On efficient methods for high-dimensional statistical estimation

Dmitry Babichev

To cite this version:

Dmitry Babichev. On efficient methods for high-dimensional statistical estimation. Machine Learning [stat.ML]. Université Paris sciences et lettres, 2019. English. NNT : 2019PSLEE032 . tel-02433016v2

HAL Id: tel-02433016
https://theses.hal.science/tel-02433016v2

Submitted on 18 Jun 2020

PSL UNIVERSITÉ PARIS

THÈSE DE DOCTORAT DE L'UNIVERSITÉ PSL

Préparée à l'École normale supérieure

On Efficient Methods for High-dimensional Statistical Estimation

Soutenue par

Babichev Dmitry

Le 22.02.2019

École doctorale nº386

Sciences Mathématiques de Paris Centre

Spécialité

Informatique

Composition du jury :

Olivier Cappé CNRS	*Président*
Arnak Dalalyan ENSAE ParisTech	*Rapporteur*
Stéphane Chrétien NPL	*Rapporteur*
Franck Iutzeler UGA	*Examinateur*
Anatoli Juditsky UGA	*Co-directeur de thèse*
Francis Bach Inria Paris	*Directeur de thèse*

ENS ÉCOLE NORMALE SUPÉRIEURE

PSL RESEARCH UNIVERSITY PARIS

The best angle from which to approach any problem is the try-angle

Author Unknown

Abstract

In this thesis we consider several aspects of parameter estimation for statistics and machine learning, and optimization techniques applicable to these problems. The goal of parameter estimation is to find the unknown hidden parameters, which govern the data, for example parameters of an unknown probability density. The construction of estimators through optimization problems is only one side of the coin, finding the optimal value of the parameter often is an optimization problem that needs to be solved, using various optimization techniques. Hopefully these optimization problems are convex for a wide class of problems, and we can exploit their structure to get fast convergence rates.

The first main contribution of the thesis is to develop moment-matching techniques for multi-index non-linear regression problems. We consider the classical non-linear regression problem, which is unfeasible in high dimensions due to the curse of dimensionality; that is why we assume a model, which states that in fact the data is a nonlinear function of several linear projections of data. We combine two existing techniques: average derivative estimator (ADE) and Sliced Inverse Regression (SIR) to develop the hybrid method (SADE) without some of the weak sides of its parents: it works both in multi-index models and with weak assumptions on the data distribution. We also extend this method to high-order moments. We provide theoretical analysis of constructed estimators both for finite sample and population cases.

In the second main contribution we use a special type of averaging for stochastic gradient descent. We consider generalized linear models for conditional exponential families (such as logistic regression), where the goal is to find the unknown value of the parameter. Classical approaches, such as Stochastic Gradient Descent (SGD) with constant step-size are known to converge only to some neighborhood of the optimal value of the parameter, even with Rupert-Polyak averaging. We propose the averaging of moment parameters, which we call prediction functions. For finite-dimensional models this type of averaging surprisingly can lead to negative error, i.e., this approach provides us with the estimator better than any linear estimator can ever achieve. For infinite-dimensional models our approach converges to the optimal prediction, while parameter averaging never does.

The third main contribution of this thesis deals with Fenchel-Young losses. We consider multi-class linear classifiers with the losses of a certain type, such that their dual conjugate has a direct product of simplices as a support. The corresponding saddle-point convex-concave formulation has a special form with a bilinear matrix term and classical approaches suffer from the time-consuming multiplication of matrices. We show, that for multi-class SVM losses, under mild regularity assumption and with smart matrix-multiplication sampling techniques, our approach has an iteration complexity which is sublinear in the size of the data. It means, that to do one iteration, we need to pay only trice $O(n + d + k)$: for number of classes k, number of features d and number of samples n, whereas all existing techniques use at least one of nd, nk or dk arithmetical operations per iteration. This is possible due to the right choice of geometries and using a mirror descent approach.

Keywords : parameter estimation, moment-matching, constant step-size SGD, conditional exponential family, Fenchel-Young loss, mirror descent.

Résumé

Dans cette thèse, nous examinons plusieurs aspects de l'estimation des paramètres pour les statistiques et les techniques d'apprentissage automatique, aussi que les méthodes d'optimisation applicables à ces problèmes. Le but de l'estimation des paramètres est de trouver les paramètres cachés inconnus qui régissent les données, par exemple les paramètres dont la densité de probabilité est inconnue. La construction d'estimateurs par le biais de problèmes d'optimisation n'est qu'une partie du problème, trouver la valeur optimale du paramètre est souvent un problème d'optimisation qui doit être résolu, en utilisant diverses techniques. Ces problèmes d'optimisation sont souvent convexes pour une large classe de problèmes, et nous pouvons exploiter leur structure pour obtenir des taux de convergence rapides.

La première contribution principale de la thèse est de développer des techniques d'appariement de moments pour des problèmes de régression non linéaire multi-index. Nous considérons le problème classique de régression non linéaire, qui est irréalisable dans des dimensions élevées en raison de la malédiction de la dimensionnalité et c'est pourquoi nous supposons que les données sont en fait une fonction non linéaire de plusieurs projections linéaires des données. Nous combinons deux techniques existantes : "average derivative estimator" (ADE) et "Sliced Inverse Regression" (SIR) pour développer la méthode hybride (SADE) sans certains des aspects faibles de ses parents : elle fonctionne à la fois dans des modèles multi-index et avec des hypothèses faibles sur la distribution des données. Nous étendons également cette méthode aux moments d'ordre élevé. Nous fournissons une analyse théorique des estimateurs construits à la fois pour les cas d'échantillons finis et les cas population.

Dans la deuxième contribution principale, nous utilisons un type particulier de calcul de la moyenne pour la descente stochastique du gradient. Nous considérons des modèles linéaires généralisés pour les familles exponentielles conditionnelles (comme la régression logistique), où l'objectif est de trouver la valeur inconnue du paramètre. Les approches classiques, telles que la descente à gradient stochastique (SGD) avec une taille de pas constante, ne convergent que vers un certain voisinage de la valeur optimale du paramètre, même avec le calcul de la moyenne de Rupert-Polyak. Nous proposons le calcul de la moyenne des paramètres de moments, que nous appelons fonctions de prédiction. Dans le cas des modèles à dimensions finies, ce type de calcul de la moyenne peut, de façon surprenante, conduire à une erreur négative, c'est-à-dire que cette approche nous fournit un estimateur meilleur que tout estimateur linéaire ne peut jamais le faire. Pour les modèles à dimensions infinies, notre approche converge vers la prédiction optimale, alors que le calcul de la moyenne des paramètres ne le fait jamais.

La troisième contribution principale de cette thèse porte sur les pertes de Fenchel-Young. Nous considérons des classificateurs linéaires multi-classes avec les pertes d'un certain type, de sorte que leur double conjugué a un produit direct de simplices comme support. La formulation convexe-concave à point-selle correspondante a une forme spéciale avec un terme de matrice bilinéaire, et les approches classiques souffrent de la multiplication des matrices qui prend beaucoup de temps. Nous montrons que pour les pertes SVM multi-classes, sous hypothèse de régularité légère et avec des techniques d'échantillonnage efficaces, notre approche a une complexité d'itération qui est sous-linéaire dans la taille des données. Cela signifie que pour faire une itération, nous n'avons besoin de payer que trois fois : $O(n + d + k)$ pour le nombre de classes k, le nombre de caractéristiques d et le nombre d'échantillons n, alors que toutes les techniques existantes utilisent au moins une des opérations arithmétiques nd, nk ou dk par itération. Ceci est possible grâce au bon choix des géométries et à l'utilisation d'une approche de descente en miroir.

Mots Clés : estimation des paramètres, méthode des moments, SGD à pas constant, famille exponentielle conditionnelle, fonction objectif du Fenchel-Young, descente en miroir.

Acknowledgements

First of all, I would like to thank my supervisor Francis Bach. His guidance and ability to explain difficult concepts in easy words is the quality, which helped me through all my Ph.D. His politics of open door, when anyone can come and answer questions when his door is open was a great support. Also, I would like to thank my co-advisor Anatoli Juditsky, with whom I had several very fruitful discussions and who instilled in me statistical thinking.

This thesis has received funding from the European Union's H2020 Framework Programme (H2020-MSCA-ITN-2014) under grant agreement n642685 MacSeNet. I would like to thank Helen Cooper for all the training, summer schools and workshops she helped to organize as a part of this framework. I am grateful to Pierre Vandergheynst, with whom I had an opportunity to work during my academic internship in Lausanne, EPFL. I am also grateful to Dave Betts, who was my mentor to the audio restoration field during my industrial internship in Cambridge.

I also would like to thank Arnak Dalalyan and Stephane Chretien for accepting to review my thesis. I would also like to express my sincere gratitude to Olivier Cappé and Franck Iutzeler for accepting to be part of the jury for my thesis.

I would like to thank all members of Willow and Sierra team, old and new ones, for all the group meetings, thanks to which I expanded my scientific outlook. Especially I want to say thank you to my friend and colleague Dmitrii Ostrovskii, with whom I was working on the 3rd project, and who was able to explain me a large amount of information in a short time.

Also, I am grateful to all my friends with whom I spent a lot of joyful hours and days and who share my hobbies. I would like to thank Andrey, Dasha, Kolya, Tolik, Anya and Katya, my teammates in the intellectual game "What? Where? When?". I would like to thank Marwa, who introduced me to the world of climbing. Thank you, Vitalic for our almost daily talks. Also, I say thank my friends Denis, Dima, and Petr for their support.

I would like to thank my parents who opened for me a wonderful world of mathematics and their faith in me. I say thank my sister Tanka for all her sister care and readiness to come to the rescue at any moment.

Last but not least, I am very grateful to my girlfriend Tatiana Shpakova for her love and who was always there in easy and difficult times.

Contents

1 Introduction 3

1.1 Principles for parameter estimation . . . 3

1.1.1 Moment matching . . . 3

1.1.2 Maximum likelihood . . . 6

1.1.3 Loss functions . . . 8

1.2 Convex optimization . . . 11

1.2.1 Euclidean geometry . . . 12

1.2.2 Non-Euclidean geometry . . . 14

1.2.3 Stochastic setup . . . 16

1.3 Saddle point optimization . . . 17

2 Sliced inverse regression with score functions 21

2.1 Introduction . . . 21

2.2 Estimation with infinite sample size . . . 25

2.2.1 SADE: Sliced average derivative estimation . . . 25

2.2.2 SPHD: Sliced principal Hessian directions . . . 28

2.2.3 Relationship between first and second order methods . . . 30

2.3 Estimation from finite sample . . . 30

2.3.1 Estimator and algorithm for SADE . . . 31

2.3.2 Estimator and algorithm for SPHD . . . 33

2.3.3 Consistency for the SADE estimator and algorithm . . . 34

2.4 Learning score functions . . . 35

2.4.1 Score matching to estimate score from data . . . 35

2.4.2 Score matching for sliced inverse regression: two-step approach 37

2.4.3 Score matching for SIR: direct approach . . . 40

2.5 Experiments . . . 41

2.5.1 Known score functions . . . 41

2.5.2 Unknown score functions . . . 44

2.6 Conclusion . . . 46

2.7 Appendix. Proofs . . . 47

2.7.1 Probabilistic lemma . . . 47

2.7.2 Proof of theorem 2.1 . . . 47

2.7.3 Proof of Theorem 2.2 . . . 52

3 Constant step-size SGD for probabilistic modeling 57
3.1 Introduction . . . 57
3.2 Constant step size stochastic gradient descent . . . 59
3.3 Warm-up: exponential families . . . 60
3.4 Conditional exponential families . . . 61
3.4.1 From estimators to prediction functions . . . 62
3.4.2 Averaging predictions . . . 63
3.4.3 Two types of averaging . . . 64
3.5 Finite-dimensional models . . . 66
3.5.1 Earlier work . . . 67
3.5.2 Averaging predictions . . . 67
3.6 Infinite-dimensional models . . . 68
3.7 Experiments . . . 70
3.7.1 Synthetic data . . . 70
3.7.2 Real data . . . 72
3.8 Conclusion . . . 73
3.9 Appendix. Explicit form of $B, \bar{B}, \bar{\bar{\bar{B}}}^w$ and $\bar{\bar{\bar{B}}}^m$ . . . 75
3.9.1 Estimation without averaging . . . 76
3.9.2 Estimation with averaging parameters . . . 76
3.9.3 Estimation with averaging predictions . . . 76

4 Sublinear Primal-Dual Algorithms for Large-Scale Classification 79
4.1 Introduction . . . 79
4.1.1 Related work . . . 83
4.2 Choice of geometry and basic routines . . . 83
4.2.1 Basic schemes: Mirror Descent and Mirror Prox . . . 85
4.2.2 Choice of the partial potentials . . . 88
4.2.3 Recap of the deterministic algorithms . . . 90
4.2.4 Accuracy bounds for the deterministic algorithms . . . 92
4.3 Sampling schemes . . . 94
4.3.1 Partial sampling . . . 95
4.3.2 Full sampling . . . 98
4.3.3 Efficient implementation of SVM with Full Sampling . . . 101
4.4 Discussion of Alternative Geometries . . . 102
4.5 Experiments . . . 107
4.6 Conclusion and perspectives . . . 109
4.7 Appendix . . . 110
4.7.1 Motivation for the multiclass hinge loss . . . 110
4.7.2 General accuracy bounds for the composite saddle-point MD . 110
4.7.3 Auxiliary lemmas . . . 116
4.7.4 Proof of Proposition 4.1 . . . 117
4.7.5 Proof of Proposition 4.2 . . . 118
4.7.6 Proof of Proposition 4.3 . . . 120
4.7.7 Correctness of subroutines in Algorithm 1 . . . 121
4.7.8 Additional remarks on Algorithm 1 . . . 125

5 Conclusion and Future Work **127**

Contributions and thesis outline

In Chapter 1 we give a brief overview of parameter estimation, concerning the method of moments, maximum likelihood and loss minimization problems.

Chapter 2 is dedicated to a special application of method of moments for non-linear regression. This chapter is based on the journal article: *Slice inverse regression with score functions*, D. Babichev, F. Bach, In Electronic Journal of Statistics [Babichev and Bach, 2018b]. The main contributions of this chapter are as follows:

- We propose score function extensions to sliced inverse regression problems, both for the first-order and second-order score functions.
- We consider the infinite sample case and show that in the population case our estimators are superior to the non-sliced versions.
- We consider also finite sample case and show their consistency given the exact score functions. We provide non-asymptotical bounds, given sub-Gaussian assumptions.
- We propose to learn the score function as well, in two steps, i.e., first learning the score function and then learning the effective dimension reduction space, or directly, by solving a convex optimization problem regularized by the nuclear norm.
- We illustrate our results on a series of experiments.

In Chapter 3 we consider special type of averaging for Stochastic SGD applied to generalized linear models. This chapter is based on the conference paper published as an UAI 2018 paper, which was accepted as an oral presentation: *Constant step size stochastic gradient descent for probabilistic modeling*, D. Babichev, F.Bach, Proceedings in Uncertainty in Artificial Intelligence [Babichev and Bach, 2018a]. The main contributions of this chapter are:

- For generalized linear models, we propose averaging moment parameters instead of natural parameters for constant step size stochastic gradient descent.
- For finite-dimensional models, we show that this can sometimes (and suprisingly) lead to better predictions than the best linear model.
- For infinite-dimensional models, we show that it always converges to optimal predictions, while averaging natural parameter never does.
- We illustrate our finding with simulations on synthetic data and classical

benchmarks with many observations.

In Chapter 4 we develop sublinear method for Fenchel-Young losses, this is joint work with Dmitrii Ostrovskii. We have submitted this work to ICML 2019 under a title *Sublinear-time training of mlticlass classifiers with Fenchel-Young losses*, Dmitry Babichev, Dmitrii Ostrovskii and Francis Bach. The main contributions of this chapter are:

- We develop efficient algorithms for solving the regularized empirical risk minimization problems via associated saddle-point problem and using : (i) sampling for computationally heavy matrix multiplication and (ii) right choice of geometry for mirror descent type algorithms.
- The less aggressive *partial sampling scheme* is applicable for any loss minimization problem, such that its dual conjugate has a direct product of simlices as a support. This leads to the cost $O(n(d+k))$ of one iteration.
- The more aggressive *full sampling scheme*, applied to multiclass hinge loss leads to the sublinear cost $O(d+n+k)$ of one iteration.
- We conclude the chapter with numerical experiments.

Chapter 1

Introduction

In this chapter we give the brief overview of two main pillars of this thesis: parameter estimation, and optimization (mostly convex minimization and convex-concave saddle point problems).

1.1 Principles for parameter estimation

Parameter estimation is a branch of statistics and machine learning, that solves problems of estimation of an unknown set of parameters given some observations. There are a big variety of different methods and models and we consider the three probably most famous of them: moment matching, maximum likelihood and risk minimization. The main application of parameter estimation is density and conditional density estimation (and more generally model estimation), which in turn are used in regression, classification and clustering problems.

1.1.1 Moment matching

Moment matching is a technique for finding the values of parameters, using several moments of the distribution, i.e., expectations of powers of random variable. The method of moments was introduced at least by Chebyshev and Pearson in the late 1800s (see for example [Casella and Berger, 2002] for a discussion). Suppose, that we have a real valued random variable X drawn from a family of distributions $\{f(\cdot\,|\,\theta)\mid\theta\in\Theta\}$.

Given a sample $(x_1,\ldots,x_n)$, the goal is to estimate the true value of the parameter θ^*. The classical approach is to consider the first k moments: $\mu_i=\mathbb{E}[X^i]=g_i(\theta)$, $i=1,\ldots,k$, and solve the non-linear system of equations to find the estimator $\hat{\theta}$:

$$\begin{cases}\hat{\mu}_1=\frac{1}{n}\sum_{i=1}^{n}x_i=g_1(\hat{\theta}),\\ \qquad\qquad\vdots\\ \hat{\mu}_k=\frac{1}{n}\sum_{i=1}^{n}x_i^k=g_k(\hat{\theta}).\end{cases}$$

Even though the method is called moment matching, it is non necessary to use moments of random variable, but in general it can use any functions $h_i(X)$ as long as the expectations $\int h_i(x) f(x \,|\, \theta) dx$ can be easily computed. This approach can be extended for the case of random vectors and cross-moments [Hansen, 1982]. Now we discuss some good and bad points of this approach. Also we illustrate the method on several examples, starting from the toy ones and finishing with state-of-the-art methods applicable to non-linear regression.

Disadvantages and advantages

The main advantage of this approach is that it is quite simple and the estimators are consistent under mild assumptions [Hansen, 1982]. Also in some cases the solutions can be found in closed form, where maximum likelihood approach may require a large computational effort.

However in some sense, this approach is inferior to the maximum likelihood approach and estimators are often biased. In some cases, especially for small samples, the results can be outside of the parameter space. Also the nonlinear set of equations may be hard to solve.

Uniform distribution example

Let us start with a simple example, where we need to estimate the parameters of the one-dimensional uniform distribution: $X \sim U[a, b]$. The first and the second moments can be evaluated as $\mu_1 = \mathbb{E}X = \frac{1}{2}(a+b)$ and $\mu_2 = \mathbb{E}X^2 = \frac{1}{3}(a^2 + ab + b^2)$. Solving the system of these two equations, given a sample $(x_1, \dots, x_n)$ and using the sample moments $\hat{\mu}_1$ and $\hat{\mu}_2$ instead of true ones, we get a formula for estimating parameters a and b:

$$(a, b) = (\hat{\mu}_1 - \sqrt{3(\hat{\mu}_2 - \hat{\mu}_1^2)}, \hat{\mu}_1 + \sqrt{3(\hat{\mu}_2 - \hat{\mu}_1^2)}).$$

Linear regression example

Consider now a simple linear regression model, where $y = x^\top b + \varepsilon$, where $y \in \mathbb{R}$, vectors $x, b \in \mathbb{R}^d$ and error ε has a zero expectation. The goal is to estimate the unknown vector of parameters b. Consider the cross moment $\mathbb{E}(x^i y) = \mathbb{E}(x^i x^\top b)$, for $i = 1, \dots, d$, where x^i is the i-th component of x. Replacing the expectations with empirical ones for the sample (x_i, y_i), we get the equation:

$$\frac{y_1 x_1 + \dots + y_n x_n}{n} = \frac{x_1 x_1^\top + \dots + x_n x_n^\top}{n} \cdot \hat{b},$$

which is the traditional normal equation [Goldberger, 1964]. Finally, arranging the vectors in the matrix $X \in \mathbb{R}^{n \times d}$ and $y \in \mathbb{R}^n$, we recover the $\hat{b} = (X^T X)^{-1} X^T y$, Hence in this particular formulation the moment matching estimator coincides with the ordinary least squares estimator.

Exponential families

Note, that for exponential families with probability density given by $f(x|\theta) = h(x)\exp(\theta^\top T(x) - A(\theta)$, moment matching is equivalent to maximum likelihood estimation [Lehmann and Casella, 2006]. We discuss this in more details in the next section.

Score functions

Consider the general non-linear regression problem

$$y = f(x) + \varepsilon, \quad x \in \mathbb{R}^d, y \in \mathbb{R}, \quad \text{error } \varepsilon \text{ independent of the data and } \mathbb{E}\varepsilon = 0.$$

The ambitious goal is to estimate the unknown function f, given samples (x_i, y_i). However it is impossible to solve this problem in this loose formulation, due to the *curse of dimensionality* and non-parametric regression setup. Indeed, classical non-parametric estimation results show that convergence rates with any relevant performance measure can decrease as $n^{-C/d}$ [Tsybakov, 2009, Györfi et al., 2002]. This means that the number of sample points n to reach some level of precision is exponential in the dimension d. A classical way to circumvent the curse of dimensionality is to impose an additional condition: the dependence on some hidden lower dimension of data. Let us start with a simple assumption:

— x is normal and $f(x) = g(w^\top x)$, with a matrix $w \in \mathbb{R}^{d\times k}$.

