsilon^2}\})$ , we get the result.

### Proof of Theorem 5.3.11

For the guarantee of  $(\epsilon, \delta)$  locally differentially private, it is due to the fact that  $x_i$  is known and each  $y_i \in [\langle x_i, \theta^* \rangle - C, \langle x_i, \theta^* \rangle + C]$  (since the random noise  $\sigma_i$  is bounded by  $C$ ). Thus, by the Gaussian Mechanism [107], we can see that it is locally differentially private.

Now we prove Theorem the upper bound.

Let  $\mathcal{S}^* = \text{supp}(\theta^*)$  denote the support of  $\theta^*$ , and  $s^* = |\mathcal{S}^*|$ . Similarly, we define  $\mathcal{S}^{t+1} = \text{supp}(\theta_{t+1})$ , and  $\mathcal{F}^t = \mathcal{S}^t \cup \mathcal{S}^{t+1} \cup \mathcal{S}^*$ . Thus, we have  $|\mathcal{F}^t| \leq 2s + s^*$ .

We let  $\tilde{\theta}_{t+\frac{1}{2}}$  denote the following

$$\tilde{\theta}_{t+\frac{1}{2}} = \theta_t - \eta \nabla_{\mathcal{F}^t} L(\theta_t),$$

where  $v_{\mathcal{F}^t}$  means keeping  $v_i$  for  $i \in \mathcal{F}^t$  and making all other terms 0. By the definition of  $\mathcal{F}^t$ , we have  $\theta'_{t+1} = \text{Trunc}(\tilde{\theta}_{t+\frac{1}{2}}, s)$ . Denote by  $\Delta_{t+1}$  the difference of  $\theta_{t+1} - \theta^*$ . We have the following

$$\|\tilde{\theta}_{t+\frac{1}{2}} - \theta^*\|_2 = \|\Delta_t - \eta \nabla_{\mathcal{F}^t} L(\theta_t)\|_2,$$

where  $\nabla_{\mathcal{F}^t} L(\theta_t) = [\frac{1}{n} \sum_{i=1}^n (f(\langle x_i, \theta_t \rangle) - \tilde{y}_i) f'(\langle x_i, \theta_t \rangle) x_i^T]_{\mathcal{F}^t}$ . Plugging  $\tilde{y}_i = f(\langle \theta^*, x_i \rangle) + \sigma_i + z_i$ , where  $z_i \sim \mathcal{N}(0, \tau^2)$ , and  $\tau^2 = \frac{32C^2 \log(1.25/\delta)}{\epsilon^2}$  into the above equality, we get

$$\begin{aligned} \|\tilde{\theta}_{t+\frac{1}{2}} - \theta^*\|_2 &\leq \|\Delta_t - \eta [\frac{1}{n} \sum_{i=1}^n (f(\langle x_i, \theta_t \rangle) - f(\langle x_i, \theta^* \rangle)) f'(\langle x_i, \theta_t \rangle) x_i^T]_{\mathcal{F}^t}\|_2 + \\ &\quad \eta \sqrt{|\mathcal{F}^t|} [|\frac{1}{n} \sum_{i=1}^n f'(\langle x_i, \theta_t \rangle) \sigma_i x_i^T|_\infty + |\frac{1}{n} \sum_{i=1}^n f'(\langle x_i, \theta_t \rangle) z_i x_i^T|_\infty]. \end{aligned}$$

Define the following terms

$$\begin{aligned} A^t &= \|\Delta_t - \eta \left[ \frac{1}{n} \sum_{i=1}^n (f(\langle x_i, \theta_t \rangle) - f(\langle x_i, \theta^* \rangle)) f'(\langle x_i, \theta_t \rangle) x_i^T \right]_{\mathcal{F}^t} \|_2 \\ B^t &= \eta \sqrt{|\mathcal{F}^t|} \left| \frac{1}{n} \sum_{i=1}^n f'(\langle x_i, \theta_t \rangle) \sigma_i x_i^T \right|_\infty, \\ C^t &= \eta \sqrt{|\mathcal{F}^t|} \left| \frac{1}{n} \sum_{i=1}^n f'(\langle x_i, \theta_t \rangle) z_i x_i^T \right|_\infty. \end{aligned}$$

We first bound  $B^t$ . Since each  $x_i \in \text{Uniform}\{+1, -1\}^p$ , which is sub-Gaussian with 1, we know that for each coordinate  $j \in [p]$ ,  $\frac{1}{n} \sum_{i=1}^n f'(\langle x_i, \theta_t \rangle) \sigma_i x_{i,j}$  is sub-Gaussian with  $\sigma^2 = \frac{1}{n^2} \sum_{i=1}^n f'^2(\langle x_i, \theta_t \rangle) \sigma_i^2 \leq \frac{b^2 C^2}{n}$ . Thus, by Lemma 5.3.2 we have

$$\Pr \left[ \left| \frac{1}{n} \sum_{i=1}^n f'(\langle x_i, \theta_t \rangle) \sigma_i x_i^T \right|_\infty \leq O \left( \frac{\sqrt{\log p} b C}{\sqrt{n}} \right) \right] \geq 1 - \frac{1}{p^c}.$$

This means that with probability at least  $1 - \frac{2}{p^c}$ , we have

$$B^t \leq O \left( \eta \sqrt{2s + s^*} \frac{\sqrt{\log p} b C}{\sqrt{n}} \right). \quad (5.103)$$

Similarly, for  $C^t$  we have that with probability at least  $1 - \frac{1}{p^c}$ , the following holds

$$\left| \frac{1}{n} \sum_{i=1}^n f'(\langle x_i, \theta_t \rangle) z_i x_i^T \right|_\infty \leq O \left( \frac{b \sqrt{\log p} \sqrt{\sum_{i=1}^n z_i^2}}{n} \right).$$

Since  $z_i$  is Gaussian with variance  $\tau^2$ , we know that  $\sum_{i=1}^n z_i^2 = \tau^2 \sum_{i=1}^n r_i^2$ , where  $\sum_{i=1}^n r_i^2$  is a  $\chi^2$ -distribution with parameter  $n$ .

By the above concentration bound for  $\chi^2$ -distribution and Lemma 5.3.6, we have  $\sum_{i=1}^n z_i^2 \leq 5\tau^2 n$  with probability at least  $1 - \exp(-n)$ . Thus,

$$C^t \leq \eta \sqrt{2s + s^*} O \left( \frac{b \sqrt{\log p} \tau}{\sqrt{n}} \right) \quad (5.104)$$

with probability at least  $1 - \frac{1}{p^c} - \exp(-n)$ .

For the term of  $A^t$ , the proof is the same as the one for  $A^{t-1}$  in the proof of Theorem 5.3.9, and thus we omit it from here.

By (5.103) and (5.104) and plugging  $\tau^2 = \frac{32C^2 \log(1.25/\delta)}{\epsilon^2}$  into (5.104), we have the following with probability at least  $1 - \frac{2}{p^c} - \exp(-n)$

$$\|\tilde{\theta}_{t+\frac{1}{2}} - \theta^*\|_2 \leq \frac{2}{7} \|\Delta_t\|_2 + O\left(\frac{\sqrt{(2s+s^*) \log p} \log(1/\delta) b C}{n\epsilon}\right).$$

Putting all together, we have the following with probability at least  $1 - \frac{2}{p^c} - \exp(-n)$ ,

$$\|\Delta_{t+1}\| \leq \frac{1}{2} \|\Delta_t\|_2 + O\left(\frac{\sqrt{s^* \log p} \log(1/\delta) b C}{n\epsilon}\right).$$

Thus, we get the bound in Theorem 5.3.11 with probability at least  $1 - \frac{2T}{p} - T \exp(-n)$ . For the linear case, since  $f' \equiv 1$ , (5.103) and (5.104) will be the same in each iteration, the probability for the linear case becomes  $1 - \frac{2}{p^c} - \exp(-n)$ .

# **Chapter 6**

## **Some Matrix Estimation Problems in Differential Privacy Model**

### **6.1 Principal Component Analysis in Local Differential Privacy Model**

Principal Component Analysis (PCA) is a fundamental technique for dimension reduction in statistics, machine learning, and signal processing. As of today, it remains as one of the most commonly used tools in applications, especially in social sciences [79], financial econometrics [6], medicine [23], and genomics [208].

With the rapid development of information technologies, big data now ubiquitously exist in our daily life, which need to be analyzed (or learned) statistically by methods like regression and PCA. However, due to the presence of sensitive data (especially those in social science, biomedicine and genomics) and their distributed nature, such data are extremely difficult to aggregate and learn from. Consider a case where health records are scattered across multiple hospitals (or even countries), it is challenging to process the whole dataset in a central server due to privacy and ownership concerns. A better solution is to use some differentially private mechanisms to conduct the aggregation and learning tasks. .

In the local model, two basic types of protocols are often used: interactive and non-interactive. [257] have recently investigated the power of non-interactive differentially private protocols. This type of protocols is more natural for the classical use cases of the local model: both projects from Google and Apple use the non-interactive model. Moreover, implementing efficient interactive protocols in such applications is more difficult due to the latency of the network. Despite being used in industry, the local model has been much less studied than the central one. Part of the reason for this is that there are intrinsic limitations in what one can do in the local model. As a consequence, many basic questions, that are well studied in the central model, have not been completely understood in the local model, yet.

In this Chapter, I study PCA under the non-interactive local differential privacy model and aim to answer the following main question.

**What are the limitations and the (near) optimal algorithms of PCA under the non-interactive local differential privacy model?**

We summarize our main contributions as follows:

1. We first study the  $k$ -subspace PCA problem in the low dimensional setting and show that the minimax risk (measured by the squared subspace distance) under  $\epsilon$  non-interactive local differential privacy (LDP) is lower bounded by  $\Omega\left(\frac{\lambda_1 \lambda_{k+1} p k}{(\lambda_k - \lambda_{k+1})^2 n \epsilon^2}\right)$ , where  $p$  is the dimensionality of the data and  $n$  is the number of data records,  $\lambda_1, \lambda_k$  and  $\lambda_{k+1}$  is the 1st,  $k$ -th and  $(k+1)$ -th eigenvalue of the population covariance matrix  $\Sigma$ , respectively. Moreover, we prove that the term  $\Omega\left(\frac{p k}{n \epsilon^2}\right)$  is optimal by showing that there is an  $(\epsilon, \delta)$ -LDP whose upper bound is  $O\left(\frac{\lambda_1^2 k p \log(1/\delta)}{(\lambda_k - \lambda_{k+1})^2 n \epsilon^2}\right)$ .
2. An undesirable issue of the above result is that the error bound could be too large in high dimensions (*i.e.*,  $p \gg n$ ). In such scenarios, a natural approach is to impose some additional structural constraints on the leading eigenvectors. A commonly used constraint is to assume that the leading eigenvectors are row sparse, which is referred as sparse PCA in the literature and has been studied intensively in recent years [292,

58, 293]. Thus, for the high dimensional case, we consider the sparse PCA under the non-interactive local model and show that the private minimax risk (measured by the squared subspace distance) is lower bounded by  $\Omega\left(\frac{\lambda_1 \lambda_{k+1}}{(\lambda_k - \lambda_{k+1})^2} \frac{ks \log p}{n\epsilon^2}\right)$ , where  $s$  is the sparsity parameter of the underlying subspace. We also give an algorithm to achieve a near optimal upper bound of  $O\left(\frac{\lambda_1^2}{(\lambda_k - \lambda_{k+1})^2} \frac{s^2 \log p}{n\epsilon^2}\right)$ . With additional assumptions on the correlation of the population covariance matrix, we further show that our private estimator is sparsistency, *i.e.* it recovers the support of the underlying parameter.

3. Finally, we provide an experimental study for our proposed algorithms on both synthetic and real world datasets, and show that the experimental results support our theoretical analysis.

### 6.1.1 Related Work

There is a vast number of papers studying PCA under differential privacy, starting from the SULQ framework [42], [106, 69, 162, 132, 122, 21]. We compare only those private PCA results in distributed settings.

For the low dimensional case, Balcan *et al.* [21] studied the private PCA problem under the interactive local differential privacy model and introduced an approach based on the noisy power method. They showed an upper bound which is suitable for general settings, while ours is mainly for statistical settings. It is worth pointing out that the output in [21] is only an  $O(k)$ -dimensional subspace, instead of an exact  $k$ -dimensional subspace; thus their result is incomparable with ours. Moreover, we provide, in this paper, a lower bound on the  $\epsilon$  non-interactive private minimax risk.

For the private high dimensional sparse PCA, the work most closely related to ours is the one by Ge *et al.* [122]. The authors in this paper proposed a noisy iterative hard thresholding power method, which is an interactive LDP algorithm and proved an upper bound of  $O\left(\frac{\lambda_1 \lambda_k}{(\lambda_k - \lambda_{k+1})^2} \frac{s(k + \log p)}{n(1 - \rho^{\frac{1}{4}})}\right)$  for their method, where  $\rho$  is a parameter related to  $\epsilon$ . Specifically, they showed that there exists some 'Privacy Free Region'. However, several

things need to be pointed out. Firstly, our method is for general  $\epsilon \in (0, 1]$  and non-interactive settings, while Ge *et al.* considered the interactive setting with more restricted  $\epsilon$ . Secondly, the assumptions in our paper are less strict than the ones in [122]. Finally, we provide a lower bound on the private minimax risk.

The optimal procedure in our paper is based on perturbing the covariance by Gaussian matrices, which has been studied in [106]. However, there are some major differences; firstly, we show the optimality of our algorithm under the non-interactive local model using subspace distance as the measurement, while [106] showed the optimality under the  $(\epsilon, \delta)$  central model using variance as the measurement. It is notable that in [106] the authors also provided an upper bound on the subspace distance. However, the lower bound is still unknown. Secondly, while the optimal algorithm for the low dimensional case is quite similar, we extend it to the high dimensional case. The optimal procedure in the high dimensional sparse case is quite different from that in [106]. Thirdly, in this paper, since we focus the statistical setting while [106] considered the general setting, the upper bound results are incomparable.

### 6.1.2 Preliminaries

Let  $X \in \mathbb{R}^p$  a random vector with mean 0 and covariance matrix  $\Sigma$ .  $k$ -dimensional PCA is to find a  $k$  dimensional subspace that optimizes the following problem:

$$\min \mathbb{E}\|(I_p - \Pi_{\mathcal{G}})X\|_2^2, \text{ s.t. } \mathcal{G} \in \mathbb{G}_{p,k},$$

where  $\mathbb{G}_{p,k}$  is the Grassmann manifold of  $k$ -dimensional subspaces of  $\mathbb{R}^p$ , and  $\Pi_{\mathcal{G}}$  is the projection of  $\mathcal{G}$ . There always exists at least one solution; consider  $\Sigma = \sum_{j=1}^p \lambda_j v_j v_j^T$ , where  $\lambda_1 \geq \lambda_2 \geq \dots, \lambda_p \geq 0$  are the eigenvalues of  $\Sigma$  and  $v_1, v_2, \dots, v_p \in \mathbb{R}^p$  are the corresponding eigenvectors. If  $\lambda_k \geq \lambda_{k+1}$ , then the  $k$ -dimensional principal subspace of  $\Sigma$ , *i.e.* the subspace  $\mathcal{S}$  spanned by  $v_1, \dots, v_k$  solves the above optimization problem, where the

orthogonal projector of  $\mathcal{S}$  is given by  $\Pi_{\mathcal{S}} = V_k V_k^T$ , where  $V_k = [v_1, \dots, v_k] \in \mathbb{V}_{p,k}$ ,  $\mathbb{V}_{p,k}$  is the set of all  $p \times k$  orthogonal matrices. For simplicity we denote  $\mathcal{S} = \text{col}(V_k)$ , where  $\text{col}(M)$  denotes the subspace spanned by the columns vectors of  $M$ .

**PCA under the non-interactive local model** In practice,  $\Sigma$  is unknown, and the only thing that we have is the set of observation data records  $\{X_1, \dots, X_n\}$ , which are i.i.d sampled from  $X$ . Thus, the problem of (non-interactively) locally differentially private PCA is to find a  $k$ -dimensional subspace  $\mathcal{S}^{\text{priv}}$  which is close to  $\mathcal{S}$ , where the algorithm that outputs  $\mathcal{S}^{\text{priv}}$  must be  $\epsilon$  (non-interactively) locally differentially private.

After obtaining a private estimator  $\mathcal{S}^{\text{priv}}$ , there are multiple ways to measure the success, such as variance guarantee [106], low rank approximation error [174], etc. In this paper, we will use the subspace distance as the measurement [106, 122].

**Subspace distance** Let  $\mathcal{S}$  and  $\mathcal{S}'$  be two  $k$ -dimensional subspaces in  $\mathbb{R}^p$ . Also denote by  $E$  and  $F$ , respectively, the orthogonal matrix corresponds to  $\mathcal{S}$  and  $\mathcal{S}'$ . That is,  $E = VV^T$  and  $F = WW^T$  for some orthogonal matrices  $V \in \mathbb{V}_{p,k}$  and  $W \in \mathbb{V}_{p,k}$ . Then, the squared subspace distance between  $\mathcal{S}$  and  $\mathcal{S}'$  is defined by the following [264]:

$$\|\sin \Theta(\mathcal{S}, \mathcal{S}')\|_F^2 = \|E - F\|_F^2 = \frac{1}{2} \|VV^T - WW^T\|_F^2,$$

where  $\|\cdot\|_F$  is the Frobenious norm. For simplicity, we will overload notation and write  $\sin \Theta(\mathcal{S}, \mathcal{S}') = \sin \Theta(V, W)$ .

### 6.1.3 Low Dimensional Case

In this section, we focus on the general case and always assume  $n \geq p$ . We first derive a lower bound of the  $\epsilon$  non-interactive private minimax risk using the squared subspace distance as the measurement. By the definition of the  $\epsilon$ -private minimax risk, it is important to select an appropriate class of distributions.

## Class of Distributions

1. We assume that the random vector  $X$  is sub-Gaussian, that is  $X = \Sigma^{\frac{1}{2}}Z$ , where  $Z \in \mathbb{R}^p$  is some random vector satisfying equations  $\mathbb{E}Z = 0$ ,  $\text{Var}(Z) = I_p$  and its sub-Gaussian norm  $\|Z\|_{\psi_2} \leq 1$ , where

$$\|Z\|_{\psi_2} := \sup_{v: \|v\|_2 \leq 1} \inf \{C > 0, \mathbb{E} \exp \left| \frac{\langle Z, v \rangle}{C} \right|^2 \leq 2\},$$

which means that all the one-dimensional marginals of  $X_i$  have sub-Gaussian tails. We need to note that this assumption on  $X$  is commonly used in many papers on PCA in statistical settings, such as [292, 122].

2. In the study of private PCA, it is always assumed that the  $\ell_2$  norm of each  $X_i$  is bounded by 1, as in [106][122]. For convenience, we relax this assumption in the following way; for the random vector  $X \in \mathbb{R}^p$ , we assume that  $\|X\|_2 \leq 1$  with a probability at least  $1 - e^{-\Omega(p)}$ .
3. Next, we give assumptions on the population covariance matrix  $\Sigma$ . Firstly, we assume that for the target  $k$ -dimensional subspace,  $\lambda_k - \lambda_{k+1} > 0$  so that the principal subspace is well defined. Next, we define the effective noise variance  $\sigma_k^2$ , which is proposed in [292] and [58]:

$$\sigma_k^2(\lambda_1, \lambda_2, \dots, \lambda_p) := \frac{\lambda_1 \lambda_{k+1}}{(\lambda_k - \lambda_{k+1})^2}. \quad (6.1)$$

For a given constant  $\sigma^2 > 0$ , we assume that  $\sigma_k^2 \leq \sigma^2$ .

We denote the collection of distributions which satisfy the previous conditions 1), 2) and 3) as  $\mathcal{P}(k, \sigma^2)$

## Main Results

The next theorem shows a lower bound of  $\epsilon$  non-interactive private minimax risk under squared subspace distance.

**Theorem 6.1.1.** Let  $\{X_i\}_{i=1}^n$  be samples from  $P \in \mathcal{P}(k, \sigma^2)$ . If  $\frac{p}{4} \leq k \leq \frac{3p}{4}$ ,  $\epsilon \in (0, \frac{1}{2}]$  and  $n \geq \Omega\left(\frac{1}{\epsilon^2} \frac{\lambda_1 \lambda_{k+1}}{(\lambda_k - \lambda_{k+1})^2} \min\{k, p - k\}\right)$ , then the  $\epsilon$  non-interactive private minimax risk in the metric of squared subspace distance satisfies:

$$\mathcal{M}_n^{\text{Nint}}(\mathcal{S}(\mathcal{P}(k, \sigma^2)), \Phi \circ \rho, \epsilon) \geq \Omega\left(\sigma^2 \frac{kp}{n\epsilon^2}\right).$$

**Remark 6.1.1.** We note that for the non-private case, the minimax risk is lower bounded by  $\Omega\left(\frac{\lambda_1 \lambda_{k+1}}{(\lambda_k - \lambda_{k+1})^2} \frac{kp}{n}\right)$  [58]. Thus, in this case, the impact of the local differential privacy is to change the number of efficient sample from  $n$  to  $n\epsilon^2$ . However, the collection of the considered distributions needs another assumption which says that  $\|X\|_2$  is bounded by 1 with high probability. This is not necessary in the non-private case [58], but needed in ours for showing the upper bound.

We also note that although Theorem 6.1.1 holds only for  $k = \Theta(p)$ , while in practice  $k$  always be a constant. As we can see from Section 6.1.6, the lower bound holds for all  $k$  if we relax the condition 2) in our collection of distributions  $\mathcal{P}(\sigma^2, k)$ . It is the same for the high dimensional sparse case.

Finally, note that in the central differential privacy model, [106] showed that the lower bound of the  $k$ -dimensional PCA is  $\tilde{\Omega}\left(\frac{kp \log(\frac{1}{\delta})}{n^2 \epsilon^2}\right)$  for  $(\epsilon, \delta)$ -differential privacy. However, this lower bound is measured by the variance of  $X = (X_1^T, X_2^T, \dots, X_n^T)^T \in \mathbb{R}^{n \times p}$ , not the squared subspace distance used in this paper. Although [106] gave an upper bound of  $O\left(\frac{kp \log(1/\delta)}{(\lambda_k^2 - \lambda_{k+1}^2)n^2 \epsilon^2}\right)$  in the general setting using the squared subspace distance as measurement, it is still unknown whether the bound is optimal. Also, their lower bound omits the parameters related to the eigenvalues. For the  $\epsilon$  differential privacy in the central model, [69] showed that the lower bound is  $\Omega\left(\frac{p^2}{n^2 \epsilon^2 (\lambda_1 - \lambda_2)^2}\right)$  in the special case of  $k = 1$ . However, it is

still unknown for the general case of  $k$ . Thus, from the above discussion, we can see that the lower bound of  $\epsilon$  non-interactively locally differentially private PCA is similar to the  $(\epsilon, \delta)$  differentially private PCA in the central model.

One of the main questions is whether the lower bound in Theorem 6.1.1 is tight. In the following, we show that the term  $\Omega(\frac{pk}{n\epsilon^2})$  is tight. By our definition of the parameter space, we know that for any  $X \sim P \in \mathcal{P}(\sigma^2, k)$ ,  $\|X\|_2 \leq 1$  with high probability. Thus, we always assume that the event of each  $\|X_i\|_2 \leq 1$  holds. Note that this assumption also appears in [122, 106, 21]. The idea is the same as in [106], where each  $X_i$  perturbs its covariance and aggregates the noisy version of covariance, see Algorithm 6.1.46 for details.

**Theorem 6.1.2.** For any  $\epsilon, \delta > 0$ , Algorithm 6.1.46 is  $(\epsilon, \delta)$  (non-interactively) locally differentially private. Furthermore, with probability at least  $1 - e^{-C_1 p} - \frac{1}{p^{C_2}}$ , the output satisfies:

$$\|\sin \Theta(\tilde{V}_k, V_k)\|_F^2 \leq O\left(\frac{\lambda_1^2 kp \log(1/\delta)}{(\lambda_k - \lambda_{k+1})^2 n \epsilon^2}\right), \quad (6.2)$$

where  $C_1, C_2$  are some universal constants.

---

**Algorithm 6.1.46** Local Gaussian Mechanism

**Input:** data records  $\{X_i\}_{i=1}^n \sim P^n$  for  $P \in \mathcal{P}(\sigma^2, k)$ , and for  $i \in [n]$ ,  $\|X_i\|_2 \leq 1$ .  $\epsilon, \delta$  are the privacy parameters.

- 1: **for** Each  $i \in [n]$  **do**
  - 2:     Denote  $\tilde{X}_i \tilde{X}_i^T = X_i X_i^T + Z_i$ , where  $Z_i \in \mathbb{R}^{p \times p}$  is a symmetric matrix where the upper triangle (including the diagonal) is i.i.d samples from  $\mathcal{N}(0, \sigma_1^2)$ ; here  $\sigma_1^2 = \frac{2 \ln(1.25/\delta)}{\epsilon^2}$ , and each lower triangle entry is copied from its upper triangle counterpart.
  - 3: **end for**
  - 4: Compute  $\tilde{S} = \frac{1}{n} \sum_{i=1}^n \tilde{X}_i \tilde{X}_i^T$ .
  - 5: Output  $\text{col}(\tilde{V}_k)$  where  $\tilde{V}_k \in \mathbb{R}^{p \times k}$  is the principal  $k$ -subspace of  $\tilde{S}$ .
- 

In Theorem 7 of [106], the authors provided a similar upper bound for the  $(\epsilon, \delta)$ -differential privacy in the central model. However, they need to assume that the eigenvalues satisfy the condition  $\lambda_k^2 - \lambda_{k+1}^2 = \omega(\sqrt{p})$ , which is not needed in our Theorem 6.1.2 where we use some recent result on Davis-Khan theorem (see Section 6.1.6 for details).

From the analysis, we can see that, to ensure non-interactive LDP, here we should add a randomized matrix to the covariance matrix, and this will cause an additional factor of  $O(\frac{1}{\epsilon^2})$  in the error compared with the non-private case.

From Theorems 6.1.1 and 6.1.2, we can see that there is still a gap of  $O(\frac{\lambda_1}{\lambda_{k+1}})$  between the lower and upper bounds. We leave it as an open problem to determine whether these bounds are tight or not.

### 6.1.4 High Dimensional Sparse Case

From Theorem 6.1.1, we can see that for the high dimensional case, *i.e.*  $p \gg n$ , the bound in (6.2) becomes trivial. Thus, to avoid this issue, we need some additional assumption on the parameter space. One of the commonly used assumption is sparsity. There are many definitions of sparsity on PCA and we use the row sparsity in this paper, which has also been studied in [292, 58, 122].

We first define the  $(p, q)$ -norm of a  $p \times k$  matrix  $A$  as the usual  $\ell_q$  norm of the vector of row-wise  $\ell_p$  norms of  $A$ :

$$\|A\|_{p,q} := \|(\|a_{1*}\|_p, \|a_{2*}\|_p, \dots, \|a_{p*}\|_p)\|_q, \quad (6.3)$$

where  $a_{j*}$  denotes the  $j$ -th row of  $A$ . Note that  $\|\cdot\|_{2,0}$  is coordinate independent, *i.e.*  $\|AO\|_{2,0} = \|A\|_{2,0}$  for any orthogonal matrix  $O \in \mathbb{R}^{k \times k}$ . We define the row sparse space as follows.

**Definition 6.1.1.** Let  $s$  be the sparsity level parameter satisfying the condition of  $k \leq s \leq p$ . The  $s$ -(row) sparse subspace is defined as follows

$$\mathcal{M}_0(s) = \{\text{col}(U), U \in \mathbb{R}^{p \times k} \text{ and orthogonal}, \|U\|_{2,0} \leq s\}.$$

We define our parameter space,  $\mathcal{P}(s, k, \sigma^2)$ , to be the same as in the previous section with

an additional condition that  $\mathcal{S} \in \mathcal{M}_0(s)$ , where  $\mathcal{S}$  is the  $k$ -dimensional principal subspace of covariance matrix  $\Sigma$ .

Below, we will first derive a lower bound of the non-interactive locally differentially private PCA in the high dimensional sparse case.

**Theorem 6.1.3.** Let  $\{X_i\}_{i=1}^n$  be the observations sampled from a distribution  $P \in \mathcal{P}(s, k, \sigma^2)$ . If the privacy parameter  $\epsilon \in (0, \frac{1}{2}]$ ,  $n \geq \Omega((s - k) \frac{\sigma^2(k + \log p)}{\epsilon^2})$ . Then for all  $k \in [p]$  satisfying the condition of  $2k \leq s - k \leq p - k$  and  $\frac{p}{4} \leq k \leq \frac{3p}{4}$ , the  $\epsilon$  non-interactive private minimax risk in the metric of squared subspace distance satisfies the following

$$\mathcal{M}_n^{\text{Nint}}(\mathcal{S}(\mathcal{P}(s, k, \sigma^2), \epsilon)) \geq \Omega\left(\sigma^2 \frac{s(k + \log p)}{n\epsilon^2}\right).$$

Note that in the non-private case, the optimal minimax risk is  $\Theta\left(\sigma^2 \frac{s(k + \log p)}{n}\right)$ . Thus, same as in the low dimensional case, the impact of the privacy constraint is to change the efficient samples from  $n$  to  $n\epsilon^2$ .

Next, we consider the upper bound. In the non-private case, the optimal procedure is to solve the following NP-hard optimization problem [292]:

$$\begin{aligned} & \max \langle S, UU^T \rangle \\ & \text{subject to } U^T U = I_k, U \in \mathbb{R}^{p \times k} \text{ and } \|U\|_{2,0} \leq s, \end{aligned} \tag{6.4}$$

where  $S$  is the empirical covariance matrix. Our upper bound is based on (6.4). However, instead of solving (6.4) on the perturbed version of the empirical covariance matrix, we perturb the covariance matrix and solve the following optimization problem on the convex hull of the constraints in (6.4), that is:

$$\begin{aligned} \hat{X} &= \arg \max \langle \tilde{S}, X \rangle - \lambda \|X\|_{1,1} \\ & \text{subject to } X \in \mathcal{F}^k := \{X : 0 \preceq X \preceq I \text{ and } \text{Tr}(X) = k\}, \end{aligned} \tag{6.5}$$

where  $\langle S, X \rangle = \text{Tr}(SX^T)$ . Note that the constraints in (6.5), which is called Fantope [36][293], is the convex hull of the constraints in (6.4). Also, since the constraints in (6.5) only guarantees that the rank of the output is  $\geq k$ , the output  $\hat{X}$  needs not to be a matrix with exact rank of  $k$ . Thus, in order to obtain a proper  $k$ -dimensional subspace, we just output the  $k$ -PCA of  $\hat{X}$ .

---

**Algorithm 6.1.47** Local Gaussian Mechanism-High Dimension

---

**Input:** data records  $\{X_i\}_{i=1}^n \sim P^n$  for  $P \in \mathcal{P}(s, \sigma^2, k)$ , and for  $i \in [n]$ ,  $\|X_i\|_2 \leq 1$ .  $\epsilon, \delta$  are privacy parameters.  $\rho >$  is a constant.

- 1: **for** Each  $i \in [n]$  **do**
  - 2:     Denote  $\tilde{X}_i \tilde{X}_i^T = X_i X_i^T + Z_i$ , where  $Z_i \in \mathbb{R}^{p \times p}$  is a symmetric matrix where the upper triangle, including the diagonal, is i.i.d samples from  $\mathcal{N}(0, \sigma^2)$ ; here  $\sigma^2 = \frac{2 \ln(1.25/\delta)}{\epsilon^2}$ , and each lower triangle entry is copied from its upper triangle counterpart.
  - 3: **end for**
  - 4: Compute  $\tilde{S} = \frac{1}{n} \sum_{i=1}^n \tilde{X}_i \tilde{X}_i^T$ .
  - 5: Get the optimal solution  $\hat{X}$  in (6.5) or do as the followings
  - 6: Setting  $Y^{(0)} = 0, U^{(0)} = 0$
  - 7: **for**  $t = 1, 2, \dots$  **do**
  - 8:      $X^{(t+1)} = \mathcal{P}_{\mathcal{F}^k}(Y^{(t)} - U^{(t)} + \frac{\tilde{S}}{\rho})$
  - 9:      $Y^{(t+1)} = \mathcal{S}_{\lambda/\rho}(X^{(t+1)} + U^{(t)})$  where  $\mathcal{S}$  is the entry-wise soft thresholding operator defined as  $\mathcal{S}_{\lambda/\rho}(x) = \text{sign}(x) \max(|x| - \lambda/\rho, 0)$ .
  - 10:      $U^{(t+1)} = U^{(t)} + X^{(t+1)} - Y^{(t+1)}$
  - 11:     Return  $Y^{(t)}$
  - 12: **end for**
  - 13: Let  $k$ -dimensional principal component of  $\hat{X}$  or  $Y^{(t)}$  be  $\tilde{V}_k$ , output  $\hat{\mathcal{S}} = \text{col}(\tilde{V}_k)$ .
- 

**Theorem 6.1.4.** For any given  $0 < \epsilon, \delta < 1$ , if  $\{X_i\}_{i=1}^n \sim P^n$  for  $P \in \mathcal{P}(s, \sigma^2, k)$  and  $\|X_i\|_2 \leq 1$  for all  $i \in [n]$ , then the solution to the optimization problem (6.5) is  $(\epsilon, \delta)$  non-interactive locally differentially private. Moreover, if let  $\hat{V}_k$  denote the  $k$ -dimensional principal component subspace of  $\hat{X}$  and set  $\lambda \leq O(\lambda_1 \sqrt{\frac{\log p}{n\epsilon^2}})$ , then with probability at least  $1 - \frac{2}{p^2} - \frac{1}{p^c}$ , the following holds

$$\|\sin \Theta(\hat{V}_k, V_k)\|_F^2 \leq O\left(\frac{\lambda_1^2}{(\lambda_k - \lambda_{k+1})^2} \frac{s^2 \log p}{n\epsilon^2}\right),$$

where  $c$  is a universal constant.

From the analysis, we can see that, to ensure non-interactive LDP, here we still need to add a randomized matrix to the covariance matrix, which is similar as in the low dimensional case. And this will cause an additional factor of  $O(\frac{1}{\epsilon^2})$  in the error compared with the non-private case in [293].

Since the optimization problem (6.5) is convex, we can follow the approach in [293] to solve it by using ADMM method (see Algorithm 6.1.47 for the details).

Comparing with the lower bound of the private minimax risk in Theorem 6.1.3, we can see that the bound in Theorem 6.1.4 is roughly larger than the optimal rate by a factor of  $O(\frac{\lambda_1}{\lambda_{k+1}} \frac{s}{k})$ . This means that the upper bound is only near optimal [293]. A remaining open problem is to determine whether it is possible to get a tighter upper bound that does not contain the term of  $\frac{s}{k}$  in the gap.

**Support recovery under local differential privacy** In the high dimensional sparse case, to ensure that an estimator  $\theta$  is consistent, we need to demonstrate that  $\rho(\theta, \theta^*) \rightarrow 0$  as  $n \rightarrow \infty$  and  $\text{supp}(\theta) = \text{supp}(\theta^*)$ . By the definition of row sparsity (6.3), we will show that the solution of (6.5) can recover the support under some reasonable assumptions. For a matrix  $V \in \mathbb{R}^{p \times k}$ , we let  $\text{supp}(V) = \text{supp}((\|V_{1*}\|_2, \|V_{2*}\|_2, \dots, \|V_{p*}\|_2)) = \text{supp}(\text{diag}(VV^T))$ .

Below we assume that the underlying covariance matrix  $\Sigma$  is limited correlated, *i.e.*, satisfies the limited correlation condition (LCC). LCC is first proposed by [193], which is an extension of the Irrepresentable Condition in [361]. Let  $J = \text{supp}(V_k)$ , and  $\Sigma$  be the following block representation:

$$\Sigma = \begin{bmatrix} \Sigma_{JJ} & \Sigma_{JJ^c} \\ \Sigma_{J^c J} & \Sigma_{J^c J^c} \end{bmatrix},$$

where  $\Sigma_{J_1 J_2}$  denotes the  $|J_1| \times |J_2|$  submatrix of  $\Sigma$  consisting of rows in  $J_1$  and columns in  $J_2$ .

**Definition 6.1.2** (LCC). A symmetric matrix  $\Sigma$  satisfies the limited correlation condition

with constant  $\alpha \in (0, 1]$ , if  $\frac{8s}{\lambda_k(\Sigma) - \lambda_{k+1}(\Sigma)} \|\Sigma_{J^c J}\|_{2,\infty} \leq 1 - \alpha$ .

Under the LCC assumption, we now show that our private estimator can recover the support  $J$  with high probability by a modified argument for the Theorem in [293].

**Theorem 6.1.5.** Under the same assumption in Theorem 6.1.4, if the covariance matrix is further assumed to satisfy the LCC condition with  $\alpha$  (Definition 6.1.2) and parameters  $(n, p, s, \lambda_1, \lambda_k, \lambda_{k+1}, \epsilon, \delta, \alpha)$  satisfy the following condition

$$n \geq \Omega\left(\frac{s^2 \lambda_1^2 \log \frac{1}{\delta} \log p(8\lambda_1 + \lambda_k - \lambda_{k+1})}{\epsilon^2 \alpha^2 (\lambda_k - \lambda_{k+1})^2}\right), \quad (6.6)$$

then by setting  $\lambda = O\left(\frac{\sqrt{\log 1/\delta}}{\alpha \epsilon} \sqrt{\frac{\log p}{n}}\right)$ , with probability at least  $1 - \frac{2}{p^2} - \frac{1}{p^c}$ , the solution  $\hat{X}$  to the optimization problem (6.5) is unique and satisfies  $\text{supp}(\text{diag}(\hat{X})) \subseteq J$ . Moreover, if either

$$\begin{aligned} \min_{j \in J} \sqrt{(V_k V_k^T)_{jj}} &\geq O\left(\frac{\lambda_1 s \sqrt{\log 1/\delta}}{\alpha \epsilon (\lambda_k - \lambda_{k+1})} \sqrt{\frac{\log p}{n}}\right) \text{ or} \\ \min_{(i,j) \in J^2} \Sigma_{ij} &\geq O\left(\frac{\sqrt{\log 1/\delta} \lambda_1}{\alpha \epsilon} \sqrt{\frac{\log p}{n}}\right), \text{rank}(\text{sign}(\Sigma_{JJ})) = 1 \end{aligned}$$

holds, then  $\text{supp}(\text{diag}(\hat{X})) = J$ .

### 6.1.5 Experiments

In this section we conduct numerical experiments on both synthetic and real world datasets to validate our theoretical results on utility and privacy tradeoff.

#### Low dimensional case

**Experimental settings** For synthetic datasets, we generate the data samples  $\{X_i\}_{i=1}^n$  independently from a multivariate Gaussian distribution  $\mathcal{N}(0, \Sigma)$ , where  $\Sigma = \frac{\lambda}{5p(\lambda+1)} VV^T + \frac{1}{5p(\lambda+1)} I_p$  for  $V \in \mathbb{V}_{p,k}$ . It can be shown that  $\|X_i\|_2 \leq 1 \forall i \in [n]$  with high probability. We

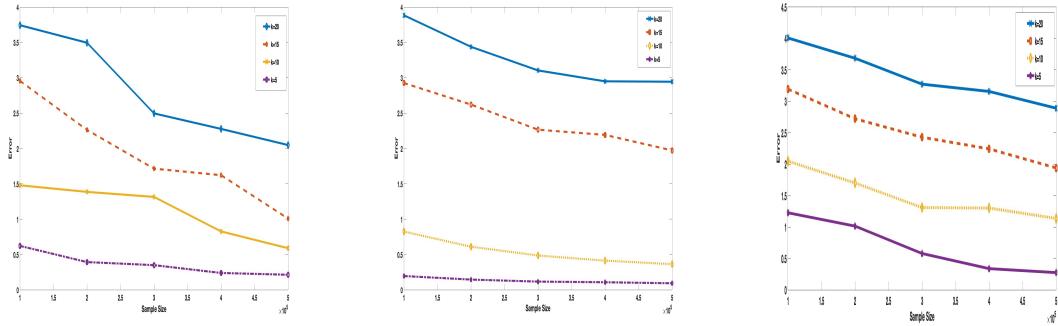


Figure 6.1: LDP-PCA in low dimensional case on real world datasets with different sample size. The left one is for Covertype. The middle one is for Buzz. The right one is for Year dataset.

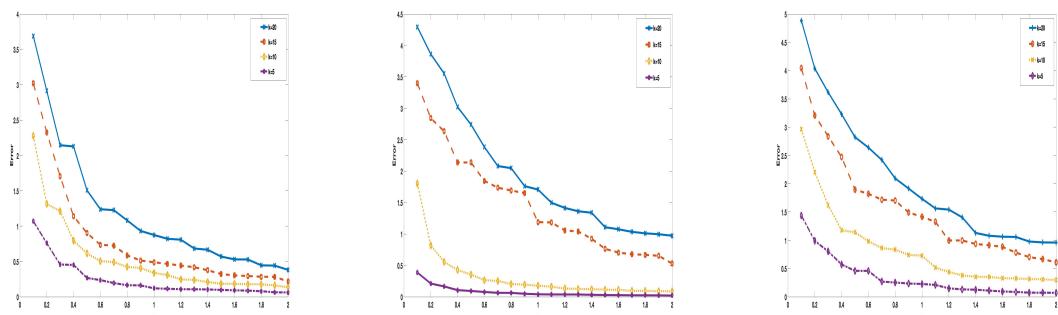


Figure 6.2: LDP-PCA in low dimensional case on real world datasets at different levels of privacy. The left one is for Covertype. The middle one is for Buzz. The right one is for Year dataset.

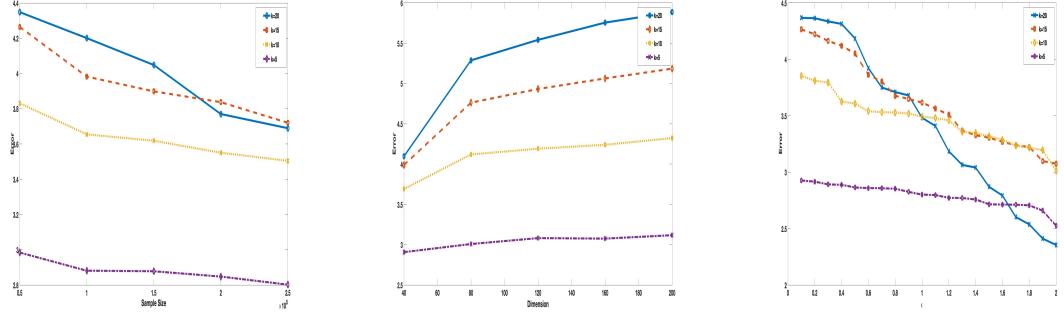


Figure 6.3: LDP-PCA in low dimensional case on synthetic datasets. The left one is for different target dimensions  $k$  over sample size  $n$  with fixed  $\epsilon = 0.5$  and  $p = 40$ . The middle one is for different dimensions with fixed  $n = 10^5$  and  $\epsilon = 0.5$ . The right one is for different level of privacy with fixed  $n = 10^5$  and  $p = 40$ .

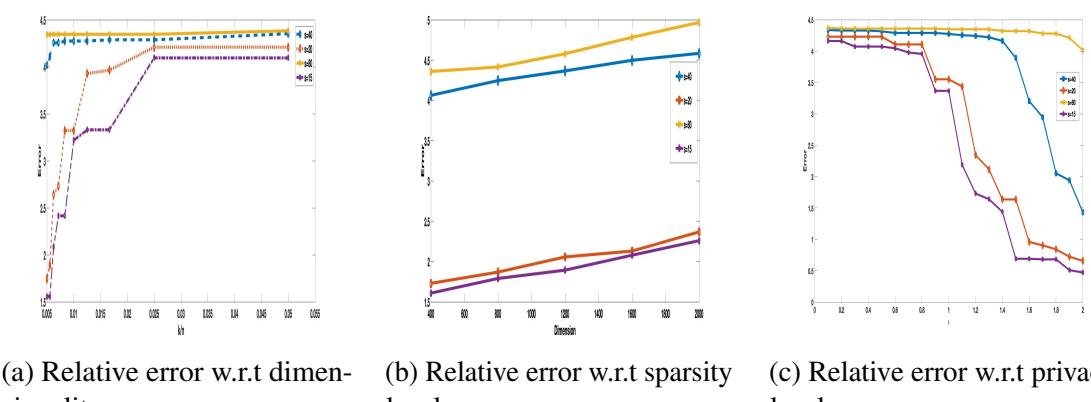


Figure 6.4: LDP-PCA in high dimensional case on synthetic datasets. The left one is for different target dimensions  $k$  over sample size  $n$  with fixed  $\epsilon = 1$  and  $p = 400$ . The middle one is for different dimensions with fixed  $n = 2000$  and  $\epsilon = 1$ . The right one is for different level of privacy with fixed  $n = 2000$  and  $p = 400$ .

choose  $n = 10^5$ ,  $p = 40$ ,  $k = \{5, 10, 15, 20\}$ ,  $\epsilon = 0.5$ ,  $\delta = 10^{-4}$ , and  $\lambda = 1$ . For real world datasets, we run Algorithm 6.1.46 on Covertype and Buzz datasets [90] with normalized rows for each dataset. The error is measured by the subspace distance  $\|\hat{V}_k \hat{V}_k^T - V_k V_k^T\|_F$ . For each experiment, we repeat 20 times and take the average as the final result.

Figure 6.1 and 6.2 are the results for the real world datasets while Figure 6.3 is for synthetic datasets. Figures indicate that 1) the error deceases as the sample size increases

or  $\epsilon$  increases (*i.e.*, becomes less private); 2) the error increases as the dimensionality  $p$  increases or the dimensionality  $k$  of the target subspace increases. All these support our theoretical analysis in Theorem 6.1.2.

### High dimensional case

**Experimental settings** For the high dimensional case, we consider the same distributions as in the low dimensional case and generate the target subspace  $V$  in the following way. For a given sparsity parameter  $s$ , we first generate a random orthogonal matrix  $\tilde{V} \in \mathbb{R}^{s \times k}$ , then pad it with rows of zeros, and finally randomly permute the matrix. We set  $k = 10$ ,  $n = 2000$ ,  $p = 400$ ,  $s = \{15, 20, 40, 80\}$  and  $\epsilon = 1$ .

Besides the synthetic datasets, we also test our algorithm on some real world datasets in [90] and [196]. We first orthogonalize each row of the datasets to 1 as the preprocessing, then run the method in [338] 50 times, and select the one with the largest variance as the optimal solution.

Figure 6.3 shows the results on the synthetic data. We can see that 1) as the term of  $\frac{k}{n}$  increases ( $n$  decreases), the error increases accordingly; 2) the error slightly increases when the dimensionality  $p$  increases, which is due to the fact that the upper bound in Theorem 6.1.4 depends only logarithmically on  $p$  (*i.e.*,  $\log p$ ); 3) the error decreases when  $\epsilon$  increases. Table 6.1 and 6.2 show the results of the error with different sparsity and privacy, respectively. We can see that these results are consistent with our theoretical analysis in Theorem 6.1.4.

### 6.1.6 Omitted Proofs

#### Proof of Theorem 6.1.1

We first prove the non-interactive case, which is based on the following lemma.

**Lemma 6.1.1** (Corollaries 2 and 4 in [97]). Let  $V$  be randomly and uniformly distributed in  $\mathcal{V}$ . Assume that given  $V = v$ ,  $X_i$  is sampled independently according to the distribution of

Dataset	Size	$s$	Error
cancer RNA-Seq	(801, 20531)	10	3.162
		20	3.381
		40	3.668
Leukemia	(72, 7128)	10	3.162
		20	3.435
		40	3.701
Colon cancer	(60, 2000)	10	2.449
		20	3.058
		40	3.228
isolet5	(1559, 617)	10	1.441
		20	2.023
		40	2.508
lung	(203, 3312)	10	2.858
		20	3.464
		40	3.901
NIPS	(11463, 5811)	10	3.643
		20	3.881
		40	4.472

Table 6.1: Results with different sparsity  $s$  for LDP-High dimensional PCA on real world datasets. For all the datasets, the target dimensions  $k$  is set to be  $k = 10$  and  $\epsilon = 2$ .

$P_{v,i}$  for  $i = 1, \dots, n$ . Then, there is a universal constant  $c < 19$  such that for any  $\alpha \in (0, \frac{23}{35}]$ ,

we have

$$I(Z_1, Z_2, \dots, Z_n; V) \leq c\epsilon^2 \sum_{i=1}^n \frac{1}{|\mathcal{V}|^2} \sum_{v, v' \in \mathcal{V}} \|P_{v,i} - P_{v',i}\|_{TV}^2.$$

The  $\epsilon$  non-interactive private minimax risk satisfies

$$\mathcal{M}_n^{\text{Nint}}(\theta(\mathcal{P}), \Phi \circ \rho, \epsilon) \geq \frac{\Phi(\delta)}{2} \left( 1 - \frac{I(Z_1, \dots, Z_n; V) + \log 2}{\log |\mathcal{V}|} \right).$$

Where  $I(\cdot; \cdot)$  is the mutual information.

For the packing set, we have the following lemma:

**Lemma 6.1.2.** [[58]] Let  $(\Theta, \rho)$  be a totally bounded metric space. For any subset  $E \subset \Theta$ , denote by  $\mathcal{N}(E, \epsilon)$  the  $\epsilon$ -covering number of  $E$ , that is, the minimal number of balls of radius  $\epsilon$  whose union contained in  $E$ . Also denote by  $\mathcal{M}(E, \epsilon)$  the  $\epsilon$ -packing number of  $E$ , that is, the maximal number of points in  $E$  whose pairwise distance is at least  $\epsilon$ . If there

Dataset	Size	$\epsilon$	Error
cancer RNA-Seq	(801, 20531)	1	3.559
		0.5	3.790
		0.1	3.967
Leukemia	(72, 7128)	1	4.375
		0.5	4.403
		0.1	4.518
Colon cancer	(60, 2000)	1	3.013
		0.5	4.237
		0.1	4.310
isolet5	(1559, 617)	1	2.884
		0.5	3.405
		0.1	3.896

Table 6.2: Results with different privacy levels  $\epsilon$  for LDP-High dimensional PCA on real world datasets. For all the datasets, the target dimensions  $k$  is set to be  $k = 10$  and  $s = 20$ .

exist  $0 \leq c_0 \leq c_1 < \infty$  and  $d > 0$  such that:

$$\left(\frac{c_0}{\epsilon}\right)^d \leq \mathcal{N}(\Theta, \epsilon) \leq \left(\frac{c_1}{\epsilon}\right)^d$$

for all  $0 < \epsilon \leq \epsilon_0$ , then for any  $1 \geq \alpha > 0$ , there exists a packing set  $\mathcal{V} = \{v_1, \dots, v_m\}$  with  $m \geq (\frac{c_0}{\alpha c_1})^d$  such that  $\alpha\epsilon \leq \rho(v_i, v_j) \leq 2\epsilon$  for each  $i \neq j$ .

Now, for the Grassmannian manifold  $\mathbb{G}_{p,k}$  we have the following lemma regarding the metric entropy (due to [267]).

**Lemma 6.1.3.** For any  $V \in \mathbb{G}_{p,k}$ , identify the subspace  $\text{span}(V)$  with its projection matrix  $VV^T$ , and define the metric on  $\mathbb{G}_{p,k}$  by  $\rho(VV^T, UU^T) = \|VV^T - UU^T\|_F$ . Then for any  $\epsilon \in (0, \sqrt{2 \min\{k, p-k\}})$ ,

$$\left(\frac{c_0}{\epsilon}\right)^{k(p-k)} \leq \mathcal{N}(\mathbb{G}_{p,k}, \epsilon) \leq \left(\frac{c_1}{\epsilon}\right)^{k(p-k)},$$

where  $c_0, c_1$  are absolute constants.

**Proof of Theorem 6.1.1** By Lemmas 6.1.3 and 6.1.2, we know that there exists a packing set  $\mathcal{V}$  with  $\log |\mathcal{V}| \geq k(p-k) \log \frac{c_0}{\alpha c_1}$  with  $2\epsilon_1 \geq \rho(VV^T, UU^T) \geq \alpha\epsilon_1$ , where  $\alpha$  and  $\epsilon_1$  will be specified later. Now we construct the collection of distributions; for each  $V \in \mathcal{V}$ , we define

$$\Sigma_V = \frac{\lambda}{5p(\lambda+1)} VV^T + \frac{1}{5p(\lambda+1)} I_p, \quad (6.7)$$

that is,  $\lambda_1 = \lambda_2 = \dots = \lambda_k = \frac{1}{5p}$  and  $\lambda_{k+1} = \dots = \lambda_p = \frac{1}{5p(\lambda+1)}$ . Then we let  $P_V$  denote the distribution  $\mathcal{N}(0, \Sigma_V)$ .

Now, we first show that the distribution is contained in our parameter space. For  $x \sim \mathcal{N}(0, \Sigma_V)$ , we know that there exists an orthogonal matrix  $M \in \mathbb{R}^{p \times p}$  which satisfies  $Mx \sim \mathcal{N}(0, \text{Diag}(\Sigma_V))$ , where

$$\text{Diag}(\Sigma_V) = \begin{bmatrix} \frac{1}{5p} & & & \\ & \frac{1}{5p} & & \\ & & \ddots & \\ & & & \frac{1}{5p(\lambda+1)} \end{bmatrix}.$$

Thus, we have  $\|x\|_2^2 = \|Mx\|_2^2 \sim \frac{1}{5p}\chi_k^2 + \frac{1}{5p(\lambda+1)}\chi_{p-k}^2$ . For the  $\chi^2$ -distribution, we have the following concentration bound:

**Lemma 6.1.4** ([189]). If  $z \sim \chi_n^2$ , then

$$\mathbb{P}[z - n \geq 2\sqrt{nx} + 2x] \leq \exp(-x).$$

By Lemma 6.1.4, we have the following with probability at least  $1 - \exp(-k) - \exp(-(p-k)) \geq 1 - 2\exp(-\frac{p}{4})$  (by our definition of  $k$ ),  $\|x\|_2^2 \leq \frac{1}{5p}5k + \frac{1}{5p(\lambda+1)}5(p-k) \leq 1$ . Thus,  $\|x\|_2 \leq 1$  with probability at least  $1 - \exp(-\Omega(p))$ , which is contained in the parameter space.

The following lemma shows that the Total Variation distance between  $P_V$  and  $P_{V'}$  can be bounded by the subspace distance between  $V$  and  $V'$ .

**Lemma 6.1.5.** For any pair of  $V, V' \in \mathcal{V}$ , by the KL-distance  $D(\cdot||\cdot)$  of two Gaussian distributions, we have that

$$D(P_V||P_{V'}) \leq \frac{\lambda^2}{2(1+\lambda)} \|\sin \Theta(V, V')\|_F^2.$$

Thus, by Pinsker's inequality that is  $\|P_V - P_{V'}\|_{TV}^2 \leq \frac{\lambda^2}{1+\lambda} \|\sin \Theta(V, V')\|_F^2$ .

*Proof of Lemma 6.1.5.*

$$\begin{aligned} D(P_V||P_{V'}) &= D(\mathcal{N}(0, \Sigma_V)||\mathcal{N}(0, \Sigma_{V'})) \\ &= \frac{1}{2} \text{trace}(\Sigma_{V'}^{-1}(\Sigma_V - \Sigma_{V'})). \end{aligned}$$

Now

$$\Sigma_{V'}^{-1} = 5p(\lambda + 1)[(1 + \lambda)^{-1}V'V'^T + (I_p - V'V'^T)]$$

and

$$\Sigma_V - \Sigma_{V'} = \frac{\lambda}{5p(\lambda + 1)}(VV^T - V'V'^T).$$

we can get

$$\text{trace}(\Sigma_{V'}^{-1}(\Sigma_V - \Sigma_{V'})) = \frac{\lambda^2}{1 + \lambda} \|\sin \Theta(V, V')\|_F^2.$$

□

By Lemmas 6.1.1, 6.1.2 and 6.1.3, we have

$$I(Z_1, Z_2, \dots, Z_n; V) \leq 4 \frac{\lambda^2}{1 + \lambda} cn\epsilon^2 \epsilon_1^2$$

and

$$\mathcal{M}_n^{\text{Nint}}(\theta(\mathcal{P}), \Phi \circ \rho, \epsilon) \geq \alpha^2 \epsilon_1^2 \left(1 - \frac{4 \frac{\lambda^2}{1 + \lambda} cn\epsilon^2 \epsilon_1^2 + \log 2}{k(p-k) \log \frac{c_0}{\alpha c_1}}\right),$$

where  $\epsilon_1 \in (0, \sqrt{2 \min\{k, p-k\}}]$ .

Let  $\alpha = \frac{c_0}{4c_1}$  and  $\epsilon_1^2 = \frac{k(p-k)}{8 \frac{\lambda^2}{1 + \lambda} cn\epsilon^2}$ . We have that if  $\epsilon_1^2 \leq 2 \min\{k, p-k\}$  (which holds

under the assumption of  $n \geq \Omega\left(\frac{1}{\epsilon^2} \frac{\lambda_1 \lambda_{k+1}}{(\lambda_k - \lambda_{k+1})^2} \min\{k, p-k\}\right)$ , then  $\mathcal{M}_n^{\text{Nint}}(\theta(\mathcal{P}), \Phi \circ \rho, \epsilon) \geq \Omega\left(\frac{(\lambda+1)k(p-k)}{\lambda^2 n \epsilon^2}\right)$ .

### Proof of Theorem 6.1.2

The following lemma is based on [183][112].

**Lemma 6.1.6.** Suppose that  $X$  and  $\{X_i\}_{i=1}^n$  are i.i.d sub-Gaussian random vectors in  $\mathcal{R}^p$  with zero mean and covariance matrix  $0 \preceq \Sigma$ . Let  $S_n = \frac{1}{n} \sum_{i=1}^n X_i X_i^T$  be the empirical covariance matrix,  $\{\lambda_i\}_{i=1}^p$  be the eigenvalues of  $\Sigma$  sorted in the descending order, and  $r = \frac{\text{Tr}(\Sigma)}{\|\Sigma\|_2}$ . Then there exist constants  $c \geq 1$  and  $C \geq 0$  such that when  $n \geq r$ , we have the following:

$$\mathbb{P}(\|S_n - \Sigma\|_2 \geq s) \leq \exp\left(-\frac{s}{c_1 \lambda_1 \sqrt{r/n}}\right), \forall s \geq 0.$$

**Proof of Theorem 6.1.2.** Instead of using Davis-Kahan sin  $-\Theta$  theorem in [85] and Weyl's inequality (which is used in [106] based on the assumption that  $\lambda_k - \lambda_{k+1} = \omega(\sqrt{p})$ ), we will use a generalized version of Davis-Kahan Theorem [354].

**Lemma 6.1.7** (Generalized Davis-Kahan Theorem). Let  $\Sigma, \hat{\Sigma} \in \mathbb{R}^{p \times p}$  be two symmetric matrices, with eigenvalues  $\lambda_1 \geq \dots, \lambda_p$  and  $\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_p$ , respectively. Fix  $1 \leq r \leq s \leq p$  and assume that  $\min(\lambda_{r-1} - \lambda_r, \lambda_s - \lambda_{s+1}) > 0$ , where  $\lambda_0 := \infty$  and  $\lambda_{p+1} := -\infty$ . Let  $d := s - r + 1$ . If  $V = (v_r, v_{r+1}, \dots, v_s) \in \mathbb{R}^{p \times d}$  and  $\hat{V} = (\hat{v}_r, \hat{v}_{r+1}, \dots, \hat{v}_s) \in \mathbb{R}^{p \times d}$  have orthogonal columns satisfying  $\Sigma v_j = \lambda_j v_j$  and  $\hat{\Sigma} \hat{v}_j = \hat{\lambda}_j \hat{v}_j$  for  $j = r, r+1, \dots, s$ , then

$$\|\sin \Theta(\hat{V}, V)\|_F \leq \frac{2 \min(\sqrt{d} \|\hat{\Sigma} - \Sigma\|_2, \|\hat{\Sigma} - \Sigma\|_F)}{\min(\lambda_{r-1} - \lambda_r, \lambda_s - \lambda_{s+1})}.$$

By taking  $r = 1, s = k$  in Lemma 6.1.7, we have

$$\|\sin \Theta(\text{col}(\tilde{V}_k), \text{col}(V_k))\|_F^2 \leq O\left(\frac{k \|\tilde{S} - \Sigma\|_2^2}{(\lambda_k - \lambda_{k+1})^2}\right).$$

Let  $S$  denote the non-noise covariance matrix  $S = \frac{1}{n} \sum_{i=1}^n X_i X_i^T$ . Then

$$\|\tilde{S} - \Sigma\|_2 \leq \|\tilde{S} - S\|_2 + \|S - \Sigma\|_2.$$

For the first term, we have  $\|\tilde{S} - S\|_2 = \|Z\|_2$ , where  $Z$  is a symmetric matrix whose upper triangle, including the diagonal, is i.i.d sample from  $\mathcal{N}(0, \sigma^2)$  with  $\sigma^2 = \frac{2 \ln(1.25/\delta)}{n\epsilon^2}$ . Thus, by Corollary 2.3.6 in [275], we have, with probability at least  $1 - \frac{1}{p^{\Omega(1)}}$ , that  $\|Z\|_2 \leq O(\sqrt{p}\sigma)$ .

For the second term, by Lemma 6.1.6, we have, with probability at least  $1 - \exp(-C_1)$ , that  $\|S - \Sigma\|_2 \leq O(\lambda_1 \sqrt{\frac{r}{n}})$ . Combining the above results, we get the proof.  $\square$

### Proof of Theorem 6.1.3

The construction of the class of distributions follows the idea presented in [292]. For self-completeness, we rephrase below some important lemmas. See [292] for the proofs.

Similar to the proof of Theorem 6.1.1, we consider the same class of distribution as in (6.7). Thus, the key step is to find a packing set in  $\mathbb{V}_{p,k}$ . The next lemma provides a general method for constructing such local packing sets.

**Lemma 6.1.8** (Local Stiefel Embedding). Let  $1 \leq d \leq k \leq p$  and the function  $A_\alpha : \mathbb{V}_{p-k,d} \mapsto \mathbb{V}_{p,k}$  be defined in block form as

$$A_\alpha(J) = \begin{bmatrix} (1 - \alpha^2)^{\frac{1}{2}} I_d & 0 \\ 0 & I_{k-d} \\ \alpha J & 0 \end{bmatrix} \quad (6.8)$$

for  $0 \leq \alpha \leq 1$ . If  $J_1, J_2 \in \mathbb{V}_{p-k,d}$ , then

$$\alpha^2(1 - \alpha^2) \|J_1 - J_2\|_F^2 \leq \|\sin \Theta(A_\alpha(J_1), A_\alpha(J_2))\|_F^2 \leq \alpha^2 \|J_1 - J_2\|_F^2.$$

By Lemmas 6.1.1 and 6.1.8, we have the following lemma.

**Lemma 6.1.9.** Let  $\alpha \in [0, 1]$ ,  $\epsilon \in (0, \frac{23}{35}]$  and  $\{J_1, \dots, J_N\} \subset \mathbb{V}_{p-k,d}$  for some  $1 \leq d \leq k \leq p$ . For each  $i \in [N]$ , let  $P_i$  be the distribution of  $\mathcal{N}(0, \Sigma_{A_\alpha(J_i)})$ , where  $\Sigma_{A_\alpha(J_i)}$  is in (6.7). If

$$\min_{i \neq j} \|J_i - J_j\|_F \geq \delta_N,$$

then the  $\epsilon$  non-interactive private minimax risk in the metric of squared subspace distance satisfies:

$$\mathcal{M}_n^{\text{Nint}}(\theta(\mathcal{P}), \Phi \circ \rho, \epsilon) \geq \frac{\delta_N^2 \alpha^2 (1 - \alpha^2)}{2} \left[ 1 - \frac{4c\alpha^2 \epsilon^2 d n \frac{\lambda^2}{1+\lambda} + \log 2}{\log N} \right].$$

For variable selection, we have the following lemma.

**Lemma 6.1.10** (Hypercube construction [214]). Let  $m$  be an integer satisfying  $e \leq m$  and  $s \in [1, m]$ . There exists a subset  $\{J_1, \dots, J_N\} \subset \mathbb{V}_{m,1}$  satisfying the following properties:

1.  $\|J_i\|_{2,0} \leq s, \forall i \in [N],$
2.  $\|J_i - J_j\|_2^2 \geq \frac{1}{4},$
3.  $\log N \geq \max\{cs[1 + \log(m/s)], \log m\}$ , where  $c \geq \frac{1}{30}$  is an absolute constant.

We choose  $d = 1$  and  $\delta_N = \frac{1}{2}$  in Lemma 6.1.9 and  $m = p - k$  in Lemma 6.1.10. Then, if set  $\alpha^2 = O(\frac{1+\lambda}{\lambda^2} \frac{s \log p}{n \epsilon^2})$ , we get  $\mathcal{M}_n^{\text{Nint}}(\theta(\mathcal{P}), \Phi \circ \rho, \epsilon) \geq \Omega(\frac{1+\lambda}{\lambda^2} \frac{s \log p}{n \epsilon^2})$ .

The following lemma shows packing sets in the Grassman manifold.

**Lemma 6.1.11** ([237]). Let  $k$  and  $s$  be integers satisfying  $1 \leq k \leq s - k$  and  $\delta > 0$ . There exists a subset  $\{J_1, \dots, J_N\} \subset \mathbb{V}_{s,k}$  satisfying the following properties:

1.  $\|\sin(J_i, J_j)\|_F \geq k\sqrt{\delta}$  for all  $i \neq j$  and
2.  $\log N \geq k(s - k) \log(\frac{c_2}{\delta})$ , where  $c_2 > 0$  is an absolute constant.

We set  $s = s - k$ ,  $m$  in Lemma 6.1.11 and  $k = d$  in Lemma 6.1.9. For each  $J_i \in \mathbb{V}_{s-k,k}$  in Lemma 6.1.11, we can turn it into a matrix in  $\mathbb{V}_{p-k,k}$  by padding additional rows with

zero entries. Thus, if taking  $\delta_N = O(\frac{\sqrt{k}}{e})$  and  $\alpha^2 = \Theta(\frac{\lambda+1}{\lambda^2} \frac{s}{n\epsilon^2})$  in Lemma 6.1.9, we have  $\mathcal{M}_n^{\text{Nint}}(\theta(\mathcal{P}), \Phi \circ \rho, \epsilon) \geq \Omega(\frac{\lambda+1}{\lambda^2} \frac{sk}{n\epsilon^2})$ . Putting everything together, we have

$$\mathcal{M}_n^{\text{Nint}}(\theta(\mathcal{P}), \Phi \circ \rho, \epsilon) \geq \Omega(\frac{1+\lambda}{\lambda^2} \max\{\frac{s \log p}{n\epsilon^2}, \frac{sk}{n\epsilon^2}\}) \geq \Omega(\frac{\lambda+1}{\lambda^2} \frac{s(k + \log p)}{n\epsilon^2}).$$

### Proof of Theorem 6.1.4

Our proof follows the framework in [293]. First, we show that the subspace distance is close to  $\|\hat{X} - V_k V_k^T\|_F^2$ , where  $V_k$  is the  $k$ -dimensional principal subspace of  $\Sigma$ .

**Lemma 6.1.12.** [[292]] Let  $A, B$  be symmetric matrices and  $V_{A,k}, V_{B,k}$  be their  $k$ -dimensional principal component subspace, respectively. Let  $\delta_{A,B} = \max\{\lambda_k(A) - \lambda_{k+1}(A), \lambda_k(B) - \lambda_{k+1}(B)\}$ . Then, we have

$$\|\sin \Theta(V_{A,k}, V_{B,k})\|_F \leq \sqrt{2} \frac{\|A - B\|_F}{\delta_{A,B}}.$$

By Lemma 6.1.12, we get the following lemma.

### Lemma 6.1.13.

$$\|\sin \Theta(\hat{V}_k, V_k)\|_F^2 \leq 2\|\hat{X} - V_k V_k^T\|_F^2.$$

Thus, we have the following bound for  $\|\hat{X} - V_k V_k^T\|_F$ .

**Lemma 6.1.14** ([293]). Let  $A$  be a symmetric matrix and  $E$  be its projection onto the subspace spanned by the eigenvectors of  $A$  corresponding to its  $k$ -largest eigenvalues  $\lambda_1 \geq \lambda_2 \geq \dots$ . If  $\delta_A = \lambda_k - \lambda_{k+1} > 0$ , then

$$\frac{\delta_A}{2} \|E - F\|_F^2 \leq \langle A, E - F \rangle$$

for all  $F$  satisfying  $0 \preceq F \preceq I$  and  $\text{Tr}(F) = k$ .

**Lemma 6.1.15.** In the optimization problem (5), if  $\lambda \geq \|\tilde{S} - \Sigma\|_{\infty, \infty}$ , then

$$\|\hat{X} - V_k V_k^T\|_F \leq \frac{4s\lambda}{\lambda_k(\Sigma) - \lambda_{k+1}(\Sigma)},$$

where  $\|A\|_{\infty, \infty} = \max_{i,j} |A_{i,j}|$  for any matrix  $A \in \mathbb{R}^{m \times n}$ .

**Proof of Theorem 6.1.4.** Since  $\hat{X}$  and  $V_k V_k^T$  are all feasible for the optimization problem (5), we have

$$0 \leq \langle \tilde{S}, \hat{X} - V_k V_k^T \rangle - \lambda(\|\hat{X}\|_{1,1} - \|V_k V_k^T\|_{1,1}).$$

In Lemma 6.1.14, taking  $A = \Sigma$  (then  $E = V_k V_k^T$ ) and  $F = \hat{X}$ , we get

$$\frac{\lambda_k - \lambda_{k+1}}{2} \|\hat{X} - V_k V_k^T\|_F^2 \leq \langle \Sigma, V_k V_k^T - \hat{X} \rangle.$$

Thus, we have

$$\frac{\lambda_k - \lambda_{k+1}}{2} \|\hat{X} - V_k V_k^T\|_F^2 \leq \langle \tilde{S} - \Sigma, \hat{X} - V_k V_k^T \rangle - \lambda(\|\hat{X}\|_{1,1} - \|V_k V_k^T\|_{1,1}).$$

Since

$$\langle \tilde{S} - \Sigma, \hat{X} - V_k V_k^T \rangle \leq \|\tilde{S} - \Sigma\|_{\infty, \infty} \|\hat{X} - V_k V_k^T\|_{1,1}$$

and  $\lambda \geq \|\tilde{S} - \Sigma\|_{\infty, \infty}$ , we have

$$\frac{\lambda_k - \lambda_{k+1}}{2} \|\hat{X} - V_k V_k^T\|_F^2 \leq \lambda(\|\hat{X} - V_k V_k^T\|_{1,1} - \|\hat{X}\|_{1,1} + \|V_k V_k^T\|_{1,1}).$$

Let  $Q$  be the subset of indices of the non-zero entries of  $v_k V_k^T$ . We have  $v_k V_k^T = (v_k V_k^T)_Q$ .

Thus,

$$\begin{aligned} & \|\hat{X} - V_k V_k^T\|_{1,1} - \|\hat{X}\|_{1,1} + \|V_k V_k^T\|_{1,1} \\ & \leq 2\|(\hat{X} - V_k V_k^T)_Q\|_{1,1}. \end{aligned}$$

Also, we have  $\|(\hat{X} - V_k V_k^T)Q\|_{1,1} \leq s\|\hat{X} - V_k V_k^T\|_F$ . This gives us the proof.  $\square$

By Lemma 6.1.15, we know that our goal is to bound the term of  $\|\tilde{S} - \Sigma\|_{\infty,\infty}$ . Note that by the definition of  $\tilde{S}$ , we have  $\tilde{S} = S + Z$ , where  $Z$  is a symmetric Gaussian matrix with covariance  $\sigma^2 = \frac{2\log 1.25/\delta}{n\epsilon^2}$ . Thus, we have

$$\|\tilde{S} - \Sigma\|_{\infty,\infty} \leq \|S - \Sigma\|_{\infty,\infty} + \|Z\|_{\infty,\infty}.$$

For the first term, we have the following lemma, since  $X$  is assumed to be sub-Gaussian.

**Lemma 6.1.16.** [[293]] Let  $S$  be the sample covariance of an i.i.d. sample of size  $n$  from a sub-Gaussian distribution with population covariance  $\Sigma$ . Then, we have

$$\max_{i,j} \mathbb{P}(|S_{ij} - \Sigma_{ij}| \geq t) \leq 2 \exp\left(-\frac{4nt^2}{(c\lambda_1)^2}\right).$$

For the second term  $\|Z\|_{\infty,\infty}$ , we have, with probability at least  $1 - 2p^2 \exp(-\frac{t^2}{\sigma^2})$ ,  $\|Z\|_{\infty,\infty} \leq t$ . Thus in total, with probability at least  $1 - \frac{2}{p^2} - \frac{1}{p^C}$  we have  $\|\tilde{S} - \Sigma\|_{\infty,\infty} \leq O(\frac{\lambda_1 \sqrt{\log p}}{\sqrt{n\epsilon}})$ . Combining this with Lemma 6.1.15, we get the proof.

### Proof of Theorem 6.1.5

The proof is based on Theorem 1 in [193], which considers the case of general symmetric matrix  $S$ .

$$\hat{X} = \arg \max \langle S, X \rangle - \lambda \|X\|_{1,1} \tag{6.9}$$

subject to  $X \in \mathcal{F}^k := \{X : 0 \preceq X \preceq I \text{ and } \text{Tr}(X) = k\}$ .

**Lemma 6.1.17** ([193]). If the parameter  $\lambda$  in (6.9) satisfies:

$$\frac{\|S - \Sigma\|_{\infty,\infty}}{\lambda} + \frac{8s}{\lambda_k(\Sigma) - \lambda_{k+1}(\Sigma)} \|\Sigma_{J^c J}\|_{2,\infty} \leq 1$$

and

$$0 \leq \lambda_k(\Sigma) - \lambda_{k+1}(\Sigma) - 4\lambda s \left(1 + \frac{8\lambda_1(\Sigma)}{\lambda_k(\Sigma) - \lambda_{k+1}(\Sigma)}\right),$$

then the solution to (6.9) is unique and satisfies  $\text{supp}(\hat{X}) \subseteq J$ . Furthermore, if either

$$\begin{aligned} \min_{j \in J} \sqrt{(V_k V_k^T)_{jj}} &\geq \frac{4\lambda s}{\lambda_k(\Sigma) - \lambda_{k+1}(\Sigma)} \text{ or} \\ \min_{(i,j) \in J^2} \Sigma_{ij} &\geq 2\lambda, \text{rank}(\text{sign}(\Sigma_{JJ})) = 1, \end{aligned}$$

then  $\text{supp}(\text{diag}(\hat{X})) = J$ .

**Proof of Theorem 6.1.5.** From the proof in Theorem 6.1.4, we know that with probability at least  $1 - \frac{2}{p^2} - \frac{1}{p^C}$ , we have  $\|\tilde{S} - \Sigma\|_{\infty, \infty} \leq O\left(\frac{\lambda_1 \sqrt{\log 1/\delta \log p}}{\sqrt{n}\epsilon}\right)$ . By the assumption of LCC and the assumption of  $n$ , we know that if taking  $\lambda = \Theta\left(\frac{\lambda_1 \sqrt{\log 1/\delta \log p}}{\sqrt{n}\epsilon\alpha}\right)$ , all the conditions in Lemma 6.1.17 are satisfied. Thus, we get the proof.  $\square$

## 6.2 Differentially Private Sparse Covariance Matrix Estimation

Estimating or studying the high dimensional datasets while keeping them (locally) differentially private could be quite challenging for many problems, such as sparse linear regression [304], sparse mean estimation [99] and selection problem [286]. However, there are also evidences showing that the loss of some problems under the privacy constraints can be quite small compared with their non-private counterparts. Examples of such nature include high dimensional sparse PCA [122], sparse inverse covariance estimation [302], and high-dimensional distributions estimation [173]. Thus, it is desirable to determine which high dimensional problem can be learned or estimated efficiently in a private manner.

In this Chapter, I try to give an answer to this question for a simple but fundamental problem in machine learning and statistics, called estimating the underlying sparse covari-

ance matrix of bounded sub-Gaussian distribution. For this problem, I propose a simple but nontrivial  $(\epsilon, \delta)$ -DP method, DP-Thresholding, and show that the squared  $\ell_w$ -norm error for any  $1 \leq w \leq \infty$  is bounded by  $O(\frac{s^2 \log p}{n\epsilon^2})$ , where  $s$  is the sparsity of each row in the underlying covariance matrix. Moreover, my method can be easily extended to the local differentially privacy model with  $O(\frac{s^2 \log p}{n\epsilon^2})$  upper bound of error. Finally I proof that this upper bound in LDP model is tight. To prove the above lower bound, I propose a framework, called **General Private Assouad Lemma**, for lower bounding the private minimax risk in the non-interactive or sequential differential privacy model. Our lemma is a generalization of the private Assoud lemma in [98], and can be viewed as a general method for locally differentially private matrix estimation problems.

Experiments on synthetic datasets confirm the theoretical claims. To our best knowledge, this is the first paper studying the problem of estimating high dimensional sparse covariance matrix under (local) differential privacy.

### 6.2.1 Related Work

Recently, there are several papers studying private distribution estimation, such as [173, 169, 176, 118, 12]. For distribution estimation under the central differential privacy model, [176] considers the 1-dimensional private mean estimation of a Gaussian distribution with (un)known variance. The work that is probably most related to ours is [173], which studies the problem of privately learning a multivariate Gaussian and product distributions. The following are the main differences with ours. Firstly, our goal is to estimate the covariance of a sub-Gaussian distribution. Even though the class of distributions considered in our paper is larger than the one in [173], it has an additional assumption which requires the  $\ell_2$  norm of a sample of the distribution to be bounded by 1. This means that it does not include the general Gaussian distribution. Secondly, although [173] also considers the high dimensional case, it does not assume the sparsity of the underlying covariance matrix. Thus, its error bound depends on the dimensionality  $p$  polynomially, which is large in the high

dimensional case ( $p \gg n$ ), while the dependence in our paper is only logarithmically (*i.e.*,  $\log p$ ). Thirdly, the error in [173] is measured by the total variation distance, while it is by  $\ell_w$ -norm in our paper. Thus, the two results are not comparable. Fourthly, the methods in [173] seem difficult to be extended to the local model. [12] recently also studies the covariance matrix estimation via iterative eigenvector sampling. However, their method is just for the low dimensional case and with Frobenious norm as the error measure.

Distribution estimation under local differential privacy has been studied in [118, 169]. However, both of them study only the 1-dimensional Gaussian distribution. Thus, it is quite different from the class of distributions in our paper.

In this paper, we mainly use Gaussian mechanism to the covariance matrix, which has been studied in [106, 122, 302]. However, as it will be shown later, simply outputting the perturbed covariance can cause big error and thus is insufficient for our problem. Compared to these problems, ours is clearly more complicated.

Using information-theoretic techniques to prove lower bounds in the local differential privacy model has also been studied in many papers, such as [99, 98, 97, 169]. [99, 98, 97] proposed several general frameworks for bounding the private minimax risk, such as the private versions of Le Cam lemma, Fano lemma, and Assouad lemma. However, none of these methods can be applied to our problem since all the previous lemmas can only be used in the one-directional case (*i.e.*, the underlying parameter is a vector), while it is a two-directional case (*i.e.*, the underlying parameter is a matrix) in our problem. Moreover, all of the previous methods need to obtain some upper bounds of some hard distribution instances under the total variation distance (or KL-divergence) while in our problem we use  $\chi^2$ -divergence, which makes our method quite different from the previous ones. The method that is the most related to ours is the private Assouad lemma proposed in [98] which can be seen as a special case of our general private Assoud lemma. Recently, [99] revisited the private Assouad lemma and proposed a general theorem with tighter lower bounds via some results in the theory of communication complexity. However, our theorems are incomparable

with theirs since we cannot use their theorem directly to our problem.

### 6.2.2 Private Sparse Covariance Estimation

Let  $x_1, x_2, \dots, x_n$  be  $n$  random samples from a  $p$ -variate distribution with covariance matrix  $\Sigma = (\sigma_{ij})_{1 \leq i,j \leq p}$ , where the dimensionality  $p$  is assumed to be high, *i.e.*,  $p \gg n \geq \text{Poly}(\log p)$ .

We define the parameter space of  $s$ -sparse covariance matrices as the following:

$$\mathcal{G}_0(s) = \{\Sigma = (\sigma_{ij})_{1 \leq i,j \leq p} : \sigma_{-j,j} \text{ is } s\text{-sparse } \forall j \in [p]\}, \quad (6.10)$$

where  $\sigma_{-j,j}$  means the  $j$ -th column of  $\Sigma$  with the entry  $\sigma_{jj}$  removed. That is, a matrix in  $\mathcal{G}_0(s)$  has at most  $s$  non-zero off-diagonal elements in each column.

We assume that each  $x_i$  is sampled from a 0-mean and sub-Gaussian distribution with parameter  $\sigma^2$ , that is,

$$\mathbb{E}[x_i] = 0, \mathbb{P}\{|v^T x_i| > t\} \leq e^{-\frac{t^2}{2\sigma^2}}, \forall t > 0 \text{ and } \|v\|_2 = 1. \quad (6.11)$$

This means that all the one-dimensional marginals of  $x_i$  have sub-Gaussian tails. We also assume that with probability 1,  $\|x_i\|_2 \leq 1$ . We note that such assumptions are quite common in the differential privacy literature, such as [122].

Let  $\mathcal{P}_d(\sigma^2, s)$  denote the set of distributions of  $x_i$  satisfying all the above conditions (*i.e.*, (6.11) and  $\|x_i\|_2 \leq 1$ ) and with the covariance matrix  $\Sigma \in \mathcal{G}_0(s)$ . The goal of private covariance estimation is to obtain an estimator  $\Sigma^{\text{priv}}$  of the underlying covariance matrix  $\Sigma$  based on  $\{x_1, \dots, x_n\} \sim P \in \mathcal{P}_d(\sigma^2, s)$  while keeping it differentially private. In this paper, we will focus on the  $(\epsilon, \delta)$ -differential privacy. We use the  $\ell_2$  norm to measure the difference between  $\Sigma^{\text{priv}}$  and  $\Sigma$ , *i.e.*,  $\|\Sigma^{\text{priv}} - \Sigma\|_2$ .

**Lemma 6.2.1.** Let  $\{x_1, \dots, x_n\}$  be  $n$  random variables sampled from Gaussian distribution

$\mathcal{N}(0, \sigma^2)$ . Then

$$\mathbb{E} \max_{1 \leq i \leq n} |x_i| \leq \sigma \sqrt{2 \log 2n}, \quad (6.12)$$

$$\mathbb{P}\left\{\max_{1 \leq i \leq n} |x_i| \geq t\right\} \leq 2ne^{-\frac{t^2}{2\sigma^2}}. \quad (6.13)$$

Particularly, if  $n = 1$ , we have  $\mathbb{P}\{|x_i| \geq t\} \leq 2e^{-\frac{t^2}{2\sigma^2}}$ .

**Lemma 6.2.2** ([60]). If  $\{x_1, x_2, \dots, x_n\}$  are sampled from a sub-Gaussian distribution in (6.11) and  $\Sigma^* = (\sigma^*)_{1 \leq i, j \leq p} = \frac{1}{n} \sum_{i=1}^n x_i x_i^T$  is the empirical covariance matrix, then there exist constants  $C_1$  and  $\gamma > 0$  such that  $\forall i, j \in [p]$

$$\mathbb{P}(|\sigma_{ij}^* - \sigma_{ij}| > t) \leq C_1 e^{-nt^2 \frac{8}{\gamma^2}} \quad (6.14)$$

for all  $|t| \leq \delta$ , where  $C_1$  and  $\gamma$  are constants and depend only on  $\sigma^2$ . Specifically,

$$\mathbb{P}\{|\sigma_{ij}^* - \sigma_{ij}| > \gamma \sqrt{\frac{\log p}{n}}\} \leq C_1 p^{-8}. \quad (6.15)$$

### 6.2.3 Main Method in Central DP Model

#### A First Approach

A direct way to obtain a private estimator is to perturb the empirical covariance matrix by symmetric Gaussian matrices, which has been used in previous work on private PCA, such as [106, 122]. However, as we can see below, this method will introduce big error.

By [106], for any give  $0 < \epsilon, \delta \leq 1$  and  $\{x_1, x_2, \dots, x_n\} \sim P \in \mathcal{P}_p(\sigma^2, s)$ , the following perturbing procedure is  $(\epsilon, \delta)$ -differentially private:

$$\tilde{\Sigma} = \Sigma^* + N = (\tilde{\sigma}_{ij})_{1 \leq i, j \leq p} = \frac{1}{n} \sum_{i=1}^n x_i x_i^T + N, \quad (6.16)$$

where  $N$  is a symmetric matrix with its upper triangle (including the diagonal) being i.i.d

samples from  $\mathcal{N}(0, \sigma_1^2)$ ; here  $\sigma_1^2 = \frac{2\ln(1.25/\delta)}{n^2\epsilon^2}$ , and each lower triangle entry is copied from its upper triangle counterpart. By [275], we know that  $\|N\|_2 \leq O(\sqrt{p}\sigma_1) = O(\frac{\sqrt{p}\sqrt{\log \frac{1}{\delta}}}{n\epsilon})$ .

We can easily get that

$$\|\tilde{\Sigma} - \Sigma\|_2 \leq \|\Sigma^* - \Sigma\|_2 + \|N\|_2 \leq O\left(\frac{\sqrt{p \log \frac{1}{\delta}}}{n\epsilon}\right), \quad (6.17)$$

where the second inequality is due to [282]. However, we can see that the upper bound of the error in (6.17) is quite large in the high dimensional case.

Another issue of the private estimator in (6.16) is that it is not clear whether it is positive-semidefinite, a property that is normally expected from an estimator.

### Post-processing via Thresholding

We note that one of the reasons that the private estimator  $\tilde{\Sigma}$  in (6.16) fails is due to the fact that some entries are quite large which make  $\|\tilde{\Sigma}_{ij} - \Sigma_{ij}\|_2$  large for some  $i, j$ . To see it more precisely, by (6.13) and (6.14) we can get the following, with probability at least  $1 - Cp^{-6}$ , for all  $1 \leq i, j \leq p$ ,

$$|\tilde{\sigma}_{ij} - \sigma_{ij}| \leq \gamma \sqrt{\frac{\log p}{n}} + \frac{4\sqrt{2\ln \frac{1.25}{\delta}}\sqrt{\log p}}{n\epsilon} = O\left(\gamma \sqrt{\frac{\log p}{n\epsilon^2}}\right). \quad (6.18)$$

Thus, to reduce the error, it is natural to think of the following way. For those  $\sigma_{ij}$  with larger values, we keep the corresponding  $\tilde{\sigma}_{ij}$  in order to make their difference less than some threshold. For those  $\sigma_{ij}$  with smaller values compared with (6.18), since the corresponding  $\tilde{\sigma}_{ij}$  may still be large, if we threshold  $\tilde{\sigma}_{ij}$  to 0, we can lower the error on  $\tilde{\sigma}_{ij} - \sigma_{ij}$ .

Following the above thinking and the thresholding methods in [60] and [38], we propose the following DP-Thresholding method, which post-processes the perturbed covariance matrix in (6.16) with the threshold  $\gamma \sqrt{\frac{\log p}{n}} + \frac{4\sqrt{2\ln 1.25/\delta}\sqrt{\log p}}{n\epsilon}$ . After thresholding, we further threshold the eigenvalues of  $\hat{\Sigma}$  in order to make it positive semi-definite. See Algorithm

6.2.48 for detail.

---

**Algorithm 6.2.48 DP-Thresholding**


---

**Input:**  $\epsilon, \delta$  are privacy parameters and  $\{x_1, x_2, \dots, x_n\} \sim P \in \mathcal{P}(\sigma^2, s)$ .

1: Compute

$$\tilde{\Sigma} = (\tilde{\sigma}_{ij})_{1 \leq i, j \leq p} = \frac{1}{n} \sum_{i=1}^n x_i x_i^T + N,$$

where  $N$  is a symmetric matrix with its upper triangle (including the diagonal) being i.i.d samples from  $\mathcal{N}(0, \sigma_1^2)$ ; here  $\sigma_1^2 = \frac{2 \ln(1.25/\delta)}{n^2 \epsilon^2}$ , and each lower triangle entry is copied from its upper triangle counterpart.

2: Define the thresholding estimator  $\hat{\Sigma} = (\hat{\sigma}_{ij})_{1 \leq i, j \leq n}$  as

$$\hat{\sigma}_{ij} = \tilde{\sigma}_{ij} \cdot I[|\tilde{\sigma}_{ij}| > \gamma \sqrt{\frac{\log p}{n}} + \frac{4\sqrt{2 \ln 1.25/\delta} \sqrt{\log p}}{n\epsilon}]. \quad (6.19)$$

3: Let the eigen-decomposition of  $\hat{\Sigma}$  as  $\hat{\Sigma} = \sum_{i=1}^p \lambda_i v_i v_i^T$ . Let  $\lambda^+ = \max\{\lambda_i, 0\}$  be the positive part of  $\lambda_i$ , then define  $\Sigma^+ = \sum_{i=1}^p \lambda^+ v_i v_i^T$ .

4: **return**  $\Sigma^+$ .

---

**Theorem 6.2.1.** For any  $0 < \epsilon, \delta \leq 1$ , Algorithm 6.2.48 is  $(\epsilon, \delta)$ -differentially private.

For the matrix  $\hat{\Sigma}$  in (6.19) after the first step of thresholding, we have the following key lemma.

**Lemma 6.2.3.** For every fixed  $1 \leq i, j \leq p$ , there exists a constant  $C_1 > 0$  such that with probability at least  $1 - C_1 p^{-\frac{9}{2}}$ , the following holds:

$$|\hat{\sigma}_{ij} - \sigma_{ij}| \leq 4 \min\{|\sigma_{ij}|, \gamma \sqrt{\frac{\log p}{n}} + \frac{4\sqrt{2 \ln 1.25/\delta} \sqrt{\log p}}{n\epsilon}\}. \quad (6.20)$$

By Lemma 6.2.3, we have the following upper bound on the  $\ell_2$ -norm error of  $\Sigma^+$ .

**Theorem 6.2.2.** The output  $\Sigma^+$  of Algorithm 6.2.48 satisfies:

$$\mathbb{E} \|\hat{\Sigma} - \Sigma\|_2^2 = O\left(\frac{s^2 \log p \log \frac{1}{\delta}}{n\epsilon^2}\right), \quad (6.21)$$

where the expectation is taken over the coins of the Algorithm and the randomness of  $\{x_1, x_2, \dots, x_n\}$ .

**Corollary 6.2.1.** For any  $1 \leq w \leq \infty$ , the matrix  $\hat{\Sigma}$  in (6.19) after the first step of thresholding satisfies

$$\|\hat{\Sigma} - \Sigma\|_w^2 \leq O(s^2 \frac{\log p \log \frac{1}{\delta}}{n\epsilon^2}), \quad (6.22)$$

where the  $w$ -norm of any matrix  $A$  is defined as  $\|A\|_w = \sup \frac{\|Ax\|_w}{\|x\|_w}$ . Specifically, for a matrix  $A = (a_{ij})_{1 \leq i,j \leq p}$ ,  $\|A\|_1 = \sup_j \sum_i |a_{ij}|$  is the maximum absolute column sum, and  $\|A\|_\infty = \sup_i \sum_j |a_{ij}|$  is the maximum absolute row sum.

Comparing the bound in the above corollary with the optimal minimax rate  $\Theta(\frac{s^2 \log p}{n})$  in [60] for the non-private case, we can see that the impact of the differential privacy is to make the number of efficient sample from  $n$  to  $n\epsilon^2$ . It is an open problem to determine whether the bound in Theorem 6.2.2 is tight.

## 6.2.4 Extension to Local Differential Privacy

One advantage of our Algorithm 6.2.48 is that it can be easily extended to the locally differentially private (LDP) model.

**Differential privacy in the local model.** In LDP, we have a data universe  $\mathcal{D}$ ,  $n$  players with each holding a private data record  $x_i \in \mathcal{D}$ , and a server that is in charge of coordinating the protocol. An LDP protocol proceeds in  $T$  rounds. In each round, the server sends a message, which sometime is called a query, to a subset of the players, requesting them to run a particular algorithm. Based on the queries, each player  $i$  in the subset selects an algorithm  $Q_i$ , runs it on her data, and sends the output back to the server.

**Definition 6.2.1.** [299] An algorithm  $Q$  is  $(\epsilon, \delta)$ -locally differentially private (LDP) if for all pairs  $x, x' \in \mathcal{D}$ , and for all events  $E$  in the output space of  $Q$ , we have  $\Pr[Q(x) \in E] \leq e^\epsilon \Pr[Q(x') \in E] + \delta$ . A multi-player protocol is  $\epsilon$ -LDP if for all possible inputs and runs of the protocol, the transcript of player  $i$ 's interaction with the server is  $\epsilon$ -LDP. If  $T = 1$ , we say that the protocol is  $(\epsilon, \delta)$  non-interactive LDP.

---

**Algorithm 6.2.49** LDP-Thresholding

---

**Input:**  $\epsilon, \delta$  are privacy parameters,  $\{x_1, x_2, \dots, x_n\} \sim P \in \mathcal{P}(\sigma^2, s)$ .

- 1: **for**  $E$  **do**  $i \in [n]$
  - 2: Denote  $\tilde{x}_i \tilde{x}_i^T = x_i x_i^T + z_i$ , where  $z_i \in \mathbb{R}^{p \times p}$  is a symmetric matrix with its upper triangle (including the diagonal) being i.i.d samples from  $\mathcal{N}(0, \sigma^2)$ ; here  $\sigma^2 = \frac{2 \ln(1.25/\delta)}{\epsilon^2}$ , and each lower triangle entry is copied from its upper triangle counterpart.
  - 3: **end for**
  - 4: Compute  $\tilde{\Sigma} = (\tilde{\sigma}_{ij})_{1 \leq i, j \leq p} = \frac{1}{n} \sum_{i=1}^n \tilde{x}_i \tilde{x}_i^T$ ,
  - 5: Define the thresholding estimator  $\hat{\Sigma} = (\hat{\sigma}_{ij})_{1 \leq i, j \leq n}$  as
- $$\hat{\sigma}_{ij} = \tilde{\sigma}_{ij} \cdot I[|\tilde{\sigma}_{ij}| > \gamma \sqrt{\frac{\log p}{n}} + \frac{4\sqrt{2 \ln 1.25/\delta} \sqrt{\log p}}{\sqrt{n}\epsilon}] \quad (6.23)$$
- 6: Let the eigen-decomposition of  $\hat{\Sigma}$  as  $\hat{\Sigma} = \sum_{i=1}^p \lambda_i v_i v_i^T$ . Let  $\lambda^+ = \max\{\lambda_i, 0\}$  be the positive part of  $\lambda_i$ , then define  $\Sigma^+ = \sum_{i=1}^p \lambda^+ v_i v_i^T$ .
  - 7: **return**  $\Sigma^+$ .
- 

Inspired by Algorithm 6.2.48, it is easy to extend our DP algorithm to the LDP model. The idea is that each  $X_i$  perturbs its covariance and aggregates the noisy version of covariance, see Algorithm 6.2.49 for detail.

The following theorem shows that the error bound of the output in Algorithm 6.2.49 is the same as the the bound in Theorem 6.2.2 asymptotically, whose proof is almost the same as in Theorem 6.2.2.

**Theorem 6.2.3.** The output  $\Sigma^+$  of Algorithm 6.2.49 satisfies:

$$\mathbb{E} \|\hat{\Sigma} - \Sigma\|_2^2 = O\left(\frac{s^2 \log p \log \frac{1}{\delta}}{n\epsilon^2}\right), \quad (6.24)$$

where the expectation is taken over the coins of the Algorithm and the randomness of  $\{x_1, x_2, \dots, x_n\}$ . Moreover,  $\hat{\Sigma}$  in (6.23) satisfies  $\|\hat{\Sigma} - \Sigma\|_w^2 = O\left(\frac{s^2 \log p \log \frac{1}{\delta}}{n\epsilon^2}\right)$ .

## 6.2.5 Lower Bound in Local Differential Privacy Model

In this section we introduce our general framework for lower bounding. Before that, we first review the classical Assouad lemma [283] and its two-directional generalization [60].

### Assouad lemma

Assouad's method works with a hypercube  $\mathcal{V} = \{-1, +1\}^r$  for some  $r \in \mathbb{N}$ . It transforms an estimation problem into multiple hypothesis testing problems using the structure of the problem in an essential way. Let  $\{P_v\}_{v \in \mathcal{V}} \in \mathcal{P}$  be a family of distributions with its corresponding parameters  $\{\theta_v\}_{v \in \mathcal{V}}$  indexed by the hypercube. Similar to the standard reduction from estimation to testing, we consider the following random process. Let  $V$  be a random vector uniformly chosen from the hypercube  $\{-1, +1\}^r$ . After that, the samples  $X_1, X_2, \dots, X_n$  are drawn from the distribution  $P_v$  conditioned on  $V = v$ . For each  $j \in [r]$ , we define the mixture of distributions

$$P_{j,+1}^n = \frac{1}{2^{r-1}} \sum_{v:v_j=1} P_v^n, \quad P_{j,-1}^n = \frac{1}{2^{r-1}} \sum_{v:v_j=-1} P_v^n, \quad (6.25)$$

where  $P_v^n$  is the product distribution of  $X_1, \dots, X_n$ . Then, Assouad lemma can be stated as follows.

**Lemma 6.2.4** (Assouad Lemma). Under the conditions stated in the above paragraph,

$$\mathcal{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \frac{\alpha}{4} \sum_{j=1}^r [1 - \|P_{j,+1}^n - P_{j,-1}^n\|_{TV}], \quad (6.26)$$

where  $\|\cdot\|_{TV}$  is the total variation distance,  $\alpha = \min_{H(v,v') \geq 1, v, v' \in \mathcal{V}} \frac{\Phi(\rho(\theta_v, \theta_{v'}))}{2H(v, v')}$ , and  $H(v, v')$  is the hamming distance between  $v$  and  $v'$ , i.e.,  $H(v, v') = \sum_{j=1}^r \mathbb{1}\{v_j \neq v'_j\}$ .

Instead of restricting to a hypercube  $\mathcal{V}$ , the general Assouad lemma in [60] works with the Cartesian product of a hypercube and the  $r$ -th power of a finite set of vectors. Specifically, for a given  $r \in \mathbb{N}$  and a finite set of  $p$ -dimensional vectors  $B \subset \mathbb{R}^p \setminus \{0_{1 \times p}\}$ , let  $\mathcal{V} = \{-1, +1\}^r$  and  $\Lambda \subseteq B^r$ . Define  $T = \mathcal{V} \otimes \Lambda = \{\tau = (v, \lambda) : v \in \mathcal{V} \text{ and } \lambda \in \Lambda\}$ . This means that one can view an element  $\lambda \in \Lambda$  as an  $r \times p$  matrix with each row coming from set  $B$ , and  $\mathcal{V}$  as a set of parameters with each row indicating whether a given row of  $\lambda$  is

present or not. Similar to Assouad lemma, we assume that there is a family of distributions in the class  $\mathcal{P}$ ,  $\{P_\tau\}_{\tau \in T}$  indexed by  $T$  and its corresponding parameters  $\{\theta_\tau\}_{\tau \in T}$ .

Let  $D_\Lambda = |\Lambda|$ . For a given  $a \in \{-1, +1\}$  and  $j \in [r]$ , we let  $T_{i,a} = \{\tau : v_i(\tau) = a\}$ , where  $v_i(\tau)$  is the  $i$ -th coordinate of the first component of  $\tau$ . It is easy to see that  $|T_{i,a}| = 2^{r-1}D_\Lambda$ . We have the following mixture of distributions

$$P_{j,a}^n = \frac{1}{2^{r-1}D_\Lambda} \sum_{\tau \in T_{j,a}} P_\tau^n, P_{j,a} = \frac{1}{2^{r-1}D_\Lambda} \sum_{\tau \in T_{j,a}} P_\tau. \quad (6.27)$$

**Lemma 6.2.5** (General Assouad's Lemma [57]). Under the conditions stated in above paragraph, we have the following

$$\mathcal{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \frac{\alpha}{4} \sum_{j=1}^r [1 - \|P_{j,+1}^n - P_{j,-1}^n\|_{TV}],$$

where  $\alpha$  satisfies

$$\alpha = \min_{H(v(\tau), v(\tau')) > 1, v(\tau), v(\tau') \in \mathcal{V}} \frac{\Phi(\rho(\theta_\tau, \theta_{\tau'}))}{2H(v(\tau), v(\tau'))},$$

and  $v(\tau)$  is the first component of  $\tau$ .

Now, we present the locally private version of Lemma 6.2.5. Suppose that we draw samples  $Z_1, \dots, Z_n$  according to  $\epsilon$ -LDP channel  $Q(\cdot | X_{1:n})$ . Then, conditioned on  $V = \tau$ , the private sample is distributed according to the marginal distribution  $M_\tau^n$ :

$$M_\tau^n(S) = \int Q^n(S | x_1, x_2, \dots, x_n) dP_\tau^n(x_1, x_2, \dots, x_n). \quad (6.28)$$

Specifically, when  $Q$  is non-interactive, we have  $M_\tau^n = (\int Q(\cdot | x) dP_\tau(x))^{\otimes n}$ . Similarly to (6.27), we can define  $M_{j,a}^n$  and  $M_{j,a}$  for  $a \in \{-1, +1\}$  and  $j \in [r]$ . Thus, combining the above with Lemma 6.2.5, we have the following theorem:

**Theorem 6.2.4.** Under the conditions given in Lemma 6.2.5, the  $\epsilon$  private minimax risk

satisfies:

$$\mathcal{M}_n(\theta(\mathcal{P}), \Phi \circ \rho, \epsilon) \geq \frac{\alpha}{4} \sum_{j=1}^r [1 - \|M_{j,+1}^n - M_{j,-1}^n\|_{TV}]. \quad (6.29)$$

For the sequential private minimax risk, we have the following general lower bound.

**Theorem 6.2.5.** Under the conditions given in Theorem 6.2.4 and further assuming that  $\epsilon \in (0, \frac{1}{2}]$ , the  $\epsilon$  sequential private minimax risk in the metric  $\Phi \circ \rho$  satisfies

$$\mathcal{M}_n^{Int}(\theta(\mathcal{P}), \Phi \circ \rho, \epsilon) \geq \frac{\alpha r}{4} [1 - \left( \frac{n\epsilon^2}{2r} \sup_{\gamma \in \mathbb{B}_\infty(\mathcal{X})} \sum_{j=1}^r \left( \int_{\mathcal{X}} \gamma(x) (dP_{j,+1} - dP_{j,-1})^2 \right)^{\frac{1}{2}} \right)], \quad (6.30)$$

where  $\mathbb{B}_\infty$  is the 1-ball of supremum norm  $\mathbb{B}_\infty = \{\gamma \in L^\infty(\mathcal{X}) \mid \|\gamma\|_\infty \leq 1\}$ , and  $L^\infty(\mathcal{X}) = \{f : \mathcal{X} \mapsto \mathbb{R} \mid \|f\|_\infty < \infty\}$  is the space of uniformly bounded functions with the supremum norm  $\|f\|_\infty = \sup_x |f(x)|$ .

Note that the lower bound in Theorem 6.2.5 reduces to the same one in the private Assouad lemma [98] when  $\Lambda$  contains only one matrix which every row is non-zero. Thus, we call Theorem 6.2.4 as the **General Private Assouad Lemma**. Particularly, if we restrict our attention only to the non-interactive LDP mechanisms, we have the following theorem bounding the private minimax risk, which will be used to prove our lower bounds in this paper.

**Theorem 6.2.6.** Under the conditions given in Theorem 6.2.4 and further assuming that  $\epsilon \in (0, \frac{\ln 2}{2}]$ , the  $\epsilon$  non-interactive private minimax risk in the metric  $\Phi \circ \rho$  satisfies

$$\mathcal{M}_n^{Nint}(\theta, \Phi \circ \rho, \epsilon) \geq \frac{r\alpha}{4} \times \min_{1 \leq j \leq r} \left( 1 - \sqrt{\frac{1}{2} (\epsilon^2 D_{\chi^2}(P_{j,+1} \| P_{j,-1}))^n} \right), \quad (6.31)$$

where  $D_{\chi^2}(\cdot \| \cdot)$  is the  $\chi^2$ -divergence, that is,  $D_{\chi^2}(P \| Q) = \int \frac{(dP - dQ)^2}{dQ}$  for distributions  $P$  and  $Q$ .

The inequality in Lemma 6.2.12 is weaker than the one in Theorem 1 of [98] in the sense that it becomes the later one if combining the inequality of  $\|M_{j,+1} - M_{j,-1}\|_{TV} \leq$

$$D_{\chi^2}(M_{j,+1} \| M_{j,-1}).$$

**Remark 6.2.1.** We note that comparing to existing general lower bounding methods on the private minimax risk, such as [98, 99, 95], Theorem 6.2.6 is quite different. Firstly, while all previous lower bounds depend only linearly on the sample size  $n$ , the lower bound in Theorem 6.2.6 depends exponentially on  $n$ . Secondly, due to the special structure of our indexing set  $T$ , Theorem 6.2.6 is more suitable for matrix estimation problems, while previous methods are more suitable for vector estimation problems. Thirdly, previous lower bounds are measured by (or derived from) the mutual information, the total variation distance, or the KL-divergence between the hard distribution instances, while in Theorem 6.2.6, the lower bound is measured by the  $\chi^2$ -divergence between distributions. This indicates that although Theorem 6.2.6 is stronger than the previous ones, as it can be seen later in the sparse covariance estimation problem, it is easier to obtain a lower bound on the  $\chi^2$ -divergence of the hard instances than other measurements. This is also the reason that existing methods cannot be applied to our problem.

From (6.31), we can see that, to obtain the lower bound, one needs to bound the terms of  $D_{\chi^2}(P_{j,+1} \| P_{j,-1})$  for all  $j$ , which are quite complicated since they are mixture distributions. To simplify the task, we fix all the other terms and consider only the  $j$ -th term, which can be seen as an  $r \times p$  matrix with all other rows fixed, except for the  $j$ -th one. Formally, for an element  $\tau \in T$ , we define the projection  $v_A(\tau) = (v_i(\tau))_{i \in A}$  for a set  $A \subseteq \{1, 2, \dots, r\}$ , and the set  $\{-j\} = [r] \setminus \{j\}$ .  $\lambda_A(\tau)$  and  $\lambda_{-i}(\tau)(\lambda_i(\tau))_{i \in A}$  can be defined similarly, where  $\lambda_i(\tau)$  is the  $i$ -th coordinate of the second component of  $\tau$ . Denote by  $\Lambda_A$  the set  $\Lambda_A = \{\lambda_A(\tau) : \tau \in T\}$ . For  $a \in \{+1, -1\}$ ,  $b \in \{-1, +1\}^{r-1}$  and  $c \in \Lambda_{-j} \subseteq B^{r-1}$ , we let

$$T_{\Lambda_j(a,b,c)} = \{\tau \in T : v_j(\tau) = a, v_{-j}(\tau) = b, \lambda_{-j}(\tau) = c\}$$

and  $D_{\Lambda_j(a,b,c)} = |T_{\Lambda_j(a,b,c)}|$ . Let  $\bar{P}_{j,a,b,c}^n$  denote the mixture distribution

$$\bar{P}_{j,a,b,c}^n = \frac{1}{D_{\Lambda_j(a,b,c)}} \sum_{\tau \in T_{\Lambda_j(a,b,c)}} P_\tau^n, \quad (6.32)$$

and  $\bar{M}_{j,a,b,c}^n$  be its corresponding marginal distribution. Similar to Theorem 6.2.6, we have the following corollary.

**Corollary 6.2.2.** Under the conditions given in Theorem 6.2.4 and further assuming that  $\epsilon \in (0, \frac{\ln 2}{2}]$ , the  $\epsilon$  non-interactive private minimax risk in the metric  $\Phi \circ \rho$  satisfies

$$\begin{aligned} \mathcal{M}_n^{Nint}(\psi(\theta), \Phi \circ \rho, \epsilon) &\geq \frac{r\alpha}{4} \times \min_{1 \leq j \leq r} \\ &\quad \left( 1 - \sqrt{\frac{\epsilon^{2n}}{2} \text{Average}_{v_{-j}, \lambda_{-j}} (D_{\chi^2}(\bar{P}_{j+1, v_{-j}, \lambda_{-j}} \| \bar{P}_{j-1, v_{-j}, \lambda_{-j}}))^n} \right), \end{aligned} \quad (6.33)$$

where the average over  $v_{-j}, \lambda_{-j}$  is induced by the uniform distribution over  $T$ .

### Lower Bound of Private Sparse Covariance Estimation

We follow the settings in [60, 312]. Let  $X_1, \dots, X_n$  be random samples from a zero-mean  $p$ -variate distribution with covariance matrix  $\Sigma = (\sigma_{ij})_{1 \leq i,j \leq p}$ . The goal of sparse covariance matrix estimation is to estimate the unknown matrix  $\Sigma$  based on samples  $\{X_1, \dots, X_n\}$ , and the locally private version is to determine a locally differentially private estimator. In this paper, we focus on the high dimensional case, that is,  $c_1 n^\beta \leq p \leq \exp(c_2 n)$  for some  $\beta > 1, c_1, c_2 > 0$ . We assume that the underlying covariance is sparse. That is,  $\Sigma \in \mathcal{G}(s)$  with

$$\mathcal{G}(s) = \{\Sigma = (\sigma_{ij})_{1 \leq i,j \leq p} : \|\sigma_{-j,j}\|_0 \leq s, \forall j \in [p]\}, \quad (6.34)$$

where  $\sigma_{-j,j}$  is the  $j$ -th column of  $\Sigma$  with  $\sigma_{j,j}$  removed, i.e., a matrix in  $\mathcal{G}(s)$  has at most  $s$ -nonzero off-diagonal elements on each column.

Moreover, we assume that each  $X_i$  is sampled from a  $\rho$ -sub-Gaussian distribution. That

is, for all  $t > 0$  and  $\|v\|_2 = 1$ ,

$$\mathbb{P}\{|\langle v, X \rangle| > t\} \leq \exp\left(\frac{-t^2}{2\rho}\right), \quad (6.35)$$

which means that all the one-dimensional marginals of  $X$  have sub-Gaussian tails.

Additionally, in private matrix-related estimation problems, it is always assumed that the  $\ell_2$  norm of each  $X_i$  are bounded by 1 [106, 122, 295, 312]. In this paper, we relax the bounded norm assumption in the following way; for the random vector  $X \in \mathbb{R}^p$ , we assume that  $\|X\|_2 \leq 1$  with probability at least  $1 - e^{-\Omega(p)}$ . This leads us to the following class of distributions  $\mathcal{P}(\tau, s)$ .

$$\begin{aligned} \mathcal{P}(\rho, s) = \{P : X \sim P \text{ satisfies (6.35) and } \|X\|_2 \leq 1 \\ \text{w.p at least } 1 - e^{-\Omega(p)}, \mathbb{E}X = 0, \Sigma = \mathbb{E}[XX^T] \in \mathcal{G}(s)\}. \end{aligned} \quad (6.36)$$

Before showing the lower bound, we first describe our construction of the hard indexing set  $T$  with their distributions  $\{P_\tau\}_{\tau \in T}$  instances, which is motivated by the ones in [60].

We first construct the parameter set, which is the same as in [60]. Let  $r = \lfloor \frac{p}{2} \rfloor$  and  $B$  be the collection of all row vectors  $b = (v_j)_{1 \leq j \leq p}$  such that  $v_j = 0$  for all  $1 \leq j \leq p - r$  and  $v_j = 0$  or 1 for  $p - r + 1 \leq j \leq p$  under the constraint that  $\|b\|_0 = k$  (where the value of  $k$  will be specified later). We can view each  $(b_1, \dots, b_r)$  as an  $r \times p$  matrix with the  $i$ -th row being  $b_i$ .

Then, we define the set  $T$  and its corresponding distributions. Define  $\Lambda \subset B^r$  to be the set of all elements in  $B^r$  such that each column is less than or equal to  $2k$ . For each matrix  $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_r) \in \Lambda$ , define a  $p \times p$  matrix  $A_m(\lambda_m)$  by making the  $m$ -th row and column of  $A_m(\lambda_m)$  be  $\lambda_m$  and the rest of entries be 0.

Next, we construct the distributions. Let  $T = \mathcal{V} \otimes \Lambda$ . For each  $\tau = (v, \lambda)$ , we define a

matrix  $P_\tau = \mathcal{N}(0, \Sigma(\tau))$  with the matrix  $\Sigma(\tau)$  having the following form

$$\Sigma(\tau) = cI_p + c\alpha_{n,p,\epsilon} \sum_{j=1}^n v_j A_j(\lambda_j), \quad (6.37)$$

where  $c > 0$  is some constant to be specified later and  $\alpha_{n,p,\epsilon} = \gamma \sqrt{\frac{\log p}{n\epsilon^2}}$  for some universal small enough constant  $\gamma$ .