If $k = 1$, this model is called a *single-index* model and a *multi-index* in the other case [Horowitz, 2012]. In this formulation, the goal is to estimate the unknown matrix of parameters $w \in \mathbb{R}^{k\times d}$ and we can use moment matching techniques. We use again cross-moments and it is not difficult to show that for a single-index, using Stein's lemma, that $\mathbb{E}(yx) \sim w_1$ [Stein, 1981, Brillinger, 1982]. Indeed, using independence of noise, the Gaussian probability density $p(x) \sim \exp(-x^2/2)$ and integration by parts:

$$\mathbb{E}(yx) = \mathbb{E}\big((f(x)+\varepsilon)x\big) = \mathbb{E}\big(g(w_1^\top x)x\big) = \int g(w_1^\top x)p(x)x dx =$$

$$= \int g(w_1^\top x)\nabla p(x)dx = \int \nabla g(w_1^\top x)p(x)dx \sim w_1.$$

The straightforward extension of this approach uses the notion of score function:

$$\mathcal{S}_1(x) = -\nabla \log p(x),$$

where $p(x)$ is the probability density of data x. We use a moment matching technique in the form $\mathbb{E}(\mathcal{S}_1(x)y) \sim w_1$ as it is done in [Stoker, 1986]. The most recent approach for multi-index models by [Janzamin et al., 2014] and [Janzamin et al., 2015] uses the notion of high-order scores $\mathcal{S}_m(x) = (-1)^m \frac{\nabla^{(m)}p(x)}{p(x)}$ (which are tensors) and cross moments $\mathbb{E}[y \cdot \mathcal{S}_m(x)]$ to train neural networks.

Contribution of this thesis

One more extension of the method of moments uses conditional moments $\mathbb{E}(x|y)$, known as Sliced Inverse Regression (SIR, [Li, 1991]) and second order moments $\mathbb{E}(xx^\top|y)$ (Principal Hessian Directions PHD, [Li, 1992]) which use a normal distribution assumption. We develop a new method, combining strong sides of Stein's lemma and SIR in Chapter 2 of this thesis. We proposed new approaches (SADE and SPHD) and develop analysis for both population and sample cases.

1.1.2 Maximum likelihood

Maximum likelihood estimation or MLE is an another classical approach to estimate the unknown parameters, maximizing the likelihood function: it means intuitively, that the selected parameter makes the data most probable. More formally, let $X = (x_1, \dots, x_n)$ be a random sample from a family of distributions $\{f(\cdot \,|\, \theta) \mid \theta \in \Theta\}$, then

$$\hat{\theta} \in \arg\max_{\theta \in \Theta} \mathcal{L}(\theta\,; X),$$

where $\mathcal{L}(\theta\,; X) = f_X(X \,|\, \theta)$ is the so-called *Likelihood function*: that is, the joint probability density for the given realization of sample.

In practice, it is often convenient to work with the negative natural logarithm of the likelihood function, called the *negative log-likelihood*: $l(\theta\,; X) = -\ln \mathcal{L}(\theta\,; X)$. If the data $(x_1, \dots, x_n)$ are independent and identically distributed, then the joint distribution density can be written as a product of densities for a single x_i and the average negative log-likelihood minimization problem takes the form:

$$\arg\min_{\theta \in \Theta} \hat{l}(\theta\,; X) = \arg\min_{\theta \in \Theta} -\frac{1}{n} \sum_{i=1}^{n} \ln f(x_i \mid \theta),$$

where $f(x_i|\theta)$ is the value of the probability density at the point x_i.

Advantages and disadvantages

Under mild assumptions the maximum likelihood estimator is consistent, asymptotically efficient and asymptotically normal [LeCam, 1953, Akaike, 1998]. Even though for a simple model, solutions can be found in closed form, for more advanced models, methods of optimization must be used to get the solution. Hopefully the optimization problem is convex for a wide class of likelihood estimators, such as exponential families and conditional exponential families [Koller and Friedman, 2009, Murphy, 2012]. On the other hand, the optimization problem could be not convex if we consider for example mixture models. Moreover, maximum likelihood estimators are robust to mis-specified data: if the real data distribution f^* does not come from the model $\{f(\cdot \,|\, \theta) \mid \theta \in \Theta\}$, we can still use the approach and the solution with infinity data will be the projection (in the Kullback-Leibler sense) of f^* to the set of allowed distributions.

Example: Exponential families

The probability density for an exponential family can be written in the following form:

$$f(x|\theta) = h(x)\exp(\theta^\top T(x) - A(\theta)),$$

where $h(x)$ is the base measure, $T(x) \in \mathbb{R}^d$ is the sufficient statistics and A the log-partition function, which is always convex. Note that we do not assume that the data distribution $p(x)$ comes from this exponential family. Then the average negative log-likelihood is equal to

$$\hat{\ell}(\theta; x) = \frac{1}{n}\sum_{i=1}^{n}\Big[A(\theta) - \theta^\top T(x_i)\Big],$$

which is a convex problem and can be solved, using any convex minimization approach. Another view to this problem is moment matching: the solution can be found as:

$$A'(\theta) = \frac{1}{n}\sum_{i=1}^{n} T(x_i),$$

where we consider the moment of the sufficient statistics $T(x)$. Note, that in a fact a big variety of classical distributions can be represented in this form: Bernoulli, normal, Poisson, exponential, Gamma, Beta, Dirichlet and many others. However, also, there are few which are not for example Student's distribution and mixtures of classical distribution are not in this family.

Example: Conditional Exponential families

Now, let us consider the classical conditional exponential families:

$$f(y|x,\theta) = h(y)\cdot\exp(y\cdot\eta_\theta(x) - a(\eta_\theta(x))).$$

Again, writing down the average negative log-likelihood, the goal is to minimize

$$\frac{1}{n}\sum_{i=1}^{n}\Big[-y_i\cdot\eta_\theta(x_i) + a(\eta_\theta(x_i))\Big].$$

These families are used for regression and classification problems and closely related to the *generalized linear models* [McCullagh and Nelder, 1989], if the natural parameter is linear combination in known basis: $\eta_\theta(x) = \theta^\top\Phi(x)$. On of the most popular choices of function $a'(\cdot)$ is the sigmoid function $a'(t) = \sigma(t) = 1/(1+e^{-t})$, which leads to the *logistic regression* model. Let us illustrate the connection of logistic regression with a Bernoulli distribution. Let p be a parameter of Bernoulli distribution with $x \in \{0,1\}$, then the probability density is written as:

$$q(x|p) = x\log p - (1-x)\log(1-p) = x\log\frac{p}{1-p} + \log(1-p)$$

$$= \exp(x \cdot \theta - \log(1 + e^{\theta})),$$

where $\theta = \log \frac{p}{1-p}$ and hence logistic model is unconstrained reformulation of Bernoulli model, and the lack of constraint is often seen as a benefit for optimization.

Another common choice is Poisson regression [McCullagh, 1984], where $a(t) = \exp(t)$, which is used, when the dependent variable is a count.

Contribution of this thesis

In Chapter 3 of this thesis we consider constant step-size stochastic gradient descent for conditional exponential families. Instead of averaging of parameters θ, we propose averaging of so-called prediction functions $\mu = a'(\theta^\top \Phi(x))$, which leads to better convergence to the optimal prediction, especially for infinite-dimensional models.

1.1.3 Loss functions

One more view to the parameter estimation problem is risk minimization, which goes beyond the maximum likelihood approach. A non-negative real-valued loss function $\ell(y, \hat{y})$ measures the performance for classification or regression problem: i.e., the difference between prediction $\hat{y}$ of a and the true outcome y. The expected loss $R(\theta) = \mathbb{E}[\ell(f(x|\theta), y)]$ is called the risk. However, in practice the true distribution $P(x, y)$ is inaccessible and the empirical risk used instead. Hence, the classical regularized *empirical risk minimization* problem is written as:

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \ell\big(f(x_i|\theta), y_i\big) + \lambda\, \Omega(\theta), \tag{1.1.1}$$

where $\Omega(\theta)$ is a regularizer, which is used to avoid overfitting. There are several classical choices of loss function, and we start with in some sense the most intuitive ones:

$0 - 1$ loss

This loss, which is also called misclassification loss is used in classification problems, where the response y is located in a finite set. The definition speaks for itself: the loss is zero, in the case of true classification and one in the other case:

$$\ell(f(x_i|\theta), y_i) = 1 \iff f(x_i|\theta) \neq y_i.$$

It is known to lead to NP-hard problems, even for linear classifiers [Feldman et al., 2012, Ben-David et al., 2003]. That is why in practice people use a convex relaxation of the $0 - 1$ loss functions (see below).

Quadratic loss

One of the other classical choices for loss function in the case of regression problems is the quadratic loss:

$$\ell(f(x_i|\theta), y_i) = (f(x_i|\theta) - y_i)^2 .$$

It has a simple form, smooth (in contradiction to the ℓ_1 loss) and convex. Thus, the empirical risk minimization problem becomes *mean squared error* minimization.

The connection with likelihood estimators can be shown, using the linear model with Gaussian noise:

$$y = x^\top \theta + \varepsilon, \text{ where } \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

Then, the negative log-likelihood is

$$l(y; x, \theta) \sim (x^\top \theta - y)^2,$$

and the empirical likelihood minimization problem for this problem is equivalent to the mean squared error minimization (but the application of least squares is not limited to Gaussian noise).

Let us illustrate also a connection of quadratic loss with regression: consider that we are looking for a function f, such that $y = f(x) + \varepsilon$, where $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$. Then, the expected loss (risk) can be written as

$$R(f) = \int \ell(f(x), y) p(x, y) \; dx \; dy.$$

Solving this functional minimization problem for the quadratic loss, we get the solution $f(x) = \mathbb{E}_y[y|x]$, which is the conditional expectation of y given x and is known as the *regression function.*

Convex surrogates

Here we consider the main examples for convex surrogates of the $0-1$ loss, which make the computation more tractable.

Hinge loss. It is written as

$$\ell(f(x_i|\theta), y_i) = \max(0, 1 - f(x_i|\theta) y_i),$$

and is used in soft-margin support vector machines (SVM) approach introduced in its modern form by [Cortes and Vapnik, 1995]. The method tries to find a hyperplane which separates the data. In practice, usually regularized problem is considered due to non-robustness, especially for separable data:

$$\left[\frac{1}{n}\sum_{i=1}^{n} \max\left(0, 1 - y_i \theta x_i\right)\right] + \lambda \|\theta\|^2.$$

The classical way of solving the minimization problem is to switch to quadratic programming problem (originally [Cortes and Vapnik, 1995], see also [Bishop, 2006]). However, now the most recent approaches are used, such as gradient descent and stochastic gradient descent types [Shalev-Shwartz et al., 2011].

Logistic loss and Exponential loss. The logistic loss is defined as $\ell(f(x_i|\theta), y_i) = \log(1 + \exp(-y \cdot f(x_i|\theta)))$ and the Exponential loss is defined as $\ell(f(x_i|\theta), y_i) = \exp(-y \cdot f(x_i|\theta))$. In fact, these two losses are dictated by maximum log-likelihood formulations: the logistic loss takes its origin in logistic regression and exponential loss is used in Poisson regression. Moreover, we can say that every negative log-likelihood minimization problem is equivalent to loss minimization problem, if we introduce the corresponding log-likelihood loss. The opposite is typically not true (for example for the hinge loss).

Graphical representation

We summarize the discussed losses in the Figure 1-1, where we renormalize some of them to pass through the point $(1, 0)$. The dashed line represents the regression formulation and solid ones are for classification problems.

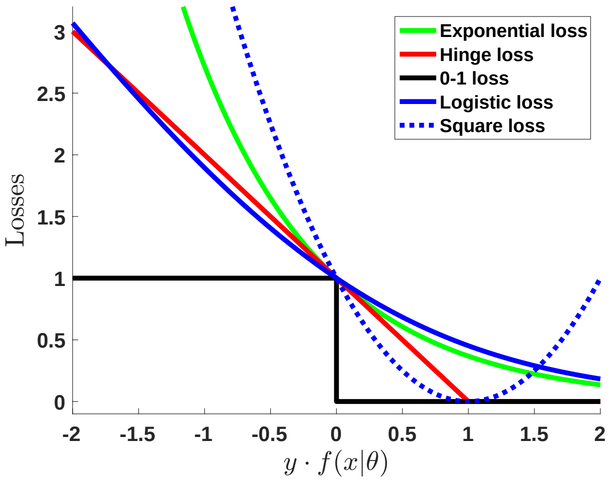

Figure 1-1 – Graphical representation of classical loss functions.

Contribution of this thesis

In Chapter 4 of this thesis we consider so-called *Fenchel-Young* losses for multiclass linear classifiers, which extend convex surrogates to the multiclass setting. This leads to saddle-point convex-concave problems with expensive matrix multiplications. Using stochastic optimization methods for non-Euclidean setup (using mirror descent) and specific variance reduction techniques we are able to reach sublinear (in the natural dimensionality of the problem) running time complexity per iteration.

1.2 Convex optimization

In this section we discuss the basics of convex optimization. Convexity is a prevailing setup for optimization problems, due to the theoretical guarantees for this class of functions. The classical convex minimization problem is the following:

$$\min_{x \in \mathcal{X}} f(x),$$

where $\mathcal{X} \in \mathbb{R}^d$ is a convex set: for any two points $x, y \in \mathcal{X}$ and for every $\alpha \in [0, 1]$ the point $\alpha x + (1 - \alpha)y$ is also in $\mathcal{X}$. This means, that for any two points inside the set, the whole segment, connecting these points is also in the set. The function f is convex as well: for any $x, y \in \mathcal{X}$ and $\alpha \in [0, 1]$: $f(\alpha x + (1 - \alpha)y) \leqslant \alpha f(x) + (1 - \alpha)f(y)$ and this means, that the epigraph of function f is a convex set: every chord lies above the graph of a function. It is known for convex minimization problems, that:

— If a local minimum exists, it is also a global minimum.

— The set of all global minima is convex (note, that global minimum is not necessarily unique).

Consider the unconstrained minimization problem, where $\mathcal{X} = \mathbb{R}^d$. The first property provides us with the criterion of global minimum: if the function f is differentiable, then

$$\nabla f(x^*) = 0 \quad \Leftrightarrow \quad x^* \text{ is global minimum.}$$

If the function f is not differentiable, we still can use the criterion, where the gradient of the function is replaced by a *sub-gradient*: a generalization of gradient for convex function which defines the cone of directions in which the function increases [Rockafellar, 2015]:

$$\partial f(x^*) \ni 0 \quad \Leftrightarrow \quad x^* \text{ is global minimum.}$$

However in practice, only for simple problems the solution can be found in closed form. For the majority of convex problems computational iterative methods are used, and the starting point for them are *gradient descent* or the steepest descent

$$x^{t+1} = x^t - \gamma_t \nabla f(x^t).$$

The idea is straightforward: we are looking for the direction in which the function increases and then descend in the opposite direction. Classical choices for stepsize are constant and decaying: $\gamma_t = Ct^{-\alpha}$, where $\alpha \in [0, 1]$.

We also consider classical assumptions, such that bounded gradients, smoothness (which requires differentiability) and strong convexity (for more details see [Nesterov, 2013, Bubeck, 2015]):

The first definition is the weakest one: bounded gradients, which does not require differentiability:

Definition 1.1. *The function f has bounded gradients (subgradients), if for any*

$x \in \mathcal{X}$ and for any $g(x) \in \partial f(x)$:

$$\|g(x)\| \leqslant B.$$

Definition 1.2. *The function f is called* smooth *with Lipschitz constant L if, for all $x, y \in \mathcal{X}$:*

$$f(y) \leqslant f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|^2,$$

which is equivalent to, when f is convex,

$$\|\nabla f(x) - \nabla f(y)\| \leqslant L\|x - y\|.$$

Definition 1.3. *The function f is called* strongly convex *with constant μ if, for all $x, y \in \mathcal{X}$:*

$$f(y) \geqslant f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|y - x\|^2.$$

The intuitive definition of Lipschitz smoothness is that at every point, the function f can be bounded above by quadratic function with coefficient $L/2$, strong convexity means, that at every point the function f is bounded below by quadratic function with coefficient $\mu/2$. A graphical representation of one dimensional case can be found in Figure 1-2.

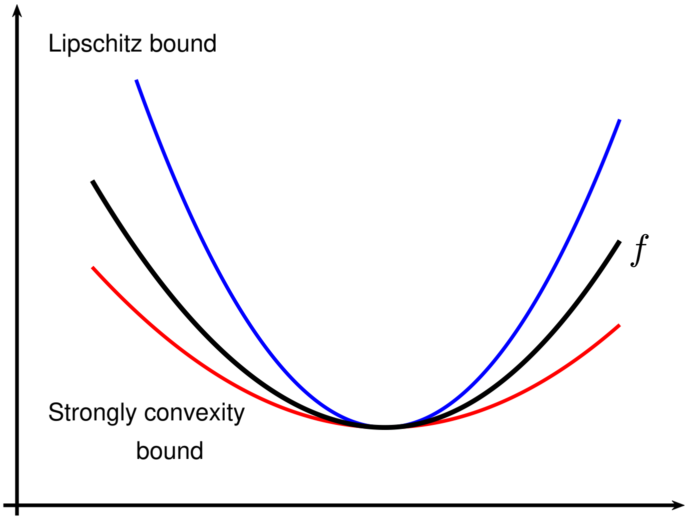

Figure 1-2 – Graphical representation of Lipschitz smoothness and strong convexity.

1.2.1 Euclidean geometry

Note, that we did not define norms for smoothness and strong convexity. If we define them as standard Euclidean norms, we reproduce the so-called *Euclidean geometry* and constants L and μ correspond to the biggest and the smallest eigenvalues of Hessians, if the function is twice differentiable.

Projected gradient descent

The gradient descent method for constrained problem is called *projected gradient descent* (see [Bertsekas, 1999] and references therein) and one iteration is the following:

$$x^{t+1} = \Pi_{\mathcal{X}}\big(x^t - \gamma_t g(x^t)\big), \quad \text{with} \quad g(x^t) \in \partial f(x^t)$$

where $\Pi_{\mathcal{X}}(x) = \arg\min_{y \in \mathcal{X}} \|x - y\|$ is the Euclidean projection of the point x to the set $\mathcal{X}$. We illustrate this approach in Figure 1-3. We can also to look at this equation through the prism of proximal approaches [Moreau, 1965, Rockafellar, 1976] as done for example in [Beck and Teboulle, 2003]:

$$x^{t+1} = \arg\min_{x \in \mathcal{X}} \left\{ \langle x, g(x^t) \rangle + \frac{1}{\gamma_t} \|x - x^t\|^2 \right\}, \quad \text{where} \quad g(x^t) \in \partial f(x^t).$$

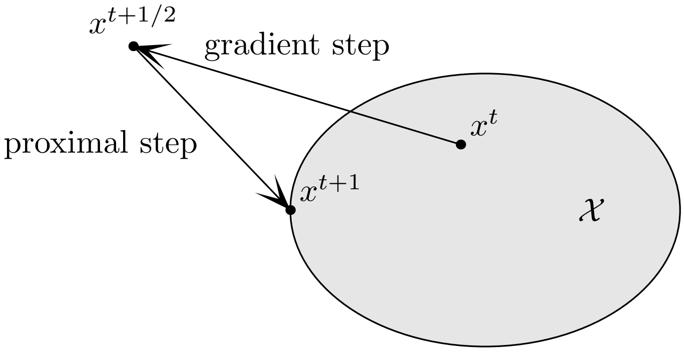

Figure 1-3 – Graphical representation of projected gradient descent: firstly we do gradient step and then project onto $\mathcal{X}$.

Note, that the proximal operator should be simple, i.e., the structure of the set $\mathcal{X}$ be simple enough to compute it in either in closed form or with a number of iterations commensurate with the computation of the gradient.

Now we formulate main results for functions with different assumpions (short proofs can be found for example in [Bubeck, 2015]):

Theorem 1.1. *Let f be convex with bounded gradients with constant B; the radius of set $\mathcal{X}$ is R, i.e., $\sup_{x,y \in \mathcal{X}} \|x - y\| = 2R$. Then the projected gradient descent with decaying stepsize $\gamma_t = \frac{R}{B\sqrt{t}}$ satisfies:*

$$f(\bar{x}^t) - f(x^*) \leqslant \frac{RB}{\sqrt{t}}, \ \textit{where } \bar{x}^t = \frac{1}{t}\sum_{i=1}^{t} x^i.$$

Hence in case of bounded gradients we get an $1/\sqrt{t}$ convergence rate.

Theorem 1.2. *Let f be convex and L-smooth on $\mathcal{X}$. Then the projected gradient descent with constant stepsize $\gamma_t = \frac{1}{L}$ satisfies:*

$$f(x^t) - f(x^*) \leqslant \frac{4L\|x^1 - x^*\|^2}{t}.$$

Hence in case of L-smooth function we get an $1/t$ convergence rate. If we add an assumption about strong convexity of the function f, we recover exponential rate:

Theorem 1.3. *Let f be μ strongly convex and L-smooth on $\mathcal{X}$. Then the projected gradient descent with constant stepsize $\gamma_t = \frac{1}{L}$ satisfies:*

$$\|x^{t+1} - x^*\|^2 \leqslant \exp\left(-t \cdot \frac{\mu}{L}\right) \|x^1 - x^*\|^2.$$

Note also that acceleration techniques, proposed by Nesterov can be used to increase the convergence rates. The first work in this direction [Nesterov, 1983] was proposed for smooth functions and unconstrained setup. It improved the rates from $1/t$ to $1/t^2$. The case of non-smooth function was considered in [Nesterov, 2005] and improved the rates from $1/\sqrt{t}$ to $1/t$, using a special smoothing technique, which can be applied to functions with explicit max-structure. Constrained problems were considered in [Nesterov, 2007] and [Beck and Teboulle, 2009] with the same convergence rates.

Note, that Theorems 1.1, 1.2 and 1.3 use Euclidean geometries: definitions of smoothness and strongly convexity are given with respect with usual Euclidean geometry. In the next section we extend these result for a more general setup.