We first choose  $c, \gamma$  and  $k$  to make the Gaussian distribution  $\mathcal{N}(0, \Sigma(\tau))$  contained in the class (6.36).

**Lemma 6.2.6.** Under the assumption of  $n \geq C \frac{s^2 \log p}{\epsilon^2}$ , if let  $c \leq \min\{\frac{\rho}{2}, \frac{1}{10p}\}$  and  $k = \max\{\lceil \frac{s}{2} \rceil - 1, 0\}$ , then there is a  $\gamma$ , which depends only on  $C$ , such that  $\mathcal{N}(0, \Sigma(\tau)) \in \mathcal{P}(\rho, s)$  for every  $\tau \in T$ , where  $T$  is the set defined in the above paragraph.

In order to use Theorem 6.2.6, we need to bound the term

$$\alpha = \min_{H(v(\tau), v(\tau')) > 1, v(\tau), v(\tau') \in \mathcal{V}} \frac{\|\Sigma(\tau) - \Sigma(\tau')\|_2^2}{2H(v(\tau), v(\tau'))},$$

which is due to the following Lemma in [60].

**Lemma 6.2.7.** Under the conditions given in Lemma 6.2.6, we have  $\alpha \geq \frac{(k\alpha_{n,p,\epsilon})^2}{p}$ .

The following key lemma gives a lower bound on the term

$$\text{Average}_{v_{-j}, \lambda_{-j}} (D_{\chi^2}(\bar{P}_{j,+1, v_{-j}, \lambda_{-j}} \| \bar{P}_{j,-1, v_{-j}, \lambda_{-j}}))^n.$$

**Lemma 6.2.8.** Under the conditions on  $T$ ,  $\Sigma(\tau)$  and the conditions of given in Lemma 6.2.6, the following holds for every  $j \in [r]$ , when  $\gamma$  is sufficiently small and  $p$  is sufficiently large

$$\text{Average}_{v_{-j}, \lambda_{-j}} (D_{\chi^2}(\bar{P}_{j,+1, v_{-j}, \lambda_{-j}} \| \bar{P}_{j,-1, v_{-j}, \lambda_{-j}}))^n \leq \frac{3}{4} \frac{1}{\epsilon^{2n}}.$$

We now have the following lemma for the term  $(\Sigma_1 - \Sigma_0)(\Sigma_2 - \Sigma_0)$ , which corresponds to the Lemma 10 in [60]:

**Lemma 6.2.9.** Let  $\Sigma_0, \Sigma_1, \Sigma_2$  be the same covariance matrices as above. Define  $J$  to be the number of overlapping  $c\alpha_{n,p,\epsilon}$ 's between  $\Sigma_1$  and  $\Sigma_2$  on the first row, and define the matrix  $Q$  as the following

$$Q = (q_{ij})_{1 \leq i,j \leq p} = (\Sigma_1 - \Sigma_0)(\Sigma_2 - \Sigma_0).$$

Then there are index subsets  $I_r$  and  $I_c$  in  $\{2, \dots, p\}$  with  $|I_r| = |I_c| = k$  and  $|I_r \cap I_c| = J$

$$q_{ij} = \begin{cases} Jc^2\alpha_{n,p,\epsilon}^2, & i = j = 1 \\ c^2\alpha_{n,p,\epsilon}^2, & i \in I_r \text{ and } j \in I_c \\ 0, & \text{otherwise} \end{cases} \quad (6.38)$$

And the matrix  $(\Sigma_0 - \Sigma_1)(\Sigma_0 - \Sigma_2)$  has rank 2 with two identical non-zero eigenvalues  $Jc^2\alpha_{n,p,\epsilon}^2$ .

Thus by Lemma 6.2.9 and the Lemma 11 in [60] we have:

**Lemma 6.2.10** (Lemma 11 in [60]). Let  $R_{1,\lambda_1,\lambda'_1}^{v-1,\lambda_{-1}}$  satisfies

$$R_{\lambda_1,\lambda'_1}^{v-1,\lambda_{-1}} = -2 \log(1 - Jc^2\alpha_{n,p,\epsilon}^2) + R_{1,\lambda_1,\lambda'_1}^{v-1,\lambda_{-1}} \quad (6.39)$$

Then uniformly over  $J$ , we have

$$\mathbb{E}_{(\lambda_1,\lambda'_1)|J} [\mathbb{E}_{(v-1,\lambda_{-1})|(\lambda_1,\lambda'_1)} [\exp(\frac{n}{2}(R_{1,\lambda_1,\lambda'_1}^{v-1,\lambda_{-1}}))] \leq \frac{3}{2}.$$

Next we will prove our lemma. By (6.82) and Lemma 6.2.10 we now have

$$\begin{aligned} & \mathbb{E}_{\lambda_1,\lambda'_1} [\mathbb{E}_{(v-1,\lambda_{-1})|(\lambda_1,\lambda'_1)} [\exp(\frac{n}{2}(R_{1,\lambda_1,\lambda'_1}^{v-1,\lambda_{-1}})) - 1]] \\ &= \mathbb{E}_J \{ \exp[-n \log(1 - Jc^2\alpha_{n,p,\epsilon}^2)] \times \\ & \quad \mathbb{E}_{(\lambda_1,\lambda'_1)|J} [\mathbb{E}_{(v-1,\lambda_{-1})|(\lambda_1,\lambda'_1)} [\exp(\frac{n}{2}(R_{1,\lambda_1,\lambda'_1}^{v-1,\lambda_{-1}}))] - 1] \} \\ &\leq \mathbb{E}_J \{ \frac{3}{2} \exp[-n \log(1 - Jc^2\alpha_{n,p,\epsilon}^2)] - 1 \} \end{aligned}$$

Recall that  $J$  is the number of overlapping  $c\alpha_{n,p,\epsilon}$ 's between  $\Sigma_1$  and  $\Sigma_2$  on the first row. Thus  $J$  has the hypergeometric distribution as  $\lambda_1, \lambda'_1$  vary in  $B$  for each given  $\lambda_{-1}$ . For  $0 \leq j \leq k$ , the same as in [60], we have

$$\mathbb{E}(\mathbb{I}\{J = j\}) = \binom{k}{j} \binom{p_{\lambda_{-1}} - k}{k - j} / \binom{p_{\lambda_{-1}}}{k} \leq \left(\frac{k^2}{p/4 - 1 - k}\right)^j.$$

Thus, we have

$$\begin{aligned} & \mathbb{E}_J \left\{ \frac{3}{2} \exp[-n \log(1 - Jc^2 \alpha_{n,p,\epsilon}^2)] - 1 \right\} \\ & \leq \sum_{j=0}^k \left(\frac{k^2}{p/4 - 1 - k}\right)^j \left\{ \frac{3}{2} \exp[-n \log(1 - jc^2 \alpha_{n,p,\epsilon}^2)] - 1 \right\} \\ & \leq \frac{1}{\epsilon^{2n}} \sum_{j=0}^k \left(\frac{k^2}{p/4 - 1 - k}\right)^j \left\{ \frac{3}{2} \exp[-n \log(1 - jc^2 \gamma^2 \frac{\log p}{n})] - 1 \right\} \end{aligned} \quad (6.40)$$

$$\begin{aligned} & \leq \frac{1}{\epsilon^{2n}} \sum_{j=0}^k \left(\frac{k^2}{p/4 - 1 - k}\right)^j \left\{ \frac{3}{2} \exp[2jc^2 \gamma^2 \log p] \right\} + \frac{1}{\epsilon^{2n}} \frac{1}{2} \\ & \leq \frac{1}{\epsilon^{2n}} \frac{3}{2} \sum_{j \geq 1} (p^{1-1/\beta} p^{-2\gamma^2 c^2})^{-j} + \frac{1}{2} \frac{1}{\epsilon^{2n}} \end{aligned} \quad (6.41)$$

$$\leq C \frac{1}{\epsilon^{2n}} \sum_{j \geq 1} (p^{1/2-1/2\beta})^{-j} + \frac{1}{2} \frac{1}{\epsilon^{2n}} \leq \frac{3}{4} \frac{1}{\epsilon^{2n}} \quad (6.42)$$

Where (6.40) is due to that, let  $a = \frac{1}{\epsilon^2}$  and  $b = jc^2 \gamma^2 \frac{\log p}{n}$ , then it is sufficient to prove

$$\begin{aligned} & -\log(1 - ab) \leq \log a - \log(1 - b) \\ & \equiv \frac{1}{1 - ab} \leq \frac{a}{1 - b} \\ & \equiv b(a + 1) \leq 1 \end{aligned}$$

The final inequality is true due to that  $b(a + 1) \leq 2ab \leq 2kc^2 \gamma^2 \frac{\log p}{n\epsilon^2} \leq 1$  when  $\gamma$  is small enough.

(6.41) is due to that  $k^2 = O(\frac{n\epsilon^2}{\log p}) = O(\frac{n}{\log p}) = O(\frac{p^{1/\beta}}{\log p})$ , and  $\gamma^2 \leq \frac{\beta-1}{54\beta}$  for sufficient large  $p$ . Combining Lemmas 6.2.6, 6.2.7 and 6.2.8 with  $r = \lfloor \frac{p}{2} \rfloor$ , by Corollary 6.2.2 we

have the following lower bound theorem.

**Theorem 6.2.7.** If  $\epsilon \in (0, \frac{\ln 2}{2}]$ ,  $n \geq C \frac{s^2 \log p}{\epsilon^2}$  and  $p \geq c_1 n^\beta$  for  $\beta > 1$ , then the  $\epsilon$  non-interactive private minimax risk in the metric of squared spectral norm satisfies the following inequality

$$\mathcal{M}_n^{Nint}(\Sigma(\mathcal{P}(s, \rho)), \Phi \circ \rho, \epsilon) \geq \Omega\left(\frac{s^2 \log p}{n\epsilon^2}\right). \quad (6.43)$$

For the upper bound, [312] recently showed that if each  $\|X_i\|_2 \leq 1$  and  $\{X_i\}_{i=1}^n \sim P$ , where  $P \in \mathcal{P}(s, \rho)$ , then by using a thresholding method on the perturbed empirical covariance matrix with some well-defined threshold, the output  $\tilde{\Sigma}$  satisfies  $\|\tilde{\Sigma} - \Sigma\|_2^2 \leq O\left(\frac{s^2 \log p}{n\epsilon^2}\right)$  with high probability. Combining this upper bound with Theorem 6.2.7, we can see that the bound  $\Theta\left(\frac{s^2 \log p}{n\epsilon^2}\right)$  is actually tight (i.e., optimal).

We note that for the non-private case, the optimal rate of minimax risk under the same measurement is  $\Theta\left(\frac{s^2 \log p}{n}\right)$  [60]. Thus, in this case, the impact of the local differential privacy is to change the number of efficient samples from  $n$  to  $n\epsilon^2$ . However, the collection of the considered distributions needs another assumption, which says that  $\|X\|_2$  is bounded by 1 with high probability. This is not necessary in the non-private case [60], but needed for showing the upper bound.

Moreover, [312] also show that there is an  $(\epsilon, \delta)$  non-interactive LDP algorithm whose output  $\tilde{\Sigma}$  satisfies  $\|\tilde{\Sigma} - \Sigma\|_w^2 \leq O\left(\frac{s^2 \log p}{n\epsilon^2}\right)$  for every  $w \in [1, \infty]$  with high probability. One natural question is whether it is optimal. The following corollary provides an affirmative answer.

**Corollary 6.2.3.** Under the assumptions given in Theorem 6.2.7, for each  $w \in [1, \infty]$ , the  $\epsilon$  non-interactive private minimax risk in the metric of squared  $\ell_w$  norm satisfies the following

$$\mathcal{M}_n^{Nint}(\Sigma(\mathcal{P}(s, \rho)), \Phi \circ \rho, \epsilon) \geq \Omega\left(\frac{s^2 \log p}{n\epsilon^2}\right), \quad (6.44)$$

where the  $\ell_w$ -norm of any matrix  $A$  is defined as  $\|A\|_w = \sup \frac{\|Ax\|_w}{\|x\|_w}$ .

There are still some open problems. Firstly, both Theorem 6.2.6 and 6.2.7 are restricted to non-interactive LDP protocols. The first open question is whether they can be extended to the sequential LDP model. Secondly, from Theorem 6.2.6 we can see that the lower bound holds under the assumption of  $\epsilon \in (0, \frac{\ln 2}{2}]$ . Thus, the second open question is whether the range of  $\epsilon$  can be enlarged, or whether better result can be achieved when  $\epsilon$  is larger, such as those in [351]? Recently, [96] extended the classical private Assouad lemma to the case where  $\epsilon \in [0, \infty)$  via some results in the theory of communication complexity. However, their theorem cannot be used in our problem. The main reason is that, in their main results (Theorem 10 and Corollary 11 in [96]), they need the two distributions  $P_1$  and  $P_{-1}$  satisfy strong data processing inequalities (SDPI), and also they should satisfy  $|\log \frac{dP_1}{dP_{-1}}|$  is bounded by a constant under the assumption that the coordinates of  $X$  are independent. However, it is quite hard to bound the term or proof the SDPI property for our distributions in (6.37) due to the facts that the coordinates of the samples are dependent and the forms of our distributions are quite complicated. Thus, to extend to general  $\epsilon \in (0, \infty)$  case we need new methods, which will be left for future work. The third open question is whether Theorem 6.2.5 and 6.2.6 can be used to other matrix-related estimation problems? We leave them for future research.

## 6.2.6 Experiments

In this section, we evaluate the performance of Algorithm 6.2.48 and 6.2.49 practically on synthetic datasets.

**Data Generation** We first generate a symmetric sparse matrix  $\tilde{U}$  with the sparsity ratio  $sr$ , that is, there are  $sr \times p \times p$  non-zero entries of the matrix. Then, we let  $U = \tilde{U} + \lambda I_p$  for some constant  $\lambda$  to make  $U$  positive semi-definite and then scale it to  $U = \frac{U}{c}$  by some constant  $c$  which makes the norm of samples less than 1 (with high probability)<sup>1</sup>. Finally,

---

<sup>1</sup>Although the distribution is not bounded by 1, actually, as we see from previous section, we can obtain the same result as long as the  $\ell_2$  norm of the samples is bounded by 1.

we sample  $\{x_1, \dots, x_n\}$  from the multivariate Gaussian distribution  $\mathcal{N}(0, U)$ . In this paper, we will use set  $\lambda = 50$  and  $c = 200$ .

**Experimental Settings** To measure the performance, we compare the  $\ell_1$  and  $\ell_2$  norm of relative error, respectively. That is,  $\frac{\|\Sigma^+ - U\|_2}{\|U\|_2}$  or  $\frac{\|\Sigma^+ - U\|_1}{\|U\|_1}$  with the sample size  $n$  in three different settings: 1) we set  $p = 100$ ,  $\epsilon = 1$ ,  $\delta = \frac{1}{n}$  and change the sparse ratio  $sr = \{0.1, 0.2, 0.3, 0.5\}$ . 2) We set  $\epsilon = 1$ ,  $\delta = \frac{1}{n}$ ,  $sr = 0.2$ , and let the dimensionality  $p$  vary in  $\{50, 100, 200, 500\}$ . 3) We fix  $p = 200$ ,  $\delta = \frac{1}{n}$ ,  $sr = 0.2$  and change the privacy level as  $\epsilon = \{0.1, 0.5, 1, 2\}$ . We run each experiment 20 times and take the average error as the final one.

**Experimental Results** Figure 6.5 and 6.6 are the results of DP-Thresholding (Algorithm 6.2.48) with  $\ell_2$  and  $\ell_1$  relative error, respectively. Figure 6.7 and 6.8 are the results of LDP-Thresholding (Algorithm 6.2.49) with  $\ell_2$  and  $\ell_1$  relative error, respectively. From the figures we can see that: 1) if the sparsity ratio is large *i.e.*, the underlying covariance matrix is more dense, the relative error will be larger, this is due to the fact showed in Theorem 6.2.2 and 6.2.3 that the error depends on the sparsity  $s$ . 2) The dimensionality only slightly affects the relative error. That is, even if we double the value of  $p$ , the error increases only slightly. This is consistent with our theoretical analysis in Theorem 6.2.2 and 6.2.3 which says that the error of our private estimators is only logarithmically depending on  $p$  (*i.e.*,  $\log p$ ). 3) With the privacy parameter  $\epsilon$  increases (which means more private), the error will become larger. This has also been showed in previous theorems.

In summary, all the experimental results support our theoretical analysis.

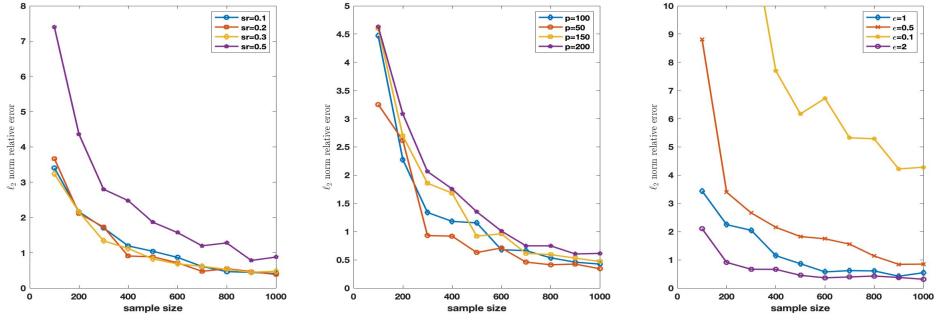


Figure 6.5: Experiment results of Algorithm 6.2.48 for  $\ell_2$ -norm relative error. The left one is for different sparsity levels, the middle one is for different dimensionality  $p$ , and the right one is for different privacy level  $\epsilon$ .

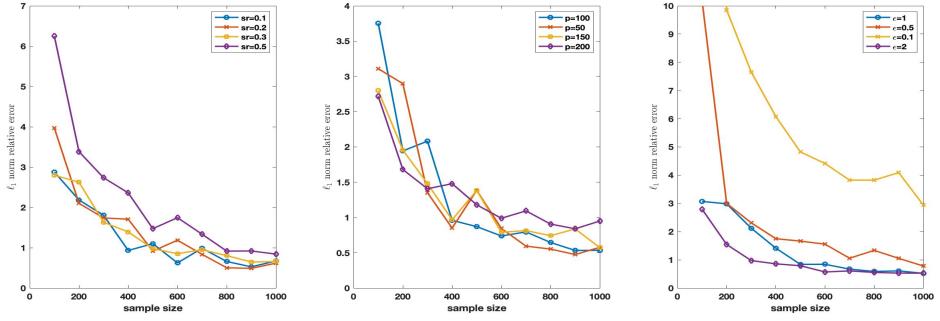


Figure 6.6: Experiment results of Algorithm 6.2.48 for  $\ell_1$ -norm relative error. The left one is for different sparsity levels, the middle one is for different dimensionality  $p$ , and the right one is for different privacy level  $\epsilon$ .

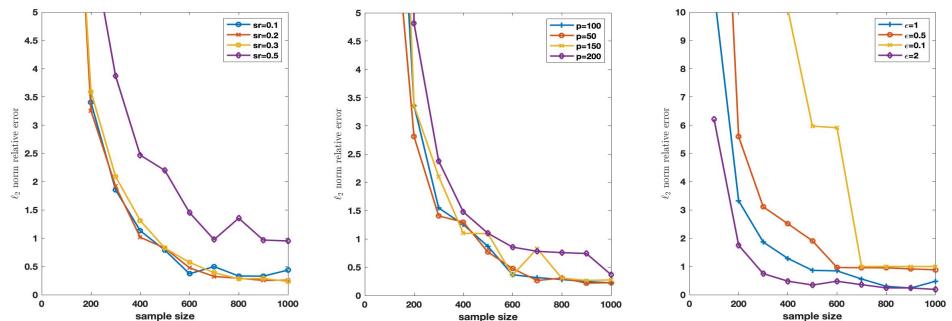


Figure 6.7: Experiment results of Algorithm 6.2.49 for  $\ell_2$ -norm relative error. The left one is for different sparsity levels, the middle one is for different dimensionality  $p$ , and the right one is for different privacy level  $\epsilon$ .

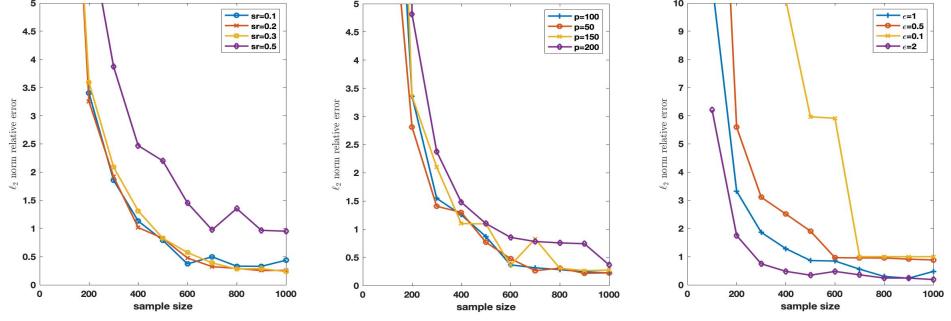


Figure 6.8: Experiment results of Algorithm 6.2.49 for  $\ell_1$ -norm relative error. The left one is for different sparsity levels, the middle one is for different dimensionality  $p$ , and the right one is for different privacy level  $\epsilon$ .

### 6.2.7 Omitted Proofs

#### Proof of Theorem 6.2.1

By [122] and [106], we know that Step 1 keeps the matrix  $(\epsilon, \delta)$ -differentially private. Thus, Algorithm 1 is  $(\epsilon, \delta)$ -differentially private due to the post-processing property of differential privacy [107].

#### Proof of Lemma 6.2.3

Let  $\Sigma^* = (\sigma_{ij}^*)_{1 \leq i,j \leq p}$  and  $N = (n_{ij})_{1 \leq i,j \leq p}$ . Define the event  $A_{ij} = \{|\tilde{\sigma}_{ij}| > \gamma \sqrt{\frac{\log p}{n}} + \frac{4\sqrt{2 \ln 1.25/\delta} \sqrt{\log p}}{n\epsilon}\}$ . We have:

$$|\hat{\sigma}_{ij} - \sigma_{ij}| = |\sigma_{ij}| \cdot I(A_{ij}^c) + |\tilde{\sigma}_{ij} - \sigma_{ij}| \cdot I(A_{ij}). \quad (6.45)$$

By the triangle inequality, it is easy to see that

$$\begin{aligned} A_{ij} &= \{|\tilde{\sigma}_{ij} - \sigma_{ij} + \sigma_{ij}| > \gamma \sqrt{\frac{\log p}{n}} + \frac{4\sqrt{2 \ln 1.25/\delta} \sqrt{\log p}}{n\epsilon}\} \\ &\subset \{|\tilde{\sigma}_{ij} - \sigma_{ij}| > \gamma \sqrt{\frac{\log p}{n}} + \frac{4\sqrt{2 \ln 1.25/\delta} \sqrt{\log p}}{n\epsilon} - |\sigma_{ij}|\} \end{aligned}$$

and

$$\begin{aligned} A_{ij}^c &= \left\{ |\tilde{\sigma}_{ij} - \sigma_{ij} + \sigma_{ij}| \leq \gamma \sqrt{\frac{\log p}{n}} + \frac{4\sqrt{2 \ln 1.25/\delta} \sqrt{\log p}}{n\epsilon} \right\} \\ &\subset \left\{ |\tilde{\sigma}_{ij} - \sigma_{ij}| > |\sigma_{ij}| - \left( \gamma \sqrt{\frac{\log p}{n}} + \frac{4\sqrt{2 \ln 1.25/\delta} \sqrt{\log p}}{n\epsilon} \right) \right\}. \end{aligned}$$

Depending on the value of  $\sigma_{ij}$ , we have the following three cases.

**Case 1**  $|\sigma_{ij}| \leq \frac{\gamma}{4} \sqrt{\frac{\log p}{n}} + \frac{\sqrt{2 \ln 1.25/\delta} \sqrt{\log p}}{n\epsilon}$ . For this case, we have

$$\mathbb{P}(A_{ij}) \leq \mathbb{P}\left(|\tilde{\sigma}_{ij} - \sigma_{ij}| > \frac{3\gamma}{4} \sqrt{\frac{\log p}{n}} + \frac{3\sqrt{2 \ln 1.25/\delta} \sqrt{\log p}}{n\epsilon}\right) \leq C_1 p^{-\frac{9}{2}} + 2p^{-\frac{9}{2}}. \quad (6.46)$$

This is due to the followings:

$$\mathbb{P}\left(|\tilde{\sigma}_{ij} - \sigma_{ij}| > \frac{3\gamma}{4} \sqrt{\frac{\log p}{n}} + \frac{3\sqrt{2 \ln 1.25/\delta} \sqrt{\log p}}{n\epsilon}\right) \quad (6.47)$$

$$\leq \mathbb{P}\left(|\sigma_{ij}^* - \sigma_{ij}| > \frac{3\gamma}{4} \sqrt{\frac{\log p}{n}} + \frac{3\sqrt{2 \ln 1.25/\delta} \sqrt{\log p}}{n\epsilon} - |n_{ij}| \right) \quad (6.48)$$

$$= \mathbb{P}\left(B_{ij} \cap \left\{ \frac{3\sqrt{2 \ln 1.25/\delta} \sqrt{\log p}}{n\epsilon} - |n_{ij}| > 0 \right\}\right) \quad (6.49)$$

$$+ \mathbb{P}\left(B_{ij} \cap \left\{ \frac{3\sqrt{2 \ln 1.25/\delta} \sqrt{\log p}}{n\epsilon} - |n_{ij}| \leq 0 \right\}\right) \quad (6.50)$$

$$\leq \mathbb{P}\left(|\sigma_{ij}^* - \sigma_{ij}| > \frac{3\gamma}{4} \sqrt{\frac{\log p}{n}}\right) + \mathbb{P}\left(\frac{2\sqrt{3 \ln 1.25/\delta} \log p}{n\epsilon} \leq |n_{ij}|\right) \leq |n_{ij}| \quad (6.51)$$

$$\leq C_1 P^{-\frac{9}{2}} + 2p^{-\frac{9}{2}}, \quad (6.52)$$

where event  $B_{ij}$  denotes  $B_{ij} = \left\{ |\sigma_{ij}^* - \sigma_{ij}| > \frac{3\gamma}{4} \sqrt{\frac{\log p}{n}} + \frac{2\sqrt{2 \ln 1.25/\delta} \log p}{n\epsilon} - |n_{ij}| \right\}$ , and the last inequality is due to (6.13) and (6.14).

Thus by (6.45), with probability at least  $1 - C_1 p^{-\frac{9}{2}} - 2p^{-\frac{9}{2}}$ , we have

$$|\hat{\sigma}_{ij} - \sigma_{ij}| = |\sigma_{ij}|,$$

which satisfies (6.20).

**Case 2**  $|\sigma_{ij}| \geq 2\gamma\sqrt{\frac{\log p}{n}} + \frac{8\sqrt{2\ln 1.25/\delta}\sqrt{\log p}}{n\epsilon}$ . For this case, we have

$$\mathbb{P}(A_{ij}^c) \leq \mathbb{P}(|\tilde{\sigma}_{ij} - \sigma_{ij}| \geq \gamma\sqrt{\frac{\log p}{n}} + \frac{4\sqrt{2\ln 1.25/\delta}\sqrt{\log p}}{n\epsilon}) \leq C_1 p^{-8} + 2p^{-8},$$

where the proof is the same as (13-17). Thus, with probability at least  $1 - C_1 p^{-\frac{9}{2}} - 2p^{-8}$ , we have

$$|\hat{\sigma}_{ij} - \sigma_{ij}| = |\tilde{\sigma}_{ij} - \sigma_{ij}|. \quad (6.53)$$

Also, by (6.18), (6.20) also holds.

**Case 3** Otherwise,

$$\frac{\gamma}{4}\sqrt{\frac{\log p}{n}} + \frac{\sqrt{2\ln 1.25/\delta}\sqrt{\log p}}{n\epsilon} \leq |\sigma_{ij}| \leq 2\gamma\sqrt{\frac{\log p}{n}} + \frac{8\sqrt{2\ln 1.25/\delta}\sqrt{\log p}}{n\epsilon}.$$

For this case, we have

$$|\hat{\sigma}_{ij} - \sigma_{ij}| = |\sigma_{ij}| \text{ or } |\tilde{\sigma}_{ij} - \sigma_{ij}|. \quad (6.54)$$

When  $|\sigma_{ij}| \leq \gamma\sqrt{\frac{\log p}{n}} + \frac{4\sqrt{2\ln 1.25/\delta}\sqrt{\log p}}{n\epsilon}$ , we can see from (6.18) that with probability at least  $1 - 2p^{-6} - C_1 p^{-8}$ ,

$$|\tilde{\sigma}_{ij} - \sigma_{ij}| \leq \gamma\sqrt{\frac{\log p}{n}} + \frac{4\sqrt{2\ln 1.25/\delta}\sqrt{\log p}}{n\epsilon} \leq 4|\sigma_{ij}|.$$

Thus, (6.20) also holds.

Otherwise when  $|\sigma_{ij}| \leq \gamma\sqrt{\frac{\log p}{n}} + \frac{4\sqrt{2\ln 1.25/\delta}\sqrt{\log p}}{n\epsilon}$ , (6.20) also holds. Thus, Lemma 6.2.3 is true.

### Proof of Theorem 6.2.2

We first show that  $\|\Sigma^+ - \Sigma\|_2 \leq 2\|\hat{\Sigma} - \Sigma\|_2$ . This is due to the following

$$\begin{aligned}\|\Sigma^+ - \Sigma\|_2 &\leq \|\Sigma^+ - \hat{\Sigma}\|_2 + \|\hat{\Sigma} - \Sigma\|_2 \leq \max_{i:\lambda_i \leq 0} |\lambda_i| + \|\hat{\Sigma} - \Sigma\|_2 \\ &\leq \max_{i:\lambda_i \leq 0} |\lambda_i - \lambda_i(\Sigma)| + \|\hat{\Sigma} - \Sigma\|_2 \leq 2\|\hat{\Sigma} - \Sigma\|_2,\end{aligned}$$

where the third inequality is due to the fact that  $\Sigma$  is positive semi-definite.

This means that we only need to bound  $\|\hat{\Sigma} - \Sigma\|_2$ . Since  $\hat{\Sigma} - \Sigma$  is symmetric, we know that  $\|\hat{\Sigma} - \Sigma\|_2 \leq \|\hat{\Sigma} - \Sigma\|_1$  [131]. Thus, it suffices to prove that the bound in (6.21) holds for  $\|\hat{\Sigma} - \Sigma\|_1$ .

We define event  $E_{ij}$  as

$$E_{ij} = \{|\hat{\sigma}_{ij} - \sigma_{ij}| \leq 4 \min\{|\sigma_{ij}|, \gamma \sqrt{\frac{\log p}{n}} + \frac{4\sqrt{2 \ln 1.25/\delta} \sqrt{\log p}}{n\epsilon}\}\}. \quad (6.55)$$

Then, by Lemma 6.2.3, we have  $\mathbb{P}(E_{ij}) \geq 1 - 2C_1 p^{-\frac{9}{2}}$ .

Let  $D = (d_{ij})_{1 \leq i,j \leq p}$ , where  $d_{ij} = (\hat{\sigma}_{ij} - \sigma_{ij}) \cdot I(E_{ij}^c)$ . Then, we have

$$\begin{aligned}\|\hat{\Sigma} - \Sigma\|_1^2 &\leq \|\hat{\Sigma} - \Sigma - D + D\|_1^2 \\ &\leq 2\|\hat{\Sigma} - \Sigma - D\|_1^2 + 2\|D\|_1^2 \\ &\leq 4(\sup_j \sum_{i \neq j} |\hat{\sigma}_{ij} - \sigma_{ij}| I(E_{ij}))^2 + 2\|D\|_1^2 + O(\frac{\log p \log \frac{1}{\delta}}{n\epsilon^2}).\end{aligned} \quad (6.56)$$

We first bound the first term of (6.56). By the definition of  $E_{ij}$  and Lemma 3, we can upper

bounded it by

$$\begin{aligned}
& \left( \sup_j \sum_{i \neq j} |\hat{\sigma}_{ij} - \sigma_{ij}| I(E_{ij}) \right)^2 \\
& \leq 16 \left( \sup_j \sum_{i \neq j} \min\{|\sigma_{ij}|, \gamma \sqrt{\frac{\log p}{n}} + \frac{4\sqrt{2 \ln 1.25/\delta} \sqrt{\log p}}{n\epsilon}\} \right)^2 \\
& \leq 16s^2 \left( \gamma \sqrt{\frac{\log p}{n}} + \frac{4\sqrt{2 \ln 1.25/\delta} \sqrt{\log p}}{n\epsilon} \right)^2 \\
& \leq O(s^2 \frac{\log p \log 1/\delta}{n\epsilon^2}),
\end{aligned} \tag{6.57}$$

where the second inequality is due to the assumption that at most  $s$  elements of  $(\sigma_{ij})_{i \neq j}$  are non-zero.

For the second term in (6.56), we have

$$\begin{aligned}
\mathbb{E} \|D\|_1^2 & \leq p \sum_{ij} d_{ij}^2 = p \mathbb{E} \sum_{ij} [(\hat{\sigma}_{ij} - \sigma_{ij})^2 I(E_{ij}^c \cap \{\hat{\sigma}_{ij} = \tilde{\sigma}_{ij}\}) \\
& \quad + (\hat{\sigma}_{ij} - \sigma_{ij})^2 I(E_{ij}^c \cap \{\hat{\sigma}_{ij} = 0\})] \\
& = p \mathbb{E} \sum_{ij} [(\tilde{\sigma}_{ij} - \sigma_{ij})^2 I(E_{ij}^c) + p \sum_{ij} \mathbb{E} \sigma_{ij}^2 I(E_{ij}^c \cap \{\hat{\sigma}_{ij} = 0\})].
\end{aligned} \tag{6.58}$$

For the first term in (6.58), we have

$$\begin{aligned}
p \sum_{ij} \mathbb{E} \{(\tilde{\sigma}_{ij} - \sigma_{ij})^2 I(E_{ij}^c)\} & \leq p \sum_{ij} [\mathbb{E}(\tilde{\sigma}_{ij} - \sigma_{ij})^6]^{\frac{1}{3}} \mathbb{P}^{\frac{2}{3}}(E_{ij}^c) \\
& \leq Cp \cdot p^2 \frac{1}{n\epsilon^2} p^{-3} = O(\frac{1}{n\epsilon^2}),
\end{aligned} \tag{6.59}$$

where the first inequality is due to Hölder inequality and the second inequality is due to the fact that  $\mathbb{E}(\tilde{\sigma}_{ij} - \sigma_{ij})^8 \leq C_3[\mathbb{E}(\sigma_{ij}^* - \sigma_{ij})^8 + \mathbb{E}n_{ij}^8]$ . Since  $n_{ij}$  is a Gaussian distribution, we have [240]  $\mathbb{E}n_{ij}^8 \leq C_4\sigma_1^8 = O(\frac{1}{n\epsilon})$ . For the first term  $\mathbb{E}(\sigma_{ij}^* - \sigma_{ij})^8$ , since  $x_i$  is sampled from a sub-Gaussian distribution (6.11), by Whittle Inequality (Theorem 2 in [343] or [60]), the quadratic form  $\sigma_{ij}^*$  satisfies  $\mathbb{E}(\sigma_{ij}^* - \sigma_{ij})^8 \leq C_5 \frac{1}{n}$  for some positive constant  $C_5 > 0$ .

For the second term of (6.58), we have

$$\begin{aligned}
& p \sum_{ij} \mathbb{E} \sigma_{ij}^2 I(E_{ij}^c \cap \{\hat{\sigma}_{ij} = 0\}) \\
&= p \sum_{ij} \mathbb{E} \sigma_{ij}^2 I(|\sigma_{ij}| > 4\gamma \sqrt{\frac{\log p}{n}} + \frac{16\sqrt{2 \ln 1.25/\delta} \sqrt{\log p}}{n\epsilon}) \\
&\quad \times I(|\tilde{\sigma}_{ij}| \leq \gamma \sqrt{\frac{\log p}{n}} + \frac{4\sqrt{2 \ln 1.25/\delta} \sqrt{\log p}}{n\epsilon}) \\
&\leq p \sum_{ij} \mathbb{E} \sigma_{ij}^2 I(|\sigma_{ij}| > 4\gamma \sqrt{\frac{\log p}{n}} + \frac{16\sqrt{2 \ln 1.25/\delta} \sqrt{\log p}}{n\epsilon}) \\
&\quad \times I(|\sigma_{ij}| - |\tilde{\sigma}_{ij} - \sigma_{ij}| \leq \gamma \sqrt{\frac{\log p}{n}} + \frac{4\sqrt{2 \ln 1.25/\delta} \sqrt{\log p}}{n\epsilon}) \\
&\leq p \sum_{ij} \sigma_{ij}^2 \mathbb{E} I(|\sigma_{ij}| > 4\gamma \sqrt{\frac{\log p}{n}} + \frac{16\sqrt{2 \ln 1.25/\delta} \sqrt{\log p}}{n\epsilon}) I(|\tilde{\sigma}_{ij} - \sigma_{ij}| \geq \frac{3}{4}|\sigma_{ij}|) \\
&\leq p \sum_{ij} \sigma_{ij}^2 \mathbb{E} I(|\sigma_{ij}| > 4\gamma \sqrt{\frac{\log p}{n}} + \frac{16\sqrt{2 \ln 1.25/\delta} \sqrt{\log p}}{n\epsilon}) I(|\sigma_{ij}^* - \sigma_{ij}| + |n_{ij}| \geq \frac{3}{4}|\sigma_{ij}|) \\
&\leq p \sum_{ij} \sigma_{ij}^2 \mathbb{P}(\{|\sigma_{ij}^* - \sigma_{ij}| \geq \frac{3}{4}|\sigma_{ij}| - |n_{ij}|\} \cap \{|\sigma_{ij}| > 4\gamma \sqrt{\frac{\log p}{n}} + \frac{16\sqrt{2 \ln 1.25/\delta} \sqrt{\log p}}{n\epsilon}\}) \\
&= p \sum_{ij} \sigma_{ij}^2 \mathbb{P}(\{|\sigma_{ij}^* - \sigma_{ij}| \geq \frac{3}{4}|\sigma_{ij}| - |n_{ij}|\} \cap \{|n_{ij}| \leq \frac{1}{4}|\sigma_{ij}|\} \cap \\
&\quad \{|\sigma_{ij}| > 4\gamma \sqrt{\frac{\log p}{n}} + \frac{16\sqrt{2 \ln 1.25/\delta} \sqrt{\log p}}{n\epsilon}\}) + p \sum_{ij} \sigma_{ij}^2 \mathbb{P}(\{|\sigma_{ij}^* - \sigma_{ij}| \geq \frac{3}{4}|\sigma_{ij}| - |n_{ij}|\} \\
&\quad \cap \{|n_{ij}| \geq \frac{1}{4}|\sigma_{ij}|\} \cap \{|\sigma_{ij}| > 4\gamma \sqrt{\frac{\log p}{n}} + \frac{16\sqrt{2 \ln 1.25/\delta} \sqrt{\log p}}{n\epsilon}\}) \tag{6.60}
\end{aligned}$$

$$\begin{aligned}
&\leq p \sum_{ij} \sigma_{ij}^2 \mathbb{P}(\{|\sigma_{ij}^* - \sigma_{ij}| \geq \frac{1}{2}|\sigma_{ij}|\} \cap \{|\sigma_{ij}| > 4\gamma \sqrt{\frac{\log p}{n}} + \frac{16\sqrt{2 \ln 1.25/\delta} \sqrt{\log p}}{n\epsilon}\}) \\
&\quad + p \sum_{ij} \sigma_{ij}^2 \mathbb{P}(\{|n_{ij}| \geq \frac{1}{4}|\sigma_{ij}|\} \cap \{|\sigma_{ij}| > 4\gamma \sqrt{\frac{\log p}{n}} + \frac{16\sqrt{2 \ln 1.25/\delta} \sqrt{\log p}}{n\epsilon}\}). \tag{6.61}
\end{aligned}$$

For the second term of (6.61), by Lemmas 1 and 2 we have

$$\begin{aligned}
& p \sum_{ij} \sigma_{ij}^2 \mathbb{P}(\{|n_{ij}| \geq \frac{1}{4} |\sigma_{ij}|\} \cap \{|\sigma_{ij}| > 4\gamma \sqrt{\frac{\log p}{n}} + \frac{16\sqrt{2 \ln 1.25/\delta} \sqrt{\log p}}{n\epsilon}\}) \\
& \leq p \sum_{ij} \sigma_{ij}^2 \mathbb{P}(|n_{ij}| \geq \gamma \sqrt{\frac{\log p}{n}} + \frac{4\sqrt{2 \ln 1.25/\delta} \log p}{n\epsilon}) \mathbb{P}(|n_{ij}| > \frac{1}{4} \sigma_{ij}) \\
& \leq Cp \sum_{ij} \sigma_{ij}^2 \cdot \exp\left(-\frac{(\gamma \sqrt{\frac{\log p}{n}} + 4\sigma_1 \sqrt{\log p})^2}{2\sigma_1^2}\right) \exp\left(-\frac{\sigma_{ij}^2}{32\sigma_1^2}\right) \\
& \leq C\sigma_1^2 p \cdot p^2 \exp\left(-\frac{\gamma^2 \log p}{2n\sigma_1^2}\right) p^{-8} \\
& \leq C\sigma_1^2 p^{-5} \frac{2n\sigma_1^2}{\gamma^2 \log p} = O\left(\frac{\log 1/\delta}{n\epsilon^2}\right).
\end{aligned}
\tag{6.62}$$

For the first term of (6.61), by Lemma 2 we have

$$\begin{aligned}
& p \sum_{ij} \sigma_{ij}^2 \mathbb{P}(\{|\sigma_{ij}^* - \sigma_{ij}| \geq \frac{1}{2} |\sigma_{ij}|\} \cap \{|\sigma_{ij}| \geq 4\gamma \sqrt{\frac{\log p}{n}}\}) \\
& \leq \frac{p}{n} \sum_{ij} n\sigma_{ij}^2 \exp\left(-n\frac{2\sigma_{ij}^2}{\gamma^2}\right) I(|\sigma_{ij}| \geq 4\gamma \sqrt{\frac{\log p}{n}}) \\
& \leq \frac{p}{n} \sum_{ij} [n\sigma_{ij}^2 \exp\left(-n\frac{\sigma_{ij}^2}{\gamma^2}\right)] \exp\left(-n\frac{\sigma_{ij}^2}{\gamma^2}\right) I(|\sigma_{ij}| \geq 4\gamma \sqrt{\frac{\log p}{n}}) \\
& \leq C \frac{p^3}{n} p^{-16} = O\left(\frac{1}{n}\right).
\end{aligned}
\tag{6.64}$$

Thus in total, we have  $\mathbb{E}\|D\|_1^2 = O\left(\frac{\log 1/\delta}{n\epsilon^2}\right)$ . This means that  $\mathbb{E}\|\hat{\Sigma} - \Sigma\|_1^2 = O\left(\frac{s^2 \log p \log 1/\delta}{n\epsilon^2}\right)$ , which completes the proof.

### Proof of Corollary 6.2.1

By Riesz-Thorin interpolation theorem [101], we have

$$\|A\|_w \leq \max\{\|A\|_1, \|A\|_2, \|A\|_\infty\}$$

for any matrix  $A$  and any  $1 \leq w \leq \infty$ . Since  $\Sigma^+ - \Sigma$  is a symmetric matrix, we have  $\|\Sigma^+ - \Sigma\|_2 \leq \|\Sigma^+ - \Sigma\|_1$  and  $\|\Sigma^+ - \Sigma\|_1 = \|\Sigma^+ - \Sigma\|_\infty$ . Thus, by the proof of Theorem 6.2.2 we get this corollary.

### Proof of Theorem 6.2.5

The proof follows the proof of Theorem 3 in [98]. We will mainly prove the following lemma

**Lemma 6.2.11.** [Theorem 3 in [98]] Under the condition in Theorem 1, for any  $\epsilon$  sequential interactive private channel  $Q$  we have

$$\begin{aligned} & \sum_{j=1}^r [D_{kl}(M_{j,+1}^n \| M_{j,-1}^n) + D_{kl}(M_{j,-1}^n \| M_{j,+1}^n)] \\ & \leq (e^\epsilon - 1)^2 n \sup_{\gamma \in \mathbb{B}_\infty(\mathcal{X})} \sum_{j=1}^r \left( \int_{\mathcal{X}} \gamma(x) (dP_{j,+1} - dP_{j,-1})^2 \right) \end{aligned}$$

By Lemma 6.2.11 we can easily get Theorem 1, which is due to the Pinsker's inequality and Cauchy-Schwartz:

$$\sum_{j=1}^r \|M_{j,+1}^n - M_{j,-1}^n\|_{TV} \leq \frac{1}{2} \sqrt{r} \left( \sum_{j=1}^r D_{kl}(M_{j,+1}^n \| M_{j,-1}^n) + D_{kl}(M_{j,-1}^n \| M_{j,+1}^n) \right)^{\frac{1}{2}}.$$

### Proof of Theorem 6.2.6

By Theorem 6.2.4, we have

$$\mathcal{M}_n^{Nint}(\theta(\mathcal{P}), \Phi \circ \rho, \epsilon) \geq \frac{r\alpha}{4} \min_{j \in [r]} (1 - \|M_{j,+1}^n - M_{j,-1}^n\|_{TV}).$$

By the non-interactivity, we have  $M_{j,a}^n = (\int Q(\cdot|x)dP_{j,a})^{\otimes n}$ . Let  $M_{j,a} = \int Q(\cdot|x)dP_{j,a}$ . By Pinsker inequality, we have the following

$$\|M_{j,+1}^n - M_{j,-1}^n\|_{TV}^2 \leq \frac{1}{2} D_{kl}(M_{j,+1}^n \| M_{j,-1}^n) \quad (6.65)$$

$$\leq \frac{1}{2} D_{\chi^2}(M_{j,+1}^n \| M_{j,-1}^n) \quad (6.66)$$

$$= \frac{1}{2} (D_{\chi^2}(M_{j,+1} \| M_{j,-1}))^n \quad (6.67)$$

$$\leq \frac{1}{2} (\min\{4, e^{2\epsilon}\} (e^\epsilon - 1)^2 \|P_{j,+1} - P_{j,-1}\|_{TV}^2)^n \quad (6.68)$$

$$\leq \frac{1}{2} (\min\{2, \frac{e^{2\epsilon}}{2}\} \epsilon^2 D_{\chi^2}(P_{j,+1} \| P_{j,-1}))^n, \quad (6.69)$$

where (6.65) is due to Pinsker inequality, (6.66) is by the relation between KL-divergence and  $\chi^2$ -divergence  $D_{kl}(P \| Q) \leq \log(1 + D_{\chi^2}(P \| Q)) \leq D_{\chi^2}(P \| Q)$  [283], (6.67) is due to the non-interactivity, (6.69) is by Pinsker inequality and inequalities  $(e^\epsilon - 1)^2 \leq 2\epsilon^2$  and  $e^{2\epsilon} \leq 2$ . Next, we prove (6.68).

### **Lemma 6.2.12.**

$$D_{\chi^2}(M_{j,+1} \| M_{j,-1}) \leq \min\{4, e^{2\epsilon}\} (e^\epsilon - 1)^2 \|P_{j,+1} - P_{j,-1}\|_{TV}^2.$$

*Proof.* W.l.o.g, we can assume that the density function of  $M_{j,a}$  is  $m_{j,a}(z) = \int q(z|x)dP_{j,a}$  and  $q(\cdot|x)$  is the density function of  $Q(\cdot|x)$ . By the definition, we have

$$D_{\chi^2}(M_{j,+1} \| M_{j,-1}) = \int \frac{(m_{j,+1}(z) - m_{j,-1}(z))^2}{m_{j,-1}(z)} dz \quad (6.70)$$

$$\leq \int \frac{c_\epsilon^2 \inf_x q^2(z|x)(e^\epsilon - 1)^2 \|P_{j,+1} - P_{j,-1}\|_{TV}^2}{\int q(z|x)dP_{j,a}} dz \quad (6.71)$$

$$\leq c_\epsilon^2 (e^\epsilon - 1)^2 \|P_{j,+1} - P_{j,-1}\|_{TV}^2 \int \inf_x q(z|x) dz \\ \leq c_\epsilon^2 (e^\epsilon - 1)^2 \|P_{j,+1} - P_{j,-1}\|_{TV}^2, \quad (6.72)$$

where  $c_\epsilon = \min\{2, e^\epsilon\}$ , (6.70) is by the definition of  $\chi^2$ -divergence, (6.71) is by Lemma 3

in [98] and (6.72) is due to the fact that  $\int \inf_x q(z|x) dz \leq 1$ .

□

### Proof of Corollary 6.2.2

The key observation is that the distributions  $P_{j,a}^n$  can be represented by a linear combination of  $\{\bar{P}_{j,a,b,c}^n\}_{b,c \in T_{-j}}$ , where the set  $T_{-j}$  is

$$\begin{aligned} T_{-j} &= \{0, 1\}^{r-1} \otimes \Lambda_{-i} \\ &= \{(b, c) : \exists \tau \in T \text{ s.t } v_{-i}(\tau) = b \text{ and } \lambda_{-i}(\tau) = c\}. \end{aligned}$$

That is,  $P_{j,a}^n = \sum_{(b,c) \in T_{-j}} w_{b,c} \bar{P}_{j,a,b,c}^n$ , where  $w_{b,c} = \frac{D_{\Lambda_j(a,b,c)}}{2^{r-1} D_\Lambda}$  (note that since  $D_{\Lambda_j(a,b,c)}$  is independent of  $a$ , we omit it). Also,  $\sum_{(b,c) \in T_{-j}} w_{b,c} = 1$ . Thus,  $P_{j,a}^n$  can be seen as an average over  $(b, c)$ . The same also holds for  $M_{j,a}^n$ .

By the convexity of total variation norm and Lemma 6.2.12, we have

$$\begin{aligned} \|M_{j,+1}^n - M_{j,-1}^n\|_{TV} &\leq \sum_{(b,c) \in T_{-j}} w_{b,c} \|\bar{M}_{j,+1,b,c}^n - \bar{M}_{j,-1,b,c}^n\|_{TV} \\ &= \text{Average}_{b,c} \|\bar{M}_{j,+1,b,c}^n - \bar{M}_{j,-1,b,c}^n\|_{TV}. \end{aligned}$$

By a similar argument given in the proof of Theorem 6.2.6, we get

$$\begin{aligned} \|\bar{M}_{j,+1,b,c}^n - \bar{M}_{j,-1,b,c}^n\|_{TV}^2 &\leq D_{\chi^2}(\bar{M}_{j,+1,v_{-j},\lambda_{-j}} \|\bar{M}_{j,-1,v_{-j},\lambda_{-j}})^n \\ &\leq \frac{1}{2} (\min\{2, \frac{e^{2\epsilon}}{2}\} \epsilon^2 D_{\chi^2}(\bar{P}_{j,+1,b,c} \|\bar{P}_{j,-1,b,c}))^n \\ &\leq \frac{1}{2} (\epsilon^2 D_{\chi^2}(\bar{P}_{j,+1,b,c} \|\bar{P}_{j,-1,b,c}))^n. \end{aligned}$$

Thus, by the inequality  $\text{Average}_{b,c} \|\bar{M}_{j,+1,b,c}^n - \bar{M}_{j,-1,b,c}^n\|_{TV}^2 \leq \text{Average}_{b,c} \|\bar{M}_{j,+1,b,c}^n - \bar{M}_{j,-1,b,c}^n\|_{TV}^2$ , we have the proof.

### Proof of Lemma 6.2.6

We first bound the term of  $\|\Sigma(\tau)\|_2$ . Note that since  $\Sigma(\tau)$  is symmetric, we have  $\|\Sigma(\tau)\|_2 \leq \|\Sigma(\tau)\|_1$ . By the construction of  $\Sigma(\tau)$ , we can see that the  $\ell_1$  norm of each column in  $\Sigma(\tau)$  is less than  $1 + 2k\alpha_{n,p,\epsilon} \leq 1 + s\gamma\sqrt{\frac{\log p}{n\epsilon^2}}$ . Thus, we have  $\|\Sigma(\tau)\|_2 \leq c + cs\gamma\sqrt{\frac{\log p}{n\epsilon^2}}$ .

We need  $\mathcal{N}(0, \Sigma(\tau))$  satisfying (6.35). By [342], we know that it is sufficient to have  $\|\Sigma(\tau)\|_2 \leq \rho$ .

Let  $\Sigma(\tau) = V^T Q V$  be the SVD decomposition and  $Q = \text{diag}(\lambda_1, \dots, \lambda_p)$ . Then, for  $X \sim \mathcal{N}(0, \Sigma(\tau))$ , we have  $VX \sim \mathcal{N}(0, Q)$ . Thus,  $\|X\|_2^2 = \|VX\|_2^2 \leq \|\Sigma(\tau)\|_2 Y$ , where  $Y$  is a  $\chi_p^2$  random variable. For the  $\chi^2$ -distribution, we have the following concentration bound.

**Lemma 6.2.13** ([189]). If  $z \sim \chi_n^2$ , then

$$\mathbb{P}[z - n \geq 2\sqrt{nx} + 2x] \leq \exp(-x).$$

Thus, with probability at least  $1 - \exp(-p)$ , we have  $Y \leq 5p$ . This means that, to ensure  $\|X\|_2 \leq 1$ , it is sufficient to have  $5p\|\Sigma(\tau)\|_2 \leq 1$ . Thus, we need that

$$c + cs\gamma\sqrt{\frac{\log p}{n\epsilon^2}} \leq \min\{\rho, \frac{1}{5p}\}. \quad (6.73)$$

Taking  $c = \min\{\rho/2, \frac{1}{10p}\}$  and choosing a small enough  $\gamma \leq \frac{\sqrt{C}}{2}$ , we can get the proof.

### Proof of Lemma 6.2.7

Let the vector  $v = (v_i)_{1 \leq i \leq p}$  be a  $p$ -vector with  $v_i = 0$  for  $1 \leq i \leq p-r$  and  $v_i = 1$  for  $p-r+1 \leq i \leq p$ . Denote  $w = (w_i)_{1 \leq i \leq p} = (\Sigma(\tau) - \Sigma(\tau'))v$ . Note that for each  $i$ , if  $|v_i(\tau) - v_i(\tau')| = 1$ , then we have  $|w_i| = k\alpha_{n,p,\epsilon}$ . Then there are at least  $H(v_i(\tau), v_i(\tau'))$

number of elements  $w_i$  with  $|w_i| = k\alpha_{n,p,\epsilon}$ , which implies

$$\|(\Sigma(\tau) - \Sigma(\tau'))v\|_2^2 \geq H(v_i(\tau), v_i(\tau'))(k\alpha_{n,p,\epsilon})^2.$$

Since  $\|v\|_2^2 \leq p$ , we have

$$\begin{aligned} \|\Sigma(\tau) - \Sigma(\tau')\|_2^2 &\geq \frac{\|(\Sigma(\tau) - \Sigma(\tau'))v\|_2^2}{\|v\|_2^2} \\ &\geq \frac{H(v_i(\tau), v_i(\tau'))(k\alpha_{n,p,\epsilon})^2}{p} \end{aligned}$$

Thus,  $\alpha \geq \frac{(k\alpha_{n,p,\epsilon})^2}{p}$ .

### Proof of Lemma 6.2.8

Our proof is similar to the proof of Lemma 6 in [60] with difference parameters. Here we only give a sketch of the proof.

Without loss of generality, we only consider the case where  $j = 1$ . And we denote the density function of  $\bar{P}_{1,a,v_{-1},\lambda_{-1}}$  be  $\bar{p}_{1,a,v_{-1},\lambda_{-1}}$ . Also, we have

$$D_{\chi^2}(\bar{P}_{1,+1,v_{-1},\lambda_{-1}} \| \bar{P}_{1,-1,v_{-1},\lambda_{-1}}) = \int \frac{\bar{p}_{1,1,v_{-1},\lambda_{-1}}^2(x)}{\bar{p}_{1,-1,v_{-1},\lambda_{-1}}(x)} dx - 1.$$

By the definition, we know that the covariance matrix of the distribution  $\bar{P}_{1,-1,v_{-1},\lambda_{-1}}$  has the form

$$\Sigma_0 = \begin{pmatrix} c & 0_{1 \times (p-1)} \\ 0_{(p-1) \times 1} & S_{(p-1) \times (p-1)} \end{pmatrix} \quad (6.74)$$

Here  $S_{(p-1) \times (p-1)} = (s_{ij})_{2 \leq i,j \leq p}$  is a symmetric matrix uniquely determined by  $(v_{-1}, \lambda_{-1})$  where for  $i \leq j$ ,

$$s_{ij} = \begin{cases} 1, & i = 1 \\ c\alpha_{n,p,\epsilon}, & v_i = \lambda_i(j) = 1 \\ 0, & \text{otherwise} \end{cases} \quad (6.75)$$

Let

$$\Lambda_1(m) = \{a \in B : \exists \tau \in T \text{ s.t. } \lambda_1(\tau) = a, \lambda_{-1} = m\}$$

which gives the rest of all possible values of the first row with the rest of the rows fixed, that  $\lambda_{-1}(\tau) = m$ . Let  $n_{\lambda_{-1}}$  be the number of columns of  $\lambda_{-1}$  with the column sum equal to  $2k$  for which the first row has no choice but to take value 0 in this column. Set  $p_{\lambda_{-1}} = r - n_{\lambda_{-1}}$ . We have  $p_{\lambda_{-1}} \geq \frac{p}{4} - 1$ . Since  $2kn_{\lambda_{-1}} \leq rk$ , the total number of 1s in the upper triangular matrix by the construction of the parameter set, we thus have  $n_{\lambda_{-1}} \leq \frac{r}{2}$ , thus  $p_{\lambda_{-1}} = r - n_{\lambda_{-1}} \geq \frac{r}{2} \geq \frac{p}{4} - 1$ . Thus we have  $|\Lambda_1(\lambda_{-1})| = \binom{p_{\lambda_{-1}}}{k}$ . Then from the definition, we have  $\bar{P}_{1,1,v_{-1},\lambda_{-1}}$  is an average of  $\binom{p_{\lambda_{-1}}}{k}$  multivariate normal distribution with the covariance matrix has the following form:

$$\begin{pmatrix} c & \mathbf{r}_{1 \times (p-1)} \\ \mathbf{r}_{(p-1) \times 1} & S_{(p-1) \times (p-1)} \end{pmatrix} \quad (6.76)$$

With  $\|\mathbf{r}\|_0 = k$  with non-zero elements of  $\mathbf{r}$  equal  $c\alpha_{n,p,\epsilon}$  and the submatrix  $S_{(p-1) \times (p-1)}$  is the same as the ones in  $\Sigma_0$  in (6.74).

We have the following lemma, given by [60]

**Lemma 6.2.14.** Let  $g_i$  be the density function of  $\mathcal{N}(0, \Sigma_i)$  for  $i = 0, 1, 2$ , then we have

$$\int \frac{g_1 g_2}{g_0} = [\det(I - \Sigma_0^{-2}(\Sigma_1 - \Sigma_0)(\Sigma_2 - \Sigma_0))]^{-\frac{1}{2}}. \quad (6.77)$$

Let  $\Sigma_0$  defined above and determined by  $v_{-1}, \lambda_{-1}$ . Let  $\Sigma_1$  and  $\Sigma_2$  be the form above with the first row  $\lambda_1, \lambda'_1$ , respectively. Set

$$R_{\lambda_1, \lambda'_1}^{v_{-1}, \lambda_{-1}} = -\log \det(I - \Sigma_0^{-2}(\Sigma_0 - \Sigma_1)(\Sigma_0 - \Sigma_2)). \quad (6.78)$$

Now we denote the average as the expectation, then we have

$$\mathbb{E}_{v_{-1}, \lambda_{-1}}(D_{\chi^2}(\bar{P}_{1,+1, v_{-1}, \lambda_{-1}} \| \bar{P}_{1,-1, v_{-1}, \lambda_{-1}}))^n \quad (6.79)$$

$$\leq \mathbb{E}_{v_{-1}, \lambda_{-1}}[\mathbb{E}_{(\lambda_1, \lambda'_1) | \lambda_{-1}}[\exp(\frac{n}{2}(R_{\lambda_1, \lambda'_1}^{v_{-1}, \lambda_{-1}} - 1))]] \quad (6.80)$$

$$\leq \mathbb{E}_{v_{-1}, \lambda_{-1}}[\mathbb{E}_{(\lambda_1, \lambda'_1) | \lambda_{-1}}[\exp(\frac{n}{2}(R_{\lambda_1, \lambda'_1}^{v_{-1}, \lambda_{-1}})) - 1]] \quad (6.81)$$

$$= \mathbb{E}_{\lambda_1, \lambda'_1}[\mathbb{E}_{(v_{-1}, \lambda_{-1}) | (\lambda_1, \lambda'_1)}[\exp(\frac{n}{2}(R_{\lambda_1, \lambda'_1}^{v_{-1}, \lambda_{-1}})) - 1]] \quad (6.82)$$

where  $\lambda_1$  and  $\lambda'_1$  are independent and uniformly distributed over  $\Lambda_1(\lambda_{-1})$  for given  $\lambda_{-1}$ , and the distribution of  $(v_{-1}, \lambda_{-1})$  given  $(\lambda_1, \lambda'_1)$  is inform over  $T_{-1}(\lambda_1, \lambda_{-1})$ , where

$$\begin{aligned} T_{-1}(a_1, a_2) = & \{-1, +1\}^{r-1} \otimes \{c \in \Lambda_{-1} : \exists \tau_i \in T, i = 1, 2 \\ & \text{s.t. } \lambda_1(\tau_i) = a_i, \lambda_{-1}(\tau_i) = v\} \end{aligned}$$

### Proof of Theorem 6.2.7

By Corollary 6.2.2, Lemma 6.2.8 and 6.2.7 we have

$$\begin{aligned} \mathcal{M}_n^{Nint}(\psi(\theta), \Phi \circ \rho, \epsilon) &\geq \frac{r\alpha}{4} \times \min_{1 \leq j \leq r} \\ &(1 - \sqrt{\frac{\epsilon^{2n}}{2} \text{Average}_{v_{-j}, \lambda_{-j}}(D_{\chi^2}(\bar{P}_{j,+1, v_{-j}, \lambda_{-j}} \| \bar{P}_{j,-1, v_{-j}, \lambda_{-j}}))^n}) \\ &\geq \frac{p}{2} \frac{k^2 \alpha_{n,p,\epsilon}^2}{p} (1 - \sqrt{\frac{\epsilon^{2n}}{2} \frac{3}{4} \frac{1}{\epsilon^{2n}}}) \\ &\geq \Omega(\frac{s^2 \log p}{n \epsilon^2}). \end{aligned}$$

### Proof of Corollary 6.2.3

First, by the Riesz-Thorin Interpolation Theorem [57], we know that for every symmetric matrix  $M$ ,  $\|M\|_2 \leq \|M\|_w$  for all  $w \in [1, \infty]$ . Thus we have under  $\ell_w$  norm, by Lemma 6 we always have  $\alpha_{n,p,\epsilon} \geq \frac{(k\alpha_{n,p,\epsilon})^2}{p}$ , also since the term  $\text{Average}_{v_{-j}, \lambda_{-j}}(D_{\chi^2}(\bar{P}_{j,+1, v_{-j}, \lambda_{-j}} \| \bar{P}_{j,-1, v_{-j}, \lambda_{-j}}))$

is independent on the norm, so we have the corollary.

## 6.3 Differentially Private Sparse Inverse Covariance Matrix Estimation

Estimating the inverse covariance matrix (also called precision matrix) in high dimensional space is a fundamental problem in statistics and finds applications in many fields such as machine learning, signal processing, computational biology, etc [355]. It provides a good way for discovering the interactions among variables in high dimensional datasets, especially those from genetics, medicine, and healthcare. The inverse covariance matrix is also a natural way for parameterizing the Gaussian graphical model. One problem that often occurs in applying such a model is how to deal with sensitive data. For example, datasets related to gene expression may contain private information of individuals. Thus, it becomes a challenge for estimating the inverse covariance while preserving privacy. In this part, we study the problem under the differential privacy model, and provide some results on this problem.

**Differentially Private Sparse Inverse Covariance Estimation** Let  $\{x_1, \dots, x_n\}$  be  $n$  instances sampled from a Gaussian distribution  $\mathcal{N}(0, \Sigma)$ , where each instance  $x_i \in \mathbb{R}^d$  for  $i \in [n]$  and  $\Sigma \in \mathbb{R}^{d \times d}$  is the covariance matrix. The inverse covariance problem is to recover  $\Sigma^{-1}$  in a high dimensional setting, where  $n \ll d$ . Note that if  $n \geq d$ , we can solve the problem by optimizing  $\Theta^* = S^{-1} = \arg \min_{\Theta \in \mathcal{S}_{++}^d} -\log \det \Theta + \langle S, \Theta \rangle$ , where  $S = \frac{1}{n} \sum_{i=1}^n x_i x_i^T$  is the empirical covariance. But in a high dimensional setting, the above optimization problem is ill-posed, since  $S$  is rank-deficient. To make it well-defined, we borrow an idea in LASSO and use an  $\ell_1$  norm regularization in the objective function, which

assumes that  $\Theta^*$  is sparse. Thus, the objective function becomes the following:

$$\Theta_\rho^* = \arg \min_{\Theta \in \mathcal{S}_{++}^d} \{-\log \det \Theta + \langle S, \Theta \rangle + \rho \|\Theta\|_1\}, \quad (6.83)$$

where  $\rho > 0$  is the penalty parameter,  $\langle S, \Theta \rangle = \text{tr}(S\Theta^T)$ , and  $\|\Theta\|_1 = \sum_{i,j} |\Theta_{i,j}|$ .

Under the differential privacy model, our problem is to obtain  $\Theta_{\text{priv}}$  so that  $\|\Theta_{\text{priv}} - \Theta_\rho^*\|_F$  is as small as possible.

Below is summary of our main contributions to this problem. Table 6.3 lists the related error bounds.

Perturbation method	Mechanism	Error upper bound	Keeping covariance matrix semidefinite?	Type of privacy
Output	Wishart	$O\left(\frac{\log dd^4}{n\epsilon\rho^2}\right)$	Yes	$\epsilon$ -DP
Covariance	Wishart	$O\left(\frac{\log dd^{\frac{3}{2}} \max\{\ \Theta_\rho^*\ _2^2, \ \Theta_{\text{priv}}\ _2^2\}}{n\epsilon}\right)$	Yes	$\epsilon$ -DP
Covariance	Laplacian	$O\left(\frac{d^2 \max\{\ \Theta_\rho^*\ _2^2, \ \Theta_{\text{priv}}\ _2^2\}}{n\epsilon}\right)$	No	$\epsilon$ -DP
Covariance	Wishart	$O\left(\frac{\max\{\ \Theta_\rho^*\ _2^2, \ \Theta_{\text{priv}}\ _2^2\} \ln(1/\delta) d^{\frac{3}{2}}}{n\epsilon^2}\right)$	Yes	$(\epsilon, \delta)$ -DP
Covariance	Gaussian	$O\left(\frac{d\sqrt{\ln(\frac{1}{\delta})} \max\{\ \Theta_\rho^*\ _2^2, \ \Theta_{\text{priv}}\ _2^2\}}{\epsilon n}\right)$	No	$(\epsilon, \delta)$ -DP
Covariance	Gaussian	$O\left(\frac{d\sqrt{\ln(\frac{1}{\delta})} \max\{\ \Theta_\rho^*\ _2^2, \ \Theta_{\text{priv}}\ _2^2\}}{\epsilon\sqrt{n}}\right)$	No	$(\epsilon, \delta)$ -Local DP

Table 6.3: The error upper bound of methods in the paper, which is measured by  $\|\Theta_{\text{priv}} - \Theta_\rho^*\|_F$ , here we assume the  $\ell_2$ -norm of each  $x_i$  is bounded by 1.

- We first present an output-perturbation algorithm (See Algorithm 6.3.50) based on the sensitivity of (6.83). Unlike the commonly used Laplacian or Gaussian mechanisms in differential privacy, we adopt the Wishart distribution to preserve the positive definite property for the resulting matrix.
- To reduce the error bound of the above algorithm, we then introduce a general method by perturbing the covariance matrix, and analyze the error upper bound for different perturbing matrices.
- We show that our covariance perturbation method can also be extended to distributed settings and the local differential privacy model. In the local differential privacy model,

each individual discloses his/her personal information through some differentially private algorithms and shares with the public only the output; a server then collects the disclosed information and analyzes it. This is quite different from the central differential privacy model, where institutions release databases of information or answer queries of such databases.

- Finally, we evaluate the performance of our algorithms using both synthetic and real world datasets. Experimental results confirm our theoretical analysis.

To the best of our knowledge, this is the first paper studying the sparse inverse covariance estimation problem under the differential privacy model.

### 6.3.1 Related Work

There is a large number of papers studying the sparse inverse covariance estimation problem from different perspectives. For example, [246, 56] investigated the issue of statistical consistency of the problem, and [209, 250, 81, 150] considered how to efficiently solve the associated optimization problem (6.83).

Perhaps, the most closely related work to ours is differentially private PCA, since it also relies on random matrices to preserve privacy. For example, [162, 158] used the Wishart mechanism to achieve  $\epsilon$ -differentially private PCA, and [68, 106, 321] adopted the Gaussian mechanism to analyze the optimal bound of PCA under the  $(\epsilon, \delta)$ -differential privacy model. Note that although our paper uses the same mechanisms (as the aforementioned results), the way for analyzing the error bound is quite different. While the above results mainly relied on techniques in linear algebra, ours is based on some optimization techniques (due to the  $\ell_1$  regularization and the positive definite requirement for the resulting matrix). Thus, existing approaches/techniques cannot be used to analyze our problem.

### 6.3.2 Preliminaries

**Notations** For a matrix  $X$ , we let  $\sigma_l(X)$  denote the  $l$ -th largest singular value of  $X$ ,  $\text{tr}(X)$  denote the trace of  $X$ ,  $\det X$  denote the determinant of  $X$ ,  $\|X\|_F$  denote the Frobenius norm, and  $\|X\|_2$  denote the spectral norm. Also, we let  $\mathcal{S}_{++}^n$  be the set of  $n \times n$  positive definite matrices. We write  $0 \preceq X$  to mean that  $X$  is positive semidefinite and  $X \preceq Y$  to mean that  $0 \preceq Y - X$ . We use  $I_d$  to denote the identity  $d \times d$  matrix. Note that we assume that the  $\ell_2$  norm of each data record  $x_i$  is bounded by 1.

Before introducing Wishart mechanism used in this part, we first introduce Wishart distribution.

**Definition 6.3.1.** A  $d \times d$  random symmetric positive definite matrix  $W$  is said to have a Wishart distribution  $W \sim \mathcal{W}_d(m, C)$  if its probability density function is

$$p(W) = \frac{(\det W)^{\frac{m-d-1}{2}}}{2^{\frac{md}{2}} (\det C)^{\frac{m}{2}} \Gamma_d(\frac{m}{2})} \exp(-\frac{1}{2} \text{tr}(C^{-1}W)), \quad (6.84)$$

where  $m > d - 1$  and  $C$  is a  $d \times d$  positive definite matrix.

One property of Wishart distribution is its multivariate extension of the  $\chi^2$ -distribution. More specifically, if  $v_1, v_2 \dots, v_m$  are i.i.d sampled from a  $d$ -dimensional multivariate Gaussian distribution  $\mathcal{N}(0, C)$ ,  $\sum_{i=1}^m v_i v_i^T \sim \mathcal{W}_d(m, C)$ . We will use this property in distributed settings.

Next we show how to select the parameters  $m$  and  $C$  to ensure differential privacy.

**Lemma 6.3.1** (( $\epsilon, \delta$ )-differential privacy [254]). Fix  $\epsilon \in (0, 1)$  and  $\delta \in (0, \frac{1}{e})$ . For a fixed constant  $B > 0$ , let  $A$  be an  $n \times d$  matrix, where each row of  $A$  has bounded  $\ell_2$ -norm  $B$ . Let  $N$  be a matrix sampled from  $\mathcal{W}(m, B^2 I_d)$  for  $m \geq d + \frac{14}{\epsilon^2} \ln(\frac{4}{\delta})$ . Then, outputting  $X = A^T A + N$  is  $(\epsilon, \delta)$ -differentially private.

**Lemma 6.3.2** ( $\epsilon$ -differential privacy [162]). Fix  $\epsilon > 0$  and let  $A$  be an  $n \times d$  matrix, where each row of  $A$  has bounded  $\ell_2$ -norm of  $B$ . Let  $N \sim \mathcal{W}_d(d+1, C)$ , where  $C = \frac{3}{2n\epsilon} B^2 I_d$ .

Then, outputting  $X = A^T A + N$  is  $\epsilon$ -differentially private.

Below we list some theorems related to the tail bound of a Wishart distribution, which will be used later in our error bound analysis.

**Lemma 6.3.3** ([254]). Fix  $\delta' \in (0, \frac{1}{e})$ , and a random matrix  $X \sim \mathcal{W}_d(m, V)$ , where  $m > (\sqrt{d} + \sqrt{2 \log \frac{2}{\delta'}})^2$ . Then, with probability at least  $1 - \delta'$ , the following holds for every  $j = 1, \dots, d$

$$\sigma_j(X) \in (\sqrt{m} \pm (\sqrt{d} + \sqrt{2 \log \frac{2}{\delta'}}))^2 \sigma_j(V). \quad (6.85)$$

For the tail bound of the noise added by an  $\epsilon$ -differential private algorithm, we have the following lemma.

**Lemma 6.3.4** ([370]). If  $X \sim \mathcal{W}_d(m, V)$ , then with probability at least  $1 - 2d \exp(-\theta)$  for any  $\theta \geq 0$ , the following holds for each  $l = 1, \dots, d$

$$|\sigma_l(\frac{1}{m}X) - \sigma_l(V)| \leq (\sqrt{\frac{2\theta k_l^2(r+1)}{m}} + \frac{2\theta k_l r}{m}) \sigma_l(V), \quad (6.86)$$

where  $r = \frac{\text{tr}(V)}{\sigma_1(V)}$  and  $k_l = \frac{\sigma_1(V)}{\sigma_l(V)}$ .

If taking  $V = B^2 I_d$ ,  $m = d + 1$ , and  $\theta = \log \frac{2d}{\delta'}$ , Lemma 6.3.4 tells us that with probability at least  $1 - \delta'$ , we have  $\sigma_l(X) \leq O(d \log \frac{d}{\delta'} B^2)$  for each  $l = 1, \dots, d$ . This means that there is a factor of  $\log d$  compared with Lemma 6.3.3.

### 6.3.3 Sparse Inverse Covariance Estimation

Before presenting our methods, we first introduce some properties of the optimization problem (6.83). For  $\rho > 0$ , the problem is strongly convex and thus has a unique optimal solution  $\Theta_\rho^*$ , which satisfies the following condition.

**Lemma 6.3.5** ([81, 209]). The solution of (6.83),  $\Theta_\rho^*$ , satisfies that  $\alpha I_d \preceq \Theta_\rho^* \preceq \beta I_d$ , for

$$\alpha = \frac{1}{\|S\|_2 + \rho d}, \beta = \min\left\{\frac{d - \alpha \text{tr}(S)}{\rho}, \gamma\right\}, \quad (6.87)$$

where  $\gamma$  has the following value

$$\gamma = \begin{cases} 2\|(S + \frac{\rho}{2}I_d)^{-1}\|_1 - \text{tr}((S + \frac{\rho}{2}I_d)^{-1}), & \text{if } S \notin \mathcal{S}_{++}^d \\ \min\{\|S^{-1}\|_1, (d - \rho\sqrt{d}\alpha)\|S^{-1}\|_2 - (d - 1)\alpha\}, & \text{else.} \end{cases} \quad (6.88)$$

**Proximal Operator** Now we consider a general optimization problem with the following form

$$\min_{x \in \mathcal{X}} F(x) = f(x) + g(x), \quad (6.89)$$

where  $\mathcal{X}$  is a Hilbert space with inner product  $\langle \cdot, \cdot \rangle$  and associated with norm  $\|\cdot\|$ ,  $f$  is a continuously differentiable, convex function, and  $g$  is a lower semi-continuous convex function. For a given lower semi-continuous convex function  $g$ , the proximity of  $g$ , denoted by  $\text{prox}_g : \mathcal{X} \rightarrow \mathcal{X}$ , is given by

$$\text{prox}_g(x) = \arg \min_{y \in \mathcal{X}} \{g(y) + \frac{1}{2}\|x - y\|^2\}.$$

One of the basic results in [77] says that for every  $\eta > 0$ ,  $x^* \in \mathcal{X}$  is an optimal solution of (6.89) if and only if

$$x^* = \text{prox}_{\eta g}(x^* - \eta \nabla f(x^*)). \quad (6.90)$$

For our problem (6.83),  $f(\Theta) = -\log \det \Theta + \langle S, \Theta \rangle$  and  $g(\Theta) = \rho\|\Theta\|_1$ . Since  $f(\Theta)$  is continuously differentiable in  $\mathcal{S}_{++}^d$ ,  $\Theta_\rho^*$  can be determined by

$$\Theta_\rho^* = \text{prox}_{\eta g}(\Theta_\rho^* - \eta(S - \Theta_\rho^{*-1})). \quad (6.91)$$

### 6.3.4 Output Perturbation Method

In this section, we present an  $\epsilon$ -differentially private algorithm based on the output perturbation strategy (see Algorithm 6.3.50 for details), and analyze the sensitivity and stability of the problem (6.83). Although the method has some undesirable features, the error bound

analysis and the guarantee of differential privacy are useful for our later methods.

---

**Algorithm 6.3.50** Output Perturbation

---

**Input:**  $D = \{x_i\}_{i=1}^n$ ,  $S = \frac{1}{n} \sum_{i=1}^n x_i x_i^T \in \mathbb{R}^{d \times d}$ , where the  $\ell_2$ -norm of each row  $x_i$  is bounded by 1,  $\rho > 0$ .  $\epsilon > 0$  is the privacy parameter.

1: Compute

$$\Theta_\rho^* = \arg \min_{\Theta \in \mathcal{S}_{++}^d} \{-\log \det \Theta + \langle S, \Theta \rangle + \rho \|\Theta\|_1\},$$

**return**  $\tilde{\Theta}_\rho^* = \Theta_\rho^* + N$ , where  $N \sim \mathcal{W}_d(d+1, C)$ ,  $C = \frac{d^{\frac{5}{2}}}{n\epsilon\rho^2} I_d$ .

---

**Theorem 6.3.1** (Privacy guarantee). For any  $\epsilon > 0$ , Algorithm 6.3.50 is  $\epsilon$ -differentially private.

By Lemma 6.3.4, we have the following error upper bound.

**Theorem 6.3.2.** For Algorithm 6.3.50, with probability at least  $1 - \delta$  for any  $0 < \delta < 1$ , we have

$$\|\tilde{\Theta}_\rho^* - \Theta_\rho^*\|_F \leq O\left(\frac{\log \frac{d}{\delta} d^4}{n\epsilon\rho^2}\right), \quad (6.92)$$

where  $\Theta_\rho^*$  is the optimal solution of the original problem (6.83).

**Remark 6.3.1.** Note that in Algorithm 6.3.50, a Wishart matrix needs to be added to the output to ensure that the resulting matrix is positive definite (as required by problem (6.83)). Since other random matrices, such as symmetric Laplacian matrices, may not be positive definite [117], adding them to the output may not yield the desired solution.

Although Algorithm 6.3.50 provides an  $\epsilon$ -differentially private algorithm for the inverse covariance estimation problem. It also leaves quite a few unresolved issues. Firstly, from Theorem 6.3.2, we know that the error bound heavily depends on the dimensionality (*i.e.*,  $d^4 \log d$ ), which could be too large for high dimensional datasets. Thus, a natural question is whether the error bound in (6.92) can be further reduced. Secondly, for many problems, the error bound of an  $(\epsilon, \delta)$ -differentially private algorithm is often lower than that of an  $\epsilon$ -differentially private algorithm (*e.g.*, Differentially Private Empirical Risk Minimization

[29, 326]). Thus, an interesting question is whether the problem considered in this paper also follows the same pattern. Thirdly, the goal of the sparse inverse covariance estimation problem is to obtain a sparse estimator. However, the output perturbation strategy in Algorithm 6.3.50 could destroy the sparsity property of the resulting estimator. Thus, it is desirable to obtain a solution which always yields a sparse private estimator. Below we will address the three issues by proposing a covariance perturbation method.

### 6.3.5 Covariance Perturbation Method

As shown in (6.99), the sensitivity of problem (6.83) is high (since  $\beta$  is often large). This means that we need to add a large amount of noise in Algorithm 6.3.50 to ensure the  $\epsilon$ -differential privacy. To deal with this problem, along with the aforementioned issues, we propose in this section a general method which perturbs the empirical covariance  $S$  (see Algorithm 6.3.51), instead of the output. This allows us to significantly reduce the amount of noise that needs to be added. Also, it can be implemented by using different kinds of random matrices  $N$ . To compare the performance for different mechanisms, we analyze the error bound for each of them. We first determine the relationship between error bound and the noise matrix  $N$ .

---

#### Algorithm 6.3.51 Covariance Perturbation

**Input:**  $D = \{x_i\}_{i=1}^n$ , where the  $\ell_2$ -norm of each row  $x_i$  is bounded by 1,  $\rho > 0$ .  $\epsilon, \delta \geq 0$  are the privacy parameters.

- 1: Let  $S = \frac{1}{n} \sum_{i=1}^n x_i x_i^T$ ; sample a symmetric matrix  $N \in \mathbb{R}^{d \times d} \sim \mathcal{P}$ , which makes  $S + N$   $\epsilon$ - or  $(\epsilon, \delta)$ -differentially private. Let  $\tilde{S} = S + N$ .
- 2: Compute and return

$$\hat{\Theta}_\rho^* = \arg \min_{\Theta \in \mathcal{S}_{++}^d} \{-\log \det \Theta + \langle \tilde{S}, \Theta \rangle + \rho \|\Theta\|_1\}.$$


---

**Theorem 6.3.3.** The output  $\hat{\Theta}_\rho^*$  of Algorithm 6.3.51 satisfies the following inequality

$$\|\hat{\Theta}_\rho^* - \Theta_\rho^*\|_F \leq \max\{\|\Theta_\rho^*\|_2^2, \|\hat{\Theta}_\rho^*\|_2^2\} \|N\|_F, \quad (6.93)$$

where  $\Theta_\rho^*$  is the optimal solution of the original problem (6.83).

From the above theorem, we can see that the error is measured by the Frobenius norm of the added random matrix. Below, we consider those random matrices that ensure  $\epsilon$ -differential privacy. The first one is due to Lemmas 6.3.2 and 6.3.4.

**Theorem 6.3.4.** In Algorithm 6.3.51, for any  $\epsilon > 0$ , if choose  $\mathcal{P} = \mathcal{W}_d(m, C)$  with  $C = \frac{3}{2\epsilon n} I_d$  and  $m = d + 1$ , it is  $\epsilon$ -differentially private for any  $\epsilon > 0$ . Moreover, with probability at least  $1 - \delta'$ , the following holds

$$\|\hat{\Theta}_\rho^* - \Theta_\rho^*\|_F \leq O\left(\frac{\log \frac{d}{\delta'} d^{\frac{3}{2}} \max\{\|\Theta_\rho^*\|_2^2, \|\hat{\Theta}_\rho^*\|_2^2\}}{n\epsilon}\right). \quad (6.94)$$

Next, we consider the case that  $N$  is sampled from a Laplacian distribution. Since the covariance matrix is symmetric, the added noise also needs to be symmetric, the following lemma is due to [275].

**Theorem 6.3.5.** In Algorithm 6.3.51, for any  $\epsilon > 0$ , if  $N$  is a symmetric Laplacian matrix  $N$  whose entries are i.i.d drawn from  $\text{Lap}(0, \frac{2d}{n\epsilon})$ , then it is  $\epsilon$ -differentially private. Moreover, with high probability, the following holds

$$\|\hat{\Theta}_\rho^* - \Theta_\rho^*\|_F \leq O\left(\frac{d^2 \max\{\|\Theta_\rho^*\|_2^2, \|\hat{\Theta}_\rho^*\|_2^2\}}{n\epsilon}\right). \quad (6.95)$$

**Remark 6.3.2.** Comparing Theorems 6.3.4 and 6.3.5, we can see that the error in (6.94) is less than that in (6.95) (if we omit the term  $\max\{\|\Theta_\rho^*\|_2^2, \|\hat{\Theta}_\rho^*\|_2^2\}$ ). Another advantage is that adding Wishart matrix not only preserves the symmetry property, but also guarantees the positive semi-definite property of the covariance matrix. Thus, for  $\epsilon$ -differential privacy, it is better to use Wishart mechanism.

Next, we consider  $(\epsilon, \delta)$ -differential privacy and also start with adding Wishart matrices. The following theorem is due to Lemmas 6.3.1 and 6.3.3.

**Theorem 6.3.6.** For any  $\epsilon \in (0, 1)$  and  $\delta \in (0, \frac{1}{\epsilon})$ , if choose  $\mathcal{P} = \mathcal{W}_d(m, C)$  with  $C = \frac{1}{n}I_d$  and  $m = d + \frac{14}{\epsilon^2} \ln(\frac{4}{\delta})$  in Algorithm 6.3.51, it is  $(\epsilon, \delta)$ -differentially private. Moreover, if  $m > (\sqrt{d} + \sqrt{2 \log \frac{2}{\delta'}})^2$  for  $0 < \delta' < 1$ , then with probability at least  $1 - \delta'$ , the following holds

$$\|\hat{\Theta}_\rho^* - \Theta_\rho^*\|_F \leq O\left(\frac{\max\{\|\Theta_\rho^*\|_2^2, \|\hat{\Theta}_\rho^*\|_2^2\} \ln(1/\delta) \ln(1/\delta') d^{\frac{3}{2}}}{n\epsilon^2}\right). \quad (6.96)$$

Now, we consider adding symmetric Gaussian matrices.

**Theorem 6.3.7.** In Algorithm 6.3.51, for any  $\epsilon > 0$  and  $0 < \delta < 1$ , if  $N$  is a symmetric Gaussian matrix  $N$  whose entries are i.i.d drawn from  $\mathcal{N}(0, \beta^2)$ , where  $\beta = \frac{\sqrt{2 \ln(\frac{1.25}{\delta})}}{n\epsilon}$ , then it is  $(\epsilon, \delta)$ -differentially private. Moreover, with high probability, the following holds

$$\|\hat{\Theta}_\rho^* - \Theta_\rho^*\|_F \leq O\left(\frac{d\sqrt{\ln(\frac{1}{\delta})} \max\{\|\Theta_\rho^*\|_2^2, \|\hat{\Theta}_\rho^*\|_2^2\}}{en}\right). \quad (6.97)$$

**Remark 6.3.3.** From the above two theorems, we can see that although the Wishart mechanism preserves the positive definite property of  $\tilde{S}$ , which is not the case for the Gaussian mechanism [117], it has an additional factor of  $\sqrt{d}$  in its error bound compared with the Gaussian mechanism (if we omit the term  $\max\{\|\Theta_\rho^*\|_2^2, \|\hat{\Theta}_\rho^*\|_2^2\}$ ). Thus, if we need a more accurate solution, Gaussian mechanism is a better choice.

Now, we address the three issues raised in last section. Firstly, for the large error bound in Theorem 6.3.2, we know from Theorem 6.3.4 that the covariance perturbation based  $\epsilon$ -differentially private algorithm always has a lower error bound than that of an output perturbation based algorithm (since  $\max\{\|\Theta_\rho^*\|_2^2, \|\hat{\Theta}_\rho^*\|_2^2\} \leq \frac{d^2}{\rho^2}$  by (6.3.5)). Secondly, if we view  $\epsilon$  as a constant and omit the term of  $\max\{\|\Theta_\rho^*\|_2^2, \|\hat{\Theta}_\rho^*\|_2^2\}$ , the error bound of the  $(\epsilon, \delta)$ -differentially private algorithm with covariance perturbation strategy is lower than it under  $\epsilon$ -differential privacy, and Gaussian mechanism achieves the lowest error bound. Thirdly, another advantage of the covariance perturbation method is that it produces a sparse solution.

By using the idea of covariance perturbation and some properties of the random matrices, we can extend our methods to other settings.

Firstly, by using the property of Wishart distribution, we can extend our method to distributed settings. In a distributed environment, we assume that there are  $n$  players and one central server; each player  $i \in [n]$  stores one data record  $x_i \in \mathbb{R}^d$  (note that it is easy to extend to the case where each server has any number of data records); the dimensionality  $d$  of the data points is assumed to be larger than the number  $n$  of players. The following algorithm (*i.e.*, Algorithm 6.3.52) is either  $(\epsilon, \delta)$ - or  $\epsilon$ -differentially private. Since it is equivalent to Algorithm 6.3.51 with Wishart matrix perturbation, the error bound is, thus, the same as in Theorem 6.3.4 and 6.3.6.

---

**Algorithm 6.3.52** Distributed Setting

---

**Input:** Each player  $i \in [n]$  has a data record  $x_i \in \mathbb{R}^d$  with its  $\ell_2$  norm bounded by 1.  $\epsilon, \delta$  are the privacy parameters.

- 1: **for** Each Player  $i \in [n]$  **do**
- 2:     Sample  $v_i \sim \mathcal{N}(0, C)$ , where  $C$  is the same as in Theorem 6.3.4 for  $\epsilon$ -differential privacy and the same as in Theorem 6.3.6 for  $(\epsilon, \delta)$ -differential privacy.
- 3:     Compute and send  $A_i = \frac{1}{n}x_i x_i^T + v_i v_i^T$  to the central server.
- 4: **end for**
- 5: In the central server, i.i.d sample  $k$  vectors  $[u_1, u_2, \dots, u_k]$ , where  $u_i \sim \mathcal{N}(0, C)$ ,  $k = m - n$ , where  $m$  is the same as in Theorem 6.3.4 for  $\epsilon$ -differential privacy and the same as in Theorem 6.3.6 for  $(\epsilon, \delta)$ -differential privacy. Let  $B = \sum_{i=1}^k u_i u_i^T$ . Compute  $\tilde{S} = \sum_{i=1}^n A_i + B$  and

$$\hat{\Theta}_\rho^* = \arg \min_{\Theta \in \mathcal{S}_{++}^d} \{-\log \det \Theta + \langle \tilde{S}, \Theta \rangle + \rho \|\Theta\|_1\}.$$


---

Based on the idea of covariance perturbation and the property of Gaussian matrices, we can easily have an  $(\epsilon, \delta)$ -LDP algorithm (see Algorithm 6.3.53, the same for  $\epsilon$ -LDP) with the following error bound.

**Theorem 6.3.8.** For any  $\epsilon > 0$  and  $0 < \delta < 1$ , Algorithm 4 is  $(\epsilon, \delta)$ -LDP. Moreover, with

high probability, the output  $\hat{\Theta}_\rho^*$  satisfies the inequality

$$\|\hat{\Theta}_\rho^* - \Theta_\rho^*\|_F \leq O\left(\frac{d\sqrt{\ln(\frac{1}{\delta})} \max\{\|\Theta_\rho^*\|_2^2, \|\hat{\Theta}_\rho^*\|_2^2\}}{\epsilon\sqrt{n}}\right).$$

---

**Algorithm 6.3.53 Local Differential Privacy**


---

**Input:** Each player  $i \in [n]$  has a data record  $x_i \in \mathbb{R}^d$  with its  $\ell_2$  norm bounded by 1.  $\epsilon, \delta$  are the privacy parameters.

- 1: **for** Each Player  $i \in [n]$  **do**
- 2:     Generate a symmetric Gaussian Matrix  $N_i$  whose entries are i.i.d drawn from  $\mathcal{N}(0, \beta^2)$ , where  $\beta = \frac{\sqrt{2\ln(\frac{1.25}{\delta})}}{n\epsilon}$
- 3:     Compute and send  $A_i = \frac{1}{n}x_i x_i^T + N_i$  to the central server.
- 4: **end for**
- 5: In the central server, compute  $\tilde{S} = \sum_{i=1}^n A_i$ , and solve the following

$$\hat{\Theta}_\rho^* = \arg \min_{\Theta \in \mathcal{S}_{++}^d} \{-\log \det \Theta + \langle \tilde{S}, \Theta \rangle + \rho \|\Theta\|_1\}.$$


---

Comparing with  $(\epsilon, \delta)$ -differential privacy in the central model, we can see that the only difference is the term of  $n$ . Also, since LDP is more rigorous than DP, we can see that although Algorithms 6.3.52 and 6.3.53 are both for distributed settings, the error bound of Algorithm 6.3.53 is worse (if we omit the term  $\max\{\|\Theta_\rho^*\|_2^2, \|\hat{\Theta}_\rho^*\|_2^2\}$ ).

### 6.3.6 Experiments

In this section, we present some numerical results on both real-world and synthetic datasets to evaluate the performance of our proposed differentially private algorithms. More experiments are left to the full paper.

We first introduce the algorithms that we are going to compare; all related methods are described in the previous section. For  $\epsilon$ -differentially private algorithm, we will compare with output perturbation, Laplace and Wishart covariance perturbation methods. For  $(\epsilon, \delta)$ -differentially private algorithm, we will compare with SULQ Framework [42], Wishart and Gaussian covariance perturbation methods.

We let  $\hat{\Theta}_\rho^*$  denote the output of the differentially private algorithm and  $\Theta_\rho^*$  denote the optimal solution of the original problem. To evaluate the performance of the proposed methods, we choose **Relative Error**, which is defined as  $\frac{\|\hat{\Theta}_\rho^* - \Theta_\rho^*\|_F}{\|\Theta_\rho^*\|_F}$ . If the relative error is greater than 200, we use NA to indicate.

For synthetic datasets, we first fix the dimensionality  $d$  and create a sparse matrix  $U$  with nonzero entries equal to -1 or 1 with equal probability. Then, we compute  $S = (U * U^T)^{-1}$  as the true covariance matrix. The inverse covariance matrix  $S^{-1} = UU^T$  is, thus, sparse. Given the inverse covariance matrix  $S^{-1} = UU^T$ , we then draw  $n = r \times d$  samples from the Gaussian distribution  $\mathcal{N}(0, S)$  to simulate the high-dimensional settings, where  $r$  denotes the ratio of  $n$  (*i.e.*, the sample size) over  $d$  (*i.e.*, the dimensionality of the samples). We test our proposed methods for  $d = 400$  and  $r = 0.5, 1.0, 1.5$ .

For real-world datasets, we use the colon cancer dataset [11] and the Parkinson's disease dataset [200] to evaluate our proposed methods. The colon cancer dataset contains information of 69 individuals with 2000 attributes. We choose 300 variables for the experiment. The size of Parkinson's disease dataset is (192, 22). The datasets are normalized before processing.

For each experiment, we choose  $\epsilon = 0.5, 1, 1.5$ , respectively. For  $(\epsilon, \delta)$ -DP, we let  $\delta = 0.01$ . To solve the optimization problem (6.83), we set  $\rho = 0.001$  and use the method in [251]. All experiments run in MATLAB.

$\epsilon$	Methods	Synthetic Datasets			Real-world Datasets	
		$r = 0.5$	$r = 1.0$	$r = 1.5$	Colon	Parkinson's
0.5	Wishart	<b>0.993</b>	<b>0.9918</b>	<b>0.9914</b>	<b>0.995</b>	<b>0.9140</b>
	Output	NA	NA	NA	NA	NA
	Laplace	101.4	52.85	35.42	190.57	9.950
1.0	Wishart	<b>0.986</b>	<b>0.9863</b>	<b>0.9856</b>	<b>0.993</b>	<b>0.8899</b>
	Output	NA	NA	NA	NA	NA
	Laplace	49.44	25.41	16.83	95.01	4.690
1.5	Wishart	<b>0.9817</b>	<b>0.9815</b>	<b>0.9806</b>	<b>0.9907</b>	<b>0.8796</b>
	Output	NA	NA	NA	NA	NA
	Laplace	32.30	16.41	10.76	63.67	3.913

Table 6.4: Performance comparisons of the  $\epsilon$ -differentially private algorithms on both synthetic and real-world datasets.

$\epsilon$	Methods	Synthetic Datasets			Real-world Datasets	
		$r = 0.5$	$r = 1.0$	$r = 1.5$	Colon	Parkinson's
0.5	Wishart	0.9999	0.9997	0.9993	1.636	1.00
	SQLU	NA	NA	NA	NA	0.7419
	Gaussian	<b>0.1285</b>	<b>0.1607</b>	<b>0.1759</b>	<b>0.3039</b>	<b>0.1527</b>
1.0	Wishart	0.9982	0.9947	0.9906	1.1155	0.990
	SQLU	NA	NA	NA	NA	0.7318
	Gaussian	<b>0.1254</b>	<b>0.1605</b>	<b>0.1737</b>	<b>0.1081</b>	<b>0.1514</b>
1.5	Wishart	0.9954	0.9895	0.9837	1.0474	0.9992
	SQLU	NA	NA	NA	NA	0.7065
	Gaussian	<b>0.1252</b>	<b>0.1595</b>	<b>0.1701</b>	<b>0.0833</b>	<b>0.1514</b>

Table 6.5: Performance comparisons of the  $(\epsilon, \delta)$ -differentially private algorithms on both synthetic and real-world datasets.

**Results Analysis** The experimental results of the  $\epsilon$ -differentially private algorithms on both synthetic and real-world datasets are shown in Table 6.4. From the table, we can see that in all the cases, Wishart and Laplacian mechanisms achieve better performance than the output perturbation method. Furthermore, Wishart mechanism is the best among the three types of methods. From Table 6.5, we can see that the Gaussian method has the lowest relative error among all  $(\epsilon, \delta)$ -differentially private algorithms. Also, Gaussian mechanism has the lowest relative error among all the compared methods. In general, the relative error becomes smaller for larger  $\epsilon$ . But in some cases, the relative error are almost same for different  $\epsilon$  values. This could be due to the fact the difference of these  $\epsilon$  values is small. In summary, all the above results are consistent with our theoretical analysis.

### 6.3.7 Omitted Proofs

#### Proof of Theorem 6.3.1

For convenience, we denote Step 1 as  $\mathcal{A}$ . That is,  $\Theta_\rho^* = \mathcal{A}(D)$ . Also, we let  $D'$  be a neighboring dataset, and  $S' = S - \frac{1}{n}vv^T + \frac{1}{n}v'v'^T$ ,  $\Theta'^*_\rho = \mathcal{A}(D')$ . Then by (6.91), we have, for any  $\eta > 0$ ,

$$\|\Theta_\rho^* - \Theta'^*_\rho\|_F = \|\text{prox}_{\eta g}(\Theta_\rho^* - \eta(S - \Theta_\rho^{*-1})) - \text{prox}_{\eta g}(\Theta'^*_\rho - \eta(S' - \Theta'^{*,-1}))\|_F.$$

Then, by the non-expansive property of the proximal operator, we have

$$\|\Theta_\rho^* - \Theta'^*_\rho\|_F \leq \|(\Theta_\rho^* - \eta(S - \Theta_\rho^{*-1})) - (\Theta'^*_\rho - \eta(S' - \Theta'^{*,-1}_\rho))\|_F.$$

If let  $f(\Theta_\rho^*) = \Theta_\rho^* + \eta\Theta_\rho^{*-1}$ , we have the following inequality:

$$\|\Theta_\rho^* - \Theta'^*_\rho\|_F \leq \|f(\Theta_\rho^*) - f(\Theta'^*_\rho)\|_F + \eta\|S - S'\|_F. \quad (6.98)$$

For the last term, we have  $\|S - S'\|_F = \|\frac{1}{n}(vv^T - v'v'^T)\|_F \leq \frac{2}{n}$ . In order to bound the first term, we need the following lemma, which has been proved in [250].

**Lemma 6.3.6** ([250]). For  $\Theta_1, \Theta_2 \in \mathcal{S}_{++}^d$ ,  $\eta > 0$ , we have

$$\|f(\Theta_1) - f(\Theta_2)\|_F \leq \max\{|1 - \frac{\eta}{a^2}|, |1 - \frac{\eta}{b^2}|\}\|\Theta_1 - \Theta_2\|_F,$$

where  $a = \max\{\sigma_{\max}(\Theta_1), \sigma_{\max}(\Theta_2)\}$  and  $b = \min\{\sigma_{\min}(\Theta_1), \sigma_{\min}(\Theta_2)\}$ .

Take  $\Theta_\rho^*, \Theta'^*_\rho$  into Lemma 6.3.6 and set  $0 < \eta < b^2$  in (6.98), we now have

$$\|\Theta_\rho^* - \Theta'^*_\rho\|_F \leq \frac{2\beta^2}{n}, \quad (6.99)$$

where  $\beta = \max\{\|\Theta_\rho^*\|_2, \|\Theta'^*_\rho\|_2\}$ . Now we will show the  $\epsilon$ -differential privacy. Since for

every  $W$ ,

$$\begin{aligned}
\frac{\Pr[\Theta_\rho^* + N = W]}{\Pr[\Theta_\rho'^* + N = W]} &= \frac{\Pr[N = W - \Theta_\rho^*]}{\Pr[N = W - \Theta_\rho'^*]} \\
&= \frac{\exp(-\frac{1}{2} \text{tr}(C^{-1}(W - \Theta_\rho^*)))}{\exp(-\frac{1}{2} \text{tr}(C^{-1}(W - \Theta_\rho'^*)))} \\
&= \exp(-\frac{1}{2} \text{tr}(C^{-1}(\Theta_\rho^* - \Theta_\rho'^*))) \\
&\leq \exp(\frac{1}{2} \|C^{-1}\|_F \|\Theta_\rho^* - \Theta_\rho'^*\|_F) \\
&\leq \exp(\frac{1}{2} \sqrt{d} \frac{n\epsilon\rho^2}{d^{\frac{5}{2}}} \frac{2\beta^2}{n}) \\
&\leq \exp(\epsilon).
\end{aligned}$$

Where the last inequality comes from (6.3.5).

### Proof of Theorem 6.3.3

The proof is the same as Theorem 6.3.1, we have  $\|\hat{\Theta}_\rho^* - \Theta_\rho^*\|_F \leq \|f(\hat{\Theta}_\rho^*) - f(\Theta_\rho^*)\|_F + \eta\|S - \tilde{S}\|_F$ . Thus by Lemma 6.3.5 and take  $0 < \eta < \min\{\sigma_{\min}^2(\hat{\Theta}_\rho^*), \sigma_{\min}^2(\Theta_\rho^*)\}$ , we have the theorem.

# **Chapter 7**

## **Some Other Machine Learning Problems**

### **7.1 Inferring Ground Truth From Crowdsourced Data Under Local Attribute Differential Privacy**

Nowadays, crowdsourcing gains an increasing popularity as it can be adopted to solve many challenging question answering tasks that are easy for humans but difficult for the computer, and it has many real-world machine learning or data mining applications. For example, patients who are taking new drugs can answer the question on whether a specific drug has a certain side-effect [235]. Also there are many commercial web service for crowdsourcing such as Amazon Mechanical Turk (AMT). In these and many more applications, crowds of users can contribute their efforts to answer questions of interest, which largely reduces the financial cost and benefits various application domains.

Due to the variety in the quality of users, the information quality of the answers given by the users varies significantly. Some users may have sufficient domain knowledge and can provide accurate answers while others may submit biased or wrong answers. This diversity of users motivates a basic and important problem in crowdsourcing: how do the server

get the accurate answers (or ground truth) via these noisy answers while also could infer the underlying ability of each user. This problem is called **Ground Truth Inference**<sup>1</sup> [365] and there is a large amount of work study this problem in both Machine Learning [358], Data Mining [366] and Theoretical Computer Science [197, 156, 93] communities.

However, in the problem of ground truth inference, collecting individual users answers may cause the privacy issue on the users. For example, individual users can report the relevance between a search query and a webpage, but their answers may leak their personal preference. Patients' reactions to drugs are valuable for physicians to discover drugs' side-effect, but these also contain sensitive information. Moreover, recently it has been reported that AMT platform was leveraged by politicians to access a large pool of Facebook profiles and collects ten of thousands of individuals demographic data [266].

Ground truth inference in Local Differential Privacy model has been first studied by [198] and was later extended by [266] to the sparse crowdsourcing data case. Although their methods are effective with tolerable accuracy loss practically, there are still some basic theoretical open problems which have not been studied or solved. First, it is still unknown what is the average error of the private estimators with respect to the underlying ground truth. Secondly, while all the previous work focus on the quality private ground truth estimator, we do not know whether we can infer the ability of each user under LDP model and what is the estimation error with respect to the underlying ability of users. Finally, previous work only shows that their methods have better performance than the private major voting algorithm through experiments on some datasets. However, there is still no theoretically result which shows the priority of their methods formally or mathematically.

In this section, I partially solve the above theoretical issues. That is, instead of considering the LDP model, in this section I will focus on one of its relaxations called local attribute differential privacy (LADP) model. This is motivated by the fact that in practice of ground truth inference, instead of keeping the each whole data record of each user private, it is

---

<sup>1</sup>Note that in the data mining community this problem is also called Truth Discovery.

always the case that only a small number of answers given by users may contain sensitive information, which means it is sufficient to protect some attributes of a vector (if we see the set of answers of each user as a vector). LADP corresponds to an adversary cannot infer a single attribute value despite he knows the values of all other attributes and thus is more suitable for ground truth inference. We study the previous issues of ground truth inference in LADP model. In particular, I propose a method called private Dawid-Skene method which outputs the private truth estimators and private ability of users. Specifically, our contributions can be summarized as the followings.

- I first show that my private Dawid-Skene method is LADP. Then I provide the result on the average error of the private truth estimators w.r.t the ground truth. I show that under some statistical assumptions of the problem and if the initial vector of the algorithm is closed enough to the ground truth, then the average error will be upper bounded by  $\exp(-n\tilde{v})$  with high probability, where  $n$  is the number of users and  $\tilde{v}$  is the term called collective private wisdom which is related to the privacy level  $\epsilon$  (see Theorem 7.1.2 for details).
- I also show that under the same assumptions, the output of private ability of users has the estimation error of  $O(\sqrt{\frac{\log m}{m\epsilon^2}})$ , where  $m$  is the number of tasks with high probability (see Theorem 7.1.3 for details).
- Finally, I compare our method with the classical private major voting algorithm. To show the priority of my method, I propose a special instance. I show that the estimation error given by the private major voting error is always greater than the error given by our algorithm, which means the private major voting is always worse than our method on this instance theoretically. See Theorem 7.1.4 for details.

### 7.1.1 Related Work

There is much attention on studying crowdsourcing system in LDP model. For example, [248] consider the problem of publishing high dimensional crowdsourced data in LDP model. [155] propose a method which could generating synthetic crowdsourced data via some Privacy-Test. However, their methods are incomparable with ours due to that there utility is different with ours and also there is no theoretical guarantees on their output.

Among all the previous work, maybe [198] and [266] are the most relevant to ours. In [198] the authors propose a two-layer perturbation mechanism based on randomized response to protect users privacy. [266] consider the case where the data is sparse and propose a private mechanism based on the formula of Matrix Factorization and randomized response. However, as we mentioned before, first, their methods can only output private estimators of the ground truth and it is unknown whether they can also estimate the ability of users. Secondly, there is no theoretical guarantees of the average error of the ground truth. Moreover, in all of these work they compared with the private major voting algorithm practically on some datasets and showed that their method have better performance. However, there is no theoretical guarantees on these comparisons. Thus, our work provides some theoretical guarantees which have not been solved in these previous work.

Our method is motivated by the classical Dawid-Skene method [86], which laid a solid foundation in the field of crowdsourcing. Extensions of the framework under a Bayesian setting were investigated by [72]. However, there is no previous study on the private version of Dawid-Skene method. Moreover, compared with the classical Dawid-Skene method, here we need some modifications such as perturbation and projection.

### 7.1.2 Preliminaries

In this section, we review the definition of ground truth estimation in crowdsourcing, local attribute differential privacy and the classical Dawid-Skene algorithm.

## Local Attribute Differential Privacy

In this section, we will mainly focus on Local Attribute Differential Privacy (LADP), which is a relaxation of LDP and has been studied in many previous papers, such as [180, 141, 146, 206]. Mathematically it can be defined as the follows.

**Definition 7.1.1** (Local Attribute Differential Privacy). A randomized algorithm  $\mathcal{A}$  is  $\epsilon$ -locally differentially private if for all  $x, x' \in \mathcal{X}$  with there is some  $i$  where  $x$  and  $x'$  differ in the  $i$ -th coordinate and all for all events  $S$  in the output of  $\mathcal{A}$ , we have

$$\mathbb{P}(\mathcal{A}(x) \in S) \leq e^\epsilon \mathbb{P}(\mathcal{A}(x') \in S).$$

We note that the only difference between LDP and LADP is in LADP we have an additional restriction on  $x, x'$ . LADP corresponds to an adversary cannot infer a single attribute value despite he knows the values of all other attributes.

## Problem Setting

We now start by formally define the problem of **Ground Truth Inference**. Conceptually, there are two parties, sever and user, are involved in the crowdsourced question answering. We assume there are  $m$  tasks and  $n$  users, each task  $j \in [m]$  is independent with other tasks and is associated with a label  $y_j^* \in \{0, 1\}$  which is called the **ground truth**. We note that in practice the number of tasks  $m$  is much larger than the number of users  $n$ , such as the Web and AdultCotent datasets [266]. The users, who represent the individual participants, provide their own answer 0 or 1 to each of these tasks and send them to the server. However, there is one main issue. Due to the quality of the users, these answers are noisy. It is more challenging that the underlying quality of the workers are also unknown. Mathematically, to model the users' quality, [86] proposed the so-called confusion matrix. The confusion

matrix for the  $i$ -the worker is denoted as

$$\begin{bmatrix} \pi_{00}^{(i)}, \pi_{01}^{(i)} \\ \pi_{10}^{(i)}, \pi_{11}^{(i)} \end{bmatrix}$$

where the number  $\pi_{kl}^{(i)}$  represents the probability for the  $i$ -th user to give answer  $l$  given the ground truth is  $k$ . In our paper, we will study a special class of the confusion matrix, where the ability of the  $i$ -th user is characterized by the probability of success  $p_i^* \in [0, 1]$  with the confusion matrix

$$\begin{bmatrix} p_i^*, 1 - p_i^* \\ 1 - p_i^*, p_i^* \end{bmatrix}.$$

Equivalently, here we will assume that for each user  $i \in [n]$ , his/her abilities are the same for all the  $m$  tasks.

After collecting the users answers, the server aggregate them to derive the final inference and estimation. The goal is not only inferring the truth labels  $\{y_j^*\}_{j=1}^m$ , but also estimating the abilities of the users, *i.e.*,  $\{p_i^*\}_{i=1}^n$ .

The main privacy concern of users is that the submitted answers many contain their sensitive information and thus users are not willing to leak these answers to other parties. This prevents users from sharing their own answers with the server. The server, who is assumed to be untrusted, may try to infer additional knowledge of users forms their submitted answers. The unfaithful behavior of server can be driven by financial incentives or other benefits. Motivated by this, it is naturally to study the problem of ground truth inference under LDP model. However, the definition if LDP might be too strong for the problem of ground truth inference. Since in the problem, it is always the case that only some of the tasks are related to users sensitive information. Thus it is sufficient of we can protect these tasks instead of the whole data record of each user in LDP model, which is just the LADP model.

Thus, motivated by the strong need to provide users with privacy protection. In the

**Private Ground Truth Inference** problem, we want to design  $\epsilon$ -LADP algorithms whose outputs  $\{y_j\}_{j=1}^n$  and  $\{p_i^*\}_{i=1}^n$  are close to  $\{y_j^*\}_{j=1}^m$  and  $\{p_i^*\}_{i=1}^n$ , respectively.

### 7.1.3 Main Method

In this section we will propose our method and analyze its theoretical performance. Before that, we first recall the classical Dawid-Skene method [86].

#### Dawid-Skene Method

Now we consider the problem of ground truth inference in the non-private case (see Section 7.1.2). We first observe that the ability of the works  $\{p_i\}_{i=1}^n$  can be easily estimated by using the frequency of success of the workers if the ground truth  $\{y_j^*\}_{j=1}^m$  is known. Motivated by this, [86] proposed to estimate  $\{p_i\}_{i=1}^n$  by maximizing the marginal likelihood function by giving the ground truth:

$$\begin{aligned}\mathbb{P}(X|y, p) &= \prod_{j \in [m]} \prod_{i \in [n]} \mathbb{P}(X_{ij}|y_j, p_i) \\ &= \prod_{j \in [m]} \prod_{i \in [n]} p_i^{\mathbb{I}(X_{ij}=y_j)} (1 - p_i)^{\mathbb{I}(X_{ij}=1-y_j)},\end{aligned}\quad (7.1)$$

where  $\mathbb{I}$  is the indicator function<sup>2</sup>. Integrating out the ground truth with a uniform prior, the marginal likelihood is

$$\mathbb{P}(X|p) = \prod_{j \in [m]} \left( \frac{1}{2} \prod_{i \in [n]} p_i^{X_{ij}} (1 - p_i)^{1-X_{ij}} + \frac{1}{2} \prod_{i \in [n]} (1 - p_i)^{X_{ij}} p_i^{1-X_{ij}} \right). \quad (7.2)$$

Thus, the maximum likelihood estimator (MLE) based on (7.2) is defined as

$$\hat{p} = \arg \max_p \log P(X|p).$$

---

<sup>2</sup>Given an event  $A$ ,  $\mathbb{I}(A) = 1$  if  $A$  happens and otherwise it is 0.