1.2.2 Non-Euclidean geometry

The ideas of non-Euclidean approach are to exploit the geometry of the set $\mathcal{X}$ and achieve the rates of projected gradient descent with constants adapting to the geometry of the set $\mathcal{X}$. Let us fix an arbitrary norm $\|\cdot\|_{\mathscr{X}}$ and a compact set $\mathcal{X} \in \mathbb{R}^d$. We introduce a definition of the dual norm:

Definition 1.4. *The norm $\|\cdot\|_{\mathscr{X}^*}$ is called the dual norm, if $\|g\|_{\mathscr{X}^*} = \sup\limits_{x \in \mathbb{R}^d : \|x\|_{\mathscr{X}} \leqslant 1} g^\top x$.*

Now we need to adapt definitions of bounded gradients, smoothness and strong convexity: (see for example [Bubeck, 2015, Beck and Teboulle, 2003]) (i) we say, that function $f(x)$ has bounded gradients, if

$$\|g\|_{\mathscr{X}^*} \leqslant B \quad \text{for any } g(x) \in \partial f(x).$$

(ii) we say, that function $f(x)$ is L-smooth with respect to norm $\|\cdot\|_{\mathscr{X}}$ if

$$\|\nabla f(x) \; - \; \nabla f(y)\|_{\mathscr{X}^*} \leqslant \|x - y\|_{\mathscr{X}},$$

where $\|\cdot\|_{\mathscr{X}^*}$ is dual norm, (iii) μ-strongly convex if

$$f(y) \geqslant f(x) + \nabla f(x)^\top (y-x) + \frac{\mu}{2}\|y-x\|^2_{\mathscr{X}}.$$

The idea of the mirror approach is to consider a so-called *potential* function $\Phi(x)$ (which is also called DGF — distance generating function or mirror map), then switch to the *mirror* space (using gradient of potential), do the gradient step in this space, return back to the main space and project the point to the set $\mathcal{X}$.

Mirror descent

More formally, one step of Mirror descent [Nemirovsky and Yudin, 1983] is

$$\nabla \Phi(x^{t+1/2}) = \nabla \Phi(x_t) - \gamma_t \nabla f(x^t),$$

$$x^{t+1} \in \Pi^{\Phi}_{\mathcal{X}}(x^{t+1/2}),$$

where $\Pi^{\Phi}_{\mathcal{X}}(x,y) = \arg\min\limits_{x\in\mathcal{X}} B_\Phi(x,y)$ is the projection in the *Bregman divergence* sense: $B_\Phi(x,y) = \Phi(x) - \Phi(y) - \nabla\Phi(y)^\top(x-y)$. However, this notation could be difficult to embrace and [Beck and Teboulle, 2003] introduced the equivalent definition without switching to the mirror space, and one step of mirror descent can be written as:

$$x^{t+1} = \arg\min_{x\in\mathcal{X}} \left\{ \langle x, \nabla f(x^t)\rangle + \frac{1}{\gamma_t} B_\phi(x, x^t) \right\}. \tag{Mirror descent}$$

Note the similarity with projected gradient descent. In practice, the potential $\Phi(x)$ should be a strictly convex and differentiable function with the following properties:

- $\Phi(x)$ is 1-strongly convex on $\mathcal{X}$ with respect to $\|\cdot\|_{\mathscr{X}}$.
- The effective square radius of set $\mathcal{X}$, which is defined as $\Omega = \sup\limits_{x\in\mathcal{X}} \Phi(x) - \inf\limits_{x\in\mathcal{X}} \Phi(x)$ should be small.
- The proximal step above should be feasible.

Finally, we can formulate convergence rates for the Mirror descent algorithm for non-smooth case (a short proof can be found for example in [Bubeck, 2015]):

Theorem 1.4. *Let Φ be a 1-strongly convex with respect to $\|\cdot\|_{\mathscr{X}}$ and f be convex with B-bounded gradients with respect to $\|\cdot\|_{\mathscr{X}}$. Then mirror descent with $\gamma_t = \frac{\sqrt{2\Omega}}{B\sqrt{t}}$ satisfies:*

$$f(\bar{x}^t) - f(x^*) \leqslant B\sqrt{\frac{2\Omega}{t}}, \qquad \textit{where} \quad \bar{x}^t = \frac{1}{t}\sum_{i=1}^{t} x^i.$$

Note, that there are various extensions of mirror descent, such as mirror prox [Nemirovski, 2004], Dual Averaging [Nesterov, 2007] and extentions to saddle point minimax problems [Juditsky and Nemirovski, 2011a,b, Nesterov and Nemirovski, 2013].

One more direction is NoLips approach, where the idea is to get rid of the Lipschitz-continuous gradient, using convexity condition which captures the geometry of the constraints [Bauschke et al., 2016].

Examples of geometries

ℓ_2 geometry. This is the simplest geometry, which actually corresponds to the Euclidean case. Indeed, in this case $\Phi(x) = \frac{1}{2}\|x\|_2^2$ is 1-strongly convex with respect to the ℓ_2 norm on the whole $\mathbb{R}^d$. The Bregman divergence, associated with this norm is $B_\phi(x,y) = \frac{1}{2}\|x-y\|_2^2$ and one step of mirror descent is reduced to one step of projected gradient descent. Note, that Theorem 1.4 is reduced to Theorem 1.1 in this case.

ℓ_1 geometry. This is more interesting choice of geometry, which is also called *simplex* setup. If we define potential as negative entropy:

$$\Phi(x) = \sum_{i=1}^{d} x_i \log x_i,$$

then, using Pinsker's inequality, we can show, that Φ is 1-strongly convex with respect to ℓ_1 norm on the simplex $\Delta_d = \{x \in \mathbb{R}_+^d : \sum\limits_{i=1}^{d} x_i = 1\}$. The Bregman divergence associated with negative entropy is so-called *he Kullback-Leibler divergence*: $B_\phi(x,y) = D_{\mathrm{KL}}(x\|y) = \sum\limits_{i=1}^{d} x_i \log \frac{x_i}{y_i}$. The effective squared radius of Δ_d is $\Omega = \log d$ and moreover the solution for one step of MD can be found in closed form and lead to so-called multiplicative updates. This implies, that if we minimize function f on the simplex Δ_d, such that $\|\nabla f\|_\infty$ are bounded, the right choice of geometry give rates as $O\left(\frac{\log d}{t}\right)$, whereas the Euclidean geometry give rates only as $O\left(\frac{d}{t}\right)$.

One more classical setup is often used: the ℓ_1 ball can be obtained from the simplex setup, by doubling the number of variables. Instead of d real values (positive or negative), we consider $2d$ positive values and transform the d-dimensional ball to the $2d$-dimensional simplex.

1.2.3 Stochastic setup

In this section we consider stochastic approaches going back to [Robbins and Monro, 1951], where we do not use the gradient $\nabla f(x)$, but evaluate the noisy version of it: namely a stochastic oracle $\widetilde{g}(x)$, such that $\mathbb{E}\widetilde{g}(x) = \nabla f(x)$. The classical setting in machine learning is when the objective function f is the sampled mean of observations f_i (probably with some regularizer term):

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) + R(x),$$

and we can choose oracle as $f_i(x)$, where i is chosen uniformly from n points. In fact, all negative log-likelihood minimization and loss minimization problems have this form. This also applies to the situation of single pass SGD where the bounds are then on the generalization error. To define the quality of an oracle $\widetilde{g}(x)$, we assume the existence of the moment

$$B^2 = \sup_{x \in \mathcal{X}} \mathbb{E}\|\widetilde{g}(x)\|^2_{\mathscr{X}^*},$$

in the non-smooth case, and assume the existence of the variance in the smooth case:

$$\sigma^2 = \sup_{x \in \mathcal{X}} \mathbb{E}\|\widetilde{g}(x) - \nabla f(x)\|^2_{\mathscr{X}^*},$$

where the norm depends on the geometry. Let us consider the general Stochastic Mirror Descent approach:

$$x^{t+1} = \arg\min_{x \in \mathcal{X}} \left\{ \langle x, \widetilde{g}(x^t) \rangle + \frac{1}{\gamma_t} B_\phi(x, x^t) \right\}. \qquad \text{(Stochastic Mirror Descent)}$$

Now we can formulate the theorem: [Juditsky et al., 2011, Lan, 2012, Xiao, 2010]

Theorem 1.5. *Let Φ be a 1-strongly convex with respect to $\|\cdot\|_{\mathcal{X}}$ and B^2 be the second moment of an oracle $\widetilde{g}(x)$. Then Stochastic Mirror Descent with stepsize $\gamma_t = \frac{\sqrt{2\Omega}}{B\sqrt{t}}$ satisfies:*

$$\mathbb{E}f(\bar{x}^t) - f(x^*) \leqslant B\sqrt{\frac{2\Omega}{t}}, \qquad \textit{where} \qquad \bar{x}_t = \frac{1}{t}\sum_{i=1}^{t} x_i.$$

Note the similarity of this result with Theorem 1.4: the only difference is that bounded gradients are replaced with the bounded moment of the oracle.

1.3 Saddle point optimization

In this section we consider saddle point optimization. Consider two convex and compact sets $\mathcal{X} \in \mathbb{R}^{d_1}$ and $\mathcal{Y} \in \mathbb{R}^{d_2}$. Let $\phi(x, y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a function, such that $\phi(\cdot, y)$ is convex and $\phi(x, \cdot)$ is concave. The goal is to find

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \phi(x, y).$$

A classical example is obtained from Fenchel duality below. We present here results from [Juditsky and Nemirovski, 2011a,b, Nesterov and Nemirovski, 2013], concerning the mirror descent approach.

Introduce the *(sub)gradient field*: $G(x, y) = \Big(\partial_x \phi(x, y), -\partial_y \phi(x, y)\Big)$ — analogue of (sub)gradients for saddle point problems. Then subgradients are given by

$$\Big(g_{\mathcal{X}}(x, y), g_{\mathcal{Y}}(x, y)\Big) \in G(x, y).$$

The analogue of bounded gradients is written as:

Definition 1.5. *Function $\phi(x, y)$ has bounded gradients, if $\|g_{\mathcal{X}}(x, y)\|_{\mathcal{X}^*} \leqslant \mathcal{L}_{\mathcal{X}}$ and $\|g_{\mathcal{Y}}(x, y)\|_{\mathcal{Y}^*} \leqslant \mathcal{L}_{\mathcal{Y}}$ for any $(x, y) \in (\mathcal{X} \times \mathcal{Y})$.*

To evaluate the quality of the point $(\widetilde{x}, \widetilde{y}) \in (\mathcal{X} \times \mathcal{Y})$, we introduce the notion of the so-called *duality gap*:

Definition 1.6. *The duality gap is $\Delta_{dual}(\widetilde{x}, \widetilde{y}) = \max\limits_{y \in \mathcal{Y}} \phi(\widetilde{x}, y) - \min\limits_{x \in \mathcal{X}} \phi(x, \widetilde{y})$.*

Observe, that the duality gap is the sum of the primal gap $\max\limits_{y \in \mathcal{Y}} \phi(\widetilde{x}, y) - \phi(x^*, y^*)$ and the dual gap $\phi(x^*, y^*) - \min\limits_{x \in \mathcal{X}} \phi(x, \widetilde{y})$. Introduce the variable $z = (x, y)$ and the set $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. The main motivation of the duality gap is that it can be controlled in the following way, similar for the convex optimization:

$$\Delta_{dual}(\widetilde{z}) \leqslant g(\widetilde{z})^\top (\widetilde{z} - z),$$

where $g(\widetilde{z})$ in the gradient field $G(\widetilde{z})$, for more details see Bertsekas [1999].

Recall, that in order to apply mirror descent, we firstly need to construct potential and choose geometries. Let $\Phi_{\mathcal{X}}(x)$ be a potential, defined for variable the x, such that it is 1-strongly convex with respect to a norm $\|\cdot\|_{\mathscr{X}}$ on $\mathcal{X}$. Similarly, let $\Phi_{\mathcal{Y}}(y)$ be a potential, defined for the variable y, such that it is 1-strongly convex with respect to a norm $\|\cdot\|_{\mathscr{Y}}$ on $\mathcal{Y}$. Let $\Omega_{\mathcal{X}}$ and $\Omega_{\mathcal{Y}}$ be the effective square radii of sets $\mathcal{X}$ and $\mathcal{Y}$ with respect to the corresponding norms.

Let us construct the composite potential $\Phi_{\mathcal{Z}}(z) = \frac{\mathcal{L}_{\mathcal{X}}}{\sqrt{\Omega_{\mathcal{X}}}} \Phi_{\mathcal{X}}(x) + \frac{\mathcal{L}_{\mathcal{Y}}}{\sqrt{\Omega_{\mathcal{Y}}}} \Phi_{\mathcal{Y}}(y)$, then one step of Saddle Point Mirror Descent (SP-MD) is the following:

$$z^{t+1} \in \arg\min \left\{ \langle g^t, z \rangle + \frac{1}{\gamma_t} B_{\Phi_{\mathcal{Z}}}(z, z^t) \right\}, \quad g^t \in G(x^t, y^t) \qquad \text{SP-MD}$$

Finally we can formulate the convergence rates:

Theorem 1.6. *Let function $\phi(x, y)$ has a bounded gradients with constants $\mathcal{L}_{\mathcal{X}}$ and $\mathcal{L}_{\mathcal{Y}}$. Then SP-MD with stepsize $\gamma_t = \sqrt{\frac{2}{t}}$ satisfies:*

$$\max_{y \in \mathcal{Y}} \phi(\bar{x}^t, y) - \min_{x \in \mathcal{X}} \phi(x, \bar{y}^t) \leqslant \left(\sqrt{\Omega_{\mathcal{X}}} \mathcal{L}_{\mathcal{X}} + \sqrt{\Omega_{\mathcal{Y}}} \mathcal{L}_{\mathcal{Y}} \right) \sqrt{\frac{2}{t}}.$$

Fenchel duality

We finish this introduction with the classical Fenchel duality result which provides us with the way to switch from convex problems to saddle-point problems. Let us firstly define the notion of the Fenchel conjugate of the convex function f:

Definition 1.7. *The function $f^*(y)$ given by $f^\star(y) := \sup\{\langle y, x\rangle - f(x) | \, x \in \mathbb{R}^n\}$ is called the Fenchel conjugate of the function f.*

Now we can formulate the theorem [see e.g. Borwein and Lewis, 2010]:

Theorem 1.7. (Fenchel's duality theorem) *Let $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{Y} \to \mathbb{R}$ be convex functions and $A : \mathcal{X} \to \mathcal{Y}$ be a linear map. Then*

$$\inf_{x \in \mathcal{X}} \Big\{ f(x) + g(Ax) \Big\} = \sup_{y \in \mathcal{Y}} \Big\{ -f^*(A^\top y) - g^*(-y) \Big\}.$$

Finally, we provide the way to switch between primal, dual and saddle point formulations of the convex problem:

$$\begin{aligned} &\text{Primal problem :} & \inf_x \Big\{ f(x) + g(Ax) \Big\}. \\ &\text{Dual problem :} & \sup_y \Big\{ -f^*(A^\top y) - g^*(-y) \Big\}. \\ &\text{Saddle problem :} & \inf_x \sup_y \Big\{ f(x) - g^*(-y) + y^\top Ax \Big\}. \end{aligned}$$

Machine learning motivation. These optimization problems are motivated by machine learning applications, where x is the parameter to estimate, matrix A the data, g the loss and f the regularizer (note similarity with regularized empirical risk minimization (1.1.1)). Dual or saddle problem formulations help to switch to an equivalent task, which in some sense has a simpler structure, like bilinear saddle-point problem with composite terms and moreover allows to control the duality gap.

Chapter 2

Sliced inverse regression with score functions

Abstract

We consider non-linear regression problems where we assume that the response depends non-linearly on a linear projection of the covariates. We propose score function extensions to sliced inverse regression problems, both for the first- order and second-order score functions. We show that they provably improve estimation in the population case over the non-sliced versions and we study finite sample estimators and their consistency given the exact score functions. We also propose to learn the score function as well, in two steps, i.e., first learning the score function and then learning the effective dimension reduction space, or directly, by solving a convex optimization problem regularized by the nuclear norm. We illustrate our results on a series of experiments.

This chapter is based on the journal article: *Slice inverse regression with score functions*, D. Babichev, F. Bach, In Electronic Journal of Statistics [Babichev and Bach, 2018b].

2.1 Introduction

Non-linear regression and related problems such as non-linear classification are core important tasks in machine learning and statistics. In this chapter, we consider a random vector $x \in \mathbb{R}^d$, a random response $y \in \mathbb{R}$, and a regression model of the form

$$y = f(x) + \varepsilon, \tag{2.1.1}$$

which we want to estimate from n independent and identically distributed (i.i.d.) observations (x_i, y_i), $i = 1, \ldots, n$. Our goal is to estimate the function f from these data. A traditional key difficulty in this general regression problem is the lack of parametric assumptions regarding the functional form of f, leading to a problem of *non-parametric* regression. This is often tackled by searching implicitly or explicitly a function f within an infinite-dimensional vector space.

While several techniques exist to estimate such a function, e.g., kernel methods, local-averaging, or neural networks [see, e.g., Györfi et al., 2002, Tsybakov, 2009], they also suffer from the *curse of dimensionality*, that is, the rate of convergence of the estimated function to the true function (with any relevant performance measure) can only decrease as a small power of n, and this power cannot be larger than a constant divided by d. In other words, the number n of observations for any level of precision is exponential in dimension.

A classical way of by-passing the curse of dimensionality is to make extra assumptions regarding the function to estimate, such as the dependence on a lower unknown low-dimensional subspace, such as done by projection pursuit or neural networks. More precisely, throughout the chapter, we make the following assumption:

(A1) For all $x \in \mathbb{R}^d$, we have $f(x) = g(w^\top x)$ for a certain matrix $w \in \mathbb{R}^{d\times k}$ and a function $g : \mathbb{R}^k \to \mathbb{R}$. Moreover, $y = f(x) + \varepsilon$ with ε independent of x with zero mean and finite variance.

The subspace of $\mathbb{R}^d$ spanned by the k columns $w_1, \ldots, w_k \in \mathbb{R}^d$ of w has dimension less than or equal to k, and is often called the *effective dimension reduction* (e.d.r.) space. The model above is often referred to as a *multiple-index model* [Yuan, 2011]. We will always make the assumption that the e.d.r. space has exactly rank k, that is the matrix w has rank k (which implies that $k \leqslant d$).

Given w, estimating g may be done by any technique in non-parametric regression, with a convergence rate which requires a number of observations n to be exponential in k, with methods based on local averaging (e.g., Nadaraya-Watson estimators) or on least-squares regression [see, e.g., Györfi et al., 2002, Tsybakov, 2009]. Given the non-linear function g, estimating w is computationally difficult because the resulting optimization problem may not be convex and thus leads to several local minima. The difficulty is often even stronger since one often wants to estimate *both* the function g and the matrix w.

Our main goal in this chapter is to estimate the matrix w, with the hope of obtaining a convergence rate where the inverse power of n will now be proportional to k and not d. Note that the matrix w is only identifiable up to a (right) linear transform, since only the subspace spanned by its column is characteristic.

Method of moments vs. optimization. This multiple-index problem and the goal of estimating w only can be tackled from two points of views: (a) the method of moments, where certain moments are built so that the effect of the unknown function g disappears [Brillinger, 1982, Li and Duan, 1989], a method that we follow here and describe in more details below. These methods rely heavily on the model being correct, and in the instances that we consider here lead to provably polynomial-time algorithms (and most often linear in the number of observations since only moments are computed). In contrast, (b) optimization-based methods use implicitly or explicitly non-parametric estimation, e.g., using local averaging methods to design an objective function that can be minimized to obtain an estimate of w [Xia et al., 2002a, Fukumizu et al., 2009]. The objective function is usually non-convex and gradient descent techniques are used to obtain a local minimum. While these procedures offer no theoretical guarantees due to the potential unknown difficulty of the optimization problem, they often work well in practice, and we have observed this in our experiments.

In this chapter, we consider and improve a specific instantiation of the method of moments, which partially circumvents the difficulty of joint estimation by estimating w directly without the knowledge of g. The starting point for this method is the work by Brillinger [1982], which shows, as a simple consequence of Stein's lemma [Stein, 1981], that if the distribution of x is Gaussian, **(A1)** is satisfied with $k = 1$ (e.d.r. of dimension one, e.g., a single-index model), and the input data have zero mean and identity covariance matrix, then the expectation $\mathbb{E}(yx)$ is proportional to w. Thus, a certain expectation, which can be easily approximated given i.i.d. observations, *simultaneously* eliminates g and reveals w.

While the result above provides a very simple algorithm to recover w, it has several strong limitations: (a) it only applies to normally distributed data x, or more generally to elliptically symmetric distributions [Cambanis et al., 1981], (b) it only applies to $k = 1$, and (c) in many situations with symmetries, the proportionality constant is equal to zero and thus we cannot recover the vector w. This has led to several extensions in the statistical literature which we now present.

Using score functions. The use of Stein's lemma with a Gaussian random variable can be directly extended using the score function $\mathcal{S}_1(x)$ defined as the negative gradient of the log-density, that is, $\mathcal{S}_1(x) = -\nabla \log p(x) = \frac{-1}{p(x)} \nabla p(x)$, which leads to the following assumption:

(A2) The distribution of x has a strictly positive density $p(x)$ which is differentiable with respect to the Lebesgue measure, and such that $p(x) \to 0$ when $\|x\| \to +\infty$.

We will need the score to be sub-Gaussian to obtain consistency results. Given Assumption **(A2)**, then Stoker [1986] showed, as a simple consequence of integration by parts, that, for $k = 1$ and if Assumption **(A1)** is satisfied, then $\mathbb{E}(y\mathcal{S}_1(x))$ is proportional to w, for all differentiable functions g, with a proportionality constant that depends on w and ∇g. This leads to the "average derivative method" (ADE) and thus replaces the Gaussian assumption by the existence of a differentiable log-density, which is much weaker. This however does not remove the restriction $k = 1$, which can be done in two ways which we now present.

Sliced inverse regression. Given a normalized Gaussian distribution for x (or any elliptically symmetric distribution), then, if **(A1)** is satisfied, almost surely in y, the conditional expectation $\mathbb{E}(x|y)$ happens to belong to the e.d.r. subspace. Given several distinct values of y, the vectors $\mathbb{E}(x|y)$ or any estimate thereof, will hopefuly span the entire e.d.r. space and we can recover the entire matrix w, leading to "slice inverse regression" (SIR), originally proposed by Li and Duan [1989], Duan and Li [1991], Li [1991]. This allows the estimation with $k > 1$, but this is still restricted to Gaussian data. In this chapter, we propose to extend SIR by the use of score functions to go beyond elliptically symmetric distributions, and we show that the new method combining SIR and score functions is formally better than the plain ADE method.

From first-order to second-order moments. Another line of extension of the simple method of Brillinger [1982] is to consider higher-order moments, namely the matrix $\mathbb{E}(yxx^\top) \in \mathbb{R}^{d\times d}$, which, with normally distributed input data x and, if **(A1)** is satisfied, will be proportional (in a particular form to be described in Section 2.2.2) to the Hessian of the function g, leading to the method of "principal Hessian directions" (PHD) from Li [1992]. Again, $k > 1$ is allowed (more than a single projection), but thus is limited to elliptically symmetric data. However Janzamin et al. [2014] proposed to used second-order score functions to go beyond this assumption. In order to define this new method, we consider the following assumption:

(A3) The distribution of x has a strictly positive density $p(x)$ which is twice differentiable with respect to the Lebesgue measure, and such that $p(x)$ and $\|\nabla p(x)\| \to 0$ when $\|x\| \to +\infty$.