After getting the MLE solution  $\hat{p} = (\hat{p}_1, \hat{p}_2, \dots, \hat{p}_n)$ , we can plug it into the Bayes formula and get an estimator for the ground truth  $y^*$ :

$$\hat{y}_j = \frac{\prod_{i \in [n]} \hat{p}_i^{X_{ij}} (1 - \hat{p}_i)^{1 - X_{ij}}}{\prod_{i \in [n]} \hat{p}_i^{X_{ij}} (1 - \hat{p}_i)^{1 - X_{ij}} + \prod_{i \in [n]} (1 - \hat{p}_i)^{X_{ij}} \hat{p}_i^{1 - X_{ij}}}, \quad (7.3)$$

Note that we implicitly use the uniform prior in the Bayes formula and the resulting estimator  $\hat{y}$  is a soft label, taking value in  $[0, 1]^m$ . Now the pair of estimator  $(\hat{p}, \hat{y})$  is the global optimizer of the following objective function.

$$\begin{aligned} F(p, y) &= \sum_i \sum_j y_j (X_{ij} \log p_i + (1 - X_{ij}) \log(1 - p_i)) + \\ &\sum_i \sum_j (1 - y_i) (X_{ij} \log(1 - p_i) + (1 - X_{ij}) \log p_i) + \sum_j (y_j \log \frac{1}{y_j} + (1 - y_j) \log \frac{1}{1 - y_j}). \end{aligned} \quad (7.4)$$

[226] showed that optimizing over  $\log \mathbb{P}(X|p)$  is equivalent as optimizing over  $F(p, y)$ , *i.e.*,  $(\hat{p}, \hat{y}) = \arg \max F(p, y)$ , while the latter one is more tractable. In order to maximize (7.4), one natural and heuristic way is to iteratively update  $p$  and  $y$ . That is, given an initial estimator  $y^{(0)}$ , the  $t$ -th step of the iterative algorithm is

$$p^{(t)} = \arg \max F(p, y^{(t-1)}), \quad y^{(t)} = \arg \max F(p^{(t)}, y). \quad (7.5)$$

Calculating (7.5) directly, we have the followings:

$$p_i^{(t)} = \frac{1}{m} \sum_{j \in [m]} ((1 - X_{ij})(1 - y_j^{(t-1)}) + X_{ij}y_j^{(t-1)}), \quad (7.6)$$

$$y_j^{(t)} \propto \prod_{i \in [n]} (p_i^{(t)})^{X_{ij}} (1 - p_i^{(t)})^{1 - X_{ij}}, \quad (7.7)$$

$$1 - y_j^{(t)} \propto \prod_{i \in [n]} (p_i^{(t)})^{1 - X_{ij}} (1 - p_i^{(t)})^{X_{ij}}. \quad (7.8)$$

Eq. (7.6)-(7.8), are given by [86] and are called Dawid-Skene method.

## Private Dawid-Skene Estimation

Now we propose the our Private Dawid-Skene method. The idea is that for each user  $i \in [n]$  who process answers  $(X_{i1}, X_{i2}, \dots, X_{ij})$ , he/she perturbs each answer by the following distribution:

$$\hat{X}_{ij} = \begin{cases} X_{ij} \text{ w.p. } \frac{e^\epsilon}{e^\epsilon + 1} \\ 1 - X_{ij} \text{ w.p. } \frac{1}{e^\epsilon + 1}. \end{cases} \quad (7.9)$$

After the server getting these perturbed answers  $\{\hat{X}_{ij}\}_{i \in [n], j \in [m]}$ , it then performs the Dawid-Skene estimator on these perturbed answers, see Algorithm 7.1.54 for details. However, we note that instead of performing (7.6) for updating the abilities of users, here we perform a projected version, that is

$$p_i^{(t)} = \Pi_{\mathcal{C}(\lambda)} \frac{1}{m} \sum_{j \in [m]} \left( (1 - X_{ij})(1 - y_j^{(t-1)}) + X_{ij}y_j^{t-1} \right). \quad (7.10)$$

Where  $\Pi_{\mathcal{C}(\lambda)}$  is the projection operator on a interval  $\mathcal{C}(\lambda) = [\lambda, 1 - \lambda]$  with some small  $\lambda > 0$ . The motivation is that in the case of the estimator  $p_i^{(t)}$  is 0 or 1 for some  $i \in [n]$  and  $t \in [T]$ ,  $p_i^{(t)}$  will be trapped in its current value, which might be a poor local optimizer. Thus, in order to avoid, we perform the projector operator to keep  $p_i^{(t)}$  be slightly away from 0 or 1. Later, we will see that an appropriate value of  $\lambda$  is crucial for the rate of convergence. We note that this operator also has been used and studied in [121].

Also, we note that, after the  $T$ -th iteration, instead of releasing the the estimators of the ability  $p_i^{(T)}$  directly, we have to post-process them via Step 7 in Algorithm 7.1.54. This is due to that,  $\{p_i^T\}_{i \in [n]}$  are some biased estimators of the underlying ability  $\{p_i^*\}_{i \in [n]}$  since the perturbation procedure in Step 2. Thus, in order to get some useful estimators we need to rescale them. We will see later for the reason of choosing these terms for rescaling.

Finally, since the terms  $\{y_j\}_{j \in [m]}$  are soft labels contained in  $[0, 1]^m$ , in order to get hard labels as final answers, we need to do a round procedure in step 8. We will show that it will

not effect the error to much.

---

**Algorithm 7.1.54** Private Dawid-Skene Method

---

**Input:**  $T$  is the number of iteration,  $\epsilon > 0$  is the privacy parameter,  $y^{(0)}$  is the initial vector, worker  $i \in [n]$  process the answers  $X_i = (X_{i1}, \dots, X_{im}) \in \{0, 1\}^m$ .

- 1: **for** Each worker  $i \in [n]$  **do**
  - 2:     Perturb each  $X_{ij}, j \in [m]$  by the distribution (7.9) and get  $\hat{X}_{ij}$ . Then send  $\{\hat{X}_{ij}\}_{j=1}^m$  to the server.
  - 3: **end for**
  - 4: **for**  $t = 1, \dots, T$  **do**
  - 5:     The server perform the updating (7.10), (7.7), (7.8) on  $\{\hat{X}_{ij}\}_{i \in [n], j \in [m]}$  and get  $\{p_i^{(t)}\}_{i \in [n]}, \{y_j^{(t)}\}_{j \in [m]}$ .
  - 6: **end for**
  - 7: For each  $i \in [n]$ , let  $\hat{p}_i^{(T)} = \frac{e^\epsilon + 1}{e^\epsilon - 1}(p_i^{(T)} - \frac{1}{e^\epsilon + 1})$ .
  - 8: For each  $j \in [m]$ , let  $\hat{y}_j^{(T)} = \mathbb{I}(y_j^{(T)} \geq \frac{1}{2})$ .
  - 9: **return**  $\hat{p}^{(T)} = \{\hat{p}_i^{(T)}\}_{i \in [n]}$  and  $\{\hat{y}_j^{(T)}\}_{j=1}^m$ .
- 

The following theorem shows that the algorithm is LADP. Not only LADP, it is also easy to see that Algorithm 7.1.54 is also  $m\epsilon$  locally differentially private.

**Theorem 7.1.1.** For any given  $\epsilon > 0$ , Algorithm 7.1.54 is  $\epsilon$ -LADP.

### 7.1.4 Theoretical Guarantees

In this section, we will give the estimation errors of the outputs  $\{\hat{p}_i^{(T)}\}_{i \in [n]}$  and  $\{\hat{y}_j^{(T)}\}_{j=1}^m$  to the underlying abilities and ground truth, respectively. Before showing the explicit result, we first introduce some critical quantities.

First, for each user  $i \in [n]$ , we define the term of **private effective ability** as

$$\hat{\mu}_i = \frac{e^\epsilon - 1}{e^\epsilon + 1} \mu_i + \frac{1}{e^\epsilon + 1}, \quad (7.11)$$

where  $\mu_i$  is the **effective ability** proposed by [121]:

$$\mu_i = p_i^* \mathbb{I}\{p_i^* \geq \frac{1}{2}\} + (1 - p_i^*) \mathbb{I}\{p_i^* < \frac{1}{2}\}. \quad (7.12)$$

Intuitively,  $\mu_i$  measures how much information we can get from the user  $i$ : when  $p_i^* > \frac{1}{2}$  it is just the underlying ability, when  $p_i < \frac{1}{2}$  then we can use the information to detect and invert the answers.  $\hat{\mu}_i$  can be thought as the private version of  $\mu_i$  due to the effect of perturbation by (7.9). When the algorithm is extremely private *i.e.*,  $\epsilon \rightarrow 0$ , we can see that  $\hat{\mu}_i \rightarrow \frac{1}{2}$ , that is all the workers become spammers. Equivalently, from (7.9) we can see that  $\mathbb{P}(\hat{X}_{ij} = 0) = \mathbb{P}(\hat{X}_{ij} = 1) = \frac{1}{2}$ , which means we cannot get any useful information from the perturbed observations. However, when the algorithm tends to be non-private, that is  $\epsilon \rightarrow \infty$ , we have  $\hat{\mu}_i \rightarrow \mu_i$ , in this case the private effective ability will be the same as the effective ability. We note that for both  $\mu_i, \hat{\mu}_i$  are in  $[\frac{1}{2}, 1]$ .

Now we define the term of **collective private wisdom**  $\hat{v}$  as

$$\hat{v} = \frac{1}{n} \sum_{i=1}^n (2\hat{\mu}_i - 1)^2. \quad (7.13)$$

$\hat{v}$  measures the proportion of experts among the crowd under the privacy constraint, when the  $\epsilon \rightarrow 0$ , then  $\hat{v} \rightarrow 0$  since in the extreme private case we cannot distinguish which one is the expert. When  $\epsilon \rightarrow \infty$ , then  $\hat{v} \rightarrow v = \frac{1}{n} \sum_i (2\mu_i - 1)^2$ , which is the collective wisdom in [121].

Note that the objective function  $F(p, y)$  in (7.4) is non-convex with the fixed  $\{\hat{X}_{ij}\}_{i \in [n], j \in [m]}$ . Thus, alternating maximization procedures (7.6) and (7.7) will only converge to some local minimum. However, in the following theorem, we will show that under the setting of  $m \gg n$ , with some appropriate initial vector  $y^{(0)}$ , the iterations  $\{y_j^{(t)}\}_{j \in [m]}$  after the first step will be in the neighborhood of the ground truth  $\{y_j^*\}_{j \in [m]}$  with high probability.

**Theorem 7.1.2.** Assume  $n$  and  $m$  are sufficiently large so that  $n \leq m \leq e^n$ ,  $\frac{\log m}{n} \leq \hat{v}$  and the initial vector  $y^{(0)}$  satisfies

$$\frac{1}{m} \sum_{j \in [m]} |y_j^{(0)} - y_j^*| \leq \sqrt{\frac{\log m}{m}}. \quad (7.14)$$

Whenever the parameter  $\lambda$  of Algorithm 7.1.54 are chose in the range

$$\frac{16}{\hat{v}} \sqrt{\frac{\log m}{m}} \leq \lambda \leq \frac{1}{8} - \frac{1}{2} \sqrt{\frac{\log m}{m}}. \quad (7.15)$$

Then for any  $y^* \in \{0, 1\}^m$ , we have

$$\frac{1}{m} \sum_{j \in [m]} |\hat{y}_j^{(T)} - y_j^*| \leq 2 \exp\left(-\frac{1}{2}n\hat{v}\right). \quad (7.16)$$

with probability at least  $1 - \frac{C'}{m}$  for some constant  $C' > 0$ .

From Theorem 7.1.2, we can see that as long as the our initial guess has the average error of  $\tilde{O}\left(\frac{1}{\sqrt{m}}\right)$ , then for some  $\lambda$  the average error will decreases to  $\exp\left(-\frac{1}{2}n\hat{v}\right)$ . We can see that when  $\epsilon$  deceases, this upper bound will increase, which means the error will be larger. Equivalently, this shows that when the algorithm is more private, the error bound will be larger. When  $\epsilon = 0$ , the upper bound becomes  $\frac{1}{2}$  and will be trivial.

The following theorem states that our algorithm not only can almost infer the ground truth, but also can estimate the users' abilities with some statistical error.

**Theorem 7.1.3.** Under the assumptions in Theorem 7.1.2. For  $0 < \epsilon \leq 1$  we have the followings with probability at least  $1 - \frac{C'}{m}$  for some  $C' > 0$ :

$$\max_{i \in [n]} |\hat{p}_i^{(T)} - p_i^*| \leq 6 \sqrt{\frac{\log m}{m\epsilon^2}}, \quad (7.17)$$

Eq. (7.17) characterize the accuracy for users' abilities from the worst-case. We know the rate of error is  $\tilde{O}\left(\frac{1}{m\epsilon^2}\right)$ , which means that it will decreases as the number of tasks increases. Moreover, when the algorithm is more private, the bound will be larger.

### 7.1.5 Comparison with Private Major Voting

In order to show the priority of our method theoretically, in this part, we will compare our algorithm with the most trivial method *i.e.*, private major voting. The algorithm of private major voting is quite simple; the steps of the user side is the same as steps 1-3 in Algorithm 7.1.54 while each user send the private answers to the server. After collecting all of the private answers, the server will do major voting and decide the output for each task, that is for all  $j \in [m]$

$$\bar{y}_j = \mathbb{I}\left(\sum_{i \in [n]} \hat{X}_{ij} \geq \frac{n}{2}\right). \quad (7.18)$$

Now we will provide a case where the upper bound (7.16) is lower than the bound of private major voting, which means our algorithm has better performance than private major voting theoretically. Formally, suppose that there are  $\lceil n^\delta \rceil$  number of experts, *i.e.*,  $p_i^* = 1$  and the left workers are spammers, *i.e.*,  $p_i^* = \frac{1}{2}$ . Here we assume that  $\delta \in (0, \frac{1}{2})$ , that is only a small proportion of workers are experts.

Next theorem show that the expected average error of the outputs  $\{\bar{y}_j\}_{j \in [m]}$  in (7.18) of private major voting is larger than the average error in Theorem 7.1.2 if  $\epsilon$  in some range.

**Theorem 7.1.4.** For any  $\epsilon > 0$ , private major voting is  $\epsilon$ -LADP. Moreover, if  $\epsilon > \ln \frac{85}{15}$  and  $n > C$  for some sufficiently large constant  $C$  (only related to  $\delta$ ), then the outputs  $\{\bar{y}_j\}_{j \in [m]}$  satisfy

$$\frac{1}{m} \sum_{j \in [m]} \mathbb{E}|\bar{y}_j - y_j^*| \geq 2 \left( \exp\left(-\frac{1}{2} \left(\frac{e^\epsilon - 1}{e^\epsilon + 1}\right)^2 \lceil n^\delta \rceil\right) \right) \geq \frac{1}{m} \sum_{j \in [m]} |\hat{y}_j^{(T)} - y_j^*|, \quad (7.19)$$

where  $\{\hat{y}_j^{(T)}\}$  are outputs of Algorithm 7.1.54.

## 7.1.6 Omitted Proofs

### Proof of Theorem 7.1.1

Now consider  $X_i, X'_i \in \{0, 1\}^m$  differ in the  $j$ -th coordinate, *i.e.*,  $X_i = (X_{i1}, X_{i2}, \dots, X_{ij}, \dots, X_{im})$  and  $X'_i = (X_{i1}, X_{i2}, \dots, X'_{ij}, \dots, X_{im})$ . For any  $S \in \{0, 1\}^m$ , by the independence and the definition of (7.9) we have

$$\frac{\mathbb{P}(\hat{X}_i \in S)}{\mathbb{P}(\hat{X}'_i \in S)} = \frac{\mathbb{P}(\hat{X}_{ij} = S_j)}{\mathbb{P}(\hat{X}'_{ij} = S_j)} \quad (7.20)$$

when  $X_{ij} = 1$  and  $X'_{ij} = 0$  and  $S_j = 1$  then (7.20) equals  $e^\epsilon$ . When  $S_j = 0$  then (7.20) equals  $\frac{1}{e^\epsilon} \leq e^\epsilon$ . The same for the case where  $X_{ij} = 0$  and  $X'_{ij} = 1$ . Thus in total we can see that (7.20) less equals than  $e^\epsilon$ , which satisfies the definition of LADP. Moreover, due to the post-processing property of differential privacy [104], we know that Algorithm 7.1.54 is LADP.

### Proof of Theorem 7.1.2

By the definition of  $\hat{X}_{ij}$  and the assumption, we can represent it as

$$\hat{X}_{ij} = y_j^* T_{ij} + (1 - y_j^*)(1 - T_{ij}) \quad (7.21)$$

where  $T_{ij}$  is a Bernoulli random variable with parameter

$$\hat{p}_i^* = \frac{e^\epsilon}{e^\epsilon + 1} p_i^* + \frac{1}{e^\epsilon + 1} (1 - p_i^*) = \frac{e^\epsilon - 1}{e^\epsilon + 1} p_i^* + \frac{1}{e^\epsilon + 1}. \quad (7.22)$$

We notice that  $T_{ij}$  means that the  $i$ -th worker answers the  $j$ -th task correctly.

We also define the projected version of  $\hat{p}_j^*$  as

$$\hat{p}_{\lambda,i}^* = \lambda \mathbb{I}(\hat{p}_i^* < \lambda) + \hat{p}_i^* \mathbb{I}(\lambda \leq \hat{p}_i^* \leq 1 - \lambda) + (1 - \lambda) \mathbb{I}(\hat{p}_i^* > 1 - \lambda).$$

To proof Theorem 7.1.2, we first proof a stronger claim that for each iteration  $t \geq 1$ ,

$\{y_j^{(t)}\}_{j \in [m]}$  satisfies Eq. (7.16) with probability at least  $1 - \frac{C'}{m}$ .

We denote the error of  $\{y_j^{(t)}\}_{j \in [m]}$  as  $r^t$ , that is

$$r^t = \frac{1}{m} \sum_{j \in [m]} |y_j^{(t)} - y_j^*|.$$

By assumption (7.14) we know  $r^0 \leq \sqrt{\frac{\log m}{m}}$ . We first prove the following lemma:

**Lemma 7.1.1.** Define the events

$$E_1 = \left\{ \max_{i \in [n]} \left| \frac{1}{m} \sum_{j \in [m]} (T_{ij} - \hat{p}_i^*) \right| \leq \sqrt{\frac{\log m}{m}} \right\}.$$

$$E_2 = \left\{ \max_{j \in [m]} \left| \sum_{i \in [n]} (T_{ij} - \hat{p}_i^*) \log \frac{\hat{p}_{\lambda,i}^*}{1 - \hat{p}_{\lambda,i}^*} \right| \leq 2 \log\left(\frac{1}{\lambda}\right) \sqrt{n \log m} \right\}.$$

Then  $\mathbb{P}(E_1 \cap E_2) \geq 1 - \frac{C'}{m}$  for some  $C' > 0$ .

*Proof of Lemma 7.1.1.* To proof this, we recall the Hoeffding's inequality

**Lemma 7.1.2** (Hoeffding's inequality). For independent bounded random variables  $\{X_i\}_{i \in [n]}$  satisfying  $X_i \in [a_i, b_i]$  for all  $i \in [n]$ , we have for any  $t \geq 0$

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{i \in [n]} (X_i - \mathbb{E}X_i)\right| > t\right) \leq 2 \exp\left(\frac{-2n^2t^2}{\sum_{i \in [n]} (b_i - a_i)^2}\right).$$

Note that for the Event  $E_1$ , by Lemma 7.1.2, we have  $\mathbb{P}(E_1) \geq 1 - \frac{C_1}{m}$  for some  $C_1 > 0$ .

For the event  $E_2$ , we note that by the definition of  $\hat{p}_{\lambda,i}^*$  we have

$$\log \frac{\hat{p}_{\lambda,i}^*}{1 - \hat{p}_{\lambda,i}^*} \leq \log \frac{1 - \lambda}{\lambda} \leq \log \frac{1}{\lambda}.$$

Thus, by Lemma 7.1.2, we know there is a  $C_2 > 0$ , where  $\mathbb{P}(E_2) \geq 1 - \frac{C_2}{m}$ .  $\square$

In the following we will always assume events  $E_1$  and  $E_2$  in Lemma 7.1.1 hold. Next we will prove the following lemma:

**Lemma 7.1.3.** Under the event  $E_1$ , as long as  $2\lambda + r^{t-1} \leq \frac{1}{4}$  and  $m \geq 9$ , we have for all  $t \geq 1$ :

$$\max_{i \in [n]} \left| \log \frac{p_i^{(t)}}{1 - p_i^{(t)}} - \log \frac{\hat{p}_{\lambda,i}^*}{1 - \hat{p}_{\lambda,i}^*} \right| \leq \frac{2}{\lambda} \sqrt{\frac{\log m}{m}} + \frac{2}{\lambda} r^{t-1}.$$

*Proof of Lemma 7.1.3.* We note that from Eq. (7.6) and Eq. (7.10) on  $\{\hat{X}_{ij}\}_{i \in [n], j \in [m]}$  we can get

$$p_i^{(t)} = \lambda \mathbb{I}(\bar{p}_i^{(t)} < \lambda) + \bar{p}_i^{(t)} \mathbb{I}(\lambda \leq \bar{p}_i^{(t)} \leq 1 - \lambda) + (1 - \lambda) \mathbb{I}(\bar{p}_i^{(t)} > 1 - \lambda).$$

Where  $\bar{p}_i^{(t)}$  is the value of (7.6), *i.e.*, the vector before projecting. By the definitions (7.21) and (7.6) we can get the following via simple calculations:

$$|\bar{p}_i^{(t)} - \hat{p}_i^*| \leq \left| \frac{1}{m} \sum_j (T_{ij} - \hat{p}_i^*) \right| + r^{t-1}. \quad (7.23)$$

To show (7.23), by definition of  $\bar{p}_i^{(t)}$  we have

$$p_i^{(t)} = \frac{1}{m} \sum_{j \in [m]} ((1 - \hat{X}_{ij})(1 - y_j^{(t-1)}) + \hat{X}_{ij}y_j^{(t-1)}). \quad (7.24)$$

Now we fix  $j \in [m]$  and assume that  $y_j^* = 1$ , then by (7.21) we have  $\hat{X}_{ij} = T_{ij}$ , we can get

$$|(1 - \hat{X}_{ij})(1 - y_j^{(t-1)}) + \hat{X}_{ij}y_j^{(t-1)} - \hat{p}_i^*| = |2T_{ij}y_j^{(t-1)} - T_{ij} - y_j^{(t-1)} + 1 - p_i^*|. \quad (7.25)$$

When  $T_{ij} = 0$ , (7.25) is  $|y_j^{(t-1)} - 1 + p_i^*| \leq |p_i^*| + |y_j^{(t-1)} - y_j^*|$ . When  $T_{ij} = 1$ , (7.25) is  $|y_j^{(t-1)} - p_i^*| \leq |1 - p_i^*| + |y_j^{(t-1)} - y_j^*|$ . Thus in total we have (7.25) less than  $|T_{ij} - p_i^*| + |y_j^{(t-1)} - y_j^*|$ . The same for the case when  $y_j^* = 0$ .

Taking the average from 1 to  $m$  we can get (7.23).

Now we have for each  $i \in [n]$

$$|\log \frac{p_i^{(t)}}{1-p_i^{(t)}} - \log \frac{\hat{p}_{\lambda,i}^*}{1-\hat{p}_{\lambda,i}^*}| \leq \frac{\frac{p_i^{(t)}}{1-p_i^{(t)}} - \frac{\hat{p}_{\lambda,i}^*}{1-\hat{p}_{\lambda,i}^*}}{\min\left\{\frac{p_i^{(t)}}{1-p_i^{(t)}}, \frac{\hat{p}_{\lambda,i}^*}{1-\hat{p}_{\lambda,i}^*}\right\}} \quad (7.26)$$

$$\leq \frac{2}{\lambda} |p_i^{(t)} - \hat{p}_{\lambda,i}^*| \quad (7.27)$$

$$\leq \frac{2}{\lambda} |\bar{p}_i^{(t)} - \hat{p}_i^*| + \frac{4}{\lambda} \mathbb{I}(|\bar{p}_i^{(t)} - \hat{p}_i^*| > 1 - 2\lambda) \quad (7.28)$$

$$\leq \frac{2}{\lambda} \left( \left| \frac{1}{m} \sum_j (T_{ij} - \hat{p}_i^*) \right| + r^{t-1} \right) + \frac{4}{\lambda} \mathbb{I} \left( \left| \frac{1}{m} \sum_j (T_{ij} - \hat{p}_i^*) \right| > \frac{3}{4} \right) \quad (7.29)$$

$$\leq \frac{2}{\lambda} \sqrt{\frac{\log m}{m}} + \frac{2}{\lambda} r^{t-1}. \quad (7.30)$$

Where the first inequality (7.26) is due to the following inequality

$$|\log x - \log y| \leq \frac{|x - y|}{\min\{x, y\}}.$$

The inequality (7.27) is due to that  $\lambda \leq p_i^{(t)}, \hat{p}_{\lambda,i}^* \leq 1 - \lambda$  and simple calculation.

The inequality (7.28) is due to the following. When  $|\bar{p}_i^{(t)} - \hat{p}_i^*| > 1 - 2\lambda$ , then  $|p_i^{(t)} - \hat{p}_{\lambda,i}^*| \leq 2$ .

Otherwise by the definition we have either  $\bar{p}_i^{(t)}$  or  $\hat{p}_i^*$  is in the interval  $[\lambda, 1 - \lambda]$ , thus we have  $|p_i^{(t)} - \hat{p}_{\lambda,i}^*| \leq |\bar{p}_i^{(t)} - \hat{p}_i^*|$  due to the property of contraction of the projection.

The inequality (7.29) is due to the following. By (7.23) we have

$$\begin{aligned} \mathbb{I}(|\bar{p}_i^{(t)} - \hat{p}_i^*| > 1 - 2\lambda) &\leq \mathbb{I} \left( \left| \frac{1}{m} \sum_j (T_{ij} - \hat{p}_i^*) \right| + r^{t-1} > 1 - 2\lambda \right) \\ &= \mathbb{I} \left( \left| \frac{1}{m} \sum_j (T_{ij} - \hat{p}_i^*) \right| > 1 - 2\lambda - r^{t-1} \right) \\ &\leq \mathbb{I} \left( \left| \frac{1}{m} \sum_j (T_{ij} - \hat{p}_i^*) \right| > \frac{3}{4} \right), \end{aligned} \quad (7.31)$$

where the last inequality is due to the assumption that  $2\lambda + r^{t-1} \leq \frac{1}{4}$ .

The inequality (7.29) is due to the assumption of the event  $E_1$  in Lemma 7.1.1 holds.

Thus, we get the proof.  $\square$

Now we back to the proof of Theorem 7.1.2.

Since we already know that  $r^0 \leq \sqrt{\frac{\log m}{m}}$  and we want to show it hold for all  $r^t$ .

We will prove it by induction, assume  $r^{t-1} \leq \sqrt{\frac{\log m}{m}}$  holds. Denote the terms

$A_j^t, B_j^t, j \in [m]$  as

$$A_j^t = \log \prod_{i \in [n]} (p_i^{(t)})^{\hat{X}_{ij}} (1 - p_i^{(t)})^{1 - \hat{X}_{ij}} = \sum_i \left( \hat{X}_{ij} \log p_i^{(t)} + (1 - \hat{X}_{ij}) \log (1 - p_i^{(t)}) \right)$$

$$B_j^t = \log \prod_{i \in [n]} (1 - p_i^{(t)})^{\hat{X}_{ij}} (p_i^{(t)})^{1 - \hat{X}_{ij}} = \sum_i \left( \hat{X}_{ij} \log (1 - p_i^{(t)}) + (1 - \hat{X}_{ij}) \log p_i^{(t)} \right).$$

Then by the definition of  $\{y_j^{(t)}\}_{j=1}^m$  we have

$$\begin{aligned} r^t &= \frac{1}{m} \sum_j |y_j^{(t)} - y_j^*| = \frac{1}{m} \sum_j \left| \frac{\exp(A_j^t)}{\exp(A_j^t) + \exp(B_j^t)} - y_j^* \right| \\ &= \frac{1}{m} \sum_j \left| \frac{\exp(A_j^t - B_j^t)}{\exp(A_j^t - B_j^t) + 1} - y_j^* \right| \\ &= \frac{1}{m} \sum_j \frac{1}{1 + \exp(\sum_i (2T_{ij} - 1) \log \frac{p_i^{(t)}}{1 - p_i^{(t)}})} \\ &\leq \frac{1}{m} \sum_j \exp(-\sum_i (2T_{ij} - 1) \log \frac{p_i^{(t)}}{1 - p_i^{(t)}}) \\ &\leq \frac{1}{m} \sum_j \exp(-\sum_i (2T_{ij} - 1) \log \frac{\hat{p}_{\lambda,i}^*}{1 - \hat{p}_{\lambda,i}^*} + \frac{4n}{\lambda} \sqrt{\frac{\log m}{m}}) \\ &\leq \frac{1}{m} \sum_j \exp(-\sum_i (2\hat{p}_i^* - 1) \log \frac{\hat{p}_{\lambda,i}^*}{1 - \hat{p}_{\lambda,i}^*}) \exp(\frac{4n}{\lambda} \sqrt{\frac{\log m}{m}} + 4 \log \frac{1}{\lambda} \sqrt{n \log m}). \end{aligned}$$

Where the equalities are followed by the direct computation. The second inequality is by Lemma 7.1.3 and the assumption of  $r^{t-1} \leq \sqrt{\frac{\log m}{m}}$ , the third inequality is due to Lemma 7.1.1.

For the exponent in the first term we have

$$\begin{aligned}
& \sum_i (2\hat{p}_i^* - 1) \log \frac{\hat{p}_{\lambda,i}^*}{1 - \hat{p}_{\lambda,i}^*} \\
&= (\sum_{i:\hat{p}_i^*<\lambda} + \sum_{i:\lambda\leq\hat{p}_i^*\leq 1-\lambda} + \sum_{i:\hat{p}_i^*>1-\lambda}) (2\hat{p}_i^* - 1) \log \frac{\hat{p}_{\lambda,i}^*}{1 - \hat{p}_{\lambda,i}^*} \\
&\geq (|\{i : \hat{p}_i^* < \lambda\}| + |\{i : \hat{p}_i^* > 1 - \lambda\}|)(1 - 2\lambda) \log \frac{1 - \lambda}{\lambda} + \sum_{i:\lambda\leq\hat{p}_i^*\leq 1-\lambda} (2\hat{p}_i^* - 1) \log \frac{\hat{p}_{\lambda,i}^*}{1 - \hat{p}_{\lambda,i}^*} \\
&\geq (|\{i : \hat{p}_i^* < \lambda\}| + |\{i : \hat{p}_i^* > 1 - \lambda\}|) + \sum_{i:\lambda\leq\hat{p}_i^*\leq 1-\lambda} (2\hat{p}_i^* - 1)^2 \\
&\geq \sum_i (2\hat{p}_i^* - 1)^2
\end{aligned}$$

In the following we will show that  $(2\hat{p}_i^* - 1)^2 = (2\hat{\mu}_i - 1)^2$ , by this we have  $\sum_i (2\hat{p}_i^* - 1)^2 = n\hat{v}$ . This is due to the following equation:

$$\hat{\mu}_i = \hat{p}_i^* \mathbb{I}(\hat{p}_i^* \geq \frac{1}{2}) + (1 - \hat{p}_i^*) \mathbb{I}(\hat{p}_i^* < \frac{1}{2}). \quad (7.32)$$

Thus, in total we have

$$r^t \leq \exp\left(\frac{4n}{\lambda} \sqrt{\frac{\log m}{m}} + 4 \log \frac{1}{\lambda} \sqrt{n \log m} - n\hat{v}\right) \quad (7.33)$$

$$\leq \exp\left(-\frac{1}{2}n\hat{v}\right) \leq \sqrt{\frac{\log m}{m}}. \quad (7.34)$$

Where the second inequality is due to the assumption on the range of  $\lambda$ .

Next, due to the rounding procedure (Step 8 of Algorithm 7.1.54) and

$$\begin{aligned}
|\mathbb{I}(y_j \geq \frac{1}{2}) - 1| &\leq 2|y_j - 1| \\
|\mathbb{I}(y_j < \frac{1}{2}) - 1| &\leq 2|y_j - 0|
\end{aligned}$$

We have  $\frac{1}{m} \sum_{j \in [m]} |\hat{y}_j^{(T)} - y_j^*| \leq 2 \exp(-\frac{1}{2}n\hat{v})$ .

### Proof of Theorem 7.1.3

For (7.17), due to (7.23) we can see for each  $j \in [n]$

$$|p_j^{(T)} - \hat{p}_j^*| \leq \sqrt{\frac{\log m}{m}} + r^{t-1} \leq 2\sqrt{\frac{\log m}{m}}.$$

By the definition of  $\hat{p}_i^*$ . We have

$$|p_j^{(T)} - \frac{e^\epsilon - 1}{e^\epsilon + 1} p_j^* + \frac{1}{e^\epsilon + 1}| \leq 2\sqrt{\frac{\log m}{m}}$$

which implies

$$|\hat{p}_j^{(T)} - p_j^*| \leq 2\frac{e^\epsilon + 1}{e^\epsilon - 1}\sqrt{\frac{\log m}{m}} \leq \frac{6}{\epsilon}\sqrt{\frac{\log m}{m}}.$$

### Proof of Theorem 7.1.4

To proof Theorem 7.1.4, we need the Berry-Essen Lemma in [255]:

**Lemma 7.1.4.** Let  $X_1, \dots, X_n$  be i.i.d random variables with mean 0 and variance  $\sigma^2$ .

Define the function  $F_n(t) = \mathbb{P}(\frac{1}{\sigma\sqrt{n}} \sum_i X_i \leq t)$ . Then we have

$$\sup_{t \in \mathbb{R}} |F_n(t) - \Phi(t)| \leq \frac{c\mathbb{E}|X_1|^3}{\sigma\sqrt{n}}, \quad (7.35)$$

where  $c < 0.4748$  and  $\Phi(t)$  is the cumulative distribution function of the standard Gaussian distribution  $\mathcal{N}(0, 1)$ .

By the definition of majority voting and the definition in (7.21) we have  $|\bar{y}_j - y_j^*| = \mathbb{I}(\frac{1}{n} \sum_i T_{ij} < \frac{1}{2})$ . To prove this, we first consider the case where  $y_j^* = 1$ . Then by (7.21) we

have

$$\begin{aligned} |\bar{y}_j - y_j^*| &= |\mathbb{I}(\sum_{i \in [n]} \hat{X}_{ij} \geq \frac{n}{2}) - 1| \\ &= |\mathbb{I}(\sum_{i \in [n]} T_{ij} \geq \frac{n}{2}) - 1| = \mathbb{I}(\frac{1}{n} \sum_i T_{ij} < \frac{1}{2}). \end{aligned}$$

The same for the case where  $y_j^* = 1$ .

Thus

$$\begin{aligned} \frac{1}{m} \sum_{j \in [m]} \mathbb{E}|\bar{y}_j - y_j^*| &= \frac{1}{m} \sum_{j \in [m]} \mathbb{P}(\frac{1}{n} \sum_i T_{ij} < \frac{1}{2}) \\ &= \mathbb{P}(\frac{1}{n} \sum_i T_i < \frac{1}{2}), \end{aligned}$$

where  $\{T_i\}_{i \in [n]}$  are independent Bernoulli random variable with parameter  $\hat{p}_i^*$  in (7.22).

By the definition (7.22), we known that if  $p_i^* = \frac{1}{2}$  then  $\hat{p}_i^* = \frac{1}{2}$ , if  $p_i^* = 1$  then  $\hat{p}_i^* = \frac{e^\epsilon}{e^\epsilon + 1}$ . W.l.o.g we assume  $p_i^* = \frac{1}{2}$  for  $i \leq n - \lceil n^\delta \rceil$ . By Lemma 7.1.4 with  $\mathbb{E}[T_i - \frac{1}{2}] = 0$ ,  $\text{Var}(T_i - \frac{1}{2}) = \frac{1}{4}$  and  $\mathbb{E}|T_i - \frac{1}{2}|^3 = \frac{1}{8}$  for  $i \leq n - \lceil n^\delta \rceil$  we have

$$\sup_t |\mathbb{P}\left\{\frac{2}{\sqrt{n - \lceil n^\delta \rceil}} \sum_{i \leq n - \lceil n^\delta \rceil} (T_i - \frac{1}{2}) \leq t\right\} - \Phi(t)| \leq \frac{(n - \lceil n^\delta \rceil)^{-\frac{1}{2}}}{16}. \quad (7.36)$$

Also by direct calculation we have

$$\begin{aligned} \mathbb{P}(\frac{1}{n} \sum_i T_i < \frac{1}{2}) &\geq \mathbb{P}\left\{\frac{2}{\sqrt{n - \lceil n^\delta \rceil}} \sum_{i \leq n - \lceil n^\delta \rceil} (T_i - \frac{1}{2}) \right. \\ &\quad \left. \geq -\frac{\lceil n^\delta \rceil}{\sqrt{n - \lceil n^\delta \rceil}}\right\} \times \mathbb{P}\{T_i = 1, \forall i > n - \lceil n^\delta \rceil\} \end{aligned}$$

Thus by (7.36) we have

$$\mathbb{P}(\frac{1}{n} \sum_i T_i < \frac{1}{2}) \geq (\frac{e^\epsilon}{e^\epsilon + 1})^{\lceil n^\delta \rceil} \times \left\{\Phi\left(-\frac{\lceil n^\delta \rceil}{\sqrt{n - \lceil n^\delta \rceil}}\right) - \frac{(n - \lceil n^\delta \rceil)^{-\frac{1}{2}}}{16}\right\}. \quad (7.37)$$

We know that since  $\delta < \frac{1}{2}$ , thus for sufficiently large  $n$  we have  $\Phi(-\frac{\lceil n^\delta \rceil}{\sqrt{n - \lceil n^\delta \rceil}}) \geq \frac{1}{4}$  and  $\frac{(n - \lceil n^\delta \rceil)^{-\frac{1}{2}}}{16} \leq \frac{1}{8}$ , which are due to that

$$\lim_{n \rightarrow \infty} \Phi(-\frac{\lceil n^\delta \rceil}{\sqrt{n - \lceil n^\delta \rceil}}) = \Phi(0) = \frac{1}{2},$$

$$\lim_{n \rightarrow \infty} \frac{(n - \lceil n^\delta \rceil)^{-\frac{1}{2}}}{16} = 0.$$

Thus

$$\mathbb{P}(\frac{1}{n} \sum_i T_i < \frac{1}{2}) \geq (\frac{e^\epsilon}{e^\epsilon + 1})^{\lceil n^\delta \rceil} \frac{1}{8}.$$

On the other side by Theorem 7.1.2 we have

$$\frac{1}{m} \sum_{j \in [m]} |\hat{y}_j^{(T)} - y_j^*| \leq 2 \left( \exp(-\frac{1}{2}(\frac{e^\epsilon - 1}{e^\epsilon + 1})^2 \lceil n^\delta \rceil) \right).$$

Now we will show that for large enough  $n$ :

$$2 \left( \exp(-\frac{1}{2}(\frac{e^\epsilon - 1}{e^\epsilon + 1})^2 \lceil n^\delta \rceil) \right) \leq (\frac{e^\epsilon}{e^\epsilon + 1})^{\lceil n^\delta \rceil} \frac{1}{8}. \quad (7.38)$$

Denote  $v = \frac{e^\epsilon}{e^\epsilon + 1} \in [0.85, 1)$ , it is equivalent to show

$$\frac{\exp(-\frac{1}{2}(2v - 1)^2 \lceil n^\delta \rceil)}{v^{\lceil n^\delta \rceil}} \leq \frac{1}{16} \quad (7.39)$$

Thus, it is sufficient if we can show the following

$$\lim_{n \rightarrow \infty} \frac{\exp(-\frac{1}{2}(2v - 1)^2 \lceil n^\delta \rceil)}{v^{\lceil n^\delta \rceil}} = 0. \quad (7.40)$$

we note LHS of (7.40) equals to  $\exp((-\frac{1}{2}(2v - 1)^2 - \log v) \lceil n^\delta \rceil)$ , we will show  $f(v) = \frac{1}{2}(2v - 1)^2 + \log v > 0$  under our assumption on  $\epsilon$ . This is due to that  $f(v)$  is an increasing function, it is easy to see that  $f(0.85) > 0$ . Thus we proof Eq. (7.40).

## 7.2 Differentially Private Expectation Maximization Algorithm

As one of the most popular techniques for estimating the maximum likelihood of mixture models or incomplete data problems, Expectation Maximization (EM) algorithm has been widely applied to many areas such as genomics [188], finance [113], and crowdsourcing [86]. EM algorithm is well-known for its convergence to an empirically good local estimator [346]. Recent studies have further revealed that it can also provide finite sample statistical guarantees [19, 369, 339, 353]. Specifically, [19] showed that classical EM and its gradient ascent variant (gradient EM) are capable of achieving the first local convergence (theory) and finite sample statistical rate of convergence. They also provided a (near) optimal minimax rate for some canonical statistical models such as Gaussian mixture model (GMM), mixture of regressions model (MRM) and linear regression with missing covariates (RMC).

The wide applications of EM also present some new challenges to this method. Particularly, due to the existence of sensitive data and their distributed nature in many applications like social science, biomedicine, and genomics, it is often challenging to preserve the privacy of such data as they are extremely difficult to aggregate and learn from. Consider a case where health records are scattered across multiple hospitals (or even countries), it is not possible to process the whole dataset in a central server due to privacy and ownership concerns. A better solution is to use some differentially private mechanisms to conduct the aggregation and learning tasks.

Thus, to be able to use (gradient) EM algorithm to learn from these sensitive data, it is urgent to design some DP versions of the (gradient) EM algorithm. [242] proposed the first DP EM algorithm which mainly focuses on the practical behaviors of the method. Their algorithm needs quite a few assumptions on the model and the data, which make it difficult to extend to some canonical models mentioned above. Furthermore, unlike the aforementioned non-private case, their algorithm does not provide any finite sample statistical guarantee on

the solution (see Related Work section for detailed comparison). Thus, it is still unknown **whether there exists any DP variant of the (gradient) EM algorithm that has finite sample statistical guarantees.**

To answer this question, I propose in this section the first  $(\epsilon, \delta)$ -DP (Gradient) EM algorithm with finite sample statistical guarantees. Specifically,

- I first show that, given an appropriate initialization  $\beta^{\text{init}}$  (*i.e.*,  $\|\beta^{\text{init}} - \beta^*\|_2 \leq \kappa \|\beta^*\|_2$  for some constant  $\kappa \in (0, 1)$ ), if the model satisfies some additional assumptions and the number of sample  $n$  is large enough, the output  $\beta^{\text{priv}}$  of our DP EM algorithm is guaranteed to have a bounded estimation error,  $\|\beta^{\text{priv}} - \beta^*\|_2 \leq \tilde{O}(\frac{d\sqrt{\tau}}{\sqrt{n}\epsilon})$ , with high probability, where  $d$  is the dimensionality and  $\tau$  is an upper bound of the second moment of each coordinate of the gradient function.
- I then apply that general framework to the three canonical models: GMM, MRM and RMC. Our private estimator achieves an estimation error that is upper bounded by  $\tilde{O}(\frac{d}{\sqrt{n}\epsilon})$ ,  $\tilde{O}(\frac{d^{\frac{3}{2}}}{\sqrt{n}\epsilon})$  and  $\tilde{O}(\frac{d^{\frac{3}{2}}}{\sqrt{n}\epsilon})$  for GMM, MRM and RMC, respectively. It is notable that they are the first statistical guarantees for MRM and RMC in the Differential Privacy model, and the error bound for GMM is near optimal in some cases. I also conduct thorough experiments on the these three models. Experimental results on these models are consistent with theoretical analysis.

### 7.2.1 Related Work

As mentioned previously, designing DP version of EM algorithm is still not well studied. To our best knowledge, the only work on DP EM algorithm is given by [242]. However, their result is incomparable with ours for the following reasons. Firstly, our work aims to achieve finite sample statistical guarantees for the DP EM algorithm, while [242] mainly focuses on designing practical DP EM algorithm that does not provide any statistical guarantees. Particularly, [242] assumed that datasets are pre-processed such that the  $\ell_2$ -norm of each

data record is less than 1. This means that their algorithm will likely introduce additional bias on the statistical guarantees. Secondly, [242] studied only the exponential family so that noise can be directly added to the sufficient statistics. However, most of the latent variable models do not satisfy such an assumption. This includes the MRM and RMC models to be considered in this section.

In this section, we implement our general framework on three specific models, and DP GMM is the only one that has been studied previously. Specifically, [233] provided the first result for the general  $k$ -GMM based on the sample-and-aggregate framework. Later on, [172] improved the result by a factor of  $\sqrt{d}/\epsilon$ , and also claimed that their sample complexity is near optimal. Compared with their result, our proposed algorithm ensures that when the error  $\alpha$  is some constant, it has the same sample complexity. Also, although their algorithm has polynomial time complexity), it is actually not very practical and thus no practical study has been conducted. Moreover, their algorithm is heavily dependent on a previous clustering algorithm; it is unclear whether it can be extended to other mixture models. From these two perspectives, our framework is more general and practical.

## 7.2.2 Preliminaries

### Expectation Maximization

Let  $Y$  and  $Z$  be two random variables taking values in the sample spaces  $\mathcal{Y}$  and  $\mathcal{Z}$ , respectively. Suppose that the pair  $(Y, Z)$  has a joint density function  $f_{\beta^*}$  that belongs to some parameterized family  $\{f_{\beta^*} | \beta^* \in \Omega\}$ . Rather than considering the whole pair of  $(Y, Z)$ , we observe only component  $Y$ . Thus, component  $Z$  can be viewed as the missing or latent structure. We assume that the term  $h_\beta(y)$  is the margin distribution over the latent variable  $Z$ , *i.e.*,  $h_\beta(y) = \int_{\mathcal{Z}} f_\beta(y, z) dz$ . Let  $k_\beta(z|y)$  be the density of  $Z$  conditional on the observed variable  $Y = y$ , that is,  $k_\beta(z|y) = \frac{f_\beta(y, z)}{h_\beta(y)}$ .

Given  $n$  observations  $y_1, y_2, \dots, y_n$  of  $Y$ , the EM algorithm is to maximize the log-

likelihood  $\max_{\beta \in \Omega} \ell_n(\beta) = \sum_{i=1}^n \log h_\beta(y_i)$ . Due to the unobserved latent variable  $Z$ , it is often difficult to directly evaluate  $\ell_n(\beta)$ . Thus, we consider the lower bound of  $\ell_n(\beta)$ . By Jensen's inequality, we have

$$\begin{aligned} \frac{1}{n} [\ell_n(\beta) - \ell_n(\beta')] &\geq \frac{1}{n} \sum_{i=1}^n \int_{\mathcal{Z}} k_{\beta'}(z|y_i) \log f_\beta(y_i, z) dz \\ &\quad - \frac{1}{n} \sum_{i=1}^n \int_{\mathcal{Z}} k_{\beta'}(z|y_i) \log f_{\beta'}(y_i, z) dz. \end{aligned} \quad (7.41)$$

Let  $Q_n(\beta; \beta') = \frac{1}{n} \sum_{i=1}^n q_i(\beta; \beta')$ , where

$$q_i(\beta; \beta') = \int_{\mathcal{Z}} k_{\beta'}(z|y_i) \log f_\beta(y_i, z) dz.^3 \quad (7.42)$$

Also, it is convenient to let  $Q(\beta; \beta')$  denote the expectation of  $Q_n(\beta; \beta')$  w.r.t  $\{y_i\}_{i=1}^n$ , that is,

$$Q(\beta; \beta') = \mathbb{E}_{y \sim h_{\beta^*}} \int_{\mathcal{Z}} k_{\beta'}(z|y) \log f_\beta(y, z) dz. \quad (7.43)$$

We can see that the second term on the right hand side of (7.41) is independent on  $\beta$ . Thus, given some fixed  $\beta'$ , we can maximize the lower bound function  $Q_n(\beta; \beta')$  over  $\beta$  to obtain sufficiently large  $\ell_n(\beta) - \ell_n(\beta')$ . Thus, in the  $t$ -th iteration of the standard EM algorithm, we can evaluate  $Q_n(\cdot; \beta^t)$  at the E-step and then perform the operation of  $\beta^{t+1} = \max_{\beta \in \Omega} Q_n(\beta; \beta^t)$  at the M-step. See [215] for more details.

In addition to the exact maximization implementation of the M-step, we add a gradient ascent implementation of the M-step, which performs an approximate maximization via a gradient descent step.

**Gradient EM Algorithm [19]** When  $Q_n(\cdot; \beta^t)$  is differentiable, the update of  $\beta^t$  to  $\beta^{t+1}$  consists of the following two steps.

- E-step: Evaluate the functions in (7.42) to compute  $Q_n(\cdot; \beta^t)$ .

---

<sup>3</sup>We use  $q(\beta; \beta')$  for general sample  $y$ .

- M-step: Update  $\beta^{t+1} = \beta^t + \eta \nabla Q_n(\beta^t; \beta^t)$ , where  $\nabla$  is the derivative of  $Q_n$  w.r.t the first component and  $\eta$  is the step size.

Next, we give some examples that use the gradient EM algorithm. Note that they are the typical examples for studying the statistical property of EM algorithm [339, 19, 353, 369].

**Gaussian Mixture Model (GMM)** Let  $y_1, \dots, y_n$  be  $n$  i.i.d samples from  $Y \in \mathbb{R}^d$  with

$$Y = Z \cdot \beta^* + V, \quad (7.44)$$

where  $Z$  is a Rademacher random variable (*i.e.*,  $\mathbb{P}(Z = +1) = \mathbb{P}(Z = -1) = \frac{1}{2}$ ), and  $V \sim \mathcal{N}(0, \sigma^2 I_d)$  is independent of  $Z$  for some known standard deviation  $\sigma$ . We have

$$\nabla q(\beta; \beta) = [2w_\beta(y) - 1] \cdot y - \beta, \quad (7.45)$$

where  $w_\beta(y) = \frac{1}{1 + \exp(-\langle \beta, y \rangle / \sigma^2)}$ .

**Mixture of (Linear) Regressions Model (MRM)** Let  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  be  $n$  samples i.i.d sampled from  $Y \in \mathbb{R}$  and  $X \in \mathbb{R}^d$  with

$$Y = Z \langle \beta^*, X \rangle + V, \quad (7.46)$$

where  $X \sim \mathcal{N}(0, I_d)$ ,  $V \sim \mathcal{N}(0, \sigma^2)$ ,  $Z$  is a Rademacher random variable, and  $X, V, Z$  are independent. In this case we have

$$\nabla q(\beta; \beta) = (2w_\beta(x, y) - 1) \cdot y \cdot x - xx^T \cdot \beta, \quad (7.47)$$

where  $w_\beta(x, y) = \frac{1}{1 + \exp(-y \langle \beta, x \rangle / \sigma^2)}$ .

**Linear Regression with Missing Covariates (RMC)** We assume that  $Y \in \mathbb{R}$  and  $X \in \mathbb{R}^d$  satisfy

$$Y = \langle X, \beta^* \rangle + V, \quad (7.48)$$

where  $X \sim \mathcal{N}(0, I_d)$  and  $V \sim \mathcal{N}(0, \sigma^2)$  are independent. Let  $x_1, x_2, \dots, x_n$  be  $n$  observations of  $X$  with each coordinate of  $x_i$  missing (unobserved) independently with probability  $p_m \in [0, 1]$ . In this case, we have

$$\nabla q(\beta; \beta) = y \cdot m_\beta(x^{\text{obs}}, y) - K_\beta(x^{\text{obs}}, y)\beta, \quad (7.49)$$

where the functions  $m_\beta(x^{\text{obs}}, y) \in \mathbb{R}^d$  and  $K_\beta(x^{\text{obs}}, y) \in \mathbb{R}^{d \times d}$  are defined as:

$$m_\beta(x^{\text{obs}}, y) = z \odot x + \frac{y - \langle \beta, z \odot x \rangle}{\sigma^2 + \|(1-z) \odot \beta\|_2^2} (1-z) \odot \beta \quad (7.50)$$

and

$$\begin{aligned} K_\beta(x^{\text{obs}}, y) &= \text{diag}(1-z) + m_\beta(x^{\text{obs}}, y) \cdot [m_\beta(x^{\text{obs}}, y)]^T \\ &\quad - [(1-z) \odot m_\beta(x^{\text{obs}}, y)] \cdot [(1-z) \odot m_\beta(x^{\text{obs}}, y)]^T, \end{aligned} \quad (7.51)$$

where vector  $z \in \mathbb{R}^d$  is defined as  $z_j = 1$  if  $x_j$  is observed and  $z_j = 0$  if  $x_j$  is missing, and  $\odot$  denotes the Hadamard product of matrices.

Next, we provide several definitions on the required properties of functions  $Q_n(\cdot; \cdot)$  and  $Q(\cdot; \cdot)$ . Note that some of them have been used in previous studies on the statistical guarantees of EM algorithm [19, 339, 369].

**Definition 7.2.1.** Function  $Q(\cdot; \beta^*)$  is self-consistent if  $\beta^* = \arg \max_{\beta \in \Omega} Q(\beta; \beta^*)$ . That is,  $\beta^*$  maximizes the lower bound of the log likelihood function.

**Definition 7.2.2** (Lipschitz-Gradient-2( $\gamma, \mathcal{B}$ )).  $Q(\cdot; \cdot)$  is called Lipschitz-Gradient-2( $\gamma, \mathcal{B}$ ), if for the underlying parameter  $\beta^*$  and any  $\beta \in \mathcal{B}$  for some set  $\mathcal{B}$ , the following holds

$$\|\nabla Q(\beta; \beta^*) - \nabla Q(\beta; \beta)\|_2 \leq \gamma \|\beta - \beta^*\|_2. \quad (7.52)$$

We note that there are some differences between the definition of Lipschitz-Gradient-2

and the Lipschitz continuity condition in the convex optimization literature [230]. Firstly, in (7.52), the gradient is w.r.t the second component, while the Lipschitz continuity is w.r.t the first component. Secondly, the property holds only for fixed  $\beta^*$  and any  $\beta$ , while the Lipschitz continuity is for all  $\beta, \beta' \in \mathcal{B}$ .

**Definition 7.2.3** ( $\mu$ -smooth).  $Q(\cdot; \beta^*)$  is  $\mu$ -smooth, that is if for any  $\beta, \beta' \in \mathcal{B}$ ,  $Q(\beta; \beta^*) \geq Q(\beta'; \beta^*) + (\beta - \beta')^T \nabla Q(\beta'; \beta^*) - \frac{\mu}{2} \|\beta' - \beta\|_2^2$ .

**Definition 7.2.4** ( $v$ -strongly concave).  $Q(\cdot; \beta^*)$  is  $v$ -strongly concave, that is if for any  $\beta, \beta' \in \mathcal{B}$ ,  $Q(\beta; \beta^*) \leq Q(\beta'; \beta^*) + (\beta - \beta')^T \nabla Q(\beta'; \beta^*) - \frac{v}{2} \|\beta' - \beta\|_2^2$ .

In the following we will propose the assumptions that will be used throughout the whole section. Note that these assumptions are commonly used in other works on statistical analysis of EM algorithm such as [20, 369, 339].

**Assumption 7.2.1.** We assume that function  $Q(\cdot; \cdot)$  in (7.43) is self-consistent, Lipschitz-Gradient-2( $\gamma, \mathcal{B}$ ),  $\mu$ -smooth,  $v$ -strongly concave over some set  $\mathcal{B}$ . Moreover, we assume that  $\forall j \in [d]$  and  $\beta \in \mathcal{B}$ , there is some known upper bound  $\tau$  on the second-order moment of the  $j$ -coordinate of  $\nabla q(\beta, \beta)$ , i.e.,  $\mathbb{E}(\nabla_j q(\beta, \beta))^2 \leq \tau$  and for each  $i \in [n]$ ,  $\nabla_j q_i(\beta, \beta)$  is independent with others.

Due to the similarity with the Gradient Descent algorithm and the simplicity of illustrating our idea compared with the original EM algorithm, we will first focus on DP Gradient EM algorithm.

### 7.2.3 Main Method

#### Main Difficulty

In the previous section, we introduced the Gradient EM algorithm, which updates the estimator via the gradient  $\nabla Q_n(\beta^t; \beta^t)$ . It is notable that this idea is quite similar to the Gradient Descent algorithm. Moreover, we know that there are several DP versions of the

(Stochastic) Gradient Descent algorithm such as [29, 328, 260, 298, 192]. The key idea of DP Gradient Descent is adding some randomized noise such as Gaussian noise to preserve DP property in each iteration, and by the composition theorem of DP ([104]), the whole algorithm will still be DP. Thus, motivated by this, to design a DP variant of Gradient EM algorithm, the most direct way is adding some Gaussian noise to the gradient  $\nabla Q_n(\beta^t; \beta^t)$  in each iteration and updating the parameter.

However, it is notable that we cannot add Gaussian noise directly to the gradient in the Gradient EM algorithm. The main reason is that all previous DP Gradient Descent algorithms need to assume that each component of the gradient (which correspond to the function  $\nabla q_i$  in (7.42)) is bounded, or the loss function is  $O(1)$ -Lipschitz, such as Logistic Regression, so that its  $\ell_2$ -norm sensitivity is bounded and thus the Gaussian mechanism can be used. However, in the Gradient EM algorithm, each component ( $\nabla q_i(\beta^t; \beta^t)$  in (7.42)) is unbounded in most of the cases. For example, we can easily show the following fact.

**Theorem 7.2.1.** Consider the GMM in (7.44), there is a case with fixed  $\beta$ , such that for each constant  $c$ , with **positive probability** w.r.t  $y$  we have  $\|\nabla q(\beta; \beta)\|_2 \geq c$ .

Thus, to design a DP (Gradient) EM algorithm, the major difficulty lies in how to process the gradient to make its sensitivity bounded. Two main approaches are used in previous work: (1) [242] assumed that datasets are pre-processed such that the  $\ell_2$  norm of each sample is bounded by 1. However, as mentioned previously, our goal is to achieve the statistical guarantees for the DP (Gradient) EM algorithm. If a similar approach is adopted in our algorithm, the (manual) normalization can easily destroy many statistical properties of the data and force the private estimator to introduce additional bias, making it inconsistent.<sup>4</sup> (2) Instead of normalizing the datasets, [1] first clipped the gradient to ensure that the  $\ell_2$ -norm of each component of the gradient is bounded by the threshold  $C$ , and then added Gaussian noise (see Algorithm 7.2.55 for more details). However, such an approach may cause two issues. First, in general clipping gradient could introduce additional bias even in statistical

---

<sup>4</sup>An estimator  $\beta_n$  is consistent if  $\lim_{n \rightarrow \infty} \|\beta_n - \beta^*\|_2 = 0$ .

estimation, which has also been pointed out in [261]. Second, the threshold  $C$  heavily affects the convergence speed and selecting the best  $C$  is quite difficult (see Experimental section for more details). Due to these two reasons, it is hard to study the statistical guarantees of Algorithm 7.2.55. Thus, we need a new approach to pre-process the gradient to ensure that it has not only bounded  $\ell_2$ -norm but also consistent statistical guarantee.

---

**Algorithm 7.2.55** Clipped DP Gradient EM

---

**Input:**  $D = \{y_i\}_{i=1}^n \subset \mathbb{R}^d$ , privacy parameters  $\epsilon, \delta$ ;  $Q_n(\cdot; \cdot)$  and its  $q(\cdot; \cdot)$ , initial parameter  $\beta^0$ , gradient norm  $C$ , step size  $\eta$  and the number of iterations  $T$ .

- 1: **for**  $t = 1, 2, \dots, T$  **do**
- 2:     For each  $i \in [n]$ , evaluate the function in (7.42) to compute  $q_i(\beta; \beta^{t-1})$ .
- 3:     Clip gradient:

$$\nabla \bar{q}_i(\beta^{t-1}; \beta^{t-1}) = \frac{\nabla q_i(\beta^{t-1}; \beta^{t-1})}{\max\left\{1, \frac{\|\nabla q_i(\beta^{t-1}; \beta^{t-1})\|_2}{C}\right\}}.$$

- 4:     Update  $\beta^t = \beta^{t-1} + \eta(\nabla \bar{Q}_n(\beta^{t-1}; \beta^{t-1}) + \mathcal{N}(0, C^2 \sigma^2 I_d))$ , where  $\nabla \bar{Q}_n(\beta^{t-1}; \beta^{t-1}) = \frac{1}{n} \sum_{i=1}^n \nabla \bar{q}_i(\beta^{t-1}; \beta^{t-1})$  and  $\sigma^2 = c \frac{T \log \frac{1}{\delta}}{n^2 \epsilon^2}$  for some constant  $c$ .
  - 5: **end for**
  - 6: Return  $\beta^T$
- 

## Our Method

In this section, we will propose our method to overcome the aforementioned difficulties.

Our method is motivated by a robust and private mean estimator for heavy-tailed distributions, which was given in [331], and it is derived from the robust mean estimator in [148]. To be self-contained, we first review their estimator. Now, we consider a 1-dimensional random variable  $x$  and assume that  $x_1, x_2, \dots, x_n$  are i.i.d. sampled from  $x$ . The estimator consists of three steps:

**Scaling and Truncation** For each sample  $x_i$ , we first re-scale it by dividing  $s$  (which will be specified later). Then, we apply the re-scaled one to some soft truncation function  $\phi$ .

Finally, we put the truncated mean back to the original scale. That is,

$$\frac{s}{n} \sum_{i=1}^n \phi\left(\frac{x_i}{s}\right) \approx \mathbb{E}X. \quad (7.53)$$

Here, we use the function given in [62],

$$\phi(x) = \begin{cases} x - \frac{x^3}{6}, & -\sqrt{2} \leq x \leq \sqrt{2} \\ \frac{2\sqrt{2}}{3}, & x > \sqrt{2} \\ -\frac{2\sqrt{2}}{3}, & x < -\sqrt{2}. \end{cases} \quad (7.54)$$

Note that a key property for  $\phi$  is that  $\phi$  is bounded, that is,  $|\phi(x)| \leq \frac{2\sqrt{2}}{3}$ .

**Noise Multiplication** Let  $\eta_1, \eta_2, \dots, \eta_n$  be random noise generated from a common distribution  $\eta \sim \chi$  with  $\mathbb{E}\eta = 0$ . We multiply each data  $x_i$  by a factor of  $1 + \eta_i$ , and then perform the scaling and truncation step on the term  $x_i(1 + \eta_i)$ . That is,

$$\tilde{x}(\eta) = \frac{s}{n} \sum_{i=1}^n \phi\left(\frac{x_i + \eta_i x_i}{s}\right). \quad (7.55)$$

**Noise Smoothing** In this final step, we smooth the multiplicative noise by taking the expectation w.r.t. the distributions. That is,

$$\hat{x} = \mathbb{E}\tilde{x}(\eta) = \frac{s}{n} \sum_{i=1}^n \int \phi\left(\frac{x_i + \eta_i x_i}{s}\right) d\chi(\eta_i). \quad (7.56)$$

Computing the explicit form of each integral in (7.56) depends on the function  $\phi(\cdot)$  and the distribution  $\chi$ . Fortunately, [62] showed that when  $\phi$  is in (7.54) and  $\chi \sim \mathcal{N}(0, \frac{1}{\beta})$  (where  $\beta$  will be specified later), we have for any  $a$  and  $b > 0$

$$\mathbb{E}_\eta \phi(a + b\sqrt{\beta}\eta) = a\left(1 - \frac{b^2}{2}\right) - \frac{a^3}{6} + C(a, b), \quad (7.57)$$

where  $C(a, b)$  is a correction form which is easy to implement and it has the following explicit form: we first define the following notations:

$$\begin{aligned} V_- &:= \frac{\sqrt{2} - a}{b}, V_+ = \frac{\sqrt{2} + a}{b} \\ F_- &:= \Phi(-V_-), F_+ := \Phi(-V_+) \\ E_- &:= \exp\left(-\frac{V_-^2}{2}\right), E_+ := \exp\left(-\frac{V_+^2}{2}\right), \end{aligned}$$

where  $\Phi$  denotes the CDF of the standard Gaussian distribution. Then

$$C(a, b) = T_1 + T_2 + \cdots + T_5,$$

where

$$\begin{aligned} T_1 &:= \frac{2\sqrt{2}}{3}(F_- - F_+) \\ T_2 &:= -(a - \frac{a^3}{6})(F_- + F_+) \\ T_3 &:= \frac{b}{\sqrt{2\pi}}(1 - \frac{a^2}{2})(E_+ - E_-) \\ T_4 &:= \frac{ab^2}{2} \left( F_+ + F_- + \frac{1}{\sqrt{2\pi}}(V_+E_+ + V_-E_-) \right) \\ T_5 &:= \frac{b^3}{6\sqrt{2\pi}} ((2 + V_-^2)E_- - (2 + V_+^2)E_+). \end{aligned}$$

[148] showed the following estimation error for the mean estimator  $\hat{x}$  after these three steps.

**Lemma 7.2.1** (Lemma 5 in [148]). Let  $x_1, x_2, \dots, x_n$  be i.i.d. samples from distribution  $x \sim \mu$ . Assume that there is some known upper bound on the second-order moment, *i.e.*,  $\mathbb{E}_\mu x^2 \leq \tau$ . For a given failure probability  $\zeta$ , if set  $\beta = 2 \log \frac{1}{\zeta}$  and  $s = \sqrt{\frac{n\tau}{2 \log \frac{1}{\zeta}}}$ , then with

probability at least  $1 - \zeta$  the following holds

$$|\hat{x} - \mathbb{E}x| \leq O\left(\sqrt{\frac{\tau \log \frac{1}{\zeta}}{n}}\right). \quad (7.58)$$

To obtain an  $(\epsilon, \delta)$ -DP estimator, the key observation is that the bounded function  $\phi$  in (7.54) also makes the integral form of (7.56) bounded by  $\frac{2\sqrt{2}}{3}$ . Thus, we know that the  $\ell_2$ -norm sensitivity is  $\frac{s}{n} \frac{4\sqrt{2}}{3}$ . Hence, the query

$$\mathcal{A}(D) = \hat{x} + Z, Z \sim \mathcal{N}(0, \sigma^2), \sigma^2 = O\left(\frac{s^2 \log \frac{1}{\delta}}{\epsilon^2 n^2}\right) \quad (7.59)$$

will be  $(\epsilon, \delta)$ -DP, which leads to the following result.

**Lemma 7.2.2** (Theorem 6 in [331]). Under the assumptions in Lemma 7.2.1, with probability at least  $1 - \zeta$  the following holds

$$|\mathcal{A}(D) - \mathbb{E}(x)| \leq O\left(\sqrt{\frac{v \log \frac{1}{\delta} \log \frac{1}{\zeta}}{n \epsilon^2}}\right). \quad (7.60)$$

It is notable that in Lemma 7.2.2 we just need to assume that  $x$  has bounded second order moment, instead of bounded norm. However, since we need weaker assumptions here, the error bound in (7.60) is larger than it for the bounded distributions [54].

Inspired by the previous private 1-dimensional mean estimation, we propose our method (Algorithm 7.2.56). In Algorithm 7.2.56, the key idea is that, in the  $t$ -th iteration of Gradient EM algorithm, we first apply the previous private estimator to each coordinate of the gradient  $\nabla Q_n(\beta^{t-1}; \beta^{t-1})$ , and then perform the M-step. We can easily show that Algorithm 7.2.56 is  $(\epsilon, \delta)$ -DP.

**Theorem 7.2.2** (Privacy guarantee). For any  $0 < \epsilon, \delta < 1$ , Algorithm 7.2.56 is  $(\epsilon, \delta)$ -DP.

In the following, we will show the statistical guarantee for the models under Assumption 7.2.1, if the initial parameter  $\beta^0$  is closed to the underlying parameter  $\beta^*$  enough.

---

**Algorithm 7.2.56** DP Gradient EM Algorithm

---

**Input:**  $D = \{y_i\}_{i=1}^n \subset \mathbb{R}^d$ , privacy parameters  $\epsilon, \delta, Q(\cdot; \cdot)$  and its  $q_i(\cdot; \cdot)$ , initial parameter  $\beta^0 \in \mathcal{B}$ ,  $\tau$  which satisfies Assumption 7.2.1, the number of iterations  $T$  (to be specified later), step size  $\eta$  and failure probability  $\zeta > 0$ .

- 1: Let  $\tilde{\epsilon} = \sqrt{\log \frac{1}{\delta} + \epsilon} - \sqrt{\log \frac{1}{\delta}}$ ,  $s = \sqrt{\frac{n\tau}{2\log \frac{d}{\zeta}}}$ ,  $\beta = \log \frac{d}{\zeta}$ .
- 2: **for**  $t = 1, 2, \dots, T$  **do**
- 3:     For each  $j \in [d]$ , calculate the robust gradient by (7.53)-(7.57) and add Gaussian noise, that is

$$g_j^{t-1}(\beta^{t-1}) = \frac{1}{n} \sum_{i=1}^n \left( \nabla_j q_i(\beta^{t-1}, \beta^{t-1}) \left( 1 - \frac{\nabla_j^2 q_i(\beta^{t-1}, \beta^{t-1})}{2s^2 \beta} \right) - \frac{\nabla_j^3 q_i(\beta^{t-1}, \beta^{t-1})}{6s^2} \right) + \frac{s}{n} \sum_{i=1}^n C \left( \frac{\nabla_j q_i(\beta^{t-1}, \beta^{t-1})}{s}, \frac{|\nabla_j q_i(\beta^{t-1}, \beta^{t-1})|}{s\sqrt{\beta}} \right) + Z_j^{t-1}, \quad (7.61)$$

where  $Z_j^{t-1} \sim \mathcal{N}(0, \sigma^2)$  with  $\sigma^2 = \frac{8\tau d T}{9\beta n \tilde{\epsilon}^2}$ .

- 4:     Let vector  $\tilde{\nabla}Q_n(\beta^{t-1}) \in \mathbb{R}^d$  denote  $\tilde{\nabla}Q_n(\beta^{t-1}) = (g_1^{t-1}(\beta^{t-1}), g_2^{t-1}(\beta^{t-1}), \dots, g_d^{t-1}(\beta^{t-1}))$ .
  - 5:     Update  $\beta^t = \beta^{t-1} + \eta \tilde{\nabla}Q_n(\beta^{t-1})$ .
  - 6: **end for**
- 

**Theorem 7.2.3** (Statistical guarantee of Algorithm 7.2.56). Let the parameter set  $\mathcal{B} = \{\beta : \|\beta - \beta^*\|_2 \leq R\}$  for  $R = \kappa \|\beta^*\|_2$  for some constant  $\kappa \in (0, 1)$ . Assume that Assumption 7.2.1 holds for parameters  $\gamma, \mathcal{B}, \mu, v, \tau$  satisfying the condition of  $1 - 2\frac{v-\gamma}{v+\mu} \in (0, 1)$ . Also, assume that  $\|\beta^0 - \beta^*\|_2 \leq \frac{R}{2}$ ,  $n$  is large enough so that

$$\tilde{\Omega}\left(\left(\frac{1}{v-\gamma}\right)^2 \frac{d^2 \tau T \log \frac{1}{\delta} \log \frac{1}{\zeta}}{\epsilon^2 R^2}\right) \leq n. \quad (7.62)$$

Then, with probability at least  $1 - 2T\zeta$ , we have, for all  $t \in [T]$ ,  $\beta^t \in \mathcal{B}$ . If it holds and if taking  $T = O\left(\frac{\mu+v}{v-\gamma} \log n\right)$  and  $\eta = \frac{2}{\mu+v}$ , we have

$$\|\beta^T - \beta^*\|_2 \leq \tilde{O}\left(R \sqrt{\frac{v+\mu}{(v-\gamma)^3}} \frac{d \log \frac{1}{\delta} \log \frac{1}{\zeta} \sqrt{\tau}}{\sqrt{n\epsilon^2}}\right), \quad (7.63)$$

where the  $\tilde{O}$ -term and  $\tilde{\Omega}$ -term omit  $\log d, \log n$  and other factors (see Appendix for the explicit form of the result).

**Remark 7.2.1.** There are several points that need to note. Firstly, the assumptions of the parameter set  $\beta$  and the initial parameter  $\beta^0$  are commonly used in other papers on statistical guarantees of (Gradient) EM algorithm such as [20, 369, 339]. Even though Theorem 7.2.3 requires that the initial estimator be close enough to the optimal one, our experiments show that the algorithm actually performs quite well for any random initialization. Secondly, in (7.62) we need to assume that  $n \propto \frac{1}{R^2}$ , where  $R$  is the radius of  $\mathcal{B}$ . This is due to that in Algorithm 7.2.56, we need to keep each  $\beta^t \in \mathcal{B}$  under perturbation. When  $R$  is small, we have to let the noise be small enough, which means that  $n$  should be large enough. Finally, for specific models,  $R, v, \mu, \gamma$  are constants, this means that the error in (7.63) is  $\tilde{O}(\frac{d\sqrt{\tau}}{\sqrt{n}\epsilon})$ . However, here  $\tau$  depends on the model, which may also depend on  $d$  and  $\|\beta^*\|_2$ .

## 7.2.4 Implications for Some Specific Models

In this section, we apply our framework (*i.e.*, Algorithm 7.2.56) to the models mentioned in the Preliminaries section. To obtain results for these models, we only need to find the corresponding  $\mathcal{B}, \gamma, k, R, v, \mu, \tau$  to ensure that Assumption 7.2.1 and the assumptions in Theorem 7.2.3 hold.

### Gaussian Mixture Model

The following lemma ensures the properties of Lipschitz-Gradient-2( $\gamma, \mathcal{B}$ ), smoothness, strongly concave and self-consistency for model (7.44).

**Lemma 7.2.3** ([19, 353]). If  $\frac{\|\beta^*\|_2}{\sigma} \geq r$ , where  $r$  is a sufficiently large constant denoting the minimum signal-to-noise ratio (SNR), then there exists an absolute constant  $C > 0$  such that the properties of self-consistent, Lipschitz-Gradient-2( $\gamma, \mathcal{B}$ ),  $\mu$ -smoothness and  $v$ -strongly concave hold for function  $Q(\cdot; \cdot)$  with  $\gamma = \exp(-Cr^2)$ ,  $\mu = v = 1$ ,  $R = k\|\beta^*\|_2$ ,  $k = \frac{1}{4}$ , and  $\mathcal{B} = \{\beta : \|\beta - \beta^*\|_2 \leq R\}$ .

We can show the following second-order moment bound for  $\nabla_j q(\beta, \beta)$ .

**Lemma 7.2.4.** With the same notations as in Lemma 7.2.3, for each  $\beta \in \mathcal{B}$ , the  $j$ -the coordinate of  $\nabla q(\beta; \beta)$  (*i.e.*,  $\nabla_j q(\beta; \beta)$ ) satisfies the following inequality

$$\mathbb{E}_y (\nabla_j q(\beta; \beta))^2 \leq O((\|\beta^*\|_\infty^2 + \sigma^2)).$$

Also, for fixed  $j \in [d]$ , each  $\nabla_j q_i(\beta; \beta)$ , where  $i \in [n]$ , is independent with others.

Combining with Lemma 7.2.3, 7.2.4 and Theorem 7.2.3 we have the following statistical guarantee for GMM.

**Theorem 7.2.4.** With the same notations as in Lemma 7.2.3, in Algorithm 7.2.56 assume that  $\|\beta^0 - \beta^*\|_2 \leq \frac{1}{8}\|\beta^*\|_2$  and  $n$  is large enough so that

$$\tilde{\Omega}\left(\frac{d^2 \sqrt{\|\beta^*\|_\infty^2 + \sigma^2} \log \frac{1}{\delta} \log \frac{1}{\zeta}}{\epsilon^2 \|\beta^*\|_2^2}\right) \leq n. \quad (7.64)$$

Moreover, if take  $T = O(\log n)$  and  $\eta = O(1)$ , then we have with probability at least  $1 - 2T\zeta$

$$\|\beta^T - \beta^*\|_2 \leq \tilde{O}\left(\|\beta^*\|_2 \frac{d \log \frac{1}{\delta} \log \frac{1}{\zeta} \sqrt{\|\beta^*\|_\infty^2 + \sigma^2}}{\sqrt{n\epsilon^2}}\right), \quad (7.65)$$

where the  $\tilde{O}, \tilde{\Omega}$  terms omit logarithmic and other factors.

**Remark 7.2.2.** Note that if we assume that  $\sigma, \|\beta^*\|_2 = O(1)$ , then the error in (7.65) is upper bounded by  $\tilde{O}\left(\frac{d}{\sqrt{n\epsilon}}\right)$ . This means that to achieve the error of  $\alpha \in (0, 1)$ , the sample complexity is  $\tilde{O}\left(\frac{d^2}{\alpha^2 \epsilon}\right)$ . It is notable that for GMM, the near optimal rate is  $\tilde{O}(d^2(\frac{1}{\alpha^2} + \frac{1}{\alpha\epsilon}))$  [172].<sup>5</sup> Thus when  $\epsilon$  is some constant, our result matches their near optimal rate. However, as mentioned in previous section, their algorithm is too complicated to be practical and it is difficult to extend their method to other Mixture models. Also, we assume that the SNR is large, which is reasonable since it has been shown that for Gaussian Mixture Model with

---

<sup>5</sup>Note that although [172] used TV distance, while we use the Euclidean distance, we can easily transfer our result to a result based on TV distance via Pinsker's inequality and the KL diatance between two Gaussian distributions.

low SNR, the variance of noise makes it harder for the algorithm to converge [210], which is the same for MRM.

### Mixture of Regressions Model

The following lemma, which was given in [19, 353], shows the properties of Lipschitz-Gradient-2( $\gamma, \mathcal{B}$ ), smoothness and strongly concave for model (7.46).

**Lemma 7.2.5** ([19, 353]). If  $\frac{\|\beta^*\|_2}{\sigma} \geq r$ , where  $r$  is a sufficiently large constant denoting the required minimal signal-to-noise ratio (SNR), then function  $Q(\cdot; \cdot)$  of the Mixture of Regressions Model has the properties of self-consistent, Lipschitz-Gradient-2( $\gamma, \mathcal{B}$ ),  $\mu$ -smoothness, and  $v$ -strongly with  $\gamma \in (0, \frac{1}{4})$ ,  $\mu = v = 1$ ,  $\mathcal{B} = \{\beta : \|\beta - \beta^*\|_2 \leq R\}$ ,  $R = k\|\beta^*\|_2$ , and  $k = \frac{1}{32}$ .

**Lemma 7.2.6.** With the same notations as in Lemma 7.2.5, for each  $\beta \in \mathcal{B}$ , the  $j$ -the coordinate of  $\nabla q_i(\beta; \beta)$ , i.e.,  $\nabla_j q(\beta; \beta)$  satisfies the following inequality

$$\mathbb{E}_y (\nabla_j q(\beta; \beta))^2 \leq O(\max\{(\|\beta^*\|_2^2 + \sigma^2)^2, d\|\beta^*\|_2^2\}).$$

Also, for fixed  $j \in [d]$ , each  $\nabla_j q_i(\beta; \beta)$  is independent with others for  $i \in [n]$ .

**Theorem 7.2.5.** With the same notations as in Lemma 7.2.5, in Algorithm 7.2.56 assume that  $\|\beta^0 - \beta^*\|_2 \leq \frac{1}{64}\|\beta^*\|_2$  and  $n$  is large enough so that

$$\tilde{\Omega}\left(\frac{d^2 \max\{(\|\beta^*\|_2^2 + \sigma^2)^2, d\|\beta^*\|_2^2\} \log \frac{1}{\delta} \log \frac{1}{\zeta}}{\epsilon^2 \|\beta^*\|_2^2}\right) \leq n.$$

Moreover, if take  $T = O(\log n)$  and  $\eta = O(1)$ , then we have, with probability at least  $1 - 2T\zeta$ ,

$$\|\beta^T - \beta^*\|_2 \leq \tilde{O}\left(\frac{d\|\beta^*\|_2 \log \frac{1}{\delta} \sqrt{\max\{\|\beta^*\|_2^2 + \sigma^2, d\|\beta^*\|_2^2\}}}{\sqrt{n\epsilon^2}}\right), \quad (7.66)$$

where the  $\tilde{O}$ -term and  $\tilde{\Omega}$ -term omit logarithmic factors.

**Remark 7.2.3.** If we assume that  $\|\beta^*\|$  and  $\sigma = O(1)$ , then the error in (7.66) is upper bounded by  $\tilde{O}(\frac{d^{\frac{3}{2}}}{\sqrt{n\epsilon}})$ , which has an additional factor of  $\sqrt{d}$  compared with the bound in (7.65) for GMM. We note that this is the first statistical result for MRM in the DP model.

### Linear Regression with Missing Covariates

**Lemma 7.2.7** ([19, 353]). If  $\frac{\|\beta^*\|_2}{\sigma} \leq r$  and  $p_m < \frac{1}{1+2b+2b^2}$ , where  $r$  is a constant denoting the required maximum signal-to-noise ratio (SNR) and  $b = r^2(1+k)^2$  for some constant  $k \in (0, 1)$ , then function  $Q(\cdot; \cdot)$  of the linear regression with missing covariates has the properties of self-consistent, Lipschitz-Gradient-2( $\gamma, \mathcal{B}$ ),  $\mu$ -smoothness and  $v$ -strongly with

$$\gamma = \frac{b + p_m(1 + 2b + 2b^2)}{1 + b} < 1, \mu = v = 1,$$

$$\mathcal{B} = \{\beta : \|\beta - \beta^*\|_2 \leq R\}, \text{ where } R = k\|\beta^*\|_2.$$

**Lemma 7.2.8.** With the same assumptions as in Lemma 7.2.7, for each  $\beta \in \mathcal{B}$  and  $j \in [d]$ ,  $\nabla_j q(\beta; \beta)$  satisfies

$$\mathbb{E}(\nabla_j q(\beta; \beta))^2 \leq O((\sqrt{d}\|\beta^*\|_2 + \sigma^2 + \|\beta^*\|_2^2)^2). \quad (7.67)$$

Also, for fixed  $j \in [d]$ , each  $\nabla_j q_i(\beta; \beta)$ , where  $i \in [n]$ , is independent with others.

**Theorem 7.2.6.** With the same notations as in Lemma 7.2.7, in Algorithm 7.2.56 assume that  $\|\beta^0 - \beta^*\|_2 \leq \frac{k}{2}\|\beta^*\|_2$  and  $n$  is large enough so that

$$\tilde{\Omega}\left(\frac{d^2(\sqrt{d}\|\beta^*\|_2 + \sigma^2 + \|\beta^*\|_2^2)^2 \log \frac{1}{\delta} \log \frac{1}{\zeta}}{\epsilon^2 \|\beta^*\|_2^2}\right) \leq n.$$

Moreover, if take  $T = O(\log n)$  and  $\eta = O(1)$ , then we have, with probability at least  $1 - 2T\zeta$ ,

$$\|\beta^T - \beta^*\|_2 \leq \tilde{O}\left(\frac{d \log \frac{1}{\delta} \log \frac{1}{\zeta} \|\beta^*\|_2 (\sqrt{d}\|\beta^*\|_2 + \sigma^2 + \|\beta^*\|_2^2)}{\sqrt{n\epsilon^2}}\right),$$

where the  $\tilde{O}, \tilde{\Omega}$  terms omit logarithmic and other factors.

Note that unlike the previous two models, we assume here that SNR is upper bounded by some constant which is unavoidable as pointed out in [201].

### 7.2.5 Statistical Guarantees of DP Expectation Maximization Algorithm

Motivated by idea of the Differentially Private version of Gradient EM algorithm in the previous section, in this section, we will propose a DP variant of EM algorithm.

Recall that compared with the Gradient EM algorithm, the main difference in EM algorithm is that, in each iteration, we will update the parameter as  $\beta^{t+1} = \arg \max_{\beta \in \Omega} Q_n(\beta; \beta^t)$ , where the  $Q_n$ -function is in (7.42). Thus, to design a DP variant, we need to post-process the parameter  $\beta^{t+1}$  via the private 1-dimensional mean estimation of heavy-tailed distribution. Just as the way we post-process the Gradient in Algorithm 7.2.56, we wish to post-process each coordinate of  $\beta^{t+1}$  to make it DP. However, unlike the Gradient EM algorithm where the  $\nabla Q_n(\beta; \beta')$  can be written as a sum of  $n$  independent components  $\frac{1}{n} \sum_{i=1}^n \nabla q_i(\beta; \beta')$ ,  $\beta^{t+1}$  in the EM algorithm may not be written as  $n$  independent components (see the Examples below), or even there is no explicit form of  $\beta^{t+1}$ . Thus, compared with the Assumption 7.2.1, we need addition assumptions on the form of  $\beta^{t+1} = \arg \max_{\beta \in \Omega} Q_n(\beta; \beta^t)$ , which may not hold for some canonical models.

**Assumption 7.2.2.** We assume that for a fixed  $\beta' \in \mathcal{B}$ , the optimal solution  $M_n(\beta') = \arg \max_{\beta \in \Omega} Q_n(\beta; \beta')$  satisfies  $M_n(\beta') = \frac{1}{n} \sum_{i=1}^n f_i(\beta')$ , where  $f_i(\cdot)$  is a function of  $y_i$ . Moreover, we assume that for each pair  $i \neq i'$ ,  $f_i(\beta'), f_{i'}(\beta')$  are independent. For any fixed  $j \in d$ , the  $j$ -th coordinate of  $f(\beta)$ <sup>6</sup> has bounded second order moment, *i.e.*,  $\mathbb{E}(f_j(\beta))^2 \leq \tau$ . We also assume that function  $Q(\cdot; \cdot)$  in (7.43) is self-consistent, Lipschitz-Gradient-2( $\gamma, \mathcal{B}$ ),  $v$ -strongly concave over some set  $\mathcal{B}$ .

---

<sup>6</sup>We denote function  $f(\cdot)$  as the function for general  $y$ .

Note that compared with Assumption 7.2.1, Assumption 7.2.2 does not need  $Q$  to be smooth. However, it needs some unnatural assumptions in the form of  $M_n(\beta')$ . To show that these assumptions are strong (especially the condition that  $f_i, f_{i'}$  are independent for each pair  $i \neq i'$ ), in the following, we will check the three canonical models in the previous section to see whether Assumption 7.2.2 holds.

**Gaussian Mixture Model** For GMM in (7.44), the  $Q$  function can be written as

$$Q_n(\beta; \beta') = -\frac{1}{2n} \sum_{i=1}^n (w_{\beta'}(y_i) \|y_i - \beta\|_2^2 + [1 - w_{\beta'}(y_i)] \|y_i + \beta\|_2^2).$$

where  $w_\beta(y) = \frac{1}{1 + \exp(-\langle \beta, y \rangle / \sigma^2)}$ . Thus, for  $M_n(\beta') = \arg \max_{\beta \in \mathbb{R}^d} Q_n(\beta; \beta')$  we have

$$M_n(\beta') = \frac{2}{n} \sum_{i=1}^n w_{\beta'}(y_i) y_i - \frac{1}{n} \sum_{i=1}^n y_i,$$

Thus

$$M_n(\beta') = \frac{1}{n} \sum_{i=1}^n f_i(\beta')$$

for  $f_i(\beta') = 2w_{\beta'}(y_i)y_i - y_i$  and for each  $i \in [n]$ ,  $f_j$  is independent with others. Later, combining with Lemma 7.2.3 we will show GMM satisfies Assumption 7.2.2.

**Mixture of Regressions Model** For MRM in (7.46), the  $Q_n$  function can be written as

$$Q_n(\beta; \beta') = \frac{1}{2n} (-w_{\beta'}(x_i, y_i)(y - \langle x_i, \beta \rangle)^2 + [1 - w_{\beta'}(x_i, y_i)](y + \langle x_i, \beta \rangle)^2),$$

where  $w_\beta(x, y) = \frac{1}{1 + \exp(-y \langle \beta, x \rangle / \sigma^2)}$ . Thus, for  $M_n(\beta') = \arg \max_{\beta \in \mathbb{R}^d} Q_n(\beta; \beta')$  we have

$$M_n(\beta') = \left( \frac{1}{n} \sum_{i=1}^n x_i x_i^T \right)^{-1} \cdot \left( \frac{1}{n} \sum_{i=1}^n [2w_{\beta'}(x_i, y_i) - 1] y_i x_i \right).$$

Thus  $M_n(\beta') = \frac{1}{n} \sum_{i=1}^n f_i(\beta')$  for  $f_i(\beta') = (\frac{1}{n} \sum_{i=1}^n x_i x_i^T)^{-1} \cdot [2w_{\beta'}(x_i, y_i) - 1]y_i x_i$ . However, we can see that due to the term of  $(\frac{1}{n} \sum_{i=1}^n x_i x_i^T)^{-1}$ , for each  $i \in [n]$ ,  $f_i$  is dependent with others. Thus, MRM **does not satisfy** Assumption 7.2.2.

**Linear Regression with Missing Covariates** For RMC in (7.48), the  $Q_n$  function can be written as

$$Q_n(\beta; \beta') = \frac{1}{n} \sum_{i=1}^n y_i \beta^T m_{\beta'}(x_i^{\text{obs}}, y_i) - \frac{1}{2n} \sum_{i=1}^n \beta^T K_{\beta'}(x_i^{\text{obs}}, y_i) \beta',$$

where the functions  $m_{\beta'}(x^{\text{obs}}, y)$ ,  $K_{\beta'}(x^{\text{obs}}, y)$  are in (7.50) and (7.51), respectively. Thus, for  $M_n(\beta') = \arg \max_{\beta \in \mathbb{R}^d} Q_n(\beta; \beta')$  we have

$$M_n(\beta') = \left( \frac{1}{n} \sum_{i=1}^n K_{\beta'}(x_i^{\text{obs}}, y_i) \right)^{-1} \cdot \left( \frac{1}{n} \sum_{i=1}^n y_i m_{\beta'}(x_i^{\text{obs}}, y_i) \right).$$

Thus  $M_n(\beta') = \frac{1}{n} \sum_{i=1}^n f_i(\beta')$  for  $f_i(\beta') = \left( \frac{1}{n} \sum_{i=1}^n K_{\beta'}(x_i^{\text{obs}}, y_i) \right)^{-1} (y_i m_{\beta'}(x_i^{\text{obs}}, y_i))$ . However, we can see that due to the term of  $\left( \frac{1}{n} \sum_{i=1}^n K_{\beta'}(x_i^{\text{obs}}, y_i) \right)^{-1}$ , for each  $i \in [n]$ ,  $f_i$  is dependent with others. Thus, RMC **does not satisfy** Assumption 7.2.2.

From the previous models, we can see that two of them do not satisfy the condition of  $f_i$  is independent with others. We note that this assumption is necessary for our analysis of statistical guarantees, since we will use the private 1-dimensional mean estimator, which needs the i.i.d assumption on the samples. Thus, from this point of view, we can see that our DP Gradient EM algorithm needs to be presented before the DP EM algorithm.

## 7.2.6 DP EM Algorithm

Next we will detail our DP EM algorithm and provide its statistical guarantee under Assumption 7.2.2, see Algorithm 7.2.57 for details. The key idea is that in each iteration, instead of post-processing the  $j$ -th coordinate of the gradient  $\nabla q_i(\beta^{t-1}, \beta^{t-1})$ , we will post-process  $j$ -th coordinate of the term  $f_i(\beta^{t-1})$ , i.e.,  $f_{i,j}(\beta^{t-1})$  via the previous private 1-dimension

mean estimator. We can easily show Algorithm 7.2.57 is  $(\epsilon, \delta)$ -DP.

---

**Algorithm 7.2.57** DP EM Algorithm

---

**Input:**  $D = \{y_i\}_{i=1}^n \subset \mathbb{R}^d$ , privacy parameters  $\epsilon, \delta, Q(\cdot; \cdot)$  and its  $f_i(\cdot)$  in Assumption 7.2.2, initial parameter  $\beta^0 \in \mathcal{B}$ ,  $\tau$  which satisfies Assumption 7.2.2, the number of iterations  $T$  (to be specified later), and failure probability  $\zeta$ .

- 1: Let  $\tilde{\epsilon} = \sqrt{\log \frac{1}{\delta} + \epsilon} - \sqrt{\log \frac{1}{\delta}}$ ,  $s = \sqrt{\frac{n\tau}{2\log \frac{d}{\zeta}}}$ ,  $\beta = \log \frac{d}{\zeta}$ .
- 2: **for**  $t = 1, 2, \dots, T$  **do**
- 3:     For each  $j \in [d]$ , calculate the robust estimator by (7.53)-(7.57) and add Gaussian noise, that is

$$g_j^{t-1}(\beta^{t-1}) = \frac{1}{n} \sum_{i=1}^n \left( f_{i,j}(\beta^{t-1}) \left( 1 - \frac{f_{i,j}^2(\beta^{t-1})}{2s^2\beta} \right) - \frac{f_{i,j}^3(\beta^{t-1})}{6s^2} \right) + \frac{1}{n} \sum_{i=1}^n C \left( \frac{f_{i,j}(\beta^{t-1})}{s}, \frac{|f_{i,j}(\beta^{t-1})|}{s\sqrt{\beta}} \right) + Z_j^{t-1}, \quad (7.68)$$

where  $f_{i,j}(\beta^{t-1})$  is the  $j$ -th coordinate of  $f_i(\beta^{t-1})$  and  $Z_j^{t-1} \sim \mathcal{N}(0, \sigma^2)$  with  $\sigma^2 = \frac{8\tau d T}{9\beta n \tilde{\epsilon}^2}$ .

- 4:     Let vector  $\tilde{f}(\beta^{t-1}) \in \mathbb{R}^d$  to denote  $\tilde{f}(\beta^{t-1}) = (g_1^{t-1}(\beta^{t-1}), g_2^{t-1}(\beta^{t-1}), \dots, g_d^{t-1}(\beta^{t-1}))$ .
  - 5:     Update  $\beta^t = \tilde{f}(\beta^{t-1})$ .
  - 6: **end for**
- 

**Theorem 7.2.7** (Privacy guarantee). For any  $0 < \epsilon, \delta < 1$ , Algorithm 7.2.56 is  $(\epsilon, \delta)$ -DP.

*Proof.* The proof is almost the same as that of Theorem 7.2.2; we thus omit it here.  $\square$

As in Theorem 7.2.3, in the following, we will show the statistical guarantee for the models under the Assumption 7.2.2, if the initial parameter  $\beta^0$  is close enough to the underlying parameter  $\beta^*$ .

**Theorem 7.2.8** (Statistical guarantee of Algorithm 7.2.57). Let the parameter set  $\mathcal{B} = \{\beta : \|\beta - \beta^*\|_2 \leq R\}$  for  $R = \kappa \|\beta^*\|_2$  for some constant  $\kappa \in (0, 1)$ . Assume that Assumption 7.2.2 holds for parameters  $\gamma, \mathcal{B}, v, \tau$  satisfying the condition of  $1 - 2\frac{v-\gamma}{v+\mu} \in (0, 1)$ . Also, assume that  $\|\beta^0 - \beta^*\|_2 \leq \frac{R}{2}$ ,  $n$  is large enough so that

$$\tilde{\Omega}\left(\left(\frac{v}{v-\gamma}\right)^2 \frac{d^2 \tau T \log \frac{1}{\delta} \log \frac{1}{\zeta}}{\epsilon^2 R^2}\right) \leq n. \quad (7.69)$$

Then with probability at least  $1 - 2T\zeta$ , we have for all  $t \in [T]$ ,  $\beta^t \in \mathcal{B}$ . If it holds and if we take  $T = O(\frac{v}{v-\gamma} \log n)$ , then we have

$$\|\beta^T - \beta^*\|_2 \leq \tilde{O}\left(R \sqrt{\frac{v}{(v-\gamma)^3}} \frac{d \log \frac{1}{\delta} \log \frac{1}{\zeta} \sqrt{\tau}}{\sqrt{n\epsilon^2}}\right), \quad (7.70)$$

where the  $\tilde{O}$ -term and  $\tilde{\Omega}$ -term omit  $\log d$ ,  $\log n$  and other factors (see Appendix for the explicit form of the result).

Comparing with Theorem 7.2.8 and Theorem 7.2.3, if we omit other factors instead of  $n, d, \epsilon, \delta$ , we can see that the two error bounds are asymptotically the same.

In the following we will apply our general framework to the GMM model in (7.44). Just the same as in Theorem 7.2.4, we will first show that  $f_j(\beta)$  has a bounded second order moment.

**Lemma 7.2.9.** Consider the function  $f(\cdot)$  in GMM. Then, for each  $j \in [d]$  we have

$$\mathbb{E} f_j^2(\beta) \leq O(\|\beta^*\|_\infty^2 + \sigma^2).$$

Thus, combining with Lemma 7.2.3, Lemma 7.2.9 and Theorem 7.2.8 we have asymptotically the same result as in Theorem 7.2.4. We omit the details here.

## 7.2.7 Experiments

In this section, we evaluate the performance of Algorithm 7.2.56 on three canonical models: GMM, MRM, and RMC. Since in the paper we mainly focus on the statistical setting and its theoretical behaviors, we only evaluate our algorithm on the synthetic data. Note that previous papers on the statistical guarantees of EM algorithm all evaluating their algorithms on synthetic data only such as [19, 353, 369]. Thus, evaluating experiments on synthetic data only is sufficient and reasonable for the paper.

**Baseline Methods** We compare our approach against two baseline algorithms. One is the

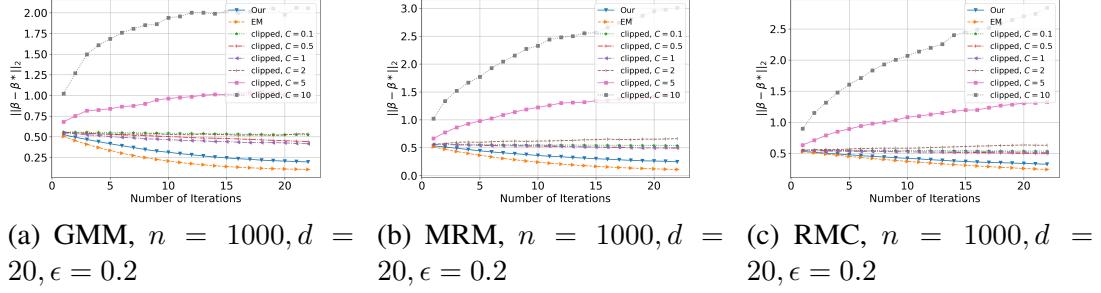


Figure 7.1: Estimation error of Algorithm 7.2.55 (clipped) v.s. iteration  $t$  under different clipping threshold  $C$

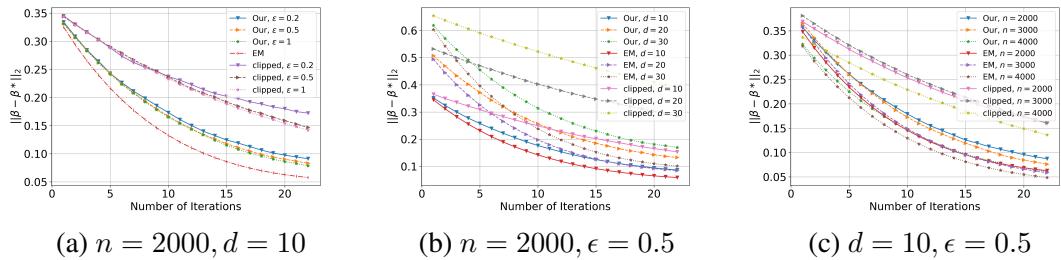


Figure 7.2: Estimation error of GMM w.r.t privacy budget  $\epsilon$ , data dimension  $d$ , data size  $n$  and iteration  $t$

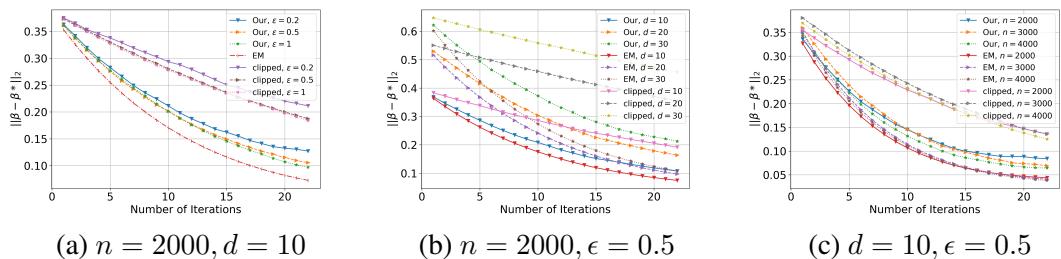


Figure 7.3: Estimation error of MRM w.r.t privacy budget  $\epsilon$ , data dimension  $d$ , data size  $n$  and iteration  $t$ .

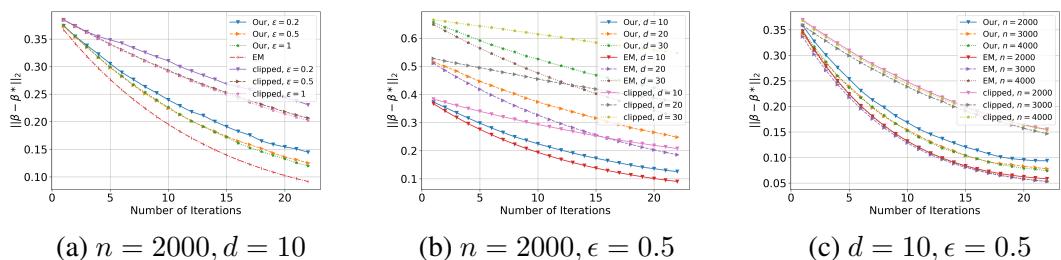


Figure 7.4: Estimation error of RMC w.r.t privacy budget  $\epsilon$ , data dimension  $d$ , data size  $n$  and iteration  $t$

gradient EM algorithm [19], namely, EM, as our non-private baseline method. The other is clipped DP Gradient EM (Algorithm 7.2.55), namely, clipped, as our private baseline method.

**Experimental Settings** For each of these models, we generate synthesized datasets according to the underlying distribution. We also utilize  $\|\beta - \beta^*\|_2$  to measure the estimation error. Instead of choosing the initial parameter  $\beta^0$  that is close to the optimal one, we consider random initialization. As we will see later, even if we select random initial parameter, the performance of our private estimator is good enough. We set signal-to-noise ratio  $\frac{\|\beta^*\|_2}{\sigma} = 3$ . For the privacy parameters, we choose  $\epsilon = \{0.2, 0.5, 1\}$  and  $\delta = \mathcal{O}(\frac{1}{n})$ .

**Experimental Results** Firstly, we will show that the performance of Algorithm 7.2.55 is heavily affected by the clipping threshold  $C$ . As shown in Figure 7.1, we conduct the algorithm on three canonical models with fixed data size  $n$ , dimension data  $d$ , and privacy budget  $\epsilon$ . If  $C$  is set to be a small value (e.g., 0.1), it significantly reduces the adding noise in each iteration but at the same time it leads much information loss in gradient estimation. Conversely, if  $C$  is set too high (e.g., 5 or 10), the noise variance becomes high, resulting in introducing too much noise to the estimation. Thus, selecting the optimal  $C$  is quite difficult since too large or too small values of  $C$  has a negative effect on the performance of Algorithm 7.2.55. Even for  $C = 1$  that achieves lowest estimation error among other threshold values, the estimation error does not decay as the number of iterations increases, whereas under the same privacy guarantee, our proposed algorithm achieves the same convergence behavior as EM, and thoroughly outperforms Algorithm 7.2.55. For fair comparison, we thus fixed  $C = 1$  for Algorithm 7.2.55 in the following experiments.

In Figure 7.2, 7.3 and 7.4, we test how the privacy budget  $\epsilon$ , data dimension  $d$  and data size  $n$  affect the estimation error  $\|\beta - \beta^*\|_2$  of all algorithms on three canonical models over iteration  $t$ . We can see that the estimation error of our proposed algorithm in each of the three models decreases when  $\epsilon$  increases,  $d$  decreases or  $n$  increases, which are consistent with our theoretical results. In these figures, our algorithm exhibits nearly the same convergence

behavior as the non-private baseline method and outperforms Algorithm 7.2.55.

In Figure 7.5, 7.6 and 7.7, we set  $T = 22$  and compute the estimation error on  $\beta = \beta^T$ . We plot  $\|\beta - \beta^*\|_2$  of all algorithm on three canonical models over data size  $n$ , data dimension  $d$  and privacy budget  $\epsilon$ . As we can see from these figures, our proposed algorithm (Algorithm 2) on the three canonical models significantly outperforms the clipped algorithm (Algorithm 1).

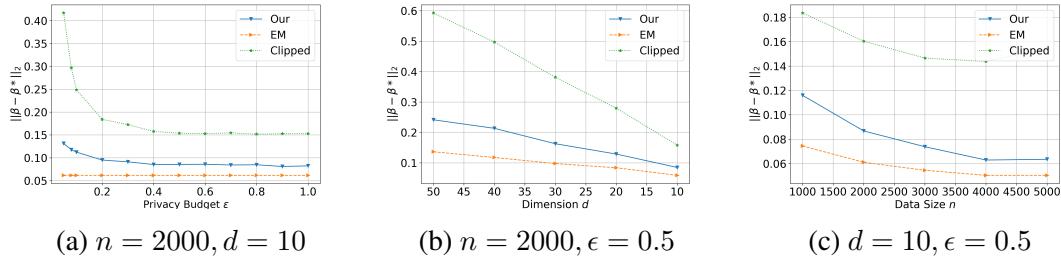


Figure 7.5: Estimation error of GMM w.r.t privacy budget  $\epsilon$ , data dimension  $d$  and data size  $n$  (we set  $\beta = \beta^T$  with  $T = 22$ )

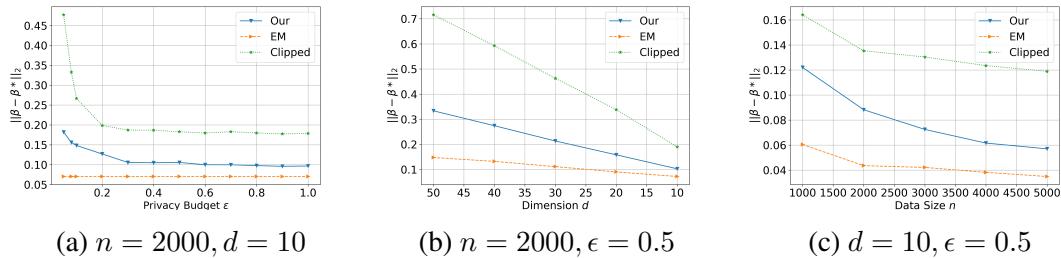


Figure 7.6: Estimation error of MRM w.r.t privacy budget  $\epsilon$ , data dimension  $d$  and data size  $n$  (we set  $\beta = \beta^T$  with  $T = 22$ )

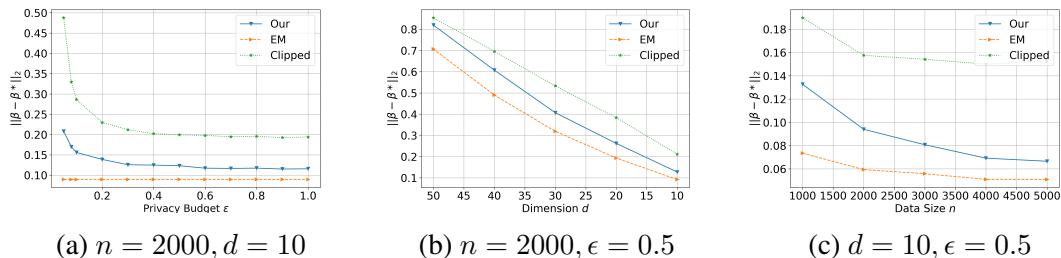


Figure 7.7: Estimation error of RMC w.r.t privacy budget  $\epsilon$ , data dimension  $d$  and data size  $n$  (we set  $\beta = \beta^T$  with  $T = 22$ )

## 7.2.8 Omitted Proofs

### Technical Lemmas

First, we will recall some definitions and lemmas on the sub-exponential and sub-Gaussian random variables. See [289] for details.

**Definition 7.2.5.** For a sub-exponential random vector  $X$ , its sub-exponential norm  $\|X\|_{\psi_1}$  is defined as

$$\|X\|_{\psi_1} = \sup_{p \geq 1} p^{-1} (\mathbb{E}|X|^p)^{\frac{1}{p}}.$$

**Definition 7.2.6** ( $\xi$ -sub-exponential). A random variable  $X$  with mean  $\mathbb{E}(X)$  is  $\xi$ -sub-exponential for  $\xi > 0$  if for all  $|t| < \frac{1}{\xi}$ ,  $\mathbb{E}\{\exp(t[X - \mathbb{E}(X)])\} \leq \exp(\frac{\xi^2 t^2}{2})$ .

**Lemma 7.2.10.** Let  $X$  be a sub-exponential random variable, then there are absolute constants  $C, c > 0$ , such that when  $|t| \leq \frac{c}{\|X\|_{\psi_1}}$ ,

$$\mathbb{E}[\exp(tX)] \leq \exp(Ct^2\|X\|_{\psi_1}^2).$$

**Lemma 7.2.11.** From Definition 7.2.5, 7.2.6 we can see that for a zero-mean sub-exponential random variable  $X$ , its second-order moment is bounded, *i.e.*,  $\mathbb{E}X^2 \leq O(\|X\|_{\psi_1}^2)$ .

**Lemma 7.2.12** (Bernstein's inequality). Let  $X_1, \dots, X_n$  be  $n$  i.i.d realizations of  $v$ -sub-exponential random variable  $X$  with mean  $\mu$ . Then,

$$\Pr\left(\left|\frac{1}{n} \sum_{i=1}^n X_i - \mu\right| \geq t\right) \leq 2 \exp\left(-n \min\left(-\frac{t^2}{v^2}, \frac{t}{2v}\right)\right).$$

**Definition 7.2.7.** A random variable  $X$  is sub-Gaussian with variance  $\sigma^2$  if for all  $t > 0$ , the following holds

$$\Pr(|X - \mathbb{E}X| \geq t) \leq 2 \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

**Definition 7.2.8.** For a sub-Gaussian random variable  $X$ , its sub-Gaussian norm  $\|X\|_{\psi_2}$  is defined as

$$\|X\|_{\psi_2} = \sup_{p \geq 1} p^{-\frac{1}{2}} (\mathbb{E}|X|^p)^{\frac{1}{p}}.$$

**Lemma 7.2.13.** If  $X$  is sub-Gaussian or sub-exponential, then  $\|X - \mathbb{E}X\|_{\psi_2} \leq 2\|X\|_{\psi_2}$  or  $\|X - \mathbb{E}X\|_{\psi_1} \leq 2\|X\|_{\psi_1}$  holds, respectively.

**Lemma 7.2.14.** For two sub-Gaussian random variables  $X_1, X_2$ ,  $X_1 \cdot X_2$  is a sub-exponential random variable with

$$\|X_1 \cdot X_2\|_{\psi_1} \leq C \max\{\|X_1\|_{\psi_2}^2, \|X_2\|_{\psi_2}^2\}.$$

**Lemma 7.2.15.** Let  $X_1, X_2, \dots, X_k$  be  $k$  independent zero-mean sub-Gaussian random variables, and  $X = \sum_{j=1}^k X_j$ . Then,  $X$  is sub-Gaussian with  $\|X\|_{\psi_2}^2 \leq C \sum_{j=1}^k \|X_j\|_{\psi_2}^2$  for some absolute constant  $C > 0$ .

Next, we provide some symmetrization results of random variables, which will be used in our proofs. See [44] for details.

**Lemma 7.2.16.** Let  $y_1, y_2, \dots, y_n$  be the  $n$  independent realizations of the random vector  $Y \in \mathcal{Y}$ , and  $\mathcal{F}$  be a function class defined on  $\mathcal{Y}$ . For any increasing convex function  $\phi(\cdot)$ , the following holds

$$\mathbb{E}\{\phi[\sup_{f \in \mathcal{F}} |\sum_{i=1}^n f(y_i) - \mathbb{E}(f(Y))|]\} \leq \mathbb{E}\{\phi[\sup_{f \in \mathcal{F}} |\sum_{i=1}^n \epsilon_i f(y_i)|]\},$$

where  $\epsilon_1, \dots, \epsilon_n$  are i.i.d Rademacher random variables that are independent of  $y_1, \dots, y_n$ .