Given **(A1)** and **(A3)**, then one can show [Janzamin et al., 2014] that $\mathbb{E}(y\mathcal{S}_2(x))$ will be proportional to the Hessian of the function g, where $\mathcal{S}_2(x) = \nabla^2 \log p(x) + \mathcal{S}_1(x)\mathcal{S}_1(x)^\top = \frac{1}{p(x)}\nabla^2 p(x)$, thus extending the Gaussian situation above where $\mathcal{S}_1$ was a linear function and $\mathcal{S}_2(x)$, up to linear terms, proportional to $xx^\top$.

In this chapter, we propose to extend the method above to allow an SIR estimator for the second-order score functions, where we condition on y, and we show that the new method is formally better than the plain method of Janzamin et al. [2014].

Learning score functions through score matching. Relying on score functions immediately raises the following question: is estimating the score function (when not available) really simpler than our original problem of non-parametric regression? Fortunately, a recent line of work [Hyvärinen, 2005] has considered this exact problem, and formulated the task of density estimation directly on score functions, which is particularly useful in our context. We may then use the data, first to learn the score, and then to use the novel score-based moments to estimate w. We will also consider a direct approach that jointly estimates the score function and the e.d.r. subspace, by regularizing by a sparsity-inducing norm.

Fighting the curse of dimensionality. Learning the score function is still a non-parametric problem, with the associated curse of dimensionality. If we first learn the score function (through score matching) and then learn the matrix w, we will not escape that curse, while our direct approach is empirically more robust.

Note that Hristache, Juditsky and Spokoiny [Hristache et al., 2001] suggested iterative improvements of the ADE method, using elliptic windows which shrink in the directions of the columns of w, stretch in all others directions and tend to flat layers orthogonal to w. Dalalyan, Juditsky and Spokoiny [Dalalyan et al., 2008] generalize the algorithm to multi-index models and proved $\sqrt{n}$-consistency of the proposed procedure in the case when the structural dimension is not larger than 4 and weaker dependence for $d > 4$. In particular, they provably avoid the curse of dimensionality. Such extensions are outside the scope of this chapter.

Contributions. In this chapter, we make the following contributions:

- We propose score function extensions to sliced inverse regression problems, both for the first-order and second-order score functions. We consider the infinite sample case in Section 2.2 and the finite sample case in Section 2.3. They provably improve estimation in the population case over the non-sliced versions, while we study in Section 2.3 finite sample estimators and their consistency given the exact score functions.
- We propose in Section 2.4 to learn the score function as well, in two steps, i.e., first learning the score function and then learning the e.d.r. space parameterized by w, or directly, by solving a convex optimization problem regularized by the nuclear norm.
- We illustrate our results in 2.5 on a series of experiments.

2.2 Estimation with infinite sample size

In this section, we focus on the population situation, where we can compute expectations and conditional expectations exactly, while we focus on finite sample estimators with known score functions in Section 2.3 with consistency results in Section 2.3.3, and with learned score functions in Section 2.4.

2.2.1 SADE: Sliced average derivative estimation

Before presenting our new moments which will lead to the novel SADE method, we consider the non-sliced method, which is based on Assumptions **(A1)** and **(A2)** and score functions (the method based on the Gaussian assumption will be derived later as corollaries). The ADE method is based on the following lemma:

Lemma 2.1 (ADE moment [Stoker, 1986]). *Assume **(A1)**, **(A2)**, the differentiability of g and the existence of expectation $\mathbb{E}(g'(w^\top x))$. Then $\mathbb{E}(\mathcal{S}_1(x)y)$ is in the e.d.r. subspace.*

Proof. Since $y = f(x) + \varepsilon$, and ε is independent of x with zero mean, we have

$$\begin{aligned}
\mathbb{E}(\mathcal{S}_1(x)y) &= \mathbb{E}(\mathcal{S}_1(x)f(x)) = \int_{\mathbb{R}^d} \frac{-\nabla p(x)}{p(x)} f(x)p(x)dx \\
&= -\int_{\mathbb{R}^d} \nabla p(x) f(x)dx = \int_{\mathbb{R}^d} p(x)\nabla f(x)dx \text{ by integration by parts,} \\
&= w \cdot \mathbb{E}(g'(w^\top x)),
\end{aligned}$$

which leads to the desired result. Note that in the integration by parts above, the decay of $p(x)$ to zero for $\|x\| \to +\infty$ is needed. □

The ADE moment above only provides a single vector in the e.d.r. subspace, which can only potentially lead to recovery for $k = 1$, and only if $\mathbb{E}(g'(w^\top x)) \neq 0$, which may not be satisfied, e.g., if x has a a symmetric distribution and g is even.

We can now present our first new lemma, the proof of which relies on similar arguments as for SIR [Li and Duan, 1989] but extended to score functions. Note that we do not require the differentiability of the function g.

Lemma 2.2 (SADE moment). *Assume **(A1)** and **(A2)**. Then, $\mathbb{E}(\mathcal{S}_1(x)|y)$ is in the e.d.r. subspace almost surely (in y).*

Proof. We consider any vector $b \in \mathbb{R}^d$ in the orthogonal complement of the subspace $\text{Span}\{w_1, \dots, w_k\}$. We need to show, that $b^\top \mathbb{E}(\mathcal{S}_1(x)|y) = 0$ with probability 1. We have by the law of total expectation

$$b^\top \mathbb{E}\big(\mathcal{S}_1(x)|y\big) = \mathbb{E}\Big(\mathbb{E}\big(b^\top \mathcal{S}_1(x)|w_1^\top x, \dots, w_k^\top x, y\big)|y\Big).$$

Because of Assumption **(A1)**, we have $y = g(w^\top x) + \varepsilon$ with ε independent of x, and thus

$$\mathbb{E}\big(b^\top \mathcal{S}_1(x)|w^\top x, y\big) = \mathbb{E}\big(b^\top \mathcal{S}_1(x)|w^\top x, \varepsilon\big) = \mathbb{E}\big(b^\top \mathcal{S}_1(x)|w^\top x\big) \text{ almost surely.}$$

This leads to

$$b^\top \mathbb{E}\big(\mathcal{S}_1(x)|y\big) = \mathbb{E}\big[\mathbb{E}\big(b^\top \mathcal{S}_1(x)|w^\top x\big)\,|y\big].$$

We now prove that almost surely $\mathbb{E}\big(b^\top \mathcal{S}_1(x)|w^\top x\big) = 0$, which will be sufficient to prove Lemma 2.2. We consider the linear transformation of coordinates: $\widetilde{x} = \widetilde{w}^\top x \in \mathbb{R}^d$, where $\widetilde{w} = (w_1, \dots, w_k, w_{k+1}, \dots, w_d)$ is a square matrix with full rank obtained by adding a basis of the subspace orthogonal to the span of the k columns of w. Then, if $\widetilde{p}$ is the density of $\widetilde{x}$, we have $p(x) = (\det \widetilde{w}) \cdot \widetilde{p}(\widetilde{x})$ and thus $\nabla p(x) = (\det \widetilde{w}) \cdot \widetilde{w} \cdot \nabla \widetilde{p}(\widetilde{x})$ and $\widetilde{b} = \widetilde{w}^\top b = (0, \dots, 0, \widetilde{b}_{k+1}, \dots, \widetilde{b}_d) \in \mathbb{R}^d$ (because $b \perp \text{Span}\{w_1, \dots, w_k\}$). The desired conditional expectation equals

$$\mathbb{E}\big(\widetilde{b}^\top \widetilde{\mathcal{S}}_1(\widetilde{x})|\widetilde{x}_1, \dots, \widetilde{x}_k\big),$$

since $\widetilde{w}\widetilde{\mathcal{S}}_1(\widetilde{x}) = \dfrac{\widetilde{w}\nabla \widetilde{p}(\widetilde{x})}{\widetilde{p}(\widetilde{x})} = \dfrac{\nabla p(x)}{p(x)} = \mathcal{S}_1(x)$ and hence $b^\top \mathcal{S}_1(x) = b^\top \widetilde{w}\ \widetilde{\mathcal{S}}_1(\widetilde{x}) = \widetilde{b}^\top \widetilde{\mathcal{S}}_1(\widetilde{x})$.

It is thus sufficient to show that $\displaystyle\int\limits_{\mathbb{R}^{d-k}} \widetilde{b}^\top \widetilde{\mathcal{S}}_1(\widetilde{x}) \widetilde{p}(\widetilde{x}_1, \dots, \widetilde{x}_d) d\widetilde{x}_{k+1} \dots d\widetilde{x}_d = 0$, for all $\widetilde{x}_1, \dots, \widetilde{x}_k$. We have

$$\int\limits_{\mathbb{R}^{d-k}} \widetilde{b}^\top \widetilde{\mathcal{S}}_1(\widetilde{x}) \widetilde{p}(\widetilde{x}_1, \dots, \widetilde{x}_d) d\widetilde{x}_{k+1} \dots d\widetilde{x}_d = \widetilde{b}^\top \int\limits_{\mathbb{R}^{d-k}} \nabla \widetilde{p}(\widetilde{x}) \cdot d\widetilde{x}_{k+1} \dots d\widetilde{x}_d$$

$$= \sum_{j=k+1}^{d} \widetilde{b}_j \cdot \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\partial \widetilde{p}(\widetilde{x})}{\partial \widetilde{x}_j} d\widetilde{x}_j \cdot \prod_{\substack{k+1 \leqslant t \leqslant d \\ t \neq j}} d\widetilde{x}_t = 0,$$

because for any $j \in \{k+1, \ldots, n\}$, $\int_{-\infty}^{\infty} \frac{\partial \widetilde{p}(\widetilde{x})}{\partial \widetilde{x}_j} d\widetilde{x}_j = 0$ by Assumption **(A2)**. This leads to the desired result. □

The key differences are now that:

— Unlike ADE, by conditioning on different values of y, we have access to *several* vectors $\mathbb{E}(\mathcal{S}_1(x)|y) \in \mathbb{R}^d$.

— Unlike SIR, SADE does not require the linearity condition from [Li, 1991] anymore and can be used with a smooth enough probability density.

In the population case, we will consider the following matrix (using the fact that $\mathbb{E}(\mathcal{S}_1(x)) = 0$):

$$\mathcal{V}_{1,\text{cov}} = \mathbb{E}\big[\mathbb{E}(\mathcal{S}_1(x)|y)\mathbb{E}(\mathcal{S}_1(x)|y)^\top\big] = \text{Cov}\big[\mathbb{E}(\mathcal{S}_1(x)|y)\big] \in \mathbb{R}^{d\times d},$$

which we will also denote $\mathbb{E}\big[\mathbb{E}(\mathcal{S}_1(x)|y)^{\otimes 2}\big]$, where for any matrix a, $a^{\otimes 2}$ denotes $aa^\top$. The matrix above is positive semi-definite, and its column space is included in the e.d.r. space. If it has rank k, then we can exactly recover the entire subspace by an eigenvalue decomposition. When $k = 1$, which is the only case where ADE may be used, the following proposition shows that if ADE allows to recover w, so is SADE.

We will also consider the other matrix (note the presence of the extra term y^2)

$$\mathcal{V}'_{1,\text{cov}} = \mathbb{E}\big[y^2 \mathbb{E}(\mathcal{S}_1(x)|y)\mathbb{E}(\mathcal{S}_1(x)|y)^\top\big],$$

because of its direct link with the non-sliced version. Note that we made the weak assumption of existence of matrices $\mathcal{V}_{1,\text{cov}}$ and $\mathcal{V}'_{1,\text{cov}}$, which is satisfied for majority of problems.

Proposition 2.1. *Assume **(A1)** and **(A2)**, with $k = 1$, as well as differentiability of g and existence of the expectation $\mathbb{E}g'(w^\top x)$. The vector w may be recovered from the ADE moment (up to scale) if and only if $\mathbb{E}g'(w^\top x) \neq 0$. If this condition is satisfied, then SADE also recovers w up to scale (i.e., $\mathcal{V}_{1,cov}$ and $\mathcal{V}'_{1,cov}$ are different from zero).*

Proof. The first statement is a consequence of the proof of Lemma 2.1. If SADE fails, that is, for almost all y, $\mathbb{E}(\mathcal{S}_1(x)|y) = 0$, then $\mathbb{E}(\mathcal{S}_1(x)y|y) = 0$ which implies that $\mathbb{E}(\mathcal{S}_1(x)y) = 0$ and thus ADE fails. Moreover, we have, using operator convexity [Donoghue, 1974]:

$$\mathcal{V}'_{1,\text{cov}} = \mathbb{E}\big[\mathbb{E}(y\mathcal{S}_1(x)|y)\mathbb{E}(y\mathcal{S}_1(x)|y)^\top\big] \succcurlyeq \big[\mathbb{E}(\mathbb{E}(y\mathcal{S}_1(x)|y))\big]\big[\mathbb{E}(\mathbb{E}(y\mathcal{S}_1(x)|y))\big]^\top =$$

$$= \big[\mathbb{E}(y\mathcal{S}_1(x))\big]\big[\mathbb{E}(y\mathcal{S}_1(x))\big]^\top,$$

showing that the new moment is dominating the ADE moment, which provides an alternative proof of the rank of $\mathcal{V}'_{1,\text{cov}}$ being larger than one if $\mathbb{E}(y\mathcal{S}_1(x)) \neq 0$. □

Elliptically symmetric distributions. If x is normally distributed with mean vector μ and covariance matrix Σ, then we have $\mathcal{S}_1(x) = \Sigma^{-1}(x-\mu)$ and we recover the result from Li and Duan [1989]. Note that the lemma then extends to all elliptical distributions of the form $\varphi(\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu))$, for a certain function $\varphi : \mathbb{R}^+ \to \mathbb{R}$. See Li and Duan [1989] for more details.

Failure modes. In some cases, slice inverse regression does not span the entire e.d.r. space, because the inverse regression curve $\mathbb{E}(\mathcal{S}_1(x)|y)$ is degenerated. For example, this can occur, if $k = 1$, $y = h(w_1^\top x) + \varepsilon$, h is an even function and $w_1^\top x$ has a symmetric distribution around 0. Then $\mathbb{E}(\mathcal{S}_1(x)|y) \equiv 0$, and thus it is a poor estimation of the desired e.d.r. directions [Cook and Weisberg, 1991].

The second drawback of SIR occurs when we have a classification task, for example, $y \in \{0, 1\}$. In this case, we have only two slices (i.e., possible values of y) and SIR can recover only one direction in the e.d.r. space [Cook and Lee, 1999].

Li [1992] suggested another way to estimate the e.d.r. space which can handle such symmetric cases: principal Hessian directions (PHD). However, this method uses the normality of the vector X. As in the SIR case, we can extend this method, using score functions to use it for any distribution, which we will refer to as SPHD, which we now present.

2.2.2 SPHD: Sliced principal Hessian directions

Before presenting our new moment which will lead to the SPHD method, we consider the non-sliced method, which is based on Assumptions **(A1)** and **(A3)** and score functions (the method based on the Gaussian assumption will be derived later as corollaries). The method of Janzamin et al. [2014], which we refer to as "PHD+" is based on the following lemma (we reproduce the proof for readability):

Lemma 2.3 (second-order score moment (PHD+) [Janzamin et al., 2014]). *Assume **(A1)**,**(A3)**,twice differentiability of function g and existence of the expectation $\mathbb{E}(\nabla^2 g(w^\top x))$. Then $\mathbb{E}(\mathcal{S}_2(x)y)$ has a column space included in the e.d.r. subspace.*

Proof. Since $y = f(x) + \varepsilon$, and ε is independent of x, we have

$$\mathbb{E}\big(y\mathcal{S}_2(x)\big) = \mathbb{E}\big(f(x)\mathcal{S}_2(x)\big) = \int \frac{\nabla^2 p(x)}{p(x)} p(x) f(x) dx =$$

$$= \int \nabla^2 p(x) \cdot f(x) dx = \int p(x) \cdot \nabla^2 f(x) dx = \mathbb{E}\big[\nabla^2 f(x)\big],$$

using integration by parts and the decay of $p(x)$ and $\nabla p(x)$ for $\|x\| \to \infty$. This leads to the desired result since $\nabla^2 f(x) = w\nabla^2 g(w^\top x)w^\top$. This was proved by Li [1992] for normal distributions. □

Failure modes. The method does not work properly if $\operatorname{rank}(\mathbb{E}(\nabla^2 g(w^\top x))) < k$. For example, if g is linear function, $\mathbb{E}(\nabla^2 g(w^\top x)) \equiv 0$ and the estimated e.d.r. space is degenerated. Moreover, the method fails in symmetric cases, for example, if g is an odd function with respect to any variable and $p(x)$ is even function, then $\operatorname{rank}(\mathbb{E}(\nabla^2 g(w^\top x))) < k$.

We can now present our second new lemma, the proof of which relies on similar arguments as for PHD [Li, 1992] but extended to score functions (again no differentiability is assumed on g):

Lemma 2.4 (SPHD moment)**.** *Assume **(A1)** and **(A3)**. Then, $\mathbb{E}(\mathcal{S}_2(x)|y)$ has a column space within the e.d.r. subspace almost surely.*

Proof. We consider any $a \in \mathbb{R}^d$ and $b \in \mathbb{R}^d$ orthogonal to the e.d.r. subspace, and prove, that $a^\top \mathbb{E}\big(\mathcal{S}_2(x)|y\big)b = 0$. We use the same transform of coordinates as in the proof of Lemma 2.2: $\widetilde{x} = \widetilde{w}^\top x \in \mathbb{R}^d$. Then $\nabla^2 p(x) = \det(\widetilde{w}) \cdot \widetilde{w}\nabla^2 \widetilde{p}(\widetilde{x})\widetilde{w}^\top$, $\widetilde{b} = \widetilde{w}^\top b = (0, \dots, 0, \widetilde{b}_{k+1}, \dots, \widetilde{b}_d)$ and we will prove, that $\mathbb{E}\big(a^\top \mathcal{S}_2(x) b | w^\top x\big) = 0$ almost surely, and $\widetilde{w}^\top \widetilde{\mathcal{S}}_2(\widetilde{x})\widetilde{w} = \dfrac{\widetilde{w}^\top \nabla^2 \widetilde{p}(\widetilde{x})\widetilde{w}}{\widetilde{p}(\widetilde{x})} = \dfrac{\nabla^2 p(x)}{p(x)} = \mathcal{S}_2(x)$. It is sufficient to show, that for all $\widetilde{x}_1, \dots, \widetilde{x}_k$:

$$\int_{\mathbb{R}^{d-k}} \widetilde{a}^\top \cdot \widetilde{\mathcal{S}_2}(\widetilde{x}) \cdot \widetilde{b} \cdot \widetilde{p}(\widetilde{x}_1, \dots, \widetilde{x}_d) d\widetilde{x}_{k+1} \dots d\widetilde{x}_d = 0.$$

We have:

$$\int_{\mathbb{R}^{d-k}} \widetilde{a}^\top \cdot \widetilde{\mathcal{S}}_2(\widetilde{x}) \cdot \widetilde{b} \cdot \widetilde{p}(\widetilde{x}_1, \dots, \widetilde{x}_d) d\widetilde{x}_{k+1} \dots d\widetilde{x}_d = \widetilde{a}^\top \cdot \left[\int_{\mathbb{R}^{d-k}} \nabla^2 \widetilde{p}(\widetilde{x}) \cdot d\widetilde{x}_{k+1} \cdots d\widetilde{x}_d \right] \cdot \widetilde{b}$$

$$= \sum_{\substack{1 \leqslant i \leqslant d \\ k+1 \leqslant j \leqslant d}} \widetilde{a}_i \widetilde{b}_j \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \left[\int_{-\infty}^{\infty} \frac{\partial^2 \widetilde{P}(\widetilde{x})}{\partial \widetilde{x}_i \partial \widetilde{x}_j} \cdot d\widetilde{x}_j \right] \cdot \prod_{\substack{k+1 \leqslant t \leqslant d \\ t \neq j}} d\widetilde{x}_t = 0,$$

because for any $j \in \{k+1, \dots, n\}$: $\displaystyle\int_{-\infty}^{\infty} \frac{\partial^2 \widetilde{P}(\widetilde{x})}{\partial \widetilde{x}_i \partial \widetilde{x}_j} \cdot d\widetilde{x}_j = 0$ due to Assumption **(A3)**, which leads to the desired result. □

In order to be able to use several values of y, we will estimate the matrix $\mathcal{V}_2 = \mathbb{E}\Big([\mathbb{E}\big(\mathcal{S}_2(x)|y\big)]^2\Big)$, which is the expectation with respect to y of the square (in the matrix multiplication sense) of the conditional expectation from Lemma 2.4, as well as $\mathcal{V}_2' = \mathbb{E}\Big(y^2[\mathbb{E}\big(\mathcal{S}_2(x)|y\big)]^2\Big)$, and consider the k largest eigenvectors (we made the weak assumption of existence of matrices $\mathcal{V}_2$ and $\mathcal{V}_2'$, which is satisfied for majority of problems). From the lemma above, this matrix has a column space included in the e.d.r. subspace, thus, if it has rank k, we get exact recovery by an eigenvalue decomposition.

Effect of slicing. We now study the effect of slicing and show that in the population case, it is superior to the non-sliced version, as it recovers the true e.d.r. subspace in more situations.

Proposition 2.2. *Assume **(A1)** and **(A3)**. The matrix w may be recovered from the moment in Lemma 2.3 (up to right linear transform) if and only if $\mathbb{E}[\nabla^2 g(w^\top x)]$ has full rank. If this condition is satisfied, then SPHD also recovers w up to scale.*

Proof. The first statement is a consequence of the proof of Lemma 2.3. Moreover, using the Lowner-Heinz theorem about operator convexity [Donoghue, 1974]:

$$\mathcal{V}_2' = \mathbb{E}\big[\mathbb{E}(y\mathcal{S}_2(x)|y)^2\big] \succcurlyeq \big[\mathbb{E}[\mathbb{E}(y\mathcal{S}_2(x)|y)]\big]^2 = \big[\mathbb{E}(y\mathcal{S}_2(x))\big]^2,$$

showing that the new moment is dominating the PHD moment, thus implying that

$$\operatorname{rank}[\mathcal{V}_2'] \geqslant \operatorname{rank}\big[\mathbb{E}(y\mathcal{S}_2(x))\big].$$

Therefore, if $\operatorname{rank}\big[\mathbb{E}(y\mathcal{S}_2(x))\big] = k$, then $\operatorname{rank}[\mathcal{V}_2'] = k$ (note that there is not such a simple proof for $\mathcal{V}_2$). □

Elliptically symmetric distributions. When x is a standard Gaussian random variable, then $\mathcal{S}_2(x) = -I + xx^\top$, and thus $\mathbb{E}\Big([\mathbb{E}\big(\mathcal{S}_2(x)|y)]^2\Big) = \mathbb{E}\Big([I - \operatorname{Cov}(x|y)]^2\Big)$, and we recover the sliced average variance estimation (SAVE) method by Cook [2000]. However, our method applies to all distributions (with known score functions).