**Lemma 7.2.17.** Let  $y_1, \dots, y_n$  be  $n$  independent realization of the random vector  $Z \in \mathcal{Z}$  and  $\mathcal{F}$  be a function class defined on  $\mathcal{Z}$ . If Lipschitz functions  $\{\phi_i(\cdot)\}_{i=1}^n$  satisfy the following for all  $v, v' \in \mathbb{R}$

$$|\phi_i(v) - \phi_i(v')| \leq L|v - v'|$$

and  $\phi_i(0) = 0$ , then for any increasing convex function  $\phi(\cdot)$ , the following holds

$$\mathbb{E}\{\phi\left[\left|\sup_{f \in \mathcal{F}} \sum_{i=1}^n \epsilon_i \phi_i(f(y_i))\right|\right]\} \leq \mathbb{E}\{\phi[2|L \sup_{f \in \mathcal{F}} \sum_{i=1}^n \epsilon_i f(y_i)|]\},$$

where  $\epsilon_1, \dots, \epsilon_n$  are i.i.d Rademacher random variables that are independent of  $y_1, \dots, y_n$ .

### Proof of Theorem 7.2.1

Note that by (7.45), we have

$$\nabla q(\beta; \beta) = \left[ \frac{2}{1 + \exp(-\langle \beta, y \rangle / \sigma^2)} - 1 \right] \cdot y - \beta.$$

W.l.o.g, we assume that  $\beta = (1, 0, \dots, 0)^T$  and  $\sigma = 1$  in the GMM model. Then, we can see that for each constant  $c \geq 0$ , if

$$\left\| \frac{y}{3} \right\|_2 \geq c + \|\beta\|_2$$

$$\langle \beta, y \rangle \geq \ln 2$$

$$y \geq 0$$

and denote the set of  $y$  satisfying the above assumptions as  $\mathcal{S}$ , we have

$$\|\nabla q(\beta; \beta)\|_2 \geq \left\| \frac{y}{3} \right\|_2 - \|\beta\|_2 \geq c.$$

The above assumptions hold if  $y = (\ln 2 + 1, 3s, a_3, a_4, \dots, a_d)$ , where  $s \geq c$  and  $a_3, \dots, a_d \geq 0$ . We can easily see that  $\mathbb{P}[y \in \mathcal{S}] > 0$  since  $y$  follows a mixture of Gaussian distributions.

### Proof of Theorem 7.2.2

We first convert  $(\epsilon, \delta)$ -DP to  $\rho$ -zCDP by using the following lemma

**Lemma 7.2.18** ([52]). Let  $M : \mathcal{X}^n \mapsto \mathcal{Y}$  be a randomized algorithm. If  $M$  is  $\rho$ -zCDP, it is  $(\rho + 2\sqrt{\rho \log \frac{1}{\delta}}, \delta)$ -DP for all  $\delta > 0$ .

Thus, it suffices to show that Algorithm 7.2.56 is  $\tilde{\epsilon}^2 = (\sqrt{\epsilon + \log \frac{1}{\delta}} - \sqrt{\log \frac{1}{\delta}})^2$ -zCDP.

The following lemma shows that adding some Gaussian noise will preserve zCDP.

**Lemma 7.2.19.** Given a function  $q : \mathcal{X}^n \rightarrow \mathbb{R}^p$ , the Gaussian Mechanism is defined as:

$\mathcal{M}_G(D, q, \epsilon) = q(D) + Y$ , where  $Y$  is drawn from a Gaussian Distribution  $\mathcal{N}(0, \sigma^2 I_p)$  is  $\frac{\Delta_2^2(q)}{2\sigma^2}$ -zCDP.  $\Delta_2(q)$  is the  $\ell_2$ -sensitivity of the function  $q$ , i.e.,  $\Delta_2(q) = \sup_{D \sim D'} \|q(D) - q(D')\|_2$ .

By Lemma 7.2.2 we know  $\Delta_2(g_j^{t-1}(\beta^{t-1})) = \frac{4\sqrt{2}}{3} \frac{s}{n}$ . By simple calculation we can show that in each iteration and each coordinate, outputting  $g_j^{t-1}(\beta^{t-1})$  will be  $\frac{\tilde{\epsilon}^2}{dT}$ -zCDP. Thus by the composition property of zCDP, we know that it is  $\tilde{\epsilon}^2$ -zCDP.

### Proof of Theorem 7.2.3

Consider  $t$ -th iteration, under the assumption that  $\beta^{t-1} \in \mathcal{B}$  we have

$$\begin{aligned} \|\beta^t - \beta^*\|_2 &= \|\beta^{t-1} + \eta \tilde{\nabla} Q_n(\beta^{t-1}) - \beta^*\|_2 \\ &\leq \|\beta^{t-1} + \eta \nabla Q(\beta^{t-1}; \beta^{t-1}) - \beta^*\|_2 + \eta \|\tilde{\nabla} Q_n(\beta^{t-1}) - \nabla Q(\beta^{t-1}; \beta^{t-1})\|_2 \end{aligned} \tag{7.71}$$

We first bound the first term of (7.71).

$$\begin{aligned} &\|\beta^{t-1} + \eta \nabla Q(\beta^{t-1}; \beta^{t-1}) - \beta^*\|_2 \\ &\leq \|\beta^{t-1} + \eta \nabla Q(\beta^{t-1}; \beta^*) - \beta^*\|_2 + \eta \|\nabla Q(\beta^{t-1}; \beta^{t-1}) - \nabla Q(\beta^{t-1}; \beta^*)\|_2 \end{aligned} \tag{7.72}$$

We then consider the first term of (7.72). We note that the self-consistent property in Definition 7.2.1 implies that

$$\beta^* = \arg \max_{\beta} Q(\beta; \beta^*), \quad (7.73)$$

which means that  $\beta^*$  is a maximizer of  $Q(\beta; \beta^*)$ . Thus, the proof follows from the convergence rate of the strongly convex and smooth functions  $Q(\beta; \beta^*)$  in [230]. For the step size  $\eta = \frac{2}{\mu+v}$ , we have

$$\|\beta^{t-1} + \eta \nabla Q(\beta^{t-1}; \beta^*) - \beta^*\|_2 \leq (\frac{\mu - v}{\mu + v}) \|\beta^{t-1} - \beta^*\|_2. \quad (7.74)$$

Thus, by the Lipschitz-Gradient-2( $\gamma, \mathcal{B}$ ) condition, we get the following of (7.72)

$$\begin{aligned} & \|\beta^{t-1} + \eta \nabla Q(\beta^{t-1}; \beta^{t-1}) - \beta^*\|_2 \\ & \leq \|\beta^{t-1} + \eta \nabla Q(\beta^{t-1}; \beta^*) - \beta^*\|_2 + \eta \|\nabla Q(\beta^{t-1}; \beta^{t-1}) - \nabla Q(\beta^{t-1}; \beta^*)\|_2 \\ & \leq (\frac{\mu - v}{\mu + v}) \|\beta^{t-1} - \beta^*\|_2 + \eta \gamma \|\beta^{t-1} - \beta^*\|_2 \\ & = (1 - 2 \frac{v - \gamma}{\mu + v}) \|\beta^{t-1} - \beta^*\|_2 \end{aligned} \quad (7.75)$$

where the the last inequality is due to taking  $\eta = \frac{2}{\mu+v}$ .

Next we bound the second term of (7.71). For convenience we denote the first sum of (7.61) (*i.e.*, the robust mean estimator ) as  $\tilde{g}_j^{t-1}(\beta^{t-1})$ . So we have

$$\|\tilde{\nabla} Q_n(\beta^{t-1}) - \nabla Q(\beta^{t-1}; \beta^{t-1})\|_2^2 = \sum_{j=1}^d (g_j^{t-1}(\beta^{t-1}) - \mathbb{E} \nabla_j q(\beta^{t-1}; \beta^{t-1}))^2 \quad (7.76)$$

$$\begin{aligned} & \leq \sum_{j=1}^d (\tilde{g}_j^{t-1}(\beta^{t-1}) - \mathbb{E} \nabla_j q(\beta^{t-1}; \beta^{t-1}))^2 + \sum_{j=1}^d |Z_j^{t-1}|^2 \\ & \end{aligned} \quad (7.77)$$

The first equality is due to Assumption 7.2.1. For the second term of (7.77), by the high

probability concentration bound of Gaussian random variable we have for fixed  $j$  with probability at least  $1 - \frac{\zeta}{d}$ ,  $|Z_j^{t-1}|^2 \leq \frac{8\tau d T \log \frac{d}{\zeta}}{9\beta n \tilde{\epsilon}^2}$ . Thus with probability at least  $1 - \zeta$  we have

$$\sum_{j=1}^d |Z_j^{t-1}|^2 \leq \frac{8\tau d^2 T \log \frac{d}{\zeta}}{9\beta n \tilde{\epsilon}^2}.$$

For the first term of (7.77), by Lemma 7.2.1 and taking  $\zeta = \frac{\zeta}{d}$ , we have for a fixed  $j \in [d]$ ,  $(\tilde{g}_j^{t-1}(\beta^{t-1}) - \mathbb{E}\nabla_j q(\beta^{t-1}; \beta^{t-1}))^2 \leq O\left(\frac{\tau \log \frac{d}{\zeta}}{n}\right)$ . Thus, with probability at least  $1 - \zeta$ , we have

$$\sum_{j=1}^d (\tilde{g}_j^{t-1}(\beta^{t-1}) - \mathbb{E}\nabla_j q(\beta^{t-1}; \beta^{t-1}))^2 \leq O\left(\frac{d\tau \log \frac{d}{\zeta}}{n}\right).$$

Hence, we have, with probability at least  $1 - 2\zeta$ , for some constant  $C_2$

$$\|\tilde{\nabla} Q_n(\beta^{t-1}) - \nabla Q(\beta^{t-1}; \beta^{t-1})\|_2 \leq C_2 \frac{d\sqrt{\tau T \log \frac{d}{\zeta}}}{\sqrt{\beta n \tilde{\epsilon}^2}}. \quad (7.78)$$

Plugging (7.78) and (7.75) into (7.71), we have, with probability  $1 - 2\zeta$  and for some constant  $C_3$ ,

$$\|\beta^t - \beta^*\|_2 \leq (1 - 2\frac{v - \gamma}{\mu + v})\|\beta^{t-1} - \beta^*\|_2 + C_3 \frac{2}{\mu + v} \cdot \frac{d\sqrt{\tau T \log \frac{d}{\zeta}}}{\sqrt{\beta n \tilde{\epsilon}^2}} \quad (7.79)$$

Next, we will show that when  $n$  is large enough, if  $\|\beta^0 - \beta^*\|_2 \leq \frac{R}{2}$  then  $\|\beta^t - \beta^*\|_2 \leq \frac{R}{2}$  holds (and thus  $\beta \in \mathcal{B}$ ) for all  $t \in [T]$  if (7.79) holds for all  $t \in [T]$  (and this hold with probability at least  $1 - 2T\zeta$ ).

We will use induction. When  $t = 1$ , by (7.79) we have

$$\begin{aligned} \|\beta^1 - \beta^*\|_2 &\leq (1 - 2\frac{v - \gamma}{\mu + v})\|\beta^0 - \beta^*\|_2 + C_3 \frac{2}{\mu + v} \cdot \frac{d\sqrt{\tau T \log \frac{d}{\zeta}}}{\sqrt{\beta n \tilde{\epsilon}^2}} \\ &\leq (1 - 2\frac{v - \gamma}{\mu + v})\frac{R}{2} + C_3 \frac{2}{\mu + v} \cdot \frac{d\sqrt{\tau T \log \frac{d}{\zeta}}}{\sqrt{\beta n \tilde{\epsilon}^2}}. \end{aligned}$$

If  $C_3 \frac{2}{\mu+v} \cdot \frac{d\sqrt{\tau T \log \frac{d}{\zeta}}}{\sqrt{\beta n \tilde{\epsilon}^2}} \leq 2 \frac{v-\gamma}{\mu+v} \cdot \frac{R}{2}$ , then we can see that  $\|\beta^1 - \beta^*\|_2 \leq \frac{R}{2}$ . This holds if

$$C_4 \left( \frac{1}{v-\gamma} \right)^2 \frac{d^2 \tau T \log \frac{d}{\zeta}}{R^2 \beta \tilde{\epsilon}^2} \leq n$$

for some constant  $C_4$ .

Next, we will assume that (7.79) holds for all  $t \in [T]$  and  $\beta \in \mathcal{B}$  for all  $t \in [T]$ . For convenience, we denote  $\iota = 1 - 2 \frac{v-\gamma}{\mu+v}$ . By (7.79), we have

$$\begin{aligned} \|\beta^T - \beta^*\|_2 &\leq (1 - 2 \frac{v-\gamma}{\mu+v})^T \|\beta^0 - \beta^*\|_2 + C_3 (1 + \iota + \iota^2 + \dots) \frac{2}{\mu+v} \cdot \frac{d\sqrt{\tau T \log \frac{d}{\zeta}}}{\sqrt{\beta n \tilde{\epsilon}^2}} \\ &\leq (1 - 2 \frac{v-\gamma}{\mu+v})^T \frac{R}{2} + C_3 \frac{1}{1-\iota} \cdot \frac{2}{\mu+v} \cdot \frac{d\sqrt{\tau T \log \frac{d}{\zeta}}}{\sqrt{\beta n \tilde{\epsilon}^2}} \\ &= (1 - 2 \frac{v-\gamma}{\mu+v})^T \frac{R}{2} + O\left(\frac{1}{v-\gamma} \frac{d\sqrt{\tau T \log \frac{d}{\zeta}}}{\sqrt{\beta n \tilde{\epsilon}^2}}\right). \end{aligned}$$

Taking  $T = O(\frac{\mu+v}{v-\gamma} \log \frac{n\tilde{\epsilon}}{d})$ , we have, with probability at least  $1 - 2T\zeta$ ,

$$\|\beta^T - \beta^*\|_2 \leq \tilde{O}(R \sqrt{\frac{\mu+v}{(v-\gamma)^3}} \frac{d\sqrt{\tau \log n \log \frac{d}{\zeta}}}{\sqrt{\beta n \tilde{\epsilon}^2}}).$$

Since  $\tilde{\epsilon} = \sqrt{\log \frac{1}{\delta}} + \epsilon - \sqrt{\log \frac{1}{\delta}}$ , by using the Taylor series of the function  $\sqrt{x+1} - \sqrt{x}$ , we have  $\tilde{\epsilon} = O\left(\frac{\epsilon}{\sqrt{\log \frac{1}{\delta}}}\right)$ . Thus, we have the proof.

### Proof of Lemma 7.2.4

To prove Lemma 7.2.4, we need a stronger lemma.

**Lemma 7.2.20.** The  $j$ -the coordinate of  $\nabla q(\beta; \beta)$  is  $\xi$ -sub-exponential with

$$\xi = C_1 \sqrt{\|\beta^*\|_\infty^2 + \sigma^2}, \quad (7.80)$$

where  $C_1$  is some absolute constant. Also, for fixed  $j \in [d]$ , each  $\nabla_j q_i(\beta; \beta)$ , where  $i \in [n]$ , is independent with others.

If Lemma 7.2.20 holds, then by Lemma 7.2.11 we can get Lemma 7.2.4.

*Proof of Lemma 7.2.20.* From (7.45) it is oblivious that each  $\nabla_j q_i(\beta; \beta)$ , where  $i \in [n], j \in [d]$ , is independent with others. Next, we prove the property of sub-exponential for each coordinate.

Note that

$$\nabla_j q(\beta; \beta) = [2w_\beta(y) - 1]y_j - \beta_j,$$

and

$$\mathbb{E}_Y \nabla_j q(\beta; \beta) = \mathbb{E}_Y (2w_\beta(Y)Y_j - Y_j) - \beta_j.$$

By the symmetrization lemma in Lemma 7.2.16, we have the following for any  $t > 0$

$$\mathbb{E}\{\exp(t|\nabla_j q(\beta; \beta) - \mathbb{E}\nabla_j q(\beta; \beta)|)\} \leq \mathbb{E}\{\exp(t|\epsilon[2w_\beta(y) - 1]y_j|)\}, \quad (7.81)$$

where  $\epsilon$  is a Rademacher random variable.

Next, we use Lemma 7.2.17 with  $f(y_j) = y_j$ ,  $\mathcal{F} = \{f\}$ ,  $\phi(v) = [2w_\beta(y) - 1]v$  and  $\phi(v) = \exp(u \cdot v)$ . It is easy to see that  $\phi$  is 1-Lipschitz. Thus, by Lemma 7.2.17 we have

$$\mathbb{E}\{\exp(t|\epsilon[2w_\beta(y) - 1]y_j|)\} \leq \mathbb{E}\{\exp[2t|\epsilon y_j|]\}. \quad (7.82)$$

By the formulation of the model, we have  $y_j = z\beta_j^* + v_j$ , where  $z$  is a Rademacher random variable and  $v_j \sim \mathcal{N}(0, \sigma^2)$ . It is easy to see that  $y_j$  is sub-Gaussian and

$$\|y_j\|_{\psi_2} = \|z \cdot \beta_j^* + v_j\|_{\psi_2} \leq C \cdot \sqrt{\|z \cdot \beta_j^*\|_{\psi_2}^2 + \|v_j\|_{\psi_2}^2} \leq C' \sqrt{|\beta_j^*|^2 + \sigma^2}, \quad (7.83)$$

for some absolute constants  $C, C'$ , where the last inequality is due to the facts that  $\|z_j \beta_j^*\|_{\psi_2} \leq |\beta_j^*|$  and  $\|v_{i,j}\|_{\psi_2} \leq C'' \sigma^2$  for some  $C'' > 0$ .

Since  $|\epsilon y_j| = |y_j|$ ,  $\|\epsilon y_j\|_{\psi_2} = \|y_j\|_{\psi_2}$  and  $\mathbb{E}(\epsilon y_j) = 0$ , by Lemma 5.5 in [289] we have that for any  $u'$  there exists a constant  $C^{(4)} > 0$  such that

$$\mathbb{E}\{\exp(u' \cdot \epsilon \cdot y_j)\} \leq \exp(u'^2 \cdot C^{(4)} \cdot (|\beta|_j^2 + \sigma^2)). \quad (7.84)$$

Thus, for any  $t > 0$  we get

$$\mathbb{E}\{\exp(2t \cdot |\epsilon \cdot y_j|)\} \leq 2 \exp(t^2 \cdot C^{(5)} \cdot (|\beta|_j^2 + \sigma^2)) \quad (7.85)$$

for some constant  $C^{(5)}$ . Therefore, in total we have the following for some constant  $C^{(6)} > 0$

$$\mathbb{E}\{\exp(t|\nabla_j q(\beta; \beta) - \mathbb{E}\nabla_j q(\beta; \beta)|)\} \leq \exp(t^2 \cdot C^{(6)} \cdot (|\beta|_j^2 + \sigma^2)) \leq \exp(t^2 \cdot C^{(6)} \cdot (\|\beta^*\|_\infty^2 + \sigma^2)). \quad (7.86)$$

Combining this with Lemma 7.2.13 and the definition, we know that  $\nabla_j q(\beta; \beta)$  is  $O(\sqrt{\|\beta^*\|_\infty^2 + \sigma^2})$ -sub-exponential.  $\square$

## Proof of Lemma 7.2.6

Just as in the proof of Lemma 7.2.4, we will show that  $\nabla_j q(\beta; \beta)$  is sub-exponential instead.

**Lemma 7.2.21.** For each  $\beta \in \mathcal{B}$ , the  $j$ -the coordinate of  $\nabla q(\beta; \beta)$  is  $\xi$ -sub-exponential with

$$\xi = C \max\{\|\beta^*\|_2^2 + \sigma^2, 1, \sqrt{d}\|\beta^*\|_2\}, \quad (7.87)$$

where  $C > 0$  is some absolute constant. Also, for fixed  $j \in [d]$ , each  $\nabla_j q_i(\beta; \beta)$ , where  $i \in [n]$ , is independent with others.

*Proof of Lemma 7.2.21.* From (7.47) it is oblivious that for fixed  $j \in [d]$ , each  $\nabla_j q_i(\beta; \beta)$ , where  $i \in [n]$ , is independent with others. Next, we prove the property of sub-exponential.

Note that  $\mathbb{E}\nabla_j q(\beta; \beta) = \mathbb{E}2w_\beta(x, y)y \cdot x_j - \beta_j$ . Thus, we have

$$\nabla_j q(\beta; \beta) - \mathbb{E}\nabla_j q(\beta; \beta) = \underbrace{2w_\beta(x, y)yx_j - \mathbb{E}[2w_\beta(x, y)yx_j]}_A + \underbrace[[xx^T\beta - \beta]_j - yx_j]_B - \underbrace[yx_j]_C. \quad (7.88)$$

For term A and any  $t > 0$ , we have

$$\mathbb{E}\{\exp(t|A|)\} \leq \mathbb{E}\{\exp[t|2w_\beta(x, y)yx_j|]\}. \quad (7.89)$$

Using Lemma 7.2.17 on  $f(yx_j) = yx_j$ ,  $\mathcal{F} = f$ ,  $\phi_i(v) = 2w_\beta(x, y)v$  and  $\phi(v) = \exp(uv)$ , we have

$$\mathbb{E}\{\exp[t|2w_\beta(x, y)yx_j]|\} \leq \mathbb{E}\{\exp[4t|\epsilon yx_j]|\}. \quad (7.90)$$

Note that since  $y = z\langle\beta^*, x\rangle + v$  and  $\|z\langle\beta^*, x\rangle\|_{\psi_2} = \|\langle\beta^*, x\rangle\|_{\psi_2} \leq C\|\beta^*\|_2$  and  $\|v\|_{\psi_2} \leq C'\sigma$  for some constants  $C, C' > 0$ , by Lemma 7.2.15 we know that there exists a constant  $C'' > 0$  such that

$$\|y\|_{\psi_2} \leq C''\sqrt{\|\beta^*\|_2^2 + \sigma^2}. \quad (7.91)$$

Thus, by Lemma 7.2.14 we have

$$\|yx_j\|_{\psi_1} \leq \max\{C''^2(\|\beta^*\|_2^2 + \sigma^2), C'''\} \leq C_4 \max\{\|\beta^*\|_2^2 + \sigma^2, 1\}. \quad (7.92)$$

For term B, we have

$$\mathbb{E}\{\exp[t|B]|\} = \mathbb{E}\{\exp[t|\sum_{k=1}^d x_j x_k \beta_k - \beta_j]|\}, \quad (7.93)$$

where  $x_j, x_k \sim \mathcal{N}(0, 1)$ . Now, by Lemma 7.2.14 we have  $\|x_j x_k \beta_k\|_{\psi_1} \leq |\beta_k|C^{(5)}$  for some constant  $C^{(5)} > 0$ . Thus, we get  $\|\sum_{k=1}^d x_j x_k \beta_k\|_{\psi_1} \leq C^{(5)}\|\beta\|_1$ .

Also, we know that  $\|\beta\|_1 \leq \sqrt{d}\|\beta\|_2$ . Furthermore, we have  $\|\beta\|_2 \leq \|\beta^*\|_2 + \|\beta^* - \beta\|_2 \leq O(\|\beta^*\|_2)$ , since  $\beta \in \mathcal{B}$  (by assumption). From Lemma 7.2.14, we get  $\|B\|_{\psi_1} \leq$

$C^{(6)}\sqrt{d}\|\beta^*\|_2$  with some constant  $C^{(6)} > 0$ .

Thus, we know that there exist some constants  $C^{(7)} > 0$  and  $C^{(8)} > 0$  such that

$$\begin{aligned}\|\nabla_j q(\beta; \beta) - \mathbb{E}\nabla_j q(\beta; \beta)\|_{\psi_1} &\leq C^{(7)} \max\{\|\beta^*\|_2^2 + \sigma^2, 1\} + C^{(8)}\sqrt{d}\|\beta^*\|_2 \\ &\leq C^{(9)} \max\{\|\beta^*\|_2^2 + \sigma^2, 1, \sqrt{d}\|\beta^*\|_2\}.\end{aligned}$$

This means that  $\nabla_j q(\beta; \beta)$  is  $O(\max\{\|\beta^*\|_2^2 + \sigma^2, 1, \sqrt{d}\|\beta^*\|_2\})$ -sub-exponential.

□

### Proof of Lemma 7.2.8

Just as in the proof of Lemma 7.2.4, we will show that  $\nabla_j q(\beta; \beta)$  is sub-exponential instead.

**Lemma 7.2.22.** For each  $\beta \in \mathcal{B}$  and  $j \in [d]$ ,  $\nabla_j q(\beta; \beta)$  is  $\xi$ -sub-exponential with

$$\xi = C[(1+k)(1+kr)^2\sqrt{d}\|\beta^*\|_2 + \max\{(1+kr)^2, \sigma^2 + \|\beta^*\|_2^2\}] = O(\sqrt{d}\|\beta^*\|_2 + \sigma^2 + \|\beta^*\|_2^2) \quad (7.94)$$

for some constant  $C > 0$ . Also, for fixed  $j \in [d]$ , each  $\nabla_j q_i(\beta; \beta)$ , where  $i \in [n]$ , is independent with others.

*Proof of Lemma 7.2.22.* From (7.49) it is oblivious that for fixed  $j \in [d]$ , each  $\nabla_j q_i(\beta; \beta)$ , where  $i \in [n]$ , is independent with others. Next, we prove the property of sub-exponential.

For simplicity, we use notations  $\bar{m} = m_\beta(x^{\text{obs}}, y)$ ,  $\bar{m} = \beta(x^{\text{obs}}, y)$ ,  $\bar{K} = K_\beta(x^{\text{obs}}, y)$ , and  $\bar{K} = K_\beta(x^{\text{obs}}, y)$ . Then, we have

$$\nabla q(\beta; \beta) - \mathbb{E}\nabla q(\beta; \beta) = \underbrace{m_\beta(x^{\text{obs}}, y)y - \mathbb{E}[m_\beta(x^{\text{obs}}, y)y]}_A + \overbrace{(\bar{K}_\beta(x^{\text{obs}}, y) - \mathbb{E}\bar{K}_\beta(x^{\text{obs}}, y))\beta}^B. \quad (7.95)$$

For the  $j$ -th coordinate of  $A$ , we have

$$A_j = \bar{m}_j y - \mathbb{E}[\bar{m}_j y]. \quad (7.96)$$

We note that  $\bar{m}_j$  is a zero-mean sub-Gaussian random variable with  $\|\bar{m}_j\|_{\psi_2} \leq C(1 + kr)$  (see Lemma B.3 in [339])

**Lemma 7.2.23.** Under the assumption of Lemma 6, for each  $j \in [d]$ ,  $\bar{m}_j$  is sub-Gaussian with mean zero and  $\|\bar{m}_j\|_{\psi_2} \leq C(1 + kr)$ .

Thus, by Lemma 7.2.14 we have

$$\|\bar{m}_j y\|_{\psi_1} \leq C \max\{\|\bar{m}_j\|_{\psi_2}^2, \|y\|_{\psi_2}^2\} \leq C' \max\{(1 + kr)^2, \sigma^2 + \|\beta^*\|_2^2\}, \quad (7.97)$$

where the last inequality is due to the fact that  $y = \langle \beta^*, x \rangle + v$ . Thus,  $\|y\|_{\psi_2}^2 \leq C_3(\|\langle \beta^*, x \rangle\|_{\psi_2}^2 + \|v\|_{\psi_2}^2)$  for some  $C_3$ .

For term B, we have

$$\bar{K}_j = \underbrace{(1 - z_j)\beta_j}_C + \underbrace{\sum_{k=1}^d \bar{m}_j \bar{m}_k \beta_k}_D - \underbrace{\sum_{k=1}^d [(1 - z_j)\bar{m}_j][(1 - z_k)\bar{m}_k] \beta_k}_E. \quad (7.98)$$

For term C, we have the following (by Example 5.8 in [289])

$$\|(1 - z_j)\beta_j\|_{\psi_2} \leq |\beta_j| \leq \|\beta\|_\infty \leq (1 + k)\sqrt{s}\|\beta^*\|_2. \quad (7.99)$$

For term D, by Lemma 7.2.23 and 7.2.14 we have

$$\left\| \sum_{k=1}^d \bar{m}_j \bar{m}_k \beta_k \right\|_{\psi_1} \leq \sum_{k=1}^d |\beta_k| \|\bar{m}_j \bar{m}_k\|_{\psi_1} \leq \sum_{k=1}^d |\beta_k| C^2 (1 + kr)^2 \leq C_4 (1 + kr)^2 \|\beta\|_1. \quad (7.100)$$

Since  $\beta \in \mathcal{B}$ , we get  $\|\beta\|_1 \leq \sqrt{d}\|\beta\|_2 \leq (1+k)\sqrt{d}\|\beta^*\|_2$ . Thus, we have

$$\left\| \sum_{k=1}^d \bar{m}_j \bar{m}_k \beta_k \right\|_{\psi_1} \leq C_4 \sqrt{s} (1+kr)^2 \|\beta^*\|_2. \quad (7.101)$$

For term E, since  $1-z \in [0, 1]$ , we have  $\|(1-z_j)\bar{m}_j\|_{\psi_2} \leq \|\bar{m}_j\|_{\psi_2} \leq C(1+kr)$ . Hence, by Lemma 7.2.14 we get

$$\begin{aligned} \left\| \sum_{k=1}^d [(1-z_j)\bar{m}_j][(1-z_k)\bar{m}_k] \beta_k \right\|_{\psi_1} &\leq \sum_{k=1}^d |\beta_k| \left\| [(1-z_j)\bar{m}_j][(1-z_k)\bar{m}_k] \right\|_{\psi_1} \\ &\leq \sum_{k=1}^d |\beta_k| C(1+kr)^2 \leq C_6 (1+kr)^2 \sqrt{s} \|\beta^*\|_2. \end{aligned} \quad (7.102)$$

This gives us

$$\|\bar{K}_j\|_{\psi_1} \leq C_7 \sqrt{s} (1+k)(1+kr)^2 \|\beta^*\|_2. \quad (7.103)$$

By Lemma 7.2.13, we get

$$\begin{aligned} &\|\nabla_j q(\beta; \beta) - \mathbb{E} \nabla_j q(\beta; \beta)\|_{\psi_1} \\ &\leq 2 \|\nabla_j q(\beta; \beta)\|_{\psi_1} \leq C_8 [(1+k)(1+kr)^2 \sqrt{s} \|\beta^*\|_2 + \max\{(1+kr)^2, \sigma^2 + \|\beta^*\|_2^2\}]. \end{aligned} \quad (7.104)$$

□

### Proof of Theorem 7.2.8

For each iteration we denote  $M(\beta^{t-1}) = \arg \max Q(\beta; \beta^{t-1})$ , by the strongly concavity of  $Q(\beta; \beta^*)$  we have

$$\langle \nabla Q(M(\beta^{t-1}); \beta^*) - \nabla Q(\beta^*; \beta^*), M(\beta^{t-1}) - \beta^* \rangle \geq v \|M(\beta^{t-1}) - \beta^*\|_2^2.$$

On the other hand, by the Lipschitz-Gradient condition and the assumption of  $M(\beta^{t-1})$ ,  $\beta^{t-1} \in \mathcal{B}$ , we have

$$\langle \nabla Q(M(\beta^{t-1}); \beta^*) - \nabla Q(M(\beta^{t-1}); \beta^{t-1}), \beta^* - M(\beta^{t-1}) \rangle \leq \gamma \|\beta^{t-1} - \beta^*\|_2 \|\beta^* - M(\beta^{t-1})\|_2.$$

Also by the optimality of  $M(\beta^{t-1})$  we have

$$\begin{aligned} \langle \nabla Q(M(\beta^{t-1}); \beta^*) - \nabla Q(\beta^*; \beta^*), M(\beta^{t-1}) - \beta^* \rangle &\leq \\ \langle \nabla Q(M(\beta^{t-1}); \beta^*) - \nabla Q(M(\beta^{t-1}); \beta^{t-1}), \beta^* - M(\beta^{t-1}) \rangle. \end{aligned}$$

Thus, we have

$$v \|M(\beta^{t-1}) - \beta^*\|_2^2 \leq \gamma \|\beta^{t-1} - \beta^*\|_2 \|\beta^* - M(\beta^{t-1})\|_2.$$

That is,  $\|M(\beta^{t-1}) - \beta^*\|_2 \leq \frac{\gamma}{v} \|\beta^{t-1} - \beta^*\|_2$ . Next, we will bound the term of  $\|\beta^t - M(\beta^{t-1})\|_2$ .

Under the assumption that  $f_i$  is independent with others, just as almost the same as in (7.76)-(7.78) via Lemma 7.2.2, we have that with probability at least  $1 - \zeta$ ,

$$\|\beta^t - M(\beta^{t-1})\|_2 = \|\tilde{f}(\beta^{t-1}) - M(\beta^{t-1})\|_2 \leq O\left(\frac{d\sqrt{\tau T \log \frac{d}{\zeta}}}{\sqrt{\beta n \tilde{\epsilon}^2}}\right).$$

Thus, we have with probability at least  $1 - \zeta$

$$\|\beta^t - \beta^*\|_2 \leq \frac{\gamma}{v} \|\beta^{t-1} - \beta^*\|_2 + O\left(\frac{d\sqrt{\tau T \log \frac{d}{\zeta}}}{\sqrt{\beta n \tilde{\epsilon}^2}}\right).$$

Since we need to make  $\beta^t \in \mathcal{B}$ , this will be true under the assumption that

$$O\left(\frac{d\sqrt{\tau T \log \frac{d}{\zeta}}}{\sqrt{\beta n \tilde{\epsilon}^2}}\right) \leq \frac{v - \gamma}{v} R.$$

If this holds, then we have with probability at least  $1 - T\zeta$

$$\|\beta^t - \beta^*\|_2 \leq \left(\frac{\gamma}{v}\right)^T R + O\left(\frac{v}{v-\gamma} \frac{d\sqrt{\tau T \log \frac{d}{\zeta}}}{\sqrt{\beta n \tilde{\epsilon}^2}}\right),$$

Taking  $T = O\left(\frac{v}{v-\gamma} \log n\right)$  and  $\tilde{\epsilon} = O\left(\frac{\epsilon}{\sqrt{\log \frac{1}{\delta}}}\right)$ , we have the result.

### Proof of Lemma 7.2.9

To prove Lemma 7.2.9, we need a stronger lemma.

**Lemma 7.2.24.** The  $j$ -the coordinate of  $f(\beta)$  is  $\xi$ -sub-exponential with

$$\xi = C_1 \sqrt{\|\beta^*\|_\infty^2 + \sigma^2}, \quad (7.105)$$

where  $C_1$  is some absolute constant. Also, for fixed  $j \in [d]$ , each  $f_{i,j}(\beta)$ , where  $i \in [n]$ , is independent with others.

If Lemma 7.2.24 holds, then by Lemma 7.2.11 we can get Lemma 7.2.9.

*Proof of Lemma 7.2.24.* The proof is almost the same as that of Lemma 7.2.4.

It is oblivious that each  $f_i(\beta)$ , where  $i \in [n]$ , is independent with others. Next, we prove the property of sub-exponential for each coordinate.

Note that

$$f_j(\beta) = [2w_\beta(y) - 1]y_j,$$

and

$$\mathbb{E}_Y f_i(\beta) = \mathbb{E}_Y (2w_\beta(Y)Y_j - Y_j).$$

By the symmetrization lemma in Lemma 7.2.16, we have the following for any  $t > 0$

$$\mathbb{E}\{\exp(t|[f_j(\beta) - \mathbb{E}f_j(\beta)]|)\} \leq \mathbb{E}\{\exp(t|\epsilon[2w_\beta(y) - 1]y_j|)\}, \quad (7.106)$$

where  $\epsilon$  is a Rademacher random variable.

Next, we use Lemma 7.2.17 with  $f(y_j) = y_j$ ,  $\mathcal{F} = \{f\}$ ,  $\phi(v) = [2w_\beta(y) - 1]v$  and  $\phi(v) = \exp(u \cdot v)$ . It is easy to see that  $\phi$  is 1-Lipschitz. Thus, by Lemma 7.2.17 we have

$$\mathbb{E}\{\exp(t|\epsilon[2w_\beta(y) - 1]y_j|)\} \leq \mathbb{E}\{\exp[2t|\epsilon y_j|]\}. \quad (7.107)$$

By the formulation of the model, we have  $y_j = z\beta_j^* + v_j$ , where  $z$  is a Rademacher random variable and  $v_j \sim \mathcal{N}(0, \sigma^2)$ . It is easy to see that  $y_j$  is sub-Gaussian and

$$\|y_j\|_{\psi_2} = \|z \cdot \beta_j^* + v_j\|_{\psi_2} \leq C \cdot \sqrt{\|z \cdot \beta_j^*\|_{\psi_2}^2 + \|v_j\|_{\psi_2}^2} \leq C' \sqrt{|\beta_j^*|^2 + \sigma^2} \quad (7.108)$$

for some absolute constants  $C, C'$ , where the last inequality is due to the facts that  $\|z_j\beta_j^*\|_{\psi_2} \leq |\beta_j^*|$  and  $\|v_i\|_{\psi_2} \leq C''\sigma^2$  for some  $C'' > 0$ .

Since  $|\epsilon y_j| = |y_j|$ ,  $\|\epsilon y_j\|_{\psi_2} = \|y_j\|_{\psi_2}$  and  $\mathbb{E}(\epsilon y_j) = 0$ , by Lemma 5.5 in [289] we have that for any  $u'$  there exists a constant  $C^{(4)} > 0$  such that

$$\mathbb{E}\{\exp(u' \cdot \epsilon \cdot y_j)\} \leq \exp(u'^2 \cdot C^{(4)} \cdot (|\beta_j^*|^2 + \sigma^2)). \quad (7.109)$$

Thus, for any  $t > 0$  we get

$$\mathbb{E}\{\exp(2t \cdot |\epsilon \cdot y_j|)\} \leq 2 \exp(t^2 \cdot C^{(5)} \cdot (|\beta_j^*|^2 + \sigma^2)) \quad (7.110)$$

for some constant  $C^{(5)}$ . Therefore, in total we have the following for some constant  $C^{(6)} > 0$

$$\mathbb{E}\{\exp(t|[f_j(\beta) - \mathbb{E}f_j(\beta)]|)\} \leq \exp(t^2 \cdot C^{(6)} \cdot (|\beta_j^*|^2 + \sigma^2)) \leq \exp(t^2 \cdot C^{(6)} \cdot (\|\beta^*\|_\infty^2 + \sigma^2)). \quad (7.111)$$

Combining this with Lemma 7.2.13 and the definition, we know that  $f_j(\beta)$  is  $O(\sqrt{\|\beta^*\|_\infty^2 + \sigma^2})$ -sub-exponential.  $\square$

# **Chapter 8**

## **Conclusion and Future Research**

### **8.1 Conclusion**

Big data has become a key resource for discovery in recent years. With the technical advancements of data acquisition in many fields, we are now generating exponentially more data in a multitude of formats. This flood of complex data poses significant opportunities to discover and understand the critical interplay among different domains. However, due to the existence of sensitive data, we are not yet able to utilize them to their full potential. This is mainly due to the fact that most of the learning models or classifiers are vulnerable to various attack techniques (e.g., model inversion attack [116] and membership attack [256]), and thus cannot protect private information satisfactorily.

An effective way to resolve this issue is to design differentially private machine learning algorithms. Differential Privacy (DP) [107], with roots in cryptography, is a strong mathematical scheme for privacy preserving. It allows for rich statistical and machine learning analysis, and is now becoming a standard for private data analysis. Despite the rapid development of DP in theory, its adoption to machine learning community remains slow. One of my research goals during my Ph.D study is to speed up this process. For this purpose, in this dissertation I have studied a number of fundamental machine learning

problems in the differential privacy model. For this purpose, I have studied a number of fundamental machine learning problems in the differential privacy model, which can be divided into three categories.

**Differentially Private Empirical Risk Minimization:** Empirical Risk Minimization (ERM) is one of the most fundamental problem in supervised learning which encompasses a large family of classical models such as linear regression, LASSO, ridge regression, SVM, logistic regression, sigmoid regression, and neural networks. Due to its importance, its differentially private version ( called DP-ERM) has become one of the core problems in both machine learning and differential privacy communities [66]. In the past few years, I have made both extensive and in-depth studies on DP-ERM. Particularly, I have explored quite a few new directions for the problem, and obtained a number of new and more practical algorithms with theoretical guarantees on the utility, which can be categorized into two classes based on their settings.

- **Central Model:** In the Central Model of Differential Privacy, there is a trusted curator that can store and compute on the entire sensitive data to produce a statistical release or synthetic data. This model has already been used in Uber and will be adopted by the United States Census Bureau for the 2020 census. For this problem, I have first studied DP-ERM with convex loss functions . Based on different assumptions, I have designed the state-of-the-art algorithms that achieve (near) optimal utility bounds. Then, I extended those techniques to the case of non-convex loss functions and adopted two ways to measure the error. I have used gradient norm to measure the error and obtained the first algorithm for population risk and the first result in high dimensions. Subsequently, I have also used the Expected Excess Empirical Risk (EEER) to measure the error, which allows us to obtain quality guaranteed solutions in a way similar to convex loss functions. I provided the first results on the bounds of excess empirical and population risks, as well as finer bounds for some special models such as robust regression and sigmoid regression. I am also the first to show that it

is possible to escape saddle points privately, which is quite useful in deep learning. Moreover, I also obtained more practical algorithms with the improved bound and the ability of escaping saddle points privately. Finally, I have extended DP-ERM with point-wise loss to pairwise loss functions and provided the first result on this topic. Specifically, I showed some theoretical results for utilities on both off-line and online settings, as well as their corresponding algorithms. Moreover, I have initiated the study of DP-ERM with heavy-tailed datasets [331] and gave the first algorithm with theoretical guarantees on the quality.

- **Local Model:** Instead of using a trusted curator, in the local differential privacy model (LDP), the curator is untrusted and each individual manages his/her own data and discloses them to a server through some differentially private mechanisms. The server collects the private data of each individual and combines them into a resulting data analysis. Based on the type of interactions between the server and each individual, there are two types of protocols: non-interactive LDP and interactive LDP. Compared with interactive LDP, non-interactive LDP (NLDP) is more repelling to the real-world applications due to its easy implementation and less influence by the network latency issue, and has been used in industries such as Apple [273], Google [109] and Microsoft [92]. DP-ERM in the NLDP model has not been well studied. Previous paper [257] gives a negative result showing that the sample complexity of the problem needs to be at least exponential in the dimensionality for general convex loss functions, which makes DP-ERM non-applicable in high dimensions.

To resolve this issue, I have conducted a series of research on this topic. I first demonstrated that the sample complexity can actually be reduced if the loss function is smooth enough. Later, I showed that the sample complexity for Lipschitz Generalized Linear loss functions can be quasi-polynomial and linear in the dimensionality. Next I considered a relaxed model of NLDP where some additional public unlabeled data are available to the coordinator. For this model, I am able to show that under

some reasonable assumptions, the sample complexity can be fully polynomial for smooth Generalized Linear Models. This finally makes DP-ERM applicable to high dimensional datasets, which is confirmed by experiments. Later, the problem I have studied is the sparse linear regression problem. I showed that it is not possible to achieve any algorithm with non-trivial bound of utility in both non-interactive and interactive LDP models if the dimensionality is high, but near optimal or even optimal solution with non-trivial utility bound is achievable if the dimensionality is low or only the labels/responses need to be private.

**Matrix Estimation Problems in Differential Privacy Model:** My work on this topic mainly focuses on the behavior of high dimensional statistical matrix estimation. I have pioneered the studies of a number of problems in the NLDP model, with most of them being their respectively first study. Specifically, I have considered the problem of PCA in the NLDP model , and provided lower bounds for both the low dimensional and the high dimensional sparse cases, along with their corresponding near optimal algorithms. I have also studied the high dimensional sparse covariance matrix estimation problem of sub-Gaussian distributions, and presented an efficient algorithm. As a by-product, I also gave a general framework for proving such type of lower bounds.

**Other Problems:** Besides the aforementioned problems, I have also studied several machine learning related problems. I have investigated the problem of crowdsourcing estimation and proposed a new method called private Dawid-Skene estimator to achieve the first theoretically guaranteed solution on the utility of the problem.

In another work, I focused on designing Differentially Private variant of Expectation Maximization algorithm with statistical guarantees. Specifically, I provided a general framework and proofed the first theoretical guarantees of Mixture Linear Regression model in DP model.

## 8.2 Future Research

Compared with the classical topics in machine learning, differentially private machine learning is far less understood, in the this Chapter I will mention some new directions and future work.

### Private Learning for Irregular Data

Compared with the datasets used in the studies of private learning, most of the sensitive data in real-world applications are quite irregular. For example, datasets in medicine and finance are often heavy-tailed, non i.i.d, follow some heterogeneous distributions and may even contain outliers. Such irregularities violate the assumptions made by most of the existing private learning algorithms, and thus can make them no longer differentially private. Recent study [259] also shows that the presence of adversarial examples or outliers in the data can cause the learning models significantly more vulnerable to privacy attacks. Thus, it is urgently needed to design private and robust algorithms for these irregular datasets.

### Making Trustworthy Algorithms Private

Most of the existing trustworthy algorithms focus only on aspects related to trust, and often do not consider the privacy issue, which could cause privacy breach. Similarly, most of the private learning algorithms are untrustworthy, and may cause other ethical issues. For example, [16] shows that existing private algorithms may cause fairness issue. Thus, one of the future directions is understanding the trade-off between these several trustworthy terminologies, such as security, privacy and accuracy, also fairness, privacy and accuracy.

### Differentially Private Deep Learning

Although there are many papers study deep learning in the differential privacy model, all of them consider the practical behaviors. Thus, it is still unclear about the theoretical behaviors

of Differentially Private (Deep) Neural Network. To solve this issue, one possible way is to start with the classical neural network (such as one-hidden layer neural network). Compared with DP-ERM with non-convex loss functions, there are many challenges on Differentially Private Deep Learning, such as in Deep Neural Network it is always over-parameterized where the number of nodes is far greater than the size of dataset. Also, some activation functions such as ReLU, is not differentiable.

### **Machine Learning Problems in Variant Privacy Models**

In this dissertation, I mainly focused on the central DP and local DP model. However, there are still other intermediate privacy, such as the Hybrid Differential Privacy Model, Central/Local DP with public but unlabeled data, Multi-party setting, Federated Learning setting and Shuffled DP model. Thus, a future direction will be designing DP algorithms for machine learning problems in these different DP models, and also understanding the gaps between these privacy models.

### **Combining with Other Privacy Enhancing Techniques**

Theoretically, in this paper we have showed some limitations of DP and LDP models, which may prevent using these two models for some machine learning problems. Moreover, recent some papers also showed that the practical behaviors of DP/LDP is bad for some problems, due to the large amount of noise these algorithms added. Thus, one direction is how to improve the practical performance of DP algorithms. One possible way is combining with other privacy-preserving techniques, such as Multiparty Secure Computation, Homomorphic Encryption, Zero-Knowledge Proof and Blockchain methods, with differential privacy to enhance the privacy-preserving ability and also learning ability of the current DP machine learning algorithms.

# Reference

- [1] Martin Abadi et al. “Deep learning with differential privacy”. In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM. 2016, pp. 308–318.
- [2] Jayadev Acharya, Clément L Canonne, and Himanshu Tyagi. “Inference under information constraints I: Lower bounds from chi-square contraction”. In: *arXiv preprint arXiv:1812.11476* (2018).
- [3] Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. “Differentially Private Assouad, Fano, and Le Cam”. In: *arXiv preprint arXiv:2004.06830* (2020).
- [4] Alekh Agarwal and Leon Bottou. “A lower bound for the optimization of finite sums”. In: *arXiv preprint arXiv:1410.0723* (2014).
- [5] Naman Agarwal et al. “Finding approximate local minima faster than gradient descent”. In: *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. ACM. 2017, pp. 1195–1199.
- [6] Yacine Ait-Sahalia and Dacheng Xiu. “Using principal component analysis to estimate a high dimensional factor model with high-frequency data”. In: *Journal of Econometrics* 201.2 (2017), pp. 384–399.
- [7] Francesco Aldà and Benjamin IP Rubinstein. “The Bernstein Mechanism: Function Release under Differential Privacy.” In: *AAAI*. 2017, pp. 1705–1711.
- [8] Zeyuan Allen-Zhu. “Katyusha: The first direct acceleration of stochastic gradient methods”. In: *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. ACM. 2017, pp. 1200–1205.
- [9] Zeyuan Allen-Zhu and Lorenzo Orecchia. “Linear Coupling: An Ultimate Unification of Gradient and Mirror Descent”. In: *Proceedings of the 8th Innovations in Theoretical Computer Science*. ITCS ’17. 2017.

- [10] Zeyuan Allen-Zhu and Yang Yuan. “Improved SVRG for Non-Strongly-Convex or Sum-of-Non-Convex Objectives”. In: *Proceedings of the 33rd International Conference on Machine Learning*. ICML ’16. 2016.
- [11] Uri Alon et al. “Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays”. In: *Proceedings of the National Academy of Sciences* 96.12 (1999), pp. 6745–6750.
- [12] Kareem Amin et al. “Private Covariance Estimation via Iterative Eigenvector Sampling”. In: *2018 NIPS workshop in Privacy-Preserving Machine Learning* (2018).
- [13] Ali Anaissi et al. “Ensemble feature learning of genomic data using support vector machine”. In: *PloS one* 11.6 (2016).
- [14] Animashree Anandkumar and Rong Ge. “Efficient approaches for escaping higher order saddle points in non-convex optimization”. In: *Conference on Learning Theory*. 2016, pp. 81–102.
- [15] Lars Backstrom, Cynthia Dwork, and Jon Kleinberg. “Wherfore art thou R3579X? Anonymized social networks, hidden patterns, and structural steganography”. In: *Proceedings of the 16th international conference on World Wide Web*. 2007, pp. 181–190.
- [16] Eugene Bagdasaryan and Vitaly Shmatikov. “Differential Privacy Has Disparate Impact on Model Accuracy”. In: *arXiv preprint arXiv:1905.12101* (2019).
- [17] Sohail Bahmani, Bhiksha Raj, and Petros T Boufounos. “Greedy sparsity-constrained optimization”. In: *Journal of Machine Learning Research* 14.Mar (2013), pp. 807–841.
- [18] Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*. Vol. 348. Springer Science & Business Media, 2013.
- [19] Sivaraman Balakrishnan, Martin J Wainwright, Bin Yu, et al. “Statistical guarantees for the EM algorithm: From population to sample-based analysis”. In: *The Annals of Statistics* 45.1 (2017), pp. 77–120.
- [20] Sivaraman Balakrishnan et al. “Computationally efficient robust sparse estimation in high dimensions”. In: *Conference on Learning Theory*. 2017, pp. 169–212.
- [21] Maria-Florina Balcan et al. “An improved gap-dependency analysis of the noisy power method”. In: *Conference on Learning Theory*. 2016, pp. 284–309.

- [22] Pierre Baldi, Peter Sadowski, and Daniel Whiteson. “Searching for exotic particles in high-energy physics with deep learning”. In: *Nature communications* 5 (2014), p. 4308.
- [23] DC Barber, PJ Howlett, and RC Smart. “Principal component analysis in medical research”. In: *Journal of Applied Statistics* 2.1 (1975), pp. 39–43.
- [24] Andrés F Barrientos et al. “Differentially private significance tests for regression coefficients”. In: *Journal of Computational and Graphical Statistics* (2019), pp. 1–24.
- [25] Raef Bassily. “Linear queries estimation with local differential privacy”. In: *The 22nd International Conference on Artificial Intelligence and Statistics*. 2019, pp. 721–729.
- [26] Raef Bassily and Anupama Nandi. “Privately Answering Classification Queries in the Agnostic PAC Model”. In: *arXiv preprint arXiv:1907.13553* (2019).
- [27] Raef Bassily and Adam Smith. “Local, private, efficient protocols for succinct histograms”. In: *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*. ACM. 2015, pp. 127–135.
- [28] Raef Bassily, Adam Smith, and Abhradeep Thakurta. “Differentially private empirical risk minimization: Efficient algorithms and tight error bounds”. In: *arXiv preprint arXiv:1405.7085* (2014).
- [29] Raef Bassily, Adam Smith, and Abhradeep Thakurta. “Private empirical risk minimization: Efficient algorithms and tight error bounds”. In: *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*. IEEE. 2014, pp. 464–473.
- [30] Raef Bassily, Abhradeep Guha Thakurta, and Om Dipakbhai Thakkar. “Model-Agnostic Private Learning”. In: *Advances in Neural Information Processing Systems*. 2018, pp. 7102–7112.
- [31] Raef Bassily et al. “Private Stochastic Convex Optimization with Optimal Rates”. In: *NeurIPS*. 2019.
- [32] Amos Beimel, Kobbi Nissim, and Eran Omri. “Distributed Private Data Analysis: Simultaneously Solving How and What.” In: *CRYPTO*. Vol. 5157. Springer. 2008, pp. 451–468.
- [33] Amos Beimel, Kobbi Nissim, and Uri Stemmer. “Private learning and sanitization: Pure vs. approximate differential privacy”. In: *APPROX*. Springer, 2013, pp. 363–378.

- [34] Garrett Bernstein and Daniel R Sheldon. “Differentially Private Bayesian Linear Regression”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 523–533.
- [35] Aditya Bhaskara et al. “Unconditional differentially private mechanisms for linear queries”. In: *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*. 2012, pp. 1269–1284.
- [36] Rajendra Bhatia. *Matrix analysis*. Vol. 169. Springer Science & Business Media, 2013.
- [37] Abhishek Bhowmick et al. “Protection Against Reconstruction and Its Applications in Private Federated Learning”. In: *arXiv preprint arXiv:1812.00984* (2018).
- [38] Peter J Bickel, Elizaveta Levina, et al. “Covariance regularization by thresholding”. In: *The Annals of Statistics* 36.6 (2008), pp. 2577–2604.
- [39] Atanu Biswas et al. *Statistical advances in the biomedical science*. Wiley Online Library, 2007.
- [40] Jaroslaw Blasiok et al. “Towards instance-optimal private query release”. In: *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics. 2019, pp. 2480–2497.
- [41] Avrim Blum, Katrina Ligett, and Aaron Roth. “A learning theory approach to noninteractive database privacy”. In: *Journal of the ACM (JACM)* 60.2 (2013), p. 12.
- [42] Avrim Blum et al. “Practical privacy: the SuLQ framework”. In: *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM. 2005, pp. 128–138.
- [43] Thomas Blumensath and Mike E Davies. “Iterative hard thresholding for compressed sensing”. In: *Applied and computational harmonic analysis* 27.3 (2009), pp. 265–274.
- [44] Stephane Boucheron, Gabor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.
- [45] David R Brillinger. “A generalized linear model with “Gaussian” regressor variables”. In: *Selected Works of David Brillinger*. Springer, 2012, pp. 589–606.
- [46] Christian Brownlees, Emilien Joly, Gábor Lugosi, et al. “Empirical risk minimization for heavy-tailed losses”. In: *The Annals of Statistics* 43.6 (2015), pp. 2507–2536.

- [47] Sébastien Bubeck et al. “Convex optimization: Algorithms and complexity”. In: *Foundations and Trends® in Machine Learning* 8.3-4 (2015), pp. 231–357.
- [48] Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.
- [49] Mark Bun, Jelani Nelson, and Uri Stemmer. “Heavy hitters and the structure of local privacy”. In: *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*. ACM. 2018, pp. 435–447.
- [50] Mark Bun, Jelani Nelson, and Uri Stemmer. “Heavy hitters and the structure of local privacy”. In: *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*. ACM. 2018, pp. 435–447.
- [51] Mark Bun and Thomas Steinke. “Average-Case Averages: Private Algorithms for Smooth Sensitivity and Mean Estimation”. In: *arXiv preprint arXiv:1906.02830* (2019).
- [52] Mark Bun and Thomas Steinke. “Concentrated differential privacy: Simplifications, extensions, and lower bounds”. In: *TCC*. Springer. 2016, pp. 635–658.
- [53] Mark Bun, Jonathan Ullman, and Salil Vadhan. “Fingerprinting codes and the price of approximate differential privacy”. In: *SIAM Journal on Computing* 47.5 (2018), pp. 1888–1938.
- [54] Mark Bun, Jonathan Ullman, and Salil Vadhan. “Fingerprinting codes and the price of approximate differential privacy”. In: *SIAM Journal on Computing* 47.5 (2018), pp. 1888–1938.
- [55] Petra Břízková. “Linear regression in genetic association studies”. In: *PLoS One* 8.2 (2013), e56976.
- [56] T Tony Cai, Weidong Liu, and Harrison H Zhou. “Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation”. In: *arXiv preprint arXiv:1212.2882* (2012).
- [57] T Tony Cai, Weidong Liu, Harrison H Zhou, et al. “Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation”. In: *The Annals of Statistics* 44.2 (2016), pp. 455–488.
- [58] T Tony Cai, Zongming Ma, Yihong Wu, et al. “Sparse PCA: Optimal rates and adaptive estimation”. In: *The Annals of Statistics* 41.6 (2013), pp. 3074–3110.