2.2.3 Relationship between first and second order methods

All considered methods have their own failure modes. The simplest one: ADE works only in single-index model and has quite a simple working condition : $\mathbb{E}[g'(w^\top x)] \neq 0$. The sliced improvement (e.g., SADE) of this algorithm has a better performance, however it still suffers from symmetric cases, when the inverse regression curve is partly degenerated. PHD+ can not work properly in linear models and symmetric cases. SPHD is stronger than PHD+ and potentially has the widest application area among four described methods. See summary in Table 2.1.

Our conditions rely on the full possible rank of certain expected covariance matrices. When the function g is selected randomly from all potential functions from $\mathbb{R}^k$ to $\mathbb{R}$, rank-deficiencies typically do not occur and it would be interesting to show that indeed they appear with probability zero for certain random function models.

2.3 Estimation from finite sample

In this section, we consider finite sample estimators for the moments we have defined in Section 2.2. Since our extensions are combinations of existing techniques

method	main equation	score	sliced
ADE	$\mathbb{E}(\mathcal{S}_1(x)y)$	first	no
SADE	$\mathbb{E}_y\left[\mathbb{E}\big(\mathcal{S}_1(x)\|y\big)^{\otimes 2}\right]$	first	yes
PHD+	$\mathbb{E}(\mathcal{S}_2(x)y)$	second	no
SPHD	$\mathbb{E}_y\left[\mathbb{E}\Big((\mathcal{S}_2(x)\|y)^2\Big)\right]$	second	yes

method	single-index	multi-index
ADE	$\mathbb{E}[g'(w^\top x)] \neq 0$	does not work
SADE	$\mathbb{E}_y\left[\|\mathbb{E}\big(\mathcal{S}_1(x)\|y\|^2\right] > 0$	rank $\mathbb{E}_y\left[\mathbb{E}\big(\mathcal{S}_1(x)\|y\big)^{\otimes 2}\right] = k$
PHD+	$\mathbb{E}[g''(w^\top x)] \neq 0$	rank $\left[\mathbb{E}[\nabla^2 g(x)]\right] = k$
SPHD	$\mathbb{E}_y\left[\mathrm{tr}\mathbb{E}\Big((\mathcal{S}_2(x)\|y)^2\Big)\right] > 0$	rank $\mathbb{E}_y\left[\mathbb{E}\Big((\mathcal{S}_2(x)\|y)^2\Big)\right] = k$

Table 2.1 – Comparison of different methods using score functions.

(using score functions and slicing) our finite-sample estimators naturally rely on existing work [Hsing and Carroll, 1992, Zhu and Ng, 1995].

In this section, we assume that the score function is known. We consider learning the score function in Section 2.4.

2.3.1 Estimator and algorithm for SADE

Our goal is to provide an estimator for $\mathcal{V}_{1,\mathrm{cov}} = \mathbb{E}\Big[\mathbb{E}\big(\mathcal{S}_1(x)|y\big)\mathbb{E}\big(\mathcal{S}_1(x)|y\big)^\top\Big] = \mathrm{Cov}\Big[\mathbb{E}\big(\mathcal{S}_1(x)|y\big)\Big]$ given a finite sample (x_i, y_i), $i = 1, \dots, n$. A similar estimator for $\mathcal{V}'_{1,\mathrm{cov}}$ could be derived. In order to estimate $\mathcal{V}_{1,\mathrm{cov}} = \mathrm{Cov}\Big[\mathbb{E}\big(\mathcal{S}_1(x)|y\big)\Big]$, we will use the identity

$$\mathrm{Cov}\big[\mathcal{S}_1(x)\big] = \mathrm{Cov}\Big[\mathbb{E}\big(\mathcal{S}_1(x)|y\big)\Big] + \mathbb{E}\Big[\mathrm{Cov}\big(\mathcal{S}_1(x)|y\big)\Big].$$

We use the natural consistent estimator $\frac{1}{n}\sum_{i=1}^{n}\mathcal{S}_1(x_i)\mathcal{S}_1(x_i)^\top$ of $\mathrm{Cov}\big[\mathcal{S}_1(x)\big]$.

In order to obtain an estimator of $\mathbb{E}\Big[\mathrm{Cov}\big(\mathcal{S}_1(x)|y\big)\Big]$, we consider slicing the real numbers in H different *slices*, $I_1, \dots, I_H$, which are contiguous intervals that form a partition of $\mathbb{R}$ (or of the range of all y_i, $i = 1, \dots, n$). We then compute an estimator of the conditional expectation $(\mathcal{S}_1)_h = \mathbb{E}(\mathcal{S}_1(x)|y \in I_h)$ with empirical averages: denoting $\hat{p}_h$ the empirical proportion of y_i, $i = 1, \dots, n$, that fall in the slice I_h (which is assumed to be strictly positive), we estimate $\mathbb{E}(\mathcal{S}_1(x)|y \in I_h)$ by

$$(\hat{\mathcal{S}}_1)_h = \frac{1}{n\hat{p}_h}\sum_{i=1}^{n} 1_{y_i \in I_h}\mathcal{S}_1(x_i).$$

We then estimate $\mathrm{Cov}\big(\mathcal{S}_1(x)|y \in I_h\big)$ by

$$(\hat{\mathcal{S}}_1)_{\mathrm{cov},h} = \frac{1}{n\hat{p}_h - 1}\sum_{i=1}^{n} 1_{y_i \in I_h}\big(\mathcal{S}_1(x_i) - (\hat{\mathcal{S}}_1)_h\big)\big(\mathcal{S}_1(x_i) - (\hat{\mathcal{S}}_1)_h\big)^\top.$$

Note that it is important here to normalize the covariance computation by $\frac{1}{n\hat{p}_h - 1}$ (usual unbiased normalization of the variance) and not $\frac{1}{n\hat{p}_h}$, to allow consistent estimation even when the number of elements per slice is small (e.g., equal to 2).

We finally use the following estimator of $\mathcal{V}_{1,\mathrm{cov}}$:

$$\hat{\mathcal{V}}_{1,\mathrm{cov}} = \frac{1}{n}\sum_{i=1}^{n}\mathcal{S}_1(x_i)\mathcal{S}_1(x_i)^\top - \sum_{h=1}^{H}\hat{p}_h \cdot (\hat{\mathcal{S}}_1)_{\mathrm{cov},h}.$$

The final SADE algorithm is thus the following:

- Divide the range of $y_1, \ldots, y_n$ into H slices $I_1, \ldots, I_H$. Let $\hat{p}_h > 0$ be the proportion of y_i, $i = 1, \ldots, n$, that fall in slice I_h.
- For each slice I_h, compute the sample mean $(\hat{\mathcal{S}}_1)_h$ and covariance $(\hat{\mathcal{S}}_1)_{\mathrm{cov},h}$: $(\hat{\mathcal{S}}_1)_h = \dfrac{1}{n\hat{p}_h}\displaystyle\sum_{i=1}^{n} 1_{y_i \in I_h}\mathcal{S}_1(x_i)$ and
$$(\hat{\mathcal{S}}_1)_{\mathrm{cov},h} = \frac{1}{n\hat{p}_h - 1}\sum_{i=1}^{n} 1_{y_i \in I_h}\big(\mathcal{S}_1(x_i) - (\hat{\mathcal{S}}_1)_h\big)\big(\mathcal{S}_1(x_i) - (\hat{\mathcal{S}}_1)_h\big)^\top.$$
- Compute $\hat{\mathcal{V}}_{1,\mathrm{cov}} = \frac{1}{n}\sum\limits_{i=1}^{n}\mathcal{S}(x_i)\mathcal{S}(x_i)^\top - \sum\limits_{h=1}^{H}\hat{p}_h \cdot (\hat{\mathcal{S}})_{\mathrm{cov},h}$.
- Find the k largest eigenvalues and let $\hat{w}_1, \ldots, \hat{w}_k$ be eigenvectors in $\mathbb{R}^d$ corresponding to these eigenvalues.

The schematic graphical representation of this method given in Figure 2-1.

Choice of slices. There are different ways to choose slices $I_1, \ldots, I_H$:

- all slices have the same length, that is we choose the maximum and the minimum of $y_1, \ldots, y_n$, and divide the range of y into H equal slices (for simplicity, we assume that n is a multiple of H),
- we can also use the distribution of y to ensure a balanced distribution of observations in each slice, and choose $I_h = (\hat{F}_y^{-1}((h-1)/H), \hat{F}_y^{-1}(h/H)]$, where $\hat{F}_y(t) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\mathbf{1}_{y_i \leqslant t}$ is the empirical distribution function of y. If n is a multiple H, there are exactly $c = n/H$ observations per slice.

Later on in our experiments, we use the second way, where every slice has $c = n/H$ points, and our consistency result applies to this situation as well. Note that there is a variation of SADE, which uses standardized data. If we denote the standardized

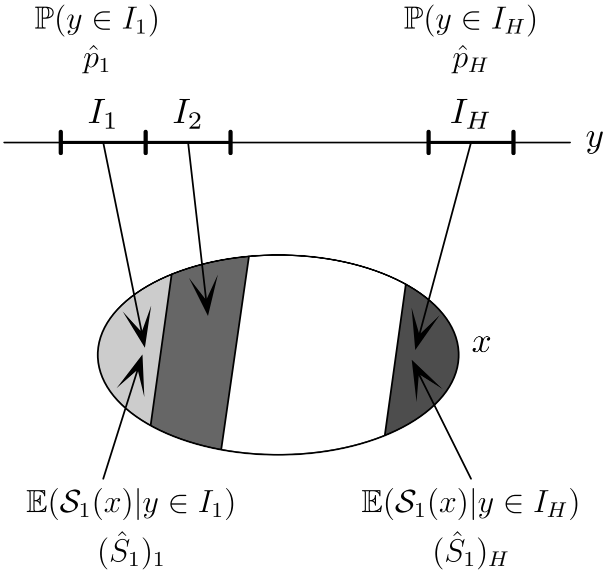

Figure 2-1 – Graphical explanation of SADE: firstly we divide y into H slices, let $\hat{p}_h$ be a proportion of y_i in slice I_h. This is an empirical estimator of $\mathbb{P}(y \in I_h)$. Then we evaluate the empirical estimator $(\hat{\mathcal{S}}_1)_h$ of $\mathbb{E}(\mathcal{S}_1(x)|y \in I_h)$. Finally, we evaluate weighted covariance matrix and find the k largest eigenvectors.

data as $\widetilde{\mathcal{S}}_1(x_i) = \Big[\frac{1}{n}\sum\limits_{j=1}^{n} \mathcal{S}_1(x_j)\mathcal{S}_1(x_j)^\top\Big]^{-1/2}\mathcal{S}_1(x_i)$ (Remark 5.3 in [Li, 1991]).

Computational complexity. The first step of the algorithm requires $O(n)$ elementary operations, the second $O(nd^2)$ operations, the third $O(Hd^2)$ and the fourth $O(kd^2)$ operations. The overall dependence on dimension d is quadratic, while the dependence on the number of observations is linear in n, as common in moment-matching methods.

Estimating the number of components. Our estimation method does not depend on k, up to the last step where the first k largest eigenvectors are selected. A simple heuristic to select k, similar to the selection of the number of components in principal component analysis, would select the largest k such that the gap between the k-th and $(k+1)$-th eigenvalue is large enough. This could be made more formal using the technique of Li [1991] for sliced inverse regression.

2.3.2 Estimator and algorithm for SPHD

We follow the same approach as for the SADE algorithm above, leading to the following algorithm, which estimates $\mathcal{V}_2 = \mathbb{E}\Big([\mathbb{E}\big(\mathcal{S}_2(x)|y\big)]^2\Big)$ and computes its principal eigenvectors. Note that $\mathbb{E}\big[\mathbb{E}(\mathcal{S}_2(x)|y \in I_h)\big] = \mathbb{E}\big[\mathcal{S}_2(x)\big] = 0$.

– Divide the range of $y_1, \dots, y_n$ into H slices $I_1, \dots, I_H$. Let $\hat{p}_h > 0$ be the proportion of y_i, $i = 1, \dots, n$, that fall in slice I_h.

– For each slice, compute the sample mean $(\hat{\mathcal{S}}_2)_h$ of $\mathcal{S}_2(x)$: $(\hat{\mathcal{S}}_2)_h = \frac{1}{n\hat{p}_h} \sum_{i=1}^{n} 1_{y_i \in I_h} \mathcal{S}_2(x_i)$.

– Compute the weighted covariance matrix $\hat{\mathcal{V}}_2 = \sum_{h=1}^{H} \hat{p}_h (\hat{\mathcal{S}}_2)_h^2$, find the k largest eigenvalues and let $\hat{w}_1, \dots, \hat{w}_k$ be eigenvectors corresponding to these eigenvalues.

The matrix $\hat{\mathcal{V}}_2$ is then an estimator of $\mathbb{E}\Big([\mathbb{E}\big(\mathcal{S}_2(x)|y\big)]^2\Big)$.

Computational complexity. The first step of the algorithm requires $O(n)$ elementary operations, the second $O(nd^2)$ operations, the third $O(Hd^3)$ and the fourth $O(kd^2)$ operations. The overall dependence on dimension d is cubic, hence the method is slower than SADE (but still linear in the number of observations n).

2.3.3 Consistency for the SADE estimator and algorithm

In this section, we prove the consistency of the SADE moment estimator and the resulting algorithm, when the score function is known. Following Hsing and Carroll [1992] and Zhu and Ng [1995], we can get $\sqrt{n}$-consistency for the SADE algorithm with very broad assumptions regarding the problem.

In this section, we focus on the simplest set of assumptions to pave the way to the analysis for the nuclear norm in future work. The key novelty compared to Hsing and Carroll [1992], Zhu and Ng [1995] is a precise *non-asymptotic* analysis with precise constants.

We make the following assumptions:

(L1) The function $m : \mathbb{R} \to \mathbb{R}^d$ such that $\mathbb{E}(\ell(x)|y) = m(y)$ is L-Lipschitz-continuous.

(L2) The random variable $y \in \mathbb{R}$ is sub-Gaussian, i.e., such that $\mathbb{E}e^{t(y-Ey)} \leqslant e^{\tau_y^2 t^2/2}$, for some $\tau_y > 0$.

(L3) The random variables $\ell_j(x) \in \mathbb{R}$ are sub-Gaussian, i.e., such that $\mathbb{E}e^{t\ell_j(x)} \leqslant e^{\tau_\ell^2 t^2/2}$ for each component $j \in \{1, \dots, d\}$, for some $\tau_\ell > 0$.

(L4) The random variables $\eta_j = \ell_j(x) - m_j(y) \in \mathbb{R}$ are sub-Gaussian, i.e., such that $\mathbb{E}e^{t\eta_j} \leqslant e^{\tau_\eta^2 t^2/2}$ for each component $j \in \{1, \dots, d\}$, for some $\tau_\eta > 0$.

Now we formulate and proof the main theorem, where $\|\cdot\|_*$ is the nuclear norm, defined as $\|A\|_* = \operatorname{tr}\big(\sqrt{A^T A}\big)$:

Theorem 2.1. *Under assumptions **(L1)** - **(L4)** we get the following bound on $\|\hat{\mathcal{V}}_{1,cov} - \mathcal{V}_{1,cov}\|_*$: for any $\delta < \frac{1}{n}$, with probability not less than $1 - \delta$:*

$$\|\hat{\mathcal{V}}_{1,cov} - \mathcal{V}_{1,cov}\|_* \leqslant \frac{d\sqrt{d}(195\tau_\eta^2 + 2\tau_\ell^2)}{\sqrt{n}} \sqrt{\log \frac{24d^2}{\delta}}$$

$$+\frac{8L^2\tau_y^2 + 16\tau_\eta\tau_y L\sqrt{d} + (157\tau_\eta^2 + 2\tau_\ell^2)d\sqrt{d}}{n}\log^2\frac{32d^2n}{\delta}.$$

The proof of the theorem can be found in Appendix 2.7.2. The leading term is proportional to $\frac{d\sqrt{d}\tau_\eta^2}{\sqrt{n}}$, with a tail which is sub-Gaussian. We thus get a $\sqrt{n}$-consistent estimator or $\mathcal{V}_1$. The dependency in d could probably be improved, in particular when using slices of sizes c that tend to infinity (as done in [Lin et al., 2018]).

2.4 Learning score functions

All previous methods can work only if we know the score function of first or second order. In practice, we do not have such information, and we have to learn score functions from sampled data. In this section, we only consider the first-order score function $\ell(x) = \mathcal{S}_1(x) = -\nabla \log p(x)$.

We first present the score matching approach of Hyvärinen [2005], and then apply it to our problem.

2.4.1 Score matching to estimate score from data

Given the true score function $\ell(x) = \mathcal{S}_1(x) = -\nabla \log p(x)$ and some i.i.d. data generated from $p(x)$, score matching aims at estimating the parameter of a model for the score function $\hat{\ell}(x)$, by minimizing a empirical quantity aiming to estimate

$$\mathcal{R}_{\text{score}}(\hat{\ell}) = \frac{1}{2}\int_{\mathbb{R}^d} p(x)\|\ell(x) - \hat{\ell}(x)\|^2 dx.$$

As is, the quantity above leads to consistent estimation (i.e., pushing $\hat{\ell}$ close to ℓ), but seems to need the knowledge of the true score $\ell(x)$. A key insight from Hyvärinen [2005] is to use integration by parts to get (assuming the integrals exist):

$$\begin{aligned}
\mathcal{R}_{\text{score}}(\hat{\ell}) &= \frac{1}{2}\int_{\mathbb{R}^d} p(x)\left[\|\ell(x)\|^2 + \|\hat{\ell}(x)\|^2 + 2\hat{\ell}(x)^\top \nabla \log p(x)\right] dx \\
&= \frac{1}{2}\int_{\mathbb{R}^d} p(x)\|\ell(x)\|^2 dx + \frac{1}{2}\int_{\mathbb{R}^d} p(x)\|\hat{\ell}(x)\|^2 dx + \int_{\mathbb{R}^d} \hat{\ell}(x)^\top \nabla p(x) dx \\
&= \frac{1}{2}\int_{\mathbb{R}^d} p(x)\|\ell(x)\|^2 dx + \frac{1}{2}\int_{\mathbb{R}^d} p(x)\|\hat{\ell}(x)\|^2 dx - \int_{\mathbb{R}^d} (\nabla\cdot\hat{\ell})(x)p(x)dx,
\end{aligned}$$

by integration by parts, where $(\nabla \cdot \hat{\ell})(x) = \sum_{i=1}^d \frac{\partial \hat{\ell}}{\partial x_i}(x)$ is the divergence of $\hat{\ell}$ at x [Arfken, 1985].

The first part of the last right hand side does not depend on $\hat{\ell}$ while the two other parts are expectations under $p(x)$ of quantities that only depend on $\hat{\ell}$. Thus is can we well approximated, up to a constant, by:

$$\hat{\mathcal{R}}_{\text{score}}(\hat{\ell}) = \frac{1}{2n}\sum_{i=1}^{n}\|\hat{\ell}(x_i)\|^2 - \frac{1}{n}\sum_{i=1}^{n}(\nabla \cdot \hat{\ell})(x_i).$$

Parametric assumption. If we assume that the score is linear combination of finitely many basis functions, we will get a consistent estimator of these parameters. That is, we make the following assumption:

(A4) The score function $\ell(x)$ is a linear combination of known basis functions $\psi^j(x)$, $j = 1, \ldots, m$, where $\psi^j : \mathbb{R}^d \to \mathbb{R}^d$, that is $\ell(x) = \sum_{j=1}^{m} \psi^j(x)\theta_j^*$, for some $\theta^* \in \mathbb{R}^m$. We assume that the score function and its derivatives are squared-integrable with respect to $p(x)$.

In this chapter, we consider for simplicity a parametric assumption for the score. In order to go towards non-parametric estimation, we would need to let the number m of basis functions to grow with n (exponentially with no added assumptions), and this is an interesting avenue for future work. In simulations in 2.5, we consider a simple set of basis function which are localized functions around observations; these can approximate reasonably well in practice most densities and led to good estimation of the e.d.r. subspace. Moreover, if we have the additional knowledge that the components of x are statistically independent (potentially after linearly transforming them using independent component analysis [Hyvärinen et al., 2004]), we can use separable functions for the scores.

We introduce the notation

$$\Psi(x) = \begin{pmatrix} \psi_1^1(x) & \cdots & \psi_d^1(x) \\ \vdots & \ddots & \vdots \\ \psi_1^m(x) & \cdots & \psi_d^m(x) \end{pmatrix} \in \mathbb{R}^{m \times d},$$

so that the score function $\hat{\ell}$ we wish to estimate has the form

$$\hat{\ell}(x) = \Psi(x)^\top \theta \in \mathbb{R}^d.$$

We also introduce the notation $(\nabla \cdot \Psi)(x) = \begin{pmatrix} (\nabla \cdot \Psi^1)(x) \\ \vdots \\ (\nabla \cdot \Psi^m)(x) \end{pmatrix} \in \mathbb{R}^m$, so that $(\nabla \cdot \hat{\ell})(x) = \theta^\top (\nabla \cdot \Psi)(x)$.

The empirical score function may then be written as:

$$\hat{\mathcal{R}}_{\text{score}}(\theta) = \frac{1}{2}\theta^\top \big(\frac{1}{n}\sum_{i=1}^{n}\Psi(x_i)\Psi(x_i)^\top\big)\theta - \theta^\top\big(\frac{1}{n}\sum_{i=1}^{n}(\nabla \cdot \Psi)(x_i)\big), \tag{2.4.1}$$

which is a quadratic function of θ and can thus be minimized by solving a linear system in running time $O(m^3 + m^2dn)$ (to form the matrix and to solve the system).