- [59] T Tony Cai, Yichen Wang, and Linjun Zhang. “The Cost of Privacy: Optimal Rates of Convergence for Parameter Estimation with Differential Privacy”. In: *arXiv preprint arXiv:1902.04495* (2019).
- [60] T Tony Cai, Harrison H Zhou, et al. “Optimal rates of convergence for sparse covariance matrix estimation”. In: *The Annals of Statistics* 40.5 (2012), pp. 2389–2420.
- [61] Qiong Cao, Zheng-Chu Guo, and Yiming Ying. “Generalization bounds for metric and similarity learning”. In: *Machine Learning* (2016).
- [62] Olivier Catoni and Ilaria Giulini. “Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression”. In: *arXiv preprint arXiv:1712.02747* (2017).
- [63] T-H Hubert Chan, Elaine Shi, and Dawn Song. “Private and continual release of statistics”. In: *ACM Transactions on Information and System Security* (2011).
- [64] Chih-Chung Chang and Chih-Jen Lin. “LIBSVM: A library for support vector machines”. In: *ACM Transactions on Intelligent Systems and Technology* 2 (3 2011). Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>, 27:1–27:27.
- [65] Kamalika Chaudhuri and Daniel Hsu. “Sample complexity bounds for differentially private learning”. In: *Proceedings of the 24th Annual Conference on Learning Theory*. 2011, pp. 155–186.
- [66] Kamalika Chaudhuri and Claire Monteleoni. “Privacy-preserving logistic regression”. In: *Advances in Neural Information Processing Systems*. 2009.
- [67] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. “Differentially private empirical risk minimization”. In: *Journal of Machine Learning Research* 12.Mar (2011), pp. 1069–1109.
- [68] Kamalika Chaudhuri, Anand Sarwate, and Kaushik Sinha. “Near-optimal differentially private principal components”. In: *Advances in Neural Information Processing Systems*. 2012, pp. 989–997.
- [69] Kamalika Chaudhuri, Anand D Sarwate, and Kaushik Sinha. “A near-optimal algorithm for differentially-private principal components”. In: *The Journal of Machine Learning Research* 14.1 (2013), pp. 2905–2943.
- [70] Changyou Chen, Nan Ding, and Lawrence Carin. “On the convergence of stochastic gradient MCMC algorithms with high-order integrators”. In: *Advances in Neural Information Processing Systems*. 2015.

- [71] Chunhui Chen and Olvi L Mangasarian. “A class of smoothing functions for non-linear and mixed complementarity problems”. In: *Computational Optimization and Applications* 5.2 (1996), pp. 97–138.
- [72] Xi Chen, Qihang Lin, and Dengyong Zhou. “Optimistic knowledge gradient policy for optimal budget allocation in crowdsourcing”. In: *International conference on machine learning*. 2013, pp. 64–72.
- [73] Yan Chen et al. “Differentially private regression diagnostics”. In: *Data Mining (ICDM), 2016 IEEE 16th International Conference on*. IEEE. 2016, pp. 81–90.
- [74] Yudong Chen, Lili Su, and Jiaming Xu. “Distributed statistical machine learning in adversarial settings: Byzantine gradient descent”. In: *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 1.2 (2017), p. 44.
- [75] Tzuu-Shuh Chiang, Chii-Ruey Hwang, and Shuenn Jyi Sheu. “Diffusion for global optimization in  $R^n$ ”. In: *SIAM Journal on Control and Optimization* 25.3 (1987), pp. 737–753.
- [76] Stéphan Cléménçon, Gábor Lugosi, Nicolas Vayatis, et al. “Ranking and empirical minimization of U-statistics”. In: *The Annals of Statistics* (2008).
- [77] Patrick L Combettes and Valerie R Wajs. “Signal recovery by proximal forward-backward splitting”. In: *Multiscale Modeling and Simulation* 4.4 (2005), pp. 1168–1200.
- [78] Andrew R Conn, Nicholas IM Gould, and Ph L Toint. *Trust region methods*. Vol. 1. Siam, 2000.
- [79] Anna B Costello and Jason W Osborne. “Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis”. In: *Practical assessment, research & evaluation* 10.7 (2005), pp. 1–9.
- [80] Robert Culkin and Sanjiv R Das. “Machine learning in finance: the case of deep learning for option pricing”. In: *Journal of Investment Management* 15.4 (2017), pp. 92–100.
- [81] Alexandre d’Aspremont, Onureena Banerjee, and Laurent El Ghaoui. “First-order methods for sparse covariance selection”. In: *SIAM Journal on Matrix Analysis and Applications* 30.1 (2008), pp. 56–66.
- [82] Yuval Dagan and Vitaly Feldman. “Interaction is necessary for distributed learning with privacy or communication constraints”. In: *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*. 2020, pp. 450–462.

- [83] Arnak Dalalyan. “Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent”. In: *Conference on Learning Theory*. 2017, pp. 678–689.
- [84] Arnak S Dalalyan and Avetik Karagulyan. “User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient”. In: *Stochastic Processes and their Applications* (2019).
- [85] Chandler Davis and William Morton Kahan. “The rotation of eigenvectors by a perturbation. III”. In: *SIAM Journal on Numerical Analysis* 7.1 (1970), pp. 1–46.
- [86] Alexander Philip Dawid and Allan M Skene. “Maximum likelihood estimation of observer error-rates using the EM algorithm”. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 28.1 (1979), pp. 20–28.
- [87] Anindya De. “Lower bounds in differential privacy”. In: *Theory of cryptography conference*. Springer. 2012, pp. 321–338.
- [88] Jan De Spiegeleer et al. “Machine learning for quantitative finance: fast derivative pricing, hedging and fitting”. In: *Quantitative Finance* 18.10 (2018), pp. 1635–1643.
- [89] Rahul C Deo. “Machine learning in medicine”. In: *Circulation* 132.20 (2015), pp. 1920–1930.
- [90] Dua Dheeru and Efi Karra Taniskidou. *UCI Machine Learning Repository*. 2017. URL: <http://archive.ics.uci.edu/ml>.
- [91] Paramveer Dhillon et al. “New subsampling algorithms for fast least squares regression”. In: *Advances in neural information processing systems*. 2013, pp. 360–368.
- [92] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. “Collecting telemetry data privately”. In: *Advances in Neural Information Processing Systems*. 2017, pp. 3571–3580.
- [93] Hu Ding, Jing Gao, and Jinhui Xu. “Finding global optimum for truth discovery: Entropy based geometric variance”. In: *Proc. 32nd International Symposium on Computational Geometry (SoCG 2016)*. 2016.
- [94] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017. URL: <http://archive.ics.uci.edu/ml>.
- [95] John Duchi and Ryan Rogers. “Lower Bounds for Locally Private Estimation via Communication Complexity”. In: *arXiv preprint arXiv:1902.00582* (2019).

- [96] John Duchi and Ryan Rogers. “Lower Bounds for Locally Private Estimation via Communication Complexity”. In: *Proceedings of the Thirty-Second Conference on Learning Theory*. Ed. by Alina Beygelzimer and Daniel Hsu. Vol. 99. Proceedings of Machine Learning Research. Phoenix, USA, 2019, pp. 1161–1191.
- [97] John C Duchi, Michael I Jordan, and Martin J Wainwright. “Local privacy and statistical minimax rates”. In: *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*. IEEE. 2013, pp. 429–438.
- [98] John C Duchi, Michael I Jordan, and Martin J Wainwright. “Minimax optimal procedures for locally private estimation”. In: *Journal of the American Statistical Association* 113.521 (2018), pp. 182–201.
- [99] John C Duchi and Feng Ruan. “The right complexity measure in locally private estimation: It is not the Fisher information”. In: *arXiv preprint arXiv:1806.05756* (2018).
- [100] John C. Duchi and Feng Ruan. “The Right Complexity Measure in Locally Private Estimation: It is not the Fisher Information”. In: *CoRR* abs/1806.05756 (2018). URL: <http://arxiv.org/abs/1806.05756>.
- [101] Nelson Dunford and Jacob T Schwartz. *Linear operators part I: general theory*. Vol. 7. Interscience publishers New York, 1958.
- [102] Pavel Dvurechensky and Alexander Gasnikov. “Stochastic intermediate gradient method for convex problems with stochastic inexact oracle”. In: *Journal of Optimization Theory and Applications* 171.1 (2016), pp. 121–145.
- [103] Cynthia Dwork and Jing Lei. “Differential privacy and robust statistics”. In: *Proceedings of the forty-first annual ACM symposium on Theory of computing*. ACM. 2009, pp. 371–380.
- [104] Cynthia Dwork, Aaron Roth, et al. “The algorithmic foundations of differential privacy”. In: *Foundations and Trends® in Theoretical Computer Science* 9.3–4 (2014), pp. 211–407.
- [105] Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. “Boosting and differential privacy”. In: *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. IEEE. 2010, pp. 51–60.
- [106] Cynthia Dwork et al. “Analyze gauss: optimal bounds for privacy-preserving principal component analysis”. In: *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*. ACM. 2014, pp. 11–20.

- [107] Cynthia Dwork et al. “Calibrating noise to sensitivity in private data analysis”. In: *Theory of cryptography conference*. Springer. 2006, pp. 265–284.
- [108] Murat A Erdogdu, Mohsen Bayati, and Lee H Dicker. “Scalable approximations for generalized linear problems”. In: *The Journal of Machine Learning Research* 20.1 (2019), pp. 231–275.
- [109] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. “Rappor: Randomized aggregatable privacy-preserving ordinal response”. In: *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*. ACM. 2014, pp. 1054–1067.
- [110] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. “Rappor: Randomized aggregatable privacy-preserving ordinal response”. In: *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*. ACM. 2014, pp. 1054–1067.
- [111] Eugene F Fama. “Mandelbrot and the stable Paretian hypothesis”. In: *The journal of business* 36.4 (1963), pp. 420–429.
- [112] Jianqing Fan et al. “Distributed Estimation of Principal Eigenspaces”. In: *arXiv preprint arXiv:1702.06488* (2017).
- [113] Susana Faria and F Goncalves. “Financial data modeling by Poisson mixture regression”. In: *Journal of Applied Statistics* 40.10 (2013), pp. 2150–2162.
- [114] Vitaly Feldman and Thomas Steinke. “Calibrating Noise to Variance in Adaptive Data Analysis”. In: *Conference On Learning Theory*. 2018, pp. 535–544.
- [115] Dylan J Foster, Ayush Sekhari, and Karthik Sridharan. “Uniform convergence of gradients for non-convex learning and optimization”. In: *Advances in Neural Information Processing Systems*. 2018, pp. 8759–8770.
- [116] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. “Model inversion attacks that exploit confidence information and basic countermeasures”. In: *CCS*.
- [117] Zoltan Furedi and Janos Komlos. “The eigenvalues of random symmetric matrices”. In: *Combinatorica* 1.3 (1981), pp. 233–241.
- [118] Marco Gaboardi, Ryan Rogers, and Or Sheffet. “Locally Private Mean Estimation: Z-test and Tight Confidence Intervals”. In: *arXiv preprint arXiv:1810.08054* (2018).
- [119] Marco Gaboardi et al. “Dual Query: Practical Private Query Release for High Dimensional Data”. In: *Proceedings of the 31th International Conference on Machine*

- Learning, ICML 2014, Beijing, China, 21-26 June 2014.* 2014, pp. 1170–1178. URL: <http://jmlr.org/proceedings/papers/v32/gaboardi14.html>.
- [120] Karan Ganju et al. “Property inference attacks on fully connected neural networks using permutation invariant representations”. In: *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. 2018, pp. 619–633.
  - [121] Chao Gao and Dengyong Zhou. “Minimax optimal convergence rates for estimating ground truth from crowdsourced labels”. In: *arXiv preprint arXiv:1310.5764* (2013).
  - [122] Jason Ge et al. “Minimax-Optimal Privacy-Preserving Sparse PCA in Distributed Systems”. In: *International Conference on Artificial Intelligence and Statistics*. 2018, pp. 1589–1598.
  - [123] Rong Ge, Chi Jin, and Yi Zheng. “No Spurious Local Minima in Nonconvex Low Rank Problems: A Unified Geometric Analysis”. In: *International Conference on Machine Learning*. 2017, pp. 1233–1242.
  - [124] Rong Ge, Jason D Lee, and Tengyu Ma. “Matrix completion has no spurious local minimum”. In: *Advances in Neural Information Processing Systems*. 2016, pp. 2973–2981.
  - [125] Rong Ge, Jason D. Lee, and Tengyu Ma. “Learning One-hidden-layer Neural Networks with Landscape Design”. In: *International Conference on Learning Representations*. 2018.
  - [126] Rong Ge et al. “Escaping from saddle points-online stochastic gradient for tensor decomposition”. In: *Conference on Learning Theory*. 2015, pp. 797–842.
  - [127] Saeed Ghadimi and Guanghui Lan. “Accelerated gradient methods for nonconvex nonlinear and stochastic programming”. In: *Mathematical Programming* 156.1-2 (2016), pp. 59–99.
  - [128] Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. “Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization”. In: *Mathematical Programming* 155.1-2 (2016), pp. 267–305.
  - [129] David Gilbarg and Neil S Trudinger. *Elliptic partial differential equations of second order*. Springer, 2015.
  - [130] Larry Goldstein, Gesine Reinert, et al. “Stein’s method and the zero bias transformation with application to simple random sampling”. In: *The Annals of Applied Probability* 7.4 (1997), pp. 935–952.

- [131] Gene H Golub and Charles F Van Loan. *Matrix computations*. Vol. 3. JHU Press, 2012.
- [132] Alon Gonen and Ram Gilad-Bachrach. “Smooth sensitivity based approach for differentially private PCA”. In: *Algorithmic Learning Theory*. 2018, pp. 438–450.
- [133] Ian Goodfellow et al. *Deep learning*. Vol. 1. MIT press Cambridge, 2016.
- [134] Nicholas IM Gould et al. “Solving the trust-region subproblem using the Lanczos method”. In: *SIAM Journal on Optimization* 9.2 (1999), pp. 504–525.
- [135] Michael Grant and Stephen Boyd. *CVX: Matlab Software for Disciplined Convex Programming, version 2.1*. <http://cvxr.com/cvx>. Mar. 2014.
- [136] Justin Grimmer. “We are all social scientists now: How big data, machine learning, and causal inference work together”. In: *PS: Political Science & Politics* 48.1 (2015), pp. 80–83.
- [137] Adam Groce, Jonathan Katz, and Arkady Yerukhimovich. “Limits of computational differential privacy in the client/server setting”. In: *Theory of Cryptography Conference*. Springer. 2011, pp. 417–431.
- [138] Anupam Gupta et al. “Privately releasing conjunctions and the statistical query barrier”. In: *SIAM Journal on Computing* 42.4 (2013), pp. 1494–1520.
- [139] Jihun Hamm, Yingjun Cao, and Mikhail Belkin. “Learning privately from multiparty data”. In: *International Conference on Machine Learning*. 2016, pp. 555–563.
- [140] Samuel Haney et al. “Utility Cost of Formal Privacy for Releasing National Employer-Employee Statistics”. In: *SIGMOD*. Chicago, Illinois, USA: ACM, 2017, pp. 1339–1354. ISBN: 978-1-4503-4197-4. DOI: 10.1145/3035918.3035940. URL: <http://doi.acm.org/10.1145/3035918.3035940>.
- [141] Moritz Hardt and Aaron Roth. “Beyond worst-case analysis in private singular vector computation”. In: *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*. ACM. 2013, pp. 331–340.
- [142] Moritz Hardt and Guy N Rothblum. “A multiplicative weights mechanism for privacy-preserving data analysis”. In: *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. IEEE. 2010, pp. 61–70.
- [143] Moritz Hardt, Guy N Rothblum, and Rocco A Servedio. “Private data release via learning thresholds”. In: *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics. 2012, pp. 168–187.

- [144] Elad Hazan, Amit Agarwal, and Satyen Kale. “Logarithmic regret algorithms for online convex optimization”. In: *Machine Learning* (2007).
- [145] JB Heaton, NG Polson, and Jan Hendrik Witte. “Deep learning for finance: deep portfolios”. In: *Applied Stochastic Models in Business and Industry* 33.1 (2017), pp. 3–12.
- [146] Christina Heinze-Deml, Brian McWilliams, and Nicolai Meinshausen. “Preserving privacy between features in distributed estimation”. In: *Stat* 7.1 (2018), e189.
- [147] Matthew Hindman. “Building better models: Prediction, replication, and machine learning in the social sciences”. In: *The ANNALS of the American Academy of Political and Social Science* 659.1 (2015), pp. 48–62.
- [148] Matthew J Holland. “Robust descent using smoothed multiplicative noise”. In: *22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*. Vol. 89. Proceedings of Machine Learning Research. 2019, pp. 703–711.
- [149] Matthew J Holland and Kazushi Ikeda. “Efficient learning with robust gradient descent”. In: *Machine Learning* (2017), pp. 1–38.
- [150] Cho-Jui Hsieh et al. “QUIC: quadratic approximation for sparse inverse covariance estimation.” In: *Journal of Machine Learning Research* 15.1 (2014), pp. 2911–2947.
- [151] Daniel Hsu and Sivan Sabato. “Loss minimization and parameter estimation with heavy tails”. In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 543–582.
- [152] Justin Hsu et al. “Private matchings and allocations”. In: *SIAM Journal on Computing* 45.6 (2016), pp. 1953–1984.
- [153] Mengdi Huai et al. “Metric Learning from Probabilistic Labels”. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*. 2018, pp. 1541–1550.
- [154] Mengdi Huai et al. “Pairwise Learning with Differential Privacy Guarantees”. In: *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York City, New York, USA, February 7-12, 2020*. 2020.
- [155] Mengdi Huai et al. “Privacy-aware Synthesizing for Crowdsourced Data”. In: *28th International Joint Conference on Artificial Intelligence (IJCAI 2019)*. 2019.
- [156] Ziyun Huang, Hu Ding, and Jinhui Xu. “Faster algorithm for truth discovery via range cover”. In: *Workshop on Algorithms and Data Structures*. Springer. 2017, pp. 461–472.

- [157] Marat Ibragimov, Rustam Ibragimov, and Johan Walden. *Heavy-tailed distributions and robustness in economics and finance*. Vol. 214. Springer, 2015.
- [158] Hafiz Imtiaz and Anand D Sarwate. “Symmetric matrix perturbation for differentially-private principal component analysis”. In: *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE. 2016, pp. 2339–2343.
- [159] Roger Iyengar et al. “Towards Practical Differentially Private Convex Optimization”. In: *Towards Practical Differentially Private Convex Optimization*. IEEE, p. 1.
- [160] Prateek Jain, Pravesh Kothari, and Abhradeep Thakurta. “Differentially private online learning”. In: *Conference on Learning Theory*. 2012, pp. 24–1.
- [161] Prateek Jain, Ambuj Tewari, and Purushottam Kar. “On iterative hard thresholding methods for high-dimensional m-estimation”. In: *Advances in Neural Information Processing Systems*. 2014, pp. 685–693.
- [162] Wuxuan Jiang, Cong Xie, and Zhihua Zhang. “Wishart Mechanism for Differentially Private Principal Components Analysis.” In: *AAAI*. 2016, pp. 1730–1736.
- [163] Chi Jin et al. “How to Escape Saddle Points Efficiently”. In: *International Conference on Machine Learning*. 2017, pp. 1724–1732.
- [164] Chi Jin et al. “On the Local Minima of the Empirical Risk”. In: *Advances in Neural Information Processing Systems*. 2018, pp. 4901–4910.
- [165] Rong Jin, Shijun Wang, and Yang Zhou. “Regularized distance metric learning: Theory and algorithm”. In: *NIPS*. 2009, pp. 862–870.
- [166] Rie Johnson and Tong Zhang. “Accelerating stochastic gradient descent using predictive variance reduction”. In: *Advances in neural information processing systems*. 2013, pp. 315–323.
- [167] IT Jolliffe and BJT Morgan. “Principal component analysis and exploratory factor analysis”. In: *Statistical methods in medical research* 1.1 (1992), pp. 69–95.
- [168] Rosie Jones et al. ““I know what you did last summer” query logs and user privacy”. In: *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. 2007, pp. 909–914.
- [169] Matthew Joseph et al. “Locally Private Gaussian Estimation”. In: *arXiv preprint arXiv:1811.08382* (2018).
- [170] Matthew Joseph et al. “The Role of Interactivity in Local Differential Privacy”. In: *arXiv preprint arXiv:1904.03564* (2019).

- [171] Anatoli Juditsky and Arkadii S Nemirovski. “Large deviations of vector-valued martingales in 2-smooth normed spaces”. In: *arXiv preprint arXiv:0809.0813* (2008).
- [172] Gautam Kamath et al. “Differentially private algorithms for learning mixtures of separated gaussians”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 168–180.
- [173] Gautam Kamath et al. “Privately Learning High-Dimensional Distributions”. In: *arXiv preprint arXiv:1805.00216* (2018).
- [174] Michael Kapralov and Kunal Talwar. “On differentially private low rank approximation”. In: *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*. SIAM. 2013, pp. 1395–1414.
- [175] Purushottam Kar et al. “On the generalization ability of online learning algorithms for pairwise loss functions”. In: *International Conference on Machine Learning*. 2013, pp. 441–449.
- [176] Vishesh Karwa and Salil Vadhan. “Finite sample differentially private confidence intervals”. In: *arXiv preprint arXiv:1711.03908* (2017).
- [177] Shiva Prasad Kasiviswanathan and Hongxia Jin. “Efficient private empirical risk minimization for high-dimensional learning”. In: *International Conference on Machine Learning*. 2016, pp. 488–497.
- [178] Shiva Prasad Kasiviswanathan et al. “What can we learn privately?” In: *SIAM Journal on Computing* 40.3 (2011), pp. 793–826.
- [179] Kenji Kawaguchi. “Deep learning without poor local minima”. In: *Advances in Neural Information Processing Systems*. 2016, pp. 586–594.
- [180] Daniel Kifer and Ashwin Machanavajjhala. “No free lunch in data privacy”. In: *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*. ACM. 2011, pp. 193–204.
- [181] Daniel Kifer, Adam Smith, and Abhradeep Thakurta. “Private convex empirical risk minimization and high-dimensional regression”. In: *Conference on Learning Theory*. 2012, pp. 25–1.
- [182] Jonas Moritz Kohler and Aurelien Lucchi. “Sub-sampled cubic regularization for non-convex optimization”. In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2017, pp. 1895–1904.
- [183] Vladimir Koltchinskii, Karim Lounici, et al. “Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance”. In: *Annales de*

*l’Institut Henri Poincaré, Probabilités et Statistiques*. Vol. 52. 4. 2016, pp. 1976–2013.

- [184] Brian Kulis and Peter L Bartlett. “Implicit online learning”. In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*.
- [185] Tejas Kulkarni, Graham Cormode, and Divesh Srivastava. “Marginal Release Under Local Differential Privacy”. In: *CoRR* abs/1711.02952 (Nov. 2017). arXiv: 1711 . 02952. URL: <http://arxiv.org/abs/1711.02952>.
- [186] Abhishek Kumar et al. “A binary classification framework for two-stage multiple kernel learning”. In: *arXiv preprint arXiv:1206.6428* (2012).
- [187] Simon Lacoste-Julien. “Convergence rate of Frank-Wolfe for non-convex objectives”. In: *arXiv preprint arXiv:1607.00345* (2016).
- [188] Nan M Laird. “The EM algorithm in genetics, genomics and public health”. In: *Statistical Science* (2010), pp. 450–457.
- [189] Beatrice Laurent and Pascal Massart. “Adaptive estimation of a quadratic functional by model selection”. In: *Annals of Statistics* (2000), pp. 1302–1338.
- [190] Nada Lavrac et al. “Intelligent data analysis for medical diagnosis: using machine learning and temporal abstraction”. In: () .
- [191] Guillaume Lecué, Matthieu Lerasle, and Timothée Mathieu. “Robust classification via MOM minimization”. In: *arXiv preprint arXiv:1808.03106* (2018).
- [192] Jaewoo Lee and Daniel Kifer. “Concentrated Differentially Private Gradient Descent with Adaptive per-Iteration Privacy Budget”. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2018.
- [193] Jing Lei and Vincent Q Vu. “Sparsistency and agnostic inference in sparse PCA”. In: *The Annals of Statistics* 43.1 (2015), pp. 299–322.
- [194] Michael KK Leung et al. “Machine learning in genomic medicine: a review of computational problems and data sets”. In: *Proceedings of the IEEE* 104.1 (2015), pp. 176–197.
- [195] Bai Li et al. “On Connecting Stochastic Gradient MCMC and Differential Privacy”. In: *Proceedings of Machine Learning Research*. Ed. by Kamalika Chaudhuri and Masashi Sugiyama. Vol. 89. 2019, pp. 557–566.
- [196] Jundong Li et al. “Feature selection: A data perspective”. In: *ACM Computing Surveys (CSUR)* 50.6 (2017), p. 94.

- [197] Shi Li, Jinhui Xu, and Minwei Ye. “Approximating global optimum for probabilistic truth discovery”. In: *International Computing and Combinatorics Conference*. Springer. 2018, pp. 96–107.
- [198] Yaliang Li et al. “An efficient two-layer mechanism for privacy-preserving truth discovery”. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM. 2018, pp. 1705–1714.
- [199] James K Lindsey and Bradley Jones. “Choosing among generalized linear models applied to medical data”. In: *Statistics in medicine* 17.1 (1998), pp. 59–68.
- [200] Max A Little et al. “Suitability of dysphonia measurements for telemonitoring of Parkinson’s disease”. In: *IEEE transactions on biomedical engineering* 56.4 (2009), pp. 1015–1022.
- [201] Po-Ling Loh and Martin J Wainwright. “High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity”. In: *Advances in Neural Information Processing Systems*. 2011, pp. 2726–2734.
- [202] Po-Ling Loh and Martin J Wainwright. “Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima”. In: *Advances in Neural Information Processing Systems*. 2013, pp. 476–484.
- [203] G.G. Lorentz. *Bernstein Polynomials*. AMS Chelsea Publishing Series. Chelsea Publishing Company, 1986. ISBN: 9780828403238.
- [204] GG Lorentz. “Metric entropy and approximation”. In: *Bulletin of the American Mathematical Society* 72.6 (1966), pp. 903–937.
- [205] Jian Lou and Yiu-ming Cheung. “Uplink Communication Efficient Differentially Private Sparse Optimization With Feature-Wise Distributed Data”. In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*. 2018.
- [206] Jian Lou and Yiu-ming Cheung. “Uplink Communication Efficient Differentially Private Sparse Optimization with Feature-Wise Distributed Data”. In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.
- [207] Aurelie C Lozano, Nicolai Meinshausen, Eunho Yang, et al. “Minimum Distance Lasso for robust high-dimensional regression”. In: *Electronic Journal of Statistics* 10.1 (2016), pp. 1296–1340.
- [208] Dongsheng Lu and Shuhua Xu. “Principal component analysis reveals the 1000 Genomes Project does not sufficiently cover the human genetic diversity in Asia”. In: *Frontiers in genetics* 4 (2013), p. 127.

- [209] Zhaosong Lu. “Smooth optimization approach for sparse covariance selection”. In: *SIAM Journal on Optimization* 19.4 (2009), pp. 1807–1827.
- [210] Jinwen Ma, Lei Xu, and Michael I Jordan. “Asymptotic convergence rate of the EM algorithm for Gaussian mixtures”. In: *Neural Computation* 12.12 (2000), pp. 2881–2907.
- [211] Benoit B Mandelbrot. “The variation of certain speculative prices”. In: *Fractals and scaling in finance*. Springer, 1997, pp. 371–418.
- [212] Leonard A Marascuilo and Ronald C Serlin. *Statistical methods for the social and behavioral sciences*. WH Freeman/Times Books/Henry Holt & Co, 1988.
- [213] Winter Mason, Jennifer Wortman Vaughan, and Hanna Wallach. *Computational social science and social computing*. 2014.
- [214] Pascal Massart. “Concentration inequalities and model selection”. In: (2007).
- [215] Geoffrey McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*. Vol. 382. John Wiley & Sons, 2007.
- [216] Audra McMillan and Anna C Gilbert. “Local differential privacy for physical sensor data and sparse recovery”. In: *Information Sciences and Systems (CISS), 2018 52nd Annual Conference on*. IEEE. 2018, pp. 1–6.
- [217] Alexander J McNeil and Jonathan P Wandin. “Bayesian inference for generalized linear mixed models of portfolio credit risk”. In: *Journal of Empirical Finance* 14.2 (2007), pp. 131–149.
- [218] Song Mei, Yu Bai, and Andrea Montanari. “The landscape of empirical risk for non-convex losses”. In: *arXiv preprint arXiv:1607.06534* (2016).
- [219] Charles Micchelli. “The saturation class and iterates of the Bernstein polynomials”. In: *Journal of Approximation Theory* 8.1 (1973), pp. 1–18.
- [220] Stanislav Minsker et al. “Geometric median and robust estimation in Banach spaces”. In: *Bernoulli* 21.4 (2015), pp. 2308–2335.
- [221] Ilya Mironov. “Renyi differential privacy”. In: *Computer Security Foundations Symposium (CSF), 2017 IEEE 30th*. IEEE. 2017, pp. 263–275.
- [222] Aryan Mokhtari, Asuman Ozdaglar, and Ali Jadbabaie. “Escaping Saddle Points in Constrained Optimization”. In: *Advances in Neural Information Processing Systems*. 2018, pp. 3633–3643.

- [223] Arvind Narayanan and Vitaly Shmatikov. “Robust de-anonymization of large sparse datasets”. In: *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE. 2008, pp. 111–125.
- [224] Milad Nasr, Reza Shokri, and Amir Houmansadr. “Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning”. In: *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2019, pp. 739–753.
- [225] Michael Natole, Yiming Ying, and Siwei Lyu. “Stochastic proximal algorithms for AUC maximization”. In: *International Conference on Machine Learning*. 2018.
- [226] Radford M Neal and Geoffrey E Hinton. “A view of the EM algorithm that justifies incremental, sparse, and other variants”. In: *Learning in graphical models*. Springer, 1998, pp. 355–368.
- [227] Sahand N Negahban et al. “A unified framework for high-dimensional analysis of  $M$ -estimators with decomposable regularizers”. In: *Statistical Science* 27.4 (2012), pp. 538–557.
- [228] Arkadi Nemirovski. “On Parallel Complexity of Nonsmooth Convex Optimization”. In: *J. Complexity* 10.4 (1994), pp. 451–463.
- [229] Yu Nesterov. “Smooth minimization of non-smooth functions”. In: *Mathematical programming* 103.1 (2005), pp. 127–152.
- [230] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*. Vol. 87. Springer Science & Business Media, 2013.
- [231] Tan Nguyen and Scott Sanner. “Algorithms for direct 0–1 loss optimization in binary classification”. In: *International Conference on Machine Learning*. 2013, pp. 1085–1093.
- [232] Kristin K Nicodemus and James D Malley. “Predictor correlation impacts machine learning algorithms: implications for genomic studies”. In: *Bioinformatics* 25.15 (2009), pp. 1884–1890.
- [233] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. “Smooth sensitivity and sampling in private data analysis”. In: *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*. ACM. 2007, pp. 75–84.
- [234] Kobbi Nissim and Uri Stemmer. “Clustering Algorithms for the Centralized and Local Models”. In: *Algorithmic Learning Theory*. 2018, pp. 619–653.

- [235] Liam O'Neill, Franklin Dexter, and Nan Zhang. “The risks to patient privacy from publishing data from clinical anesthesia studies”. In: *Anesthesia & Analgesia* 122.6 (2016), pp. 2017–2027.
- [236] Ziad Obermeyer and Ezekiel J Emanuel. “Predicting the future—big data, machine learning, and clinical medicine”. In: *The New England journal of medicine* 375.13 (2016), p. 1216.
- [237] Alain Pajor. “Metric entropy of the Grassmann manifold”. In: *Convex Geometric Analysis* 34 (1998), pp. 181–188.
- [238] Nicolas Papernot et al. “Scalable private learning with PATE”. In: *arXiv preprint arXiv:1802.08908* (2018).
- [239] Nicolas Papernot et al. “Semi-supervised knowledge transfer for deep learning from private training data”. In: *arXiv preprint arXiv:1610.05755* (2016).
- [240] Athanasios Papoulis. “Probability, random variables, and stochastic processes”. In: (1965).
- [241] E Pardoux and A Yu Veretennikov. “On the Poisson equation and diffusion approximation. I”. In: *Annals of probability* (2001), pp. 1061–1085.
- [242] Mijung Park et al. “DP-EM: Differentially Private Expectation Maximization”. In: *Artificial Intelligence and Statistics*. 2017, pp. 896–904.
- [243] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. “Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis”. In: *Proceedings of the 2017 Conference on Learning Theory*. 2017.
- [244] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. “Minimax rates of estimation for high-dimensional linear regression over  $\ell_q$ -balls”. In: *IEEE transactions on information theory* 57.10 (2011), pp. 6976–6994.
- [245] Holger Rauhut. “Compressive sensing and structured random matrices”. In: *Theoretical foundations and numerical methods for sparse recovery* 9 (2010), pp. 1–92.
- [246] Pradeep Ravikumar et al. “High-dimensional covariance estimation by minimizing  $\ell_1$ -penalized log-determinant divergence”. In: *Electronic Journal of Statistics* 5 (2011), pp. 935–980.
- [247] Sashank J Reddi et al. “Stochastic frank-wolfe methods for nonconvex optimization”. In: *Communication, Control, and Computing (Allerton), 2016 54th Annual Allerton Conference on*. IEEE. 2016, pp. 1244–1251.

- [248] Xuebin Ren et al. “LoPub: High-Dimensional Crowdsourced Data Publication with Local Differential Privacy”. In: *IEEE Transactions on Information Forensics and Security* 13.9 (2018), pp. 2151–2166.
- [249] Andrej Risteski and Yuanzhi Li. “Algorithms and matching lower bounds for approximately-convex optimization”. In: *Advances in Neural Information Processing Systems*. 2016, pp. 4745–4753.
- [250] Benjamin Rolfs et al. “Iterative thresholding algorithm for sparse inverse covariance estimation”. In: *Advances in Neural Information Processing Systems*. 2012, pp. 1574–1582.
- [251] Katya Scheinberg, Shiqian Ma, and Donald Goldfarb. “Sparse inverse covariance selection via alternating linearization methods”. In: *Advances in neural information processing systems*. 2010, pp. 2101–2109.
- [252] Shai Shalev-Shwartz et al. “Stochastic Convex Optimization.” In: *COLT*. 2009.
- [253] Or Sheffet. “Differentially Private Ordinary Least Squares”. In: *International Conference on Machine Learning*. 2017, pp. 3105–3114.
- [254] Or Sheffet. “Private approximations of the 2nd-moment matrix using existing techniques in linear regression”. In: *arXiv preprint arXiv:1507.00056* (2015).
- [255] Irina Shevtsova. “On the absolute constants in the Berry-Esseen type inequalities for identically distributed summands”. In: *arXiv preprint arXiv:1111.6554* (2011).
- [256] Reza Shokri et al. “Membership inference attacks against machine learning models”. In: *Security and Privacy (SP), 2017 IEEE Symposium on*. IEEE. 2017, pp. 3–18.
- [257] Adam Smith, Abhradeep Thakurta, and Jalaj Upadhyay. “Is Interaction Necessary for Distributed Private Learning?” In: *IEEE Symposium on Security and Privacy*. 2017.
- [258] Congzheng Song, Thomas Ristenpart, and Vitaly Shmatikov. “Machine learning models that remember too much”. In: *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 2017, pp. 587–601.
- [259] Liwei Song, Reza Shokri, and Prateek Mittal. “Privacy Risks of Securing Machine Learning Models against Adversarial Examples”. In: *arXiv preprint arXiv:1905.10291* (2019).
- [260] Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. “Stochastic gradient descent with differentially private updates”. In: *2013 IEEE Global Conference on Signal and Information Processing*. IEEE. 2013, pp. 245–248.

- [261] Shuang Song, Om Thakkar, and Abhradeep Thakurta. “Characterizing Private Clipped Gradient Descent on Convex Generalized Linear Problems”. In: *arXiv preprint arXiv:2006.06783* (2020).
- [262] Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. “Smoothness, low noise and fast rates”. In: *Advances in neural information processing systems*. 2010, pp. 2199–2207.
- [263] Manfred Stapff and Sarah Hilderbrand. “First-line treatment of essential hypertension: A real-world analysis across four antihypertensive treatment classes”. In: *The Journal of Clinical Hypertension* 21.5 (2019), pp. 627–634.
- [264] G. W. Stewart. *Matrix Perturbation Theory*. 1990.
- [265] Dong Su et al. “Differentially private k-means clustering”. In: *Proceedings of the sixth ACM conference on data and application security and privacy*. ACM. 2016, pp. 26–37.
- [266] Haipei Sun et al. “Truth Inference on Sparse Crowdsourcing Data with Local Differential Privacy”. In: *2018 IEEE International Conference on Big Data (Big Data)*. IEEE. 2018, pp. 488–497.
- [267] Stanislaw J Szarek. “Nets of Grassmann manifold and orthogonal group”. In: *Proceedings of research workshop on Banach space theory (Iowa City, Iowa, 1981)*. Vol. 169. 1982, p. 185.
- [268] Yasuaki Takada et al. “A generalized linear model for decomposing cis-regulatory, parent-of-origin, and maternal effects on allele-specific gene expression”. In: *G3: Genes, Genomes, Genetics* 7.7 (2017), pp. 2227–2234.
- [269] Kunal Talwar, Abhradeep Thakurta, and Li Zhang. “Private empirical risk minimization beyond the worst case: The effect of the constraint set geometry”. In: *arXiv preprint arXiv:1411.5417* (2014).
- [270] Kunal Talwar, Abhradeep Guha Thakurta, and Li Zhang. “Nearly optimal private lasso”. In: *Advances in Neural Information Processing Systems*. 2015, pp. 3025–3033.
- [271] Conghui Tan et al. “Barzilai-Borwein step size for stochastic gradient descent”. In: *Advances in Neural Information Processing Systems*. 2016, pp. 685–693.
- [272] Jiaxi Tang and Ke Wang. “Ranking Distillation: Learning Compact Ranking Models With High Performance for Recommender System”. In: *Proc. of the 24th SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2018.

- [273] Jun Tang et al. “Privacy Loss in Apple’s Implementation of Differential Privacy on MacOS 10.12”. In: *CoRR* abs/1709.02753 (2017).
- [274] Terence Tao. “Topics in random matrix theory”. In: *Graduate Studies in Mathematics* 132 (2011).
- [275] Terence Tao. “Topics in random matrix theory”. In: *Graduate studies in Mathematics* 132 (2012), pp. 46–47.
- [276] Abhradeep Guha Thakurta and Adam Smith. “(Nearly) optimal algorithms for private online learning in full-information and bandit settings”. In: *Advances in Neural Information Processing Systems*. 2013, pp. 2733–2741.
- [277] Abhradeep Guha Thakurta and Adam Smith. “Differentially private feature selection via stability arguments, and the robustness of the lasso”. In: *COLT*. 2013, pp. 819–850.
- [278] Justin Thaler, Jonathan Ullman, and Salil Vadhan. “Faster algorithms for privately releasing marginals”. In: *International Colloquium on Automata, Languages, and Programming*. Springer. 2012, pp. 810–821.
- [279] Florian Tramer et al. “Stealing machine learning models via prediction apis”. In: *25th USENIX Security Symposium (USENIX Security 16)*. 2016, pp. 601–618.
- [280] Lloyd N Trefethen. *Approximation theory and approximation practice*. Vol. 128. Siam, 2013.
- [281] Robert R Trippi and Jae K Preface By-Lee. *Artificial intelligence in finance and investing: state-of-the-art technologies for securities selection and portfolio management*. McGraw-Hill, Inc., 1995.
- [282] Joel A Tropp et al. “An introduction to matrix concentration inequalities”. In: *Foundations and Trends® in Machine Learning* 8.1-2 (2015), pp. 1–230.
- [283] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. 1st. Springer Publishing Company, Incorporated, 2008.
- [284] Belinda Tzen, Tengyuan Liang, and Maxim Raginsky. “Local Optimality and Generalization Guarantees for the Langevin Algorithm via Empirical Metastability”. In: *Conference On Learning Theory*. 2018, pp. 857–875.
- [285] Jonathan Ullman. “Private multiplicative weights beyond linear queries”. In: *Proceedings of the 34th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*. ACM. 2015, pp. 303–312.

- [286] Jonathan Ullman. “Tight Lower Bounds for Locally Differentially Private Selection”. In: *arXiv preprint arXiv:1802.02638* (2018).
- [287] A Yu Veretennikov. “On polynomial mixing bounds for stochastic differential equations”. In: *Stochastic processes and their applications* 70.1 (1997), pp. 115–127.
- [288] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge University Press, 2018.
- [289] Roman Vershynin. “Introduction to the non-asymptotic analysis of random matrices”. In: *arXiv preprint arXiv:1011.3027* (2010).
- [290] Jeffrey S. Vitter. “Random Sampling with a Reservoir”. In: *ACM Trans. Math. Softw.* 11.1 (Mar. 1985), pp. 37–57. ISSN: 0098-3500. DOI: 10.1145/3147.3165. URL: <http://doi.acm.org/10.1145/3147.3165>.
- [291] Sebastian J Vollmer, Konstantinos C Zygalakis, and Yee Whye Teh. “Exploration of the (non-) asymptotic bias and variance of stochastic gradient langevin dynamics”. In: *Journal of Machine Learning Research* 17.159 (2016), pp. 1–48.
- [292] Vincent Q Vu, Jing Lei, et al. “Minimax sparse principal subspace estimation in high dimensions”. In: *The Annals of Statistics* 41.6 (2013), pp. 2905–2947.
- [293] Vincent Q Vu et al. “Fantope projection and selection: A near-optimal convex relaxation of sparse PCA”. In: *Advances in neural information processing systems*. 2013, pp. 2670–2678.
- [294] Binghui Wang and Neil Zhenqiang Gong. “Stealing hyperparameters in machine learning”. In: *2018 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2018, pp. 36–52.
- [295] D Wang, M Huai, and J Xu. “Differentially private sparse inverse covariance estimation”. In: *2018 IEEE Global Conference on Signal and Information Processing, GlobalSIP*. 2018, pp. 26–29.
- [296] Di Wang, Changyou Chen, and Jinhui Xu. “Differentially Private Empirical Risk Minimization with Non-convex Loss Functions”. In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. 2019, pp. 6526–6535.
- [297] Di Wang, Changyou Chen, and Jinhui Xu. “Differentially Private Empirical Risk Minimization with Non-convex Loss Functions”. In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, Long Beach, California, USA, June 9-15, 2019*. 2019.

- [298] Di Wang, Changyou Chen, and Jinhui Xu. “Differentially Private Empirical Risk Minimization with Non-convex Loss Functions”. In: *International Conference on Machine Learning*. 2019, pp. 6526–6535.
- [299] Di Wang, Marco Gaboardi, and Jinhui Xu. “Empirical Risk Minimization in Non-interactive Local Differential Privacy Revisited”. In: *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, 3-8 December 2018, Montreal, QC, Canada* (2018). arXiv: 1802 . 04085. URL: <http://arxiv.org/abs/1802.04085>.
- [300] Di Wang, Marco Gaboardi, and Jinhui Xu. “Empirical risk minimization in non-interactive local differential privacy revisited”. In: *Advances in Neural Information Processing Systems*. 2018, pp. 965–974.
- [301] Di Wang, Marco Gaboardi, and Jinhui Xu. “Empirical risk minimization in non-interactive local differential privacy revisited”. In: *Advances in Neural Information Processing Systems*. 2018, pp. 965–974.
- [302] Di Wang, Mengdi Huai, and Jinhui Xu. “Differentially Private Sparse Inverse Covariance Estimation”. In: *2018 IEEE Global Conference on Signal and Information Processing, GlobalSIP 2018, Anaheim, CA, USA, November 26-29, 2018*.
- [303] Di Wang, Mengdi Huai, and Jinhui Xu. “Differentially private sparse inverse covariance estimation”. In: *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE. 2018, pp. 1139–1143.
- [304] Di Wang, Adam Smith, and Jinhui Xu. “High Dimensional Sparse Linear Regression under Local Differential Privacy: Power and Limitations”. In: *2018 NIPS workshop in Privacy-Preserving Machine Learning* (2018).
- [305] Di Wang, Adam Smith, and Jinhui Xu. “Noninteractive Locally Private Learning of Linear Models via Polynomial Approximations”. In: *Algorithmic Learning Theory*. 2019, pp. 897–902.
- [306] Di Wang, Adam Smith, and Jinhui Xu. “Noninteractive locally private learning of linear models via polynomial approximations”. In: *Algorithmic Learning Theory*. 2019, pp. 897–902.
- [307] Di Wang, Adam Smith, and Jinhui Xu. “Noninteractive locally private learning of linear models via polynomial approximations”. In: *Algorithmic Learning Theory*. 2019, pp. 897–902.
- [308] Di Wang and Jinhui Xu. “Differentially Private Empirical Risk Minimization with Smooth Non-Convex Loss Functions: A Non-Stationary View”. In: *The Thirty-Third*

*AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019.* 2019, pp. 1182–1189.

- [309] Di Wang and Jinhui Xu. “Differentially Private Empirical Risk Minimization with Smooth Non-convex Loss Functions: A Non-stationary View”. In: (2019).
- [310] Di Wang and Jinhui Xu. “Differentially Private Empirical Risk Minimization with Smooth Non-convex Loss Functions: A Non-stationary View”. In: *Thirty-Third AAAI Conference on Artificial Intelligence, (AAAI-19), Honolulu, Hawaii, USA, January 27-February 1, 2019* (2019).
- [311] Di Wang and Jinhui Xu. “Differentially private empirical risk minimization with smooth non-convex loss functions: A non-stationary view”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 2019, pp. 1182–1189.
- [312] Di Wang and Jinhui Xu. “Differentially Private High Dimensional Sparse Covariance Matrix Estimation”. In: *CoRR* abs/1901.06413 (2019). URL: <http://arxiv.org/abs/1901.06413>.
- [313] Di Wang and Jinhui Xu. “Escaping Saddle Points of Empirical Risk Privately and Scalably via DP-Trust Region Method”. In: *The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, ECML-PKDD 2020, 14-18 September 2020, Virtual Conference*.
- [314] Di Wang and Jinhui Xu. “Faster constrained linear regression via two-step preconditioning”. In: *Neurocomputing* 364 (2019), pp. 280–296. DOI: 10.1016/j.neucom.2019.07.070. URL: <https://doi.org/10.1016/j.neucom.2019.07.070>.
- [315] Di Wang and Jinhui Xu. “Large Scale Constrained Linear Regression Revisited: Faster Algorithms via Preconditioning”. In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.
- [316] Di Wang and Jinhui Xu. “Lower Bound of Locally Differentially Private Sparse Covariance Matrix Estimation”. In: *28th International Joint Conference on Artificial Intelligence (IJCAI 2019)*. 2019.
- [317] Di Wang and Jinhui Xu. “On Sparse Linear Regression in the Local Differential Privacy Model”. In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, Long Beach, California, USA, June 9-15, 2019*. 2019.
- [318] Di Wang and Jinhui Xu. “On Sparse Linear Regression in the Local Differential Privacy Model”. In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, Long Beach, California, USA, June 9-15, 2019*. 2019.

- [319] Di Wang and Jinhui Xu. “On Sparse Linear Regression in the Local Differential Privacy Model”. In: *International Conference on Machine Learning*. 2019, pp. 6628–6637.
- [320] Di Wang and Jinhui Xu. “Principal Component Analysis in the Local Differential Privacy Model”. In: *28th International Joint Conference on Artificial Intelligence (IJCAI 2019)*. 2019.
- [321] Di Wang and Jinhui Xu. “Principal component analysis in the local differential privacy model”. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI)*. 2019, pp. 4795–4801.
- [322] Di Wang and Jinhui Xu. “Principal component analysis in the local differential privacy model”. In: *Theoretical Computer Science* 809 (2020), pp. 296–312.
- [323] Di Wang and Jinhui Xu. “Tight lower bound of sparse covariance matrix estimation in the local differential privacy model”. In: *Theoretical Computer Science* (2020).
- [324] Di Wang, Minwei Ye, and Jinhui Xu. “Differentially Private Empirical Risk Minimization Revisited: Faster and More General”. In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*. 2017, pp. 2719–2728.
- [325] Di Wang, Minwei Ye, and Jinhui Xu. “Differentially Private Empirical Risk Minimization Revisited: Faster and More General”. In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*. 2017, pp. 2719–2728.
- [326] Di Wang, Minwei Ye, and Jinhui Xu. “Differentially Private Empirical Risk Minimization Revisited: Faster and More General”. In: *NIPS-2017*. 2017.
- [327] Di Wang, Minwei Ye, and Jinhui Xu. “Differentially private empirical risk minimization revisited: Faster and more general”. In: *Advances in Neural Information Processing Systems*. 2017, pp. 2722–2731.
- [328] Di Wang, Minwei Ye, and Jinhui Xu. “Differentially private empirical risk minimization revisited: Faster and more general”. In: *Advances in Neural Information Processing Systems*. 2017, pp. 2722–2731.
- [329] Di Wang et al. “Estimating Smooth GLMs in Non-interactive Local Differential Privacy Model with Public Unlabeled Data”. In: *Submission*. 2019.
- [330] Di Wang et al. “Estimating stochastic linear combination of non-linear regressions efficiently and scalably”. In: *Neurocomputing* 399 (2020), pp. 129–140.

- [331] Di Wang et al. “On Differentially Private Stochastic Optimization with Heavy-tailed Data”. In: *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 12-18 July 2020, Virtual Conference*. 2020.
- [332] Di Wang et al. “Scalable Estimating Stochastic Linear Combination of Non-linear Regressions”. In: *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York City, New York, USA, February 7-12, 2020*. 2020.
- [333] Rui Wang et al. “Learning your identity and disease from research papers: information leaks in genome wide association study”. In: *Proceedings of the 16th ACM conference on Computer and communications security*. 2009, pp. 534–544.
- [334] Yu-Xiang Wang. “Revisiting differentially private linear regression: optimal and adaptive prediction & estimation in unbounded domain”. In: *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*. 2018, pp. 93–103.
- [335] Yu-Xiang Wang, Stephen Fienberg, and Alex Smola. “Privacy for free: Posterior sampling and stochastic gradient monte carlo”. In: *International Conference on Machine Learning*. 2015.
- [336] Yu-Xiang Wang, Jing Lei, and Stephen E Fienberg. “Learning with differential privacy: Stability, learnability and the sufficiency and necessity of ERM principle”. In: *Journal of Machine Learning Research* 17.183 (2016), pp. 1–40.
- [337] Yining Wang, Yu-Xiang Wang, and Aarti Singh. “Differentially private subspace clustering”. In: *Advances in Neural Information Processing Systems*. 2015, pp. 1000–1008.
- [338] Zhaoran Wang, Huanran Lu, and Han Liu. “Tighten after relax: Minimax-optimal sparse PCA in polynomial time”. In: *Advances in neural information processing systems*. 2014, pp. 3383–3391.
- [339] Zhaoran Wang et al. “High dimensional em algorithm: Statistical optimization and asymptotic normality”. In: *Advances in neural information processing systems*. 2015, pp. 2521–2529.
- [340] Ziteng Wang et al. “Differentially private data releasing for smooth queries”. In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 1779–1820.
- [341] Russell T. Warne. *Statistics for the Social Sciences: A General Linear Model Approach*. Cambridge University Press, 2017. DOI: 10.1017/9781316442715.
- [342] Jon Wellner et al. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 2013.

- [343] Peter Whittle. “Bounds for the moments of linear and quadratic forms in independent variables”. In: *Theory of Probability & Its Applications* 5.3 (1960), pp. 302–305.
- [344] Blake E Woodworth et al. “Graph oracle models, lower bounds, and gaps for parallel stochastic optimization”. In: *Advances in Neural Information Processing Systems*. 2018, pp. 8505–8515.
- [345] Robert F Woolson and William R Clarke. *Statistical methods for the analysis of biomedical data*. Vol. 371. John Wiley & Sons, 2011.
- [346] CF Jeff Wu et al. “On the convergence properties of the EM algorithm”. In: *The Annals of statistics* 11.1 (1983), pp. 95–103.
- [347] Lin Xiao and Tong Zhang. “A proximal stochastic gradient method with progressive variance reduction”. In: *SIAM Journal on Optimization* 24.4 (2014), pp. 2057–2075.
- [348] Pan Xu et al. “Global convergence of Langevin dynamics based algorithms for nonconvex optimization”. In: *Advances in Neural Information Processing Systems*. 2018, pp. 3126–3137.
- [349] Yi Xu, Jing Rong, and Tianbao Yang. “First-order stochastic algorithms for escaping from saddle points in almost linear time”. In: *Advances in Neural Information Processing Systems*. 2018, pp. 5535–5545.
- [350] Zhuoran Yang et al. “Sparse nonlinear regression: Parameter estimation under nonconvexity”. In: *International Conference on Machine Learning*. 2016, pp. 2472–2481.
- [351] Min Ye and Alexander Barg. “Optimal schemes for discrete distribution estimation under locally differential privacy”. In: *IEEE Transactions on Information Theory* 64.8 (2018), pp. 5662–5676.
- [352] Samuel Yeom et al. “Privacy risk in machine learning: Analyzing the connection to overfitting”. In: *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*. IEEE. 2018, pp. 268–282.
- [353] Xinyang Yi and Constantine Caramanis. “Regularized em algorithms: A unified framework and statistical guarantees”. In: *Advances in Neural Information Processing Systems*. 2015, pp. 1567–1575.
- [354] Yi Yu, Tengyao Wang, and Richard J Samworth. “A useful variant of the Davis–Kahan theorem for statisticians”. In: *Biometrika* 102.2 (2014), pp. 315–323.
- [355] Ming Yuan and Yi Lin. “Model selection and estimation in the Gaussian graphical model”. In: *Biometrika* 94.1 (2007), pp. 19–35.

- [356] Jiaqi Zhang et al. “Efficient Private ERM for Smooth Objectives”. In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*. 2017, pp. 3922–3928.
- [357] Kaiqing Zhang, Zhuoran Yang, and Zhaoran Wang. “Nonlinear Structured Signal Estimation in High Dimensions via Iterative Hard Thresholding”. In: *International Conference on Artificial Intelligence and Statistics*. 2018, pp. 258–268.
- [358] Yuchen Zhang et al. “Spectral methods meet em: A provably optimal algorithm for crowdsourcing”. In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 3537–3580.
- [359] Zhikun Zhang et al. “Calm: Consistent adaptive local marginal for marginal release under local differential privacy”. In: *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. ACM. 2018, pp. 212–229.
- [360] Peilin Zhao et al. “Online AUC maximization”. In: *ICML*. 2011, pp. 233–240.
- [361] Peng Zhao and Bin Yu. “On model selection consistency of Lasso”. In: *Journal of Machine learning research* 7.Nov (2006), pp. 2541–2563.
- [362] Kai Zheng, Wenlong Mou, and Liwei Wang. “Collect at Once, Use Effectively: Making Non-interactive Locally Private Learning Possible”. In: *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*. 2017, pp. 4130–4139.
- [363] Kai Zheng, Wenlong Mou, and Liwei Wang. “Collect at once, use effectively: Making non-interactive locally private learning possible”. In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2017, pp. 4130–4139.
- [364] Tianhang Zheng et al. “Towards Assessment of Randomized Mechanisms for Certifying Adversarial Robustness”. In: *CoRR* abs/2005.07347 (2020). URL: <https://arxiv.org/abs/2005.07347>.
- [365] Yudian Zheng et al. “Truth inference in crowdsourcing: Is the problem solved?” In: *Proceedings of the VLDB Endowment* 10.5 (2017), pp. 541–552.
- [366] Shi Zhi et al. “Dynamic Truth Discovery on Numerical Data”. In: *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE. 2018, pp. 817–826.
- [367] Dongruo Zhou, Pan Xu, and Quanquan Gu. “Stochastic Variance-Reduced Cubic Regularized Newton Method”. In: *International Conference on Machine Learning*. 2018, pp. 5985–5994.

- [368] Kaiwen Zhou, Fanhua Shang, and James Cheng. “A Simple Stochastic Variance Reduced Algorithm with Fast Convergence Rates”. In: *International Conference on Machine Learning*. 2018, pp. 5975–5984.
- [369] Rongda Zhu et al. “High-dimensional variance-reduced stochastic gradient expectation-maximization algorithm”. In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2017, pp. 4180–4188.
- [370] Shenghuo Zhu. “A short note on the tail bound of wishart distribution”. In: *arXiv preprint arXiv:1212.5860* (2012).
- [371] Martin Zinkevich. “Online convex programming and generalized infinitesimal gradient ascent”. In: *Proc. of the 20th International Conference on Machine Learning (ICML-03)*. 2003, pp. 928–936.