Given standard results regarding the convergence of $\frac{1}{n}\sum_{i=1}^n \Psi(x_i)\Psi(x_i)^\top \in \mathbb{R}^{m\times m}$ to its expectation and of $\frac{1}{n}\sum_{i=1}^n (\nabla\cdot\hat{\ell})(x_i) \in \mathbb{R}^m$ to its expectation, we get a $\sqrt{n}$-consistent estimation of θ^* under simple assumptions (see Theorem 2.2).

2.4.2 Score matching for sliced inverse regression: two-step approach

We can now combine our linear parametrization of the score with the SIR approach outlined in Section 2.3. The true conditional expectation is

$$\mathbb{E}\big(\ell(x)|y\big) = \sum_{j=1}^m \mathbb{E}\big(\psi^j(x)|y\big)\theta_j^*,$$

and belongs to the e.d.r. subspace. We consider H different slices $I_1,\dots,I_H$, and the following estimator, which simply replaces the true score by the estimated score (i.e., θ^* by θ), and highlights the dependence in θ.

The estimator $\hat{\mathcal{V}}_{1,\text{cov}}$ can be rewritten as

$$\hat{\mathcal{V}}_{1,\text{cov}} = \sum_{i=1}^n\sum_{j=1}^n \frac{\alpha_{i,j}}{n(|I_h(i,j)|-1)}\ell(x_i)\ell(x_j)^\top,$$

where

$$\alpha_{i,j} = \begin{cases} 1 & \text{if } i\neq j \text{ in the same slice} \\ 0 & \qquad\qquad \text{otherwise} \end{cases}$$

Using linear property $\ell(x) = \Psi(x)^\top\theta$:

$$\hat{\mathcal{V}}_{1,\text{cov}}(\theta) = \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n \frac{\alpha_{i,j}}{|I_h(i,j)|-1}\Psi(x_i)^\top\theta\theta^\top\Psi(x_i). \tag{2.4.2}$$

In the two-step approach, we first solve the score matching optimization problem to obtain an estimate for the optimal parameters θ^* and then use them to get the k largest eigenvectors of covariance matrix $\hat{\mathcal{V}}_1$. This approach works well in low dimensions when the score function can be well approximated; otherwise, we may suffer from the curse of dimensionality: if we want a good approximation of the score function, we would need an exponential number of basis functions. In Section 2.4.3, we consider a direct approach aiming at improving robustness.

Consistency. Let also provide the result of consistency in the case of unknown score function under Assumption **(A4)** for the two-step algorithm. We will make the additional assumptions:

(M1) The random variables $\Psi_i^j(x) \in \mathbb{R}$ are sub-Gaussian, i.e., such that $\mathbb{E}e^{t(\Psi_i^j(x)-E\Psi_i^j(x))} \leqslant e^{\tau_\Psi^2 t^2/2}$, for some $\tau_\Psi > 0$.

(M2) The random variables $(\nabla\cdot\Psi^i)(x)$ are sub-Gaussian, i.e., such that $\mathbb{E}e^{t((\nabla\cdot\Psi^i)(x)-E(\nabla\cdot\Psi^i)(x)} \leqslant e^{\tau_{\nabla\Phi}^2 t^2/2}$, for some $\tau_{\nabla\Psi} > 0$.

(M3) The matrix $\mathbb{E}\big[\Psi(x)\Psi(x)^\top\big]$ is not degenerated and we let λ_0 denote its minimal eigenvalue.

Theorem 2.2. *Let $\hat\theta$ be the estimated θ, obtained on the first step of the algorithm. Under Assumptions **(A1)**, **(A2)**, **(A4)**, **(L1)** - **(L4)** and **(M1)**, **(M2)**, **(M3)**, for $\delta \leqslant 1/n$ and n large enough, that is, $n > c_1 + c_2 \log\frac{1}{\delta}$ for some positive constants c_1 and c_2 not depending on n and δ:*

$$\|\mathcal{V}_{1,cov}(\theta^*) - \hat{\mathcal{V}}_{1,cov}(\hat\theta)\|_* \leqslant \frac{1}{\sqrt{n}}\sqrt{2\log\frac{48m^2d}{\delta}} \times$$

$$\left[d\sqrt{d}(195\tau_\eta^2 + 2\tau_\ell^2) + \frac{9m}{2}\mathbb{E}\|\Psi(x)\|_2^2 \cdot \|\theta^*\| \left(\frac{2\tau_{\nabla\Psi}\sqrt{m}}{\lambda_0} + \frac{4\|b\|\tau_\Psi^2\sqrt{m^3d}}{\lambda_0}\right)\right] + O\left(\frac{1}{n}\right).$$

with probability not less than $1-\delta$.

Now, let us formulate and proof non-asymptotic result about the real and the estimated e.d.r. spaces. We need to define a notion of distance between subspaces spanned by two sets of vectors. We use the square trace error $R^2(w,\hat w)$ [Hooper, 1959]:

$$R^2(w,\hat w) = 1 - \frac{1}{k}\mathrm{tr}\big[w \cdot \hat w\big], \tag{2.4.3}$$

where $w \in \mathbb{R}^{k\times d}$ and $\hat w \in \mathbb{R}^{k\times d}$ both have orthonormal columns. It is always between zero and one, and equal to zero if and only if $w = \hat w$. This distance is closely related to *principal angles* notion:

$$\sin\Theta(\hat w^\top w) = \mathrm{diag}(\sin(\cos^{-1}\sigma_1), \ldots, \sin(\cos^{-1}\sigma_d),$$

where $\sigma_1, \ldots, \sigma_d$ are the singular values of the matrix $\hat w^\top w$. Actually, $R(w,\hat w)\cdot\sqrt{k} = \|\sin\Theta(\hat w^\top w)\|_F$.

We use Davis-Kahan "$\sin\theta$ theorem" [Stewart and Sun, 1990, Theorem V.3.6] in the following form [Yu et al., 2015, Theorem 2]:

Theorem 2.3. *Let Σ and $\hat\Sigma \in \mathbb{R}^{d\times d}$ be symmetric, with eigenvalues $\lambda_1 \geqslant \cdots \geqslant \lambda_d$ and $\hat\lambda_1, \ldots, \hat\lambda_d$ respectively. Fix $1 \leqslant r \leqslant s \leqslant d$, let $k = s - r + 1$, and let $U = (u_r, u_{r+1}, \ldots, u_s) \in \mathbb{R}^{d\times k}$ and $\hat U = (\hat u_r, \hat u_{r+1}, \ldots, \hat u_s) \in \mathbb{R}^{d\times k}$ have orthonormal columns satisfying $\Sigma u_j = \lambda_j u_j$ and $\hat\Sigma \hat u_j = \hat\lambda_j \hat u_j$ for $j = r, \ldots, s$. Let $\delta = \inf\{|\hat\lambda - \lambda| : \lambda \in [\lambda_s, \lambda_r],\ \hat\lambda \in (-\infty, \hat\lambda_{s+1}] \cup [\hat\lambda_{r-1}, +\infty)$, where $\hat\lambda_0 = +\infty$ and $\hat\lambda_{p+1} = -\infty$. Assume, that $\delta > 0$, then:*

$$R(U,\hat U) \leqslant \frac{\|\hat\Sigma - \Sigma\|_F}{\delta\sqrt{k}}. \tag{2.4.4}$$

Using Corollary 4.1 and its discussion by Vu and Lei [2013], we can derive, that $R(w,\hat w) \leqslant \dfrac{\|\mathcal{V}_1(\theta^*) - \hat{\mathcal{V}}_1(\hat\theta)\|_F}{|\lambda_k - \hat\lambda_{k+1}|\sqrt{k}}$, where w and $\hat w$ are the real and the estimated e.d.r. spaces respectively. Using Weyl's inequality [Stewart and Sun, 1990] : if $\|\mathcal{V}_1(\theta^*) -$

$\hat{\mathcal{V}}_1(\hat{\theta})\|_2 < \varepsilon \ \Rightarrow |\hat{\lambda}_i - \lambda_i| < \varepsilon$ for all $i = 1, \dots, d$ and taking $\varepsilon = (\lambda_k - \lambda_{k+1})/2 = \lambda_k/2$ we get:

$$R(w, \hat{w}) \leqslant \frac{2\|\mathcal{V}_1(\theta^*) - \hat{\mathcal{V}}_1(\hat{\theta})\|_*}{\lambda_k \sqrt{k}}, \text{ if } \|\mathcal{V}_1(\theta^*) - \hat{\mathcal{V}}_1(\hat{\theta})\|_F < \lambda_k/2,$$

because $\|\cdot\|_F \leqslant \|\cdot\|_*$. We can now formulate our main theorem for the analysis of the SADE algorithm:

Theorem 2.4. *Consider assumptions **(A1)**, **(A2)**, **(A4)**, **(L1)**- **(L4)**, **(M1)**, **(M2)**, **(M3)** and:*

(K1) *The matrix $\mathcal{V}_1$ has a rank k: $\lambda_k > 0$.*

For $\delta \leqslant 1/n$ and n large enough: $n > c_1 + c_2 \log \frac{1}{\delta}$ for some positive constants c_1 and c_2 not depending on n and δ:

$$R(\hat{w}, w) \leqslant \frac{2}{\sqrt{n}\lambda_k\sqrt{k}} \sqrt{2 \log \frac{48m^2 d}{\delta}} \times$$

$$\left[d\sqrt{d}(195\tau_\eta^2 + 2\tau_\ell^2) + \frac{9m}{2}\mathbb{E}\|\Psi(x)\|_2^2 \cdot \|\theta^*\| \left(\frac{2\tau_{\nabla\Psi}\sqrt{m}}{\lambda_0} + \frac{4\|b\|\tau_\Psi^2\sqrt{m^3 d}}{\lambda_0} \right) \right] + O\left(\frac{1}{n}\right).$$

with probability not less than $1 - \delta$.

We thus obtain a convergence rate in $1/\sqrt{n}$, with an explicit dependence on all constants of the problem. The dependence on dimension d and number of basis functions m is of the order (forgetting logarithmic terms): $O(\frac{d^{3/2}}{n^{1/2}} + \frac{m^{5/2}}{n^{1/2}}\|\theta^*\|)$. For $k > 1$, our dependence on the sample size n is improved compared to existing work such as [Dalalyan et al., 2008] (while it matches the dependence for $k = 1$ with [Hristache et al., 2001]), but this comes at the expense of assuming that the score functions satisfy a parametric model.

For a fixed number of basis functions m, we thus escape the curse of dimensionality as we get a polynomial dependence on d (which we believe could be improved). However, when no assumptions are made on score functions, the number m and potentially the norm $\|\theta^*\|$ has to grow with the number of observations n. We are indeed faced with a traditional non-parametric estiamation problem, and the number m will need to grow when n grows depending on the smoothness assumptions we are willing to make on the score function, with an effect on θ^* (and probably λ_0, which we will neglect in the discussion below). While a precise analysis is out of the scope of the chapter and left for future work, we can make an informal argument as follows: in order to approximate the score with precision ε with a set of basis function where $\|\theta^*\|$ is bounded, we need $m(\varepsilon)$ basis functions. Thus, we end up with two sources of errors, an approximation error ε and an estimation error of order $O(m(\varepsilon)^{5/2}/n^{1/2})$. Typically, if we assume that the score has a number of bounded derivatives proportional to dimension or if we assume the input variables are independent and we can thus estimate the score *independently* for each dimension, $m(\varepsilon)$ is of the order $1/\varepsilon^{1/r}$, where r is independent of the dimension, leading to an overall rate which is independent of the dimension.

2.4.3 Score matching for SIR: direct approach

We can also try to combine these two steps to try to avoid the "curse of dimensionality". Our estimation of the score, i.e., of the parameter θ is done only to be used within the SIR approach where we expect the matrix $\hat{\mathcal{V}}_{1,\text{cov}}$ to have rank k. Thus when estimating θ by minimizing $\hat{\mathcal{R}}_{\text{score}}(\theta)$, we may add a regularization that penalizes large ranks for $\hat{\mathcal{V}}_{1,\text{cov}}(\theta) = \frac{1}{n}\sum_{i,j=1}^{n} \frac{\alpha_{i,j}}{|I_h(i,j)|-1} \Psi(x_i)^\top \theta\theta^\top \Psi(x_i)$, where we highlight the dependence on $\theta \in \mathbb{R}^m$. By enforcing the low-rank constraint, our aim is to circumvent a potential poor estimation of the score function, which could be enough for the task of estimating the e.d.r. space (we see a better behavior in our simulations in 2.5).

Introduce matrix $\mathcal{L}(\theta) = (\Psi(x_i)^\top\theta, \ldots, \Psi(x_n)^\top\theta) \in \mathbb{R}^{d\times n}$ and $A \in \mathbb{R}^{n\times n}$ with $A_{i,j} = \frac{\alpha_{i,j}}{n\cdot(|I_h(i,j)|-1)}$ and $\mathcal{A}(\theta) = \mathcal{L}\cdot A^{1/2} \in \mathbb{R}^{d\times n}$.

We may then penalize the nuclear norm of $\mathcal{A}(\theta)$, or potentially consider norms that take into account that we look for a rank k (e.g., the k-support norm on the spectrum of $\mathcal{A}(\theta)$ [McDonald et al., 2014]). We have,

$$\|\mathcal{A}(\theta)\|_* = \operatorname{tr}(\mathcal{A}(\theta)\mathcal{A}(\theta)^\top)^{1/2} = \operatorname{tr}\big[\hat{\mathcal{V}}_{1,\text{cov}}(\theta)^{1/2}\big].$$

Combining two penalties, we have a convex optimization task:

$$\hat{\mathcal{R}}(\theta) = \hat{\mathcal{R}}_{\text{score}}(\theta) + \lambda\cdot\operatorname{tr}\big[\hat{\mathcal{V}}_{1,\text{cov}}(\theta)^{1/2}\big]. \tag{2.4.5}$$

Efficient algorithm. Following Argyriou et al. [2008], we consider reweighted least-squares algorithms. The trace norm admits the variational form:

$$\|W\|_* = \frac{1}{2}\inf_{D\succ 0} \operatorname{tr}(W^T D^{-1} W + D).$$

The optimization problem (2.4.5) can be reformulated in the following way:

$$\theta \leftarrow \arg\min_{\theta}\ \hat{\mathcal{R}}_{\text{score}}(\theta) + \frac{\lambda}{2}\operatorname{tr}\Big(\mathcal{A}(\theta)^\top D^{-1}\mathcal{A}(\theta)\Big) \quad \text{and}$$
$$D \leftarrow \big(\mathcal{A}(\theta)\mathcal{A}(\theta)^\top + \varepsilon I_d\big)^{1/2}.$$

Note that the objective function is a quadratic function of θ. Decompose matrix $\mathcal{A}$ in the form:

$$\mathcal{A}(\theta) = \sum_{k=1}^{m}\theta_k\mathcal{A}_k, \quad \text{where } \mathcal{A}_k \in \mathbb{R}^{d\times n}.$$

Rearrange the regularizer term:

$$\operatorname{tr}\Big(\mathcal{A}(\theta)^\top D^{-1}\mathcal{A}(\theta)\Big) = \sum_{k=1}^{m}\sum_{l=1}^{m}\operatorname{tr}\Big[\theta_k\mathcal{A}_k^\top D^{-1}\mathcal{A}_l\theta_l\Big] = \theta^\top\mathcal{Y}\theta,$$

where

$$\mathcal{Y}_{k,l} = \operatorname{tr}\big[\mathcal{A}_k^\top D^{-1}\mathcal{A}_l\big].$$

Introduce notation:

$$\widetilde{\mathcal{A}} = \Big(\operatorname{vect}(\mathcal{A}_1), \ldots, \operatorname{vect}(\mathcal{A}_m)\Big) \in \mathbb{R}^{dn\times m}$$

$$\widetilde{\mathcal{B}} = \Big(\operatorname{vect}(D^{-1}\mathcal{A}_1), \ldots, \operatorname{vect}(D^{-1}\mathcal{A}_m)\Big) \in \mathbb{R}^{dn\times m},$$

then

$$\mathcal{Y} = \widetilde{\mathcal{A}}^\top \widetilde{\mathcal{B}}.$$

We can now estimate the complexity of this algorithm. Firstly we need to evaluate the quadratic form in $\hat{\mathcal{R}}_{\text{score}}(\theta)$: it is a summation of n multiplications of $m \times d$ and $d \times m$ matrices. Complexity of this step is $O(nm^2d)$. Secondly, we need to evaluate matrix matrices D^{-1}, $\widetilde{\mathcal{A}}$ and $\widetilde{\mathcal{B}}$ using $O(d^3)$, $O(dHm)$ and $O(m \times d^2H)$ operations respectively. Next, we need to evaluate matrix $\mathcal{Y}$, using $O(dnm^2)$ operations. Finally, evaluating $\mathcal{A}(\theta)$ requires $O(dnm)$ operations, $\mathcal{A}(\theta)\mathcal{A}(\theta)^\top$ requires $O(d^2n)$ operations and evaluation of D requires $O(d^3)$ operations. Combining complexities, we get $O(md^2n + d^3 + nm^2d)$ operations, which is still linear in n.

2.5 Experiments

In this section we provide numerical experiments for SADE, PHD+ and SPHD on different functions f. We denote the true and estimated e.d.r. subspaces as $\mathcal{E}$ and $\hat{\mathcal{E}}$ respectively, defined from w and $\hat{w}$.

2.5.1 Known score functions

Consider a Gaussian mixture model with 2 components in $\mathbb{R}^d$:

$$p(x) = \sum_{i=1}^{2} \theta_i \cdot \frac{1}{(2\pi)^{d/2} \cdot |\Sigma_i|^{1/2}} \cdot \exp\left\{ -\frac{1}{2}(X-\mu_i)^\top \Sigma_i^{-1}(X-\mu_i)\right\}, \tag{2.5.1}$$

where $\theta = (6/10, 4/10)$, $\mu_1 = (\underbrace{-1,\ldots,-1}_{d})$, $\mu_2 = (\underbrace{1,\ldots,1}_{d})$, $\Sigma_1 = I_d$, $\Sigma_2 = 2 \cdot I_d$. Contour lines of this distribution, when $d = 2$ are shown in Figure 2-2.

The error ε has a standard normal distribution. To estimate the effectiveness of an estimated e.d.r. subspace, we use the square trace error $R^2(w, \hat{w})$

$$R^2(w, \hat{w}) = 1 - \frac{1}{k}\operatorname{tr}\big[P \cdot \hat{P}\big],$$

where w and $\hat{w}$ are the real and the estimated e.d.r. vectors respectively and P and $\hat{P}$ are projectors, corresponding to these matrices. Note, that (2.4.3) is the special case of this formula with orthonormal matrices.

To show the dominance of SADE over SIR (which should only work for elliptically

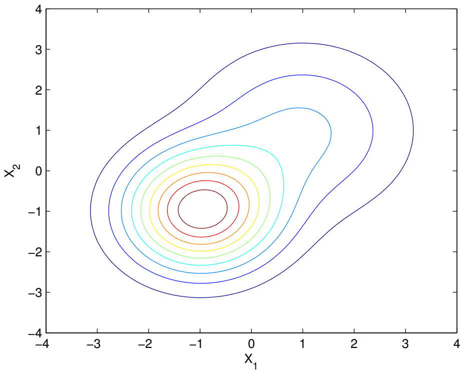

Figure 2-2 – Contour lines of Gaussian mixture pdf in 2D case.

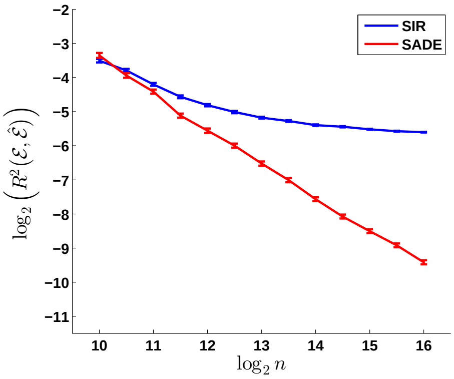

Figure 2-3 – Mean and standard deviation of $R^2(\mathcal{E},\hat{\mathcal{E}})$ divided by 10 for the rational function (2.5.2).

symmetric distributions), we consider the rational multi-index model of the form

$$d = 10; \quad y = \frac{x_1}{1/2 + (x_2+2)^2} + \sigma \cdot \varepsilon. \tag{2.5.2}$$

Here the real e.d.r. space is generated by 2 specific vectors: $(1, 0, 0, \ldots, 0)$ and $(0, 1, 0, \ldots, 0)$. The number of slices is $H = 10$, and we consider numbers of observations n from 2^{10} to 2^{16} equally spaced in logarithmic scale, and we conduct 100 replicates to obtain means and standard deviations of the logarithms of square trace errors divided by $\sqrt{100} = 10$ (to assess significance of differences) as shown in Figure 2-3. Even in this simple model, the ordinary SIR algorithm does not work properly, because the distribution of the inputs x has no elliptical symmetry. When $n \to \infty$, the squared trace error tends to some nonzero constant depending on the properties of the density function, whereas SADE shows good performance with slope -1 (corresponding to a $\sqrt{n}$-consistent estimator).

Now, we compare the moments methods SADE, PHD+, SPHD. Although the goal of this chapter is to compare moment matching techniques, we compare them with the state-of-the-art MAVE method [Xia et al., 2002b, Wang and Xia, 2007, 2008]. It is worth noting two properties which we have already discussed concerning these methods:

1. Sliced methods have a wider application area than unsliced ones: SADE is stronger than ADE and SPHD is stronger than PHD+.
2. SADE can not recover the entire e.d.r. space in several cases, for example, a classification task or symmetric cases. For those cases, we should use second-order methods (i.e., methods based on $\mathcal{S}_2$).
3. MAVE works better, but the goal of this chapter is to compare moment-matching techniques. In Figure 2-12, we provide examples where MAVE suffers from the curse of dimensionality and performs worse than moment-matching methods.

We conduct 3 experiments, where $H = 10$, $d = 10$, the error term ε has a normal standard distribution; numbers of observation n from 2^{10} to 2^{16} equally spaced in logarithmic scale and we made 10 replicas to evaluate sample means and variations:

— Rational model of the form:

$$y = \sum_{i=1}^{k} \tanh(8x_i - 16) \cdot i + \tanh(8x_i + 16) \cdot (k + 1 - i) + \varepsilon/4, \qquad (2.5.3)$$

where the effective reduction subspace dimension is $k = 2$.
Results are shown of Figure 2-4 and we can see, that first-order method SADE works better, than second-order SPHD + and SPHD works better, than PHD+, that is slicing make the method more robust.

— Classification problem of the form

$$y = 1_{x_1^2+2x_2^2>4} + \varepsilon/4. \qquad (2.5.4)$$

We can see, that the error of SADE is close to 0.5 (Figure 2-5). This means that the method finds only one direction in the e.d.r. space. Moreover, SPHD gave better results than PHD+.

— Quadratic model of the form

$$d = 10; \quad y = x_1(x_1 + x_2 - 3) + \varepsilon. \qquad (2.5.5)$$

Both SADE and SPHD show a good performance (Figure 2-6), while PHD+ can not recover the desired projection due to linearity of the function g.

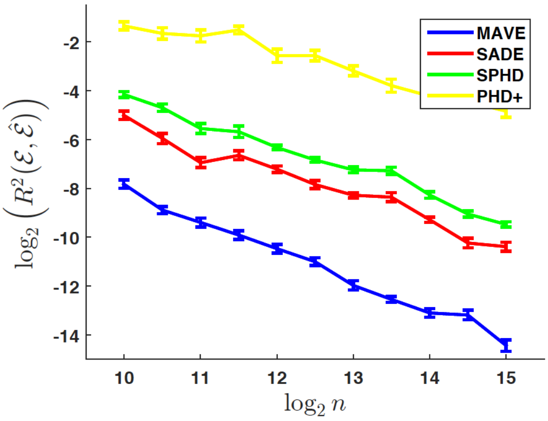

Figure 2-4 – Mean and standard deviation of $R^2(\mathcal{E}, \hat{\mathcal{E}})$ divided by $\sqrt{10}$ for the rational function (2.5.3) with $\sigma = 1/4$.

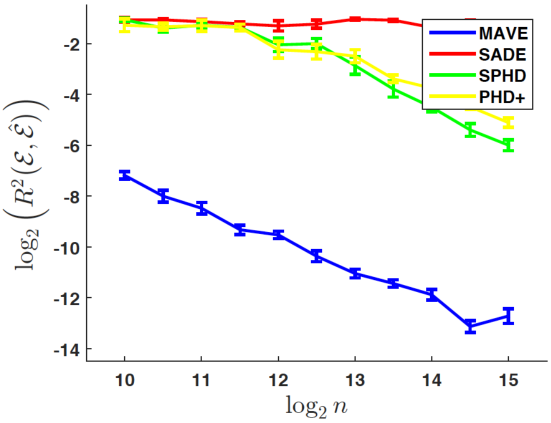

Figure 2-5 – Mean and standard deviation of $R^2(\mathcal{E}, \hat{\mathcal{E}})$ divided by $\sqrt{10}$ for the classification problem (2.5.4).

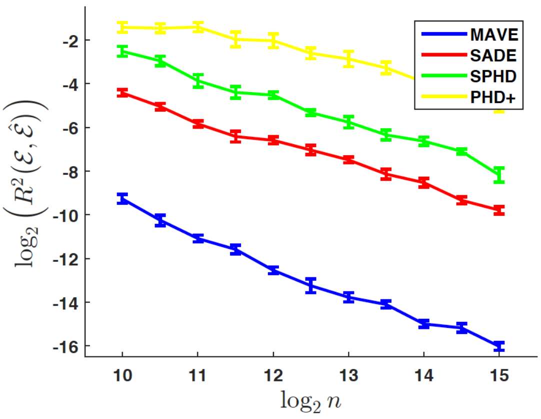

Figure 2-6 – Mean and standard deviation of $R^2(\mathcal{E},\hat{\mathcal{E}})$ divided by $\sqrt{10}$ for the quadratic model (2.5.2) with $\sigma = 1/4$.

2.5.2 Unknown score functions

Now, we conduct numerical experiments for SADE with unknown score functions. We consider a "toy" experiment and first examine results of the 2-step algorithm. We consider again a Gaussian mixture model (2.5.1) with 2 components in $\mathbb{R}^2$, where $\theta = (0.6, 0.4)$, $\mu_1 = (-0.5, 0.5), \mu_2 = (0.5, 0.5)$, $\Sigma_1 = 0.3 \cdot I_2$, $\Sigma_2 = 0.4 \cdot I_2$, with $y = \sin(x_1 + x_2) + \varepsilon$, where error ε has a standard normal distribution. We choose these parameters of Gaussian mixture to make sure, that the probability of the vector X to be in the square $[-2, 2]^2$ is close to 1.

We chose 100 Gaussian kernels as basis functions:

$$\psi_{i,j}(x) = \nabla \exp\left\{ -\frac{\|X - X_{i,j}\|^2}{2h^2} \right\}, 1 \leqslant i, j \leqslant 10,$$

where $X_{i,j}$ form the uniform grid on the square $[-2, 2]^2$ (note that this does not imply any notion of Gaussianity for the underlying density, and the Gaussian kernel here could be replaced by any differentiable local function).

In practice, $\hat{\Psi}$ in (2.4.1) is close to be degenerated, and we use a regularized estimator $\theta^* = -\left(\hat{\Psi} + \alpha I_T\right)^{-1} \cdot \hat{\Phi}$ instead.

We choose $\alpha = 0.01$, $\sigma = 1$ and conduct 100 replicates with numbers of observations n from 2^{10} to 2^{16} equally spaced in logarithmic scale. The results of the experiments are presented in Figure 2-7: the square trace error tends to zero as n increases.

In high dimensions, we can not use a uniform grid because of the curse of dimensionality. Instead of this, we will use n Gaussian kernels, centered in the sample points X_i. Note here that we can only recover an approximate score function, but our goal is the estimation of the e.d.r. subspace.

We use our reweighted least-squares algorithm to solve the convex problem (2.4.5). We conduct several experiments on functions from the previous section.

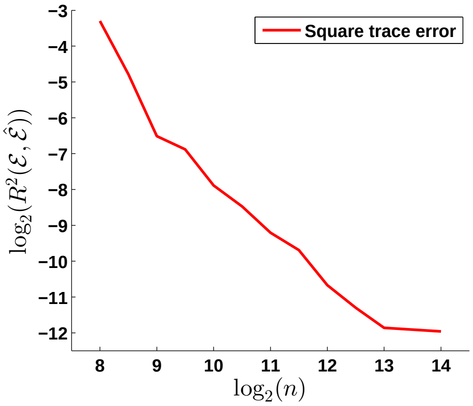

Figure 2-7 – Square trace error $R(\mathcal{E},\hat{\mathcal{E}})$ for different sample sizes.

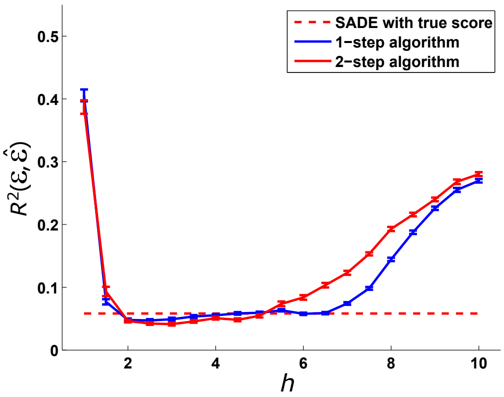

Figure 2-8 – Quadratic model, $d = 10$, $n = 1000$.

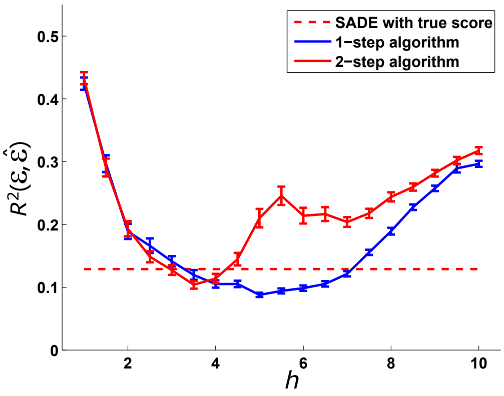

Figure 2-9 – Rational model, $d = 10$, $n = 1000$.

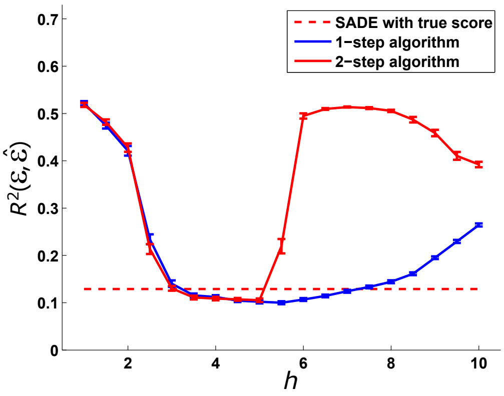

Figure 2-10 – Rational model, $d = 20$, $n = 2000$.

We plot the performance as a function of the kernel bandwidth h to assess the robustness of the methods. On Figure 2-8 we provide the relationship between the square trace error and h for quadratic model (2.5.5) for $d = 10$ and $n = 1000$. In Figures 2-9, 2-10, 2-11, we provide the relationship between the square trace error and σ for rational model (2.5.2) for $d = 10$, $n = 1000$; $d = 20, n = 2000$ and $d = 50, n = 5000$. We see that for large d, the one-step algorithm is more robust than the two-step algorithm. Moreover, the experiments show that even with weak score functions, correct estimation of the e.d.r. space can be performed.

Comparison with MAVE. We consider the rational model (2.5.3) with $n = 10000$, $k = 10$ and dimension d of data in the range $[20, 150]$. Probability density $p(x)$ has independent components, each of which has a form of mixture of 2 Gaussians with weights $(6/10, 4/10)$, means $(-1, 1)$ and standard variations $(1, 2)$. The results

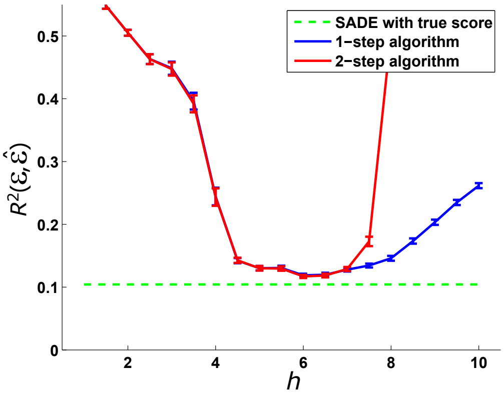

Figure 2-11 – Rational model, $d = 50$, $n = 5000$.

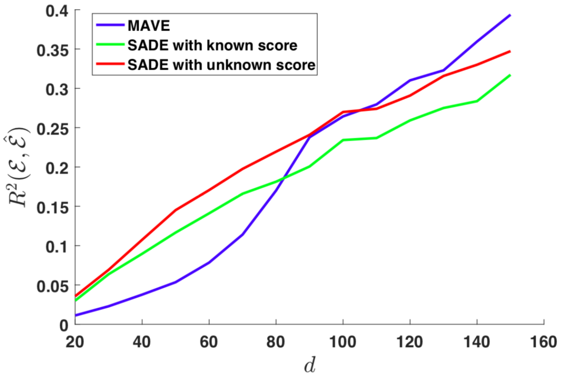

Figure 2-12 – Increasing dimensions, with rational model 2.5.3, $n = 10000$, $k = 10$, $d = 20$ to 150.

for MAVE, SADE with known score and for SADE with unknown score are shown on the Figure 2-12. We can see, that both SADE with known and unknown scores lose in case of low dimension of data, but more resistant to the curse of dimensionality (as expected as MAVE relies on non-parametric estimation in a space of dimension k, which is here larger). Moreover, the complexity of our moment-matching technique is linear in the number of observations n, while MAVE is superlinear.

2.6 Conclusion

In this chapter we consider a general non-linear regression model and the dependence on a unknown k-dimensional subspace assumption. Our goal was direct estimation of this unknown k-dimensional space, which is often called the effective dimension reduction or e.d.r. space. We proposed new approaches (SADE and SPHD), combining two existing techniques (sliced inverse regression (SIR) and score function-based estimation). We obtained consistent estimation for $k > 1$ only using the first-order score and proposed explicit approaches to learn the score from data.

It would be interesting to extend our sliced extensions to learning neural networks: indeed, our work focused on the subspace spanned by w and cannot identify individual columns, while Janzamin et al. [2015] showed that by using proper tensor decomposition algorithms and second-order score functions, columns of w can be consistently estimated (in polynomial time).

2.7 Appendix. Proofs

In this appendix, we provide proofs which were omitted in the chapter.

2.7.1 Probabilistic lemma

Let us first formulate and proof an auxiliary lemma about tail inequalities:

Lemma 2.5. *Let X be a non-negative random variable such that for some positive constants A and B, and all p:*

$$\mathbb{E}X^p \leqslant (A\sqrt{p} + Bn^{1/p}p^2)^p.$$

Then, for $t \geqslant (\log n)/2$:

$$\mathbb{P}(X \geqslant 3A\sqrt{t} + 3Bn^{1/t}t^2) \leqslant 3e^{-t}.$$

Proof. By Markov's inequality, for every non-negative integer p:

$$\mathbb{P}(X \geqslant 3A\sqrt{p} + 3Bn^{1/p}p^2) = \mathbb{P}(X^p \geqslant [3A\sqrt{p} + 3Bn^{1/p}p^2]^p) \leqslant$$

$$\leqslant \frac{[A\sqrt{p} + Bn^{1/p}p^2]^p}{[3A\sqrt{p} + 3Bn^{1/p}p^2]^p} \leqslant e^{-p}.$$

Consider any $t > (\log n)/2$ and $p = [t]$, then:

$$\mathbb{P}(X \geqslant 3A\sqrt{t} + 3Bn^{1/t}t^2) \leqslant \mathbb{P}(X \geqslant 3A\sqrt{p} + 3Bn^{1/p}p^2) \leqslant e^{-p} \leqslant 3e^{-t},$$

because function $f(t) = 3A\sqrt{t} + 3Bn^{1/t}t^2$ increases on $[(\log n)/2; +\infty)$. □

2.7.2 Proof of theorem 2.1

Proof. Let $\big(y_{(i)}, x_{(i)}\big)$, $i = 1, \dots, n$ be the ordered data set, where $y_{(1)} \leqslant y_{(2)} \leqslant \dots \leqslant y_{(n)}$. The $x_{(i)}$ are called the *concomitants* of order statistics by Yang [1977]. Introduce double subscripts: $\ell_{(h,1)} = \ell\big(x_{(2h-1)}\big)$, $\ell_{(h,2)} = \ell\big(x_{(2h)}\big)$, $y_{(h,1)} = y_{(2h-1)}$ and $y_{(h,2)} = y_{(2h)}$.

Introduce firstly alternative matrix (under the weak assumptoin of its existance):

$$\mathcal{V}_{1,\mathbb{E}} = \mathbb{E}\big[\mathrm{Cov}(\mathcal{S}_1(x)|y)\big].$$

We have the following estimator for this matrix, for $c = 2$, and $H = n/c = n/2$:

$$\hat{\mathcal{V}}_{1,\mathbb{E}} = \frac{1}{n}\sum_{h=1}^{H}\big(\ell_{(h,1)} - \ell_{(h,2)}\big)\big(\ell_{(h,1)} - \ell_{(h,2)}\big)^\top,$$

Firstly, we estimate a norm of $\hat{\mathcal{V}}_{1,\mathbb{E}} - \mathcal{V}_{1,\mathbb{E}}$ and afterwards a norm of $\hat{\mathcal{V}}_{1,\mathrm{cov}} - \mathcal{V}_{1,\mathrm{cov}}$.

Thus, the deviation from the population version, may be split into four terms as follows:

$$\begin{aligned}
\hat{\mathcal{V}}_{1,\mathbb{E}} - \mathcal{V}_{1,\mathbb{E}} &= \frac{1}{n}\sum_{h=1}^{H}\big(\ell_{(h,1)} - \ell_{(h,2)}\big)\big(\ell_{(h,1)} - \ell_{(h,2)}\big)^\top - \mathbb{E}\eta\eta^\top \\
&= \frac{1}{n}\sum_{h=1}^{H}\big(\eta_{(h,1)} - \eta_{(h,2)}\big)\big(\eta_{(h,1)} - \eta_{(h,2)}\big)^\top - \mathbb{E}\eta\eta^\top \\
&\quad + \frac{1}{n}\sum_{h=1}^{H}\big(m(y_{(h,1)}) - m(y_{(h,2)})\big)\big(\eta_{(h,1)} - \eta_{(h,2)}\big)^\top \\
&\quad + \frac{1}{n}\sum_{h=1}^{H}\big(\eta_{(h,1)} - \eta_{(h,2)}\big)\big(m(y_{(h,1)}) - m(y_{(h,2)})\big)^\top \\
&\quad + \frac{1}{n}\sum_{h=1}^{H}\big(m(y_{(h,1)}) - m(y_{(h,2)})\big)\big(m(y_{(h,1)}) - m(y_{(h,2)})\big)^\top \\
&= T_4 + T_3 + T_2 + T_1.
\end{aligned}$$

We now bound each term separately.

Bounding T_1. We have

$$\begin{aligned}
T_1 &\preccurlyeq \frac{1}{n}\sum_{h=1}^{H} L^2 (y_{(h,1)} - y_{(h,2)})(y_{(h,1)} - y_{(h,2)})^\top \\
\mathrm{tr} T_1 = \|T_1\|_* &\leqslant \frac{1}{n}\sum_{h=1}^{H} L^2 \cdot \mathrm{diameter}(y_1, \dots, y_n)|y_{(h,1)} - y_{(h,2)}| \\
&\leqslant \frac{1}{n} L^2 \cdot \mathrm{diameter}(y_1, \dots, y_n)^2.
\end{aligned}$$

The range cannot grow too much, i.e., as $\log n$. Indeed, assuming without loss of generality that $\mathbb{E}y = 0$, we have $\max\{y_1, \dots, y_n\} \leqslant u/2$ and $\min\{y_1, \dots, y_n\} \geqslant -u/2$ implies that the range is less than u, and thus, $\mathbb{P}(\mathrm{diameter}(y_1, \dots, y_n) \geqslant u) \leqslant \mathbb{P}(\max\{y_1, \dots, y_n\} \geqslant u/2) + \mathbb{P}(\min\{y_1, \dots, y_n\} \leqslant -u/2) \leqslant n\mathbb{P}(y > u/2) + n\mathbb{P}(y < -u/2) \leqslant 2n\exp(-u^2/8\tau_y^2)$ by using sub-Gaussianity. Then, by selecting $u^2/8\tau_y^2 = \log(2n) + \log(8/\delta)$, with get with probability greater then $1 - \delta/8$ that $\mathrm{diameter}(y_1, \dots, y_n) \leqslant 2\sqrt{2}\tau_y\sqrt{\log(2n) + \log(8/\delta)}$.

Bounding T_2 and T_3. We also have

$$\begin{aligned}
\max\{\|T_2\|_*, \|T_3\|_*\} &\leqslant \frac{1}{n}\sum_{h=1}^{H} L|y_{(h,1)}y_{(h,2)}| \cdot \mathrm{diameter}(\eta_1, \dots, \eta_n) \\
&\leqslant \frac{1}{n} L \cdot \mathrm{diameter}(y_1, \dots, y_n) \cdot \mathrm{diameter}(\eta_1, \dots, \eta_n).
\end{aligned}$$

Like for T_1, the ranges cannot grow too much, i.e., as $\log n$. Similarly

$$\mathbb{P}(\text{diameter}((\eta_j)_1, \ldots, (\eta_j)_n) \geqslant u) \leqslant$$

$$\leqslant n\mathbb{P}((\eta_j) > u/2) + n\mathbb{P}((\eta_j) < -u/2) \leqslant 2n \exp(-u^2/8\tau_\eta^2).$$

We thus with get with probability greater then $1 - \delta/(8d)$ that

$$\max_{j \in \{1,\ldots,d\}} \text{diameter}((\eta_j)_1, \ldots, (\eta_j)_n) \leqslant 2\sqrt{2}\tau_\eta \sqrt{\log(2n) + \log(8d/\delta)}.$$

Thus combining the two terms above, with probability greater than $1 - \delta/4$,

$$\|T_1\|_* + \|T_2\|_* + \|T_3\|_* \leqslant \frac{8L(L\tau_y^2 + 2\tau_\eta\tau_y\sqrt{d})}{n}\big(\log(2n) + \log(8d) + \log(1/\delta)\big).$$

Note the term in $\sqrt{d}$, which corresponds to the definition of the diameter$(\eta_1, \ldots, \eta_n)$ in terms of the ℓ_2-norm.

Bounding T_4. We have:

$$\begin{aligned} T_4 &= \frac{1}{n}\sum_{h=1}^{H}\big\{\eta_{(h,1)}\eta_{(h,1)}^\top + \eta_{(h,2)}\eta_{(h,2)}^\top - \eta_{(h,2)}\eta_{(h,1)}^\top - \eta_{(h,1)}\eta_{(h,2)}^\top\big\} - \mathbb{E}\eta\eta^\top \\ &= \frac{1}{n}\sum_{h=1}^{H}\big\{ - \eta_{(h,2)}\eta_{(h,1)}^\top - \eta_{(h,1)}\eta_{(h,2)}^\top\big\} + \frac{1}{n}\sum_{i=1}^{n}\eta_i\eta_i^\top - \mathbb{E}\eta\eta^\top = T_{4,1} + T_{4,2}. \end{aligned}$$

For the second term $T_{4,2}$ above, if we select any element indexed by a, b, then

$$\frac{1}{n}\sum_{i=1}^{n}(\eta_i)_a(\eta_i)_b - \mathbb{E}\eta_a\eta_b.$$

Using [Boucheron et al., 2013, Theorem 2.1], we get

$$\mathbb{E}[(\eta_i)_a(\eta_i)_b]^2 \leqslant \sqrt{\mathbb{E}(\eta_i)_a^4\mathbb{E}(\eta_i)_b^4} \leqslant 4(2\tau_\eta^2)^2 = 16\tau_\eta^4,$$

and

$$\mathbb{E}[|(\eta_i)_a(\eta_i)_b|^q] \leqslant \sqrt{\mathbb{E}(\eta_i)_a^{2q}\mathbb{E}(\eta_i)_b^{2q}} \leqslant 2q!(2\tau_\eta^2)^q = \frac{q!}{2}(2\tau_\eta^2)^{q-2}16\tau_\eta^4.$$

We can then use Bernstein's inequality [Boucheron et al., 2013, Theorem 2.10], to get that with probability less than e^{-t} then

$$\frac{1}{n}\sum_{i=1}^{n}(\eta_i)_a(\eta_i)_b - \mathbb{E}\eta_a\eta_b \geqslant 2\frac{\tau_\eta^2}{n}t + \sqrt{32\tau_\eta^4}\sqrt{t}/\sqrt{n}.$$

We can also get upper bound for this quantity, using $-\eta_a$ instead of η_a. Thus, with

$t = \log \frac{8d^2}{\delta}$, we get that all $d(d+1)/2$ absolute deviations are less than $2\tau_\eta^2\biggl(\frac{\log \frac{8d^2}{\delta}}{n} + \sqrt{\frac{2\log\frac{8d^2}{\delta}}{n}}\biggr)$, with probability greater than $1-\delta/4$. This implies that the nuclear norm of the second term is less than $2d\sqrt{d}\tau_\eta^2\biggl(\frac{\log \frac{8d^2}{\delta}}{n} + \sqrt{\frac{2\log\frac{8d^2}{\delta}}{n}}\biggr)$, because for any matrix $K \in \mathbb{R}^{d\times d}:\ \|K\|_* \leqslant d\sqrt{d}\|K\|_\infty$, where $\|K\|_\infty = \max_{j,k}|K_{jk}|$.

For the first term, we consider $Z = \sum_{h=1}^{H}(\eta_{(h,2)})_a(\eta_{(h,1)})_b$, and consider conditioning on $\mathbf{Y} = (y_1, \dots, y_n)$. A key result from the theory of order statistics is that the n random variables $\eta_{(h,2)}$, $\eta_{(h,1)}$, $h \in \{1, \dots, n/2\}$ are independent given Y [Yang, 1977]. This allows us to compute expectations.

Using Rosenthal's inequality [Boucheron et al., 2013, Theorem 15.11] conditioned on $\mathbf{Y}$, for which we have $\mathbb{E}((\eta_{(h,2)})_a(\eta_{(h,1)})_b|\mathbf{Y}) = 0$, we get:

$$\bigl[\mathbb{E}(|Z|^p|\mathbf{Y})\bigr]^{1/p} \leqslant$$

$$\begin{aligned}
&\leqslant \sqrt{8p}\bigl[\sum_h \mathbb{E}\bigl[((\eta_{(h,2)})_a^2(\eta_{(h,1)})_b^2|\mathbf{Y}\bigr]\bigr]^{1/2} + p\cdot 2\bigl[\mathbb{E}\max_h\bigl[((\eta_{(h,2)})_a^p(\eta_{(h,1)})_b^p|\mathbf{Y}\bigr]\bigr]^{1/p}\\
&\leqslant \sqrt{8p}\bigl[\sum_h \mathbb{E}\bigl[((\eta_{(h,2)})_a^2(\eta_{(h,1)})_b^2|\mathbf{Y}\bigr]\bigr]^{1/2} + p\cdot 2\bigl[\sum_h \mathbb{E}\bigl[((\eta_{(h,2)})_a^p(\eta_{(h,1)})_b^p|\mathbf{Y}\bigr]\bigr]^{1/p}.
\end{aligned}$$

By taking the p-th power, we get:

$$\begin{aligned}
\mathbb{E}(|Z|^p|\mathbf{Y}) &\leqslant 2^{p-1}\sqrt{8p}^p\bigl[\sum_h \mathbb{E}\bigl[((\eta_{(h,2)})_a^2(\eta_{(h,1)})_b^2|\mathbf{Y}\bigr]\bigr]^{p/2}\\
&+ 2^{p-1}p^p\cdot 2^p\sum_h \mathbb{E}\bigl[((\eta_{(h,2)})_a^p(\eta_{(h,1)})_b^p|\mathbf{Y}\bigr].
\end{aligned}$$

By now taking expectations with respect to $\mathbf{Y}$, we get, using Jensen's inequality:

$$\begin{aligned}
\mathbb{E}|Z|^p &\leqslant 2^{p-1}\sqrt{8p}^p\mathbb{E}\Bigl(\bigl[\sum_h(\eta_{(h,2)})_a^2(\eta_{(h,1)})_b^2\bigr]^{p/2}\Bigr)\\
&+ 2^{p-1}p^p\cdot 2^p\sum_h \mathbb{E}\bigl[((\eta_{(h,2)})_a^p(\eta_{(h,1)})_b^p\bigr]\\
&\leqslant 2^{p-1}\sqrt{8p}^p\mathbb{E}\Bigl(\bigl[\frac{1}{2}\sum_h(\eta_{(h,2)})_a^4 + (\eta_{(h,1)})_b^4\bigr]^{p/2}\Bigr)\\
&+ 2^{p-1}p^p\cdot 2^p\cdot\frac{1}{2}\sum_h \mathbb{E}\bigl[((\eta_{(h,2)})_a^{2p} + (\eta_{(h,1)})_b^{2p}\bigr]\\
&\leqslant 2^{p-1}\sqrt{8p}^p\mathbb{E}\Bigl(\bigl[\sum_i(\eta_i)_a^4 + (\eta_i)_b^4\bigr]^{p/2}\Bigr)\\
&+ 2^{p-1}p^p\cdot 2^p\sum_i \mathbb{E}\bigl[((\eta_i)_a^{2p} + (\eta_i)_b^{2p}\bigr].
\end{aligned}$$

Because summing over all order statistics is equivalent to summing over all elements. Thus, using the bound on moments of $(\eta_i)_b^2$, we get:

$$\begin{aligned}\mathbb{E}|Z|^p &\leqslant 2^{p-1}\sqrt{8p}^p 2^{p/2-1}\mathbb{E}\Big(\big[\sum_i (\eta_i)_a^4\big]^{p/2}\Big)\\ &+ 2^{p-1}\sqrt{8p}^p 2^{p/2-1}\mathbb{E}\Big(\big[\sum_i (\eta_i)_b^4\big]^{p/2}\Big) + 2^{p-1}p^p \cdot 2^p n \cdot 4p!(2\tau_\eta^2)^p\end{aligned}$$

We can now use [Boucheron et al., 2013, Theorem 15.10], to get

$$\begin{aligned}\Big[\mathbb{E}\Big(\big[\sum_i (\eta_i)_a^4\big]^{p/2}\Big)\Big]^{2/p} &\leqslant 2\mathbb{E}\big[\sum_i (\eta_i)_a^4\big] + \frac{p}{2}\big(\mathbb{E}\big[\max_i((\eta_i)_a^4)^{p/2}\big]\big)^{2/p}\\ &\leqslant 2\mathbb{E}\big[\sum_i (\eta_i)_a^4\big] + \frac{p}{2}\big(\mathbb{E}\big[\sum_i((\eta_i)_a^4)^{p/2}\big]\big)^{2/p}\\ &\leqslant 2n \times 4(2\tau_\eta^2)^2 + \frac{p}{2}n^{2/p}\mathbb{E}\eta_i^{2p}\\ &\leqslant \big(32n + n^{2/p}\frac{p}{2}(2p!)^{2/p}\big)\tau_\eta^4\\ &\leqslant \big(32n + n^{2/p}p^3\big)\tau_\eta^4.\end{aligned}$$

Thus

$$\begin{aligned}\mathbb{E}|Z|^p &\leqslant 2^p\sqrt{8p}^p 2^{p/2-1}\big(32n + n^{2/p}p^3\big)^{p/2}\tau_\eta^{2p} + 2^{p-1}p^p \cdot 2^p n \cdot 4p!(2\tau_\eta^2)^p\\ &\leqslant 2^{3p-1}p^{p/2} \cdot 2^{p/2-1}\big((32n)^{p/2} + n \cdot p^{3p/2}\big)\tau_\eta^{2p} + 2^{3p+1}\tau_\eta^{2p}np^{2p}\\ &\leqslant \big(2^{6p-2}p^{p/2}n^{p/2} + np^{2p}[2^{7p/2-2} + 2^{3p+1}]\big)\tau_\eta^{2p}\\ &\leqslant \Big(64^p \cdot p^{p/2}n^{p/2} + 19^p \cdot np^{2p}\Big)\tau_\eta^{2p}.\end{aligned}$$

Thus

$$(\mathbb{E}|Z|^p)^{1/p} \leqslant \Big(64 \cdot \sqrt{p}n^{1/2} + 19 \cdot n^{1/p}p^2\Big)\tau_\eta^2.$$

Thus, for any $\delta \leqslant 1/n$, using Lemma 2.5 for random variable Z/n with $t = \log(\frac{12d^2}{\delta}) > (\log n)/2$ and we obtain:

$$\mathbb{P}\left[\left|\frac{Z}{n}\right| \geqslant \frac{192\tau_\eta^2}{\sqrt{n}}\sqrt{t} + \frac{57\tau_\eta^2}{n}n^{1/t}t^2\right] \leqslant 3e^{-t} \Rightarrow$$

$$\mathbb{P}\left[\left|\frac{Z}{n}\right| \geqslant \frac{192\tau_\eta^2}{\sqrt{n}}\sqrt{\log(\tfrac{12d^2}{\delta})} + \frac{57\tau_\eta^2}{n}n^{1/\log(\frac{12d^2}{\delta})}\log^2(\tfrac{12d^2}{\delta})\right] \leqslant \frac{\delta}{4d^2} \Rightarrow$$

$$\mathbb{P}\left[\left|\frac{Z}{n}\right| \geqslant \frac{192\tau_\eta^2}{\sqrt{n}}\sqrt{\log(\tfrac{12d^2}{\delta})} + \frac{155\tau_\eta^2}{n}\log^2(\tfrac{12d^2}{\delta})\right] \leqslant \frac{\delta}{4d^2}.$$

Combining all terms $T_1, T_2, T_3, T_{4,1}$ and $T_{4,2}$ we get with probability not less than

$1 - \delta$:

$$\|\hat{\mathcal{V}}_{1,\mathbb{E}} - \mathcal{V}_{1,\mathbb{E}}\|_* \leqslant \frac{1}{n} \cdot \left[8L(L\tau_y^2 + 2\tau_\eta\tau_y\sqrt{d}) \cdot \log(\tfrac{16dn}{\delta}) + 2d\sqrt{d}\tau_\eta^2 \log\tfrac{8d^2}{\delta} + 155\tau_\eta^2 d\log^2(\tfrac{12d^2}{\delta})\right]$$
$$+\frac{1}{\sqrt{n}}\left[2d\sqrt{d}\tau_\eta^2\sqrt{2\log\tfrac{8d^2}{\delta}} + 192\tau_\eta^2 d\sqrt{\log\tfrac{12d^2}{\delta}}\right].$$

Rearranging terms and replacing δ by $\delta/2$, with probability not less, than $1 - \delta/2$:

$$\|\hat{\mathcal{V}}_{1,\mathbb{E}} - \mathcal{V}_{1,\mathbb{E}}\|_* \leqslant \frac{195d\sqrt{d}\tau_\eta^2}{\sqrt{n}}\sqrt{\log\frac{24d^2}{\delta}} + \frac{8L^2\tau_y^2 + 16\tau_\eta\tau_y L\sqrt{d} + 157\tau_\eta^2 d\sqrt{d}}{n}\log^2\frac{32d^2n}{\delta}.$$

Using expression:

$$\mathcal{V}_{1,\text{cov}} + \mathcal{V}_{1,\mathbb{E}} = \text{cov}[\ell_1(x)],$$

we can suggest estimator for $\mathcal{V}_{1,\text{cov}}$ as $\hat{\mathcal{V}}_{1,\text{cov}} = \frac{1}{n}\sum\limits_{i=1}^{n}\ell(x_i)\ell(x_i)^\top - \hat{\mathcal{V}}_{1,\mathbb{E}}$.
Applying the triangle inequality:

$$\|\hat{\mathcal{V}}_{1,\text{cov}} - \mathcal{V}_{1,\text{cov}}\|_* \leqslant \|\hat{\mathcal{V}}_{1,\mathbb{E}} - \mathcal{V}_{1,\mathbb{E}}\|_* + \|\frac{1}{n}\sum_{i=1}^{n}\ell(x_i)\ell(x_i)^\top - \mathbb{E}(\ell(x)\ell(x)^\top)\|_*.$$

To estimate the second term, we can use the same arguments as for bounding $T_{4,2}$: with probability greater than $1 - \delta/2$, $\|\frac{1}{n}\sum\limits_{i=1}^{n}\ell(x_i)\ell(x_i)^\top - \mathbb{E}(\ell(x)\ell(x)^\top)\|_*$ is less than $2d\sqrt{d}\tau_\ell^2\left(\frac{\log\frac{4d^2}{\delta}}{n} + \sqrt{\frac{2\log\frac{4d^2}{\delta}}{n}}\right)$. Finally, combining this bound with bound for $\|\hat{\mathcal{V}}_{1,\text{cov}} - \mathcal{V}_{1,\text{cov}}\|_*$:

$$\|\hat{\mathcal{V}}_{1,\text{cov}} - \mathcal{V}_{1,\text{cov}}\|_* \leqslant \frac{d\sqrt{d}(195\tau_\eta^2 + 2\tau_\ell^2)}{\sqrt{n}}\sqrt{\log\frac{24d^2}{\delta}} +$$
$$\frac{8L^2\tau_y^2 + 16\tau_\eta\tau_y L\sqrt{d} + (157\tau_\eta^2 + 2\tau_\ell^2)d\sqrt{d}}{n}\log^2\frac{32d^2n}{\delta}.$$

□

2.7.3 Proof of Theorem 2.2

Proof. Using triangle inequality, we get:

$$\|\mathcal{V}_{1,\text{cov}}(\theta^*) - \hat{\mathcal{V}}_{1,\text{cov}}(\hat{\theta})\|_* \leqslant$$
$$\|\mathcal{V}_{1,\text{cov}}(\theta^*) - \hat{\mathcal{V}}_{1,\text{cov}}(\theta^*)\|_* + \|\hat{\mathcal{V}}_{1,\text{cov}}(\theta^*) - \hat{\mathcal{V}}_{1,\text{cov}}(\hat{\theta})\|_* = \mathcal{F}_1 + \mathcal{F}_2. \quad (2.7.1)$$

Theorem 2.1 supplies us with non-asymptotic analysis for the $\mathcal{F}_1$ term: with probability not less then $1-\delta/2$:

$$\begin{aligned}\|\hat{\mathcal{V}}_{1,\mathrm{cov}}(\theta^*) - \mathcal{V}_{1,\mathrm{cov}}(\theta^*)\|_* \leqslant \frac{d\sqrt{d}(195\tau_\eta^2 + 2\tau_\ell^2)}{\sqrt{n}}\sqrt{\log\frac{48d^2}{\delta}} + \\ \frac{8L^2\tau_y^2 + 16\tau_\eta\tau_y L\sqrt{d} + (157\tau_\eta^2 + 2\tau_\ell^2)d\sqrt{d}}{n}\log^2\frac{64d^2n}{\delta}.\end{aligned} \tag{2.7.2}$$

For the second term, let us firstly analyse the norm $\|\theta^* - \hat{\theta}\|$. For simplicity, introduce notation: $\hat{C} = \frac{1}{n}\sum_{i=1}^n \Psi(x_i)\Psi(x_i)^\top \in \mathbb{R}^{m\times m}$, $C = \mathbb{E}\big[\Psi(x)\Psi(x)^\top\big] \in \mathbb{R}^{m\times m}$, $\hat{b} = \frac{1}{n}\sum_{i=1}^n (\nabla\cdot\Psi)(x_i) \in \mathbb{R}^m$ and $b = \mathbb{E}\big[(\nabla\cdot\Psi)(x)\big] \in \mathbb{R}^m$

Let us estimate $\|\hat{C} - C\|_F$ and $\|\hat{b} - b\|$. Introduce notation:

$$C_{a,b}^c = \frac{1}{n}\sum_{i=1}^n \Psi_c^a(x_i)\cdot\Psi_c^b(x_i) - \mathbb{E}\big[\Psi_c^a(x)\cdot\Psi_c^b(x)\big].$$

It can be shown like in the bounding of $T_{4,2}$ term in the Theorem's 2.1 proof, using [Boucheron et al., 2013, Theorem 2.1] and Bernstein's inequality [Boucheron et al., 2013, Theorem 2.10] that with probability less then e^{-t}:

$$C_{a,b}^c \geqslant 2\frac{\tau_\Psi^2}{n}t + \sqrt{32\tau_\Psi^4}\sqrt{t}/\sqrt{n}.$$

Taking $t = \log\frac{2m^2d}{\delta}$ we show that:

$$\mathbb{P}\left[\|\hat{C} - C\|_F \geqslant 2\tau_\Psi^2\sqrt{m^3d}\left(\frac{\log\frac{2m^2d}{\delta}}{n} + \sqrt{\frac{2\log\frac{2m^2d}{\delta}}{n}}\right)\right] < \delta.$$

According to assumption **(M2)** and Hoeffding bound:

$$\mathbb{P}\left[\|\hat{b}_i - b_i\| \geqslant \frac{t}{n}\right] \leqslant e^{\frac{-t^2}{2\tau_{\nabla\Psi}^2 n}},$$

combining inequalities for all m components of b, we have:

$$\mathbb{P}\left[\|b - \hat{b}\| \geqslant \frac{\tau_{\nabla\Psi}\sqrt{m}}{\sqrt{n}}\sqrt{\log\frac{m^2}{\delta^2}}\right] \leqslant \delta.$$

Now, let estimate $\|\theta^* - \hat{\theta}\|$:

$$\theta^* - \hat{\theta} = \hat{C}^{-1}\hat{b} - C^{-1}b = \hat{C}^{-1}(\hat{b} - b) + (\hat{C}^{-1} - C^{-1})b \Rightarrow$$

$$\|\theta^* - \hat{\theta}\| \leqslant \frac{\|\hat{b} - b\|}{\lambda_{\min}(\hat{C})} + \frac{\|b\|_2\cdot\|C - \hat{C}\|_{\mathrm{op}}}{\lambda_{\min}(C)\cdot\lambda_{\min}(\hat{C})}.$$

For n large enough (for some constants c_1 and c_2, not depending on n and δ: $n > c_1 + c_2 \log \frac{1}{\delta}$) $\|\hat{C} - C\| \leqslant \dfrac{\lambda_{\min}}{2}$ with probability more, than $1 - \delta/12$ hence for this n: $\lambda_{\min}(\hat{C}) \geqslant \dfrac{\lambda_{\min}(C)}{2}$, hence, combining 3 estimations for $\|\hat{b} - b\|$, $\|\hat{C} - C\|$ and for $\lambda_{\min}(\hat{C})$, we obtain estimation for $\|\theta^* - \hat{\theta}\|$: with probability not less, than $1 - \delta/4$:

$$\|\theta^* - \hat{\theta}\| \leqslant$$

$$\frac{2\tau_{\nabla\Psi}\sqrt{m}}{\sqrt{n}\lambda_{\min}(C)}\sqrt{2\log\frac{12m}{\delta}} + \frac{2\|b\|}{\lambda^2_{\min}(C)} \cdot 2\tau^2_\Psi\sqrt{m^3 d}\left(\frac{\log\frac{24m^2d}{\delta}}{n} + \sqrt{\frac{2\log\frac{24m^2d}{\delta}}{n}}\right) \tag{2.7.3}$$

Consider now (a, b) element of $\hat{\mathcal{V}}_{1,\text{cov}}(\theta^*) - \hat{\mathcal{V}}_{1,\text{cov}}(\hat{\theta})$ (using 2.4.2):

$$\left[\hat{\mathcal{V}}_{1,\text{cov}}(\theta^*) - \hat{\mathcal{V}}_{1,\text{cov}}(\hat{\theta})\right]_{a,b} = \frac{1}{n}\sum_{i,j=1}^{n}\frac{\alpha_{i,j}}{|I_h(i,j)|-1}\sum_{\alpha,\beta=1}^{m}\Psi(x_i)^\alpha_a\Psi(x_j)^\beta_b \cdot |\theta^*_\alpha\theta^*_\beta - \hat{\theta}_\alpha\hat{\theta}_\beta| \leqslant$$

$$\frac{1}{2n}\sum_{i,j=1}^{n}\frac{\alpha_{i,j}}{|I_h(i,j)|-1}\sum_{\alpha,\beta=1}^{m}\left[\left[\Psi(x_i)^\alpha_a\right]^2 + \left[\Psi(x_j)^\beta_b\right]^2\right] \cdot |\theta^*_\alpha\theta^*_\beta - \hat{\theta}_\alpha\hat{\theta}_\beta| \leqslant$$

$$\sum_{\alpha,\beta=1}^{m}\frac{1}{2n}\left[\sum_{i=1}^{n}[\Psi(x_i)^\alpha_a]^2 + \sum_{i=1}^{n}\left[\Psi(x_i)^\beta_b\right]^2\right] \cdot \left|\theta^*_\alpha\theta^*_\beta - \hat{\theta}_\alpha\hat{\theta}_\beta\right|,$$

because every row of binary matrix $\left\{\alpha_{i,j}\right\}_{i,j=\overline{1,n}}$ has exactly $|I_h(i,j)| - 1$ non-zero elements.

Now, let us estimate desired norm:

$$\|\hat{\mathcal{V}}_{1,\text{cov}}(\theta^*) - \hat{\mathcal{V}}_{1,\text{cov}}(\hat{\theta})\|_* \leqslant m \cdot \sum_{k=1}^{n}\sum_{l=1}^{d}\frac{\sum_{i=1}^{n}|\Psi^k_l(x_i)|^2}{n} \cdot (2\|\theta^*\| + \|\hat{\theta} - \theta^*\|) \cdot \|\hat{\theta} - \theta^*\| \tag{2.7.4}$$

Using again [Boucheron et al., 2013, Theorem 2.1] and Bernstein's inequality [Boucheron et al., 2013, Theorem 2.10] with probability not less then $1 - \delta/8$:

$$\sum_{k=1}^{n}\sum_{l=1}^{d}\frac{\sum_{i=1}^{n}|\Psi^k_l(x_i)|^2}{n} \leqslant \mathbb{E}\|\Psi(x)\|^2_2 + 2md\tau^2_\psi\left(\frac{\log\frac{16md}{\delta}}{n} + \sqrt{\frac{2\log\frac{16md}{\delta}}{n}}\right) \tag{2.7.5}$$

For n large enough (for some constants c_1 and c_2, not depending on n and δ: $n >$

$c_1 + c_2 \log \frac{1}{\delta}$):

$$\sum_{k=1}^{n} \sum_{l=1}^{d} \frac{\sum\limits_{i=1}^{n} |\Psi_l^k(x_i)|^2}{n} \leqslant \frac{3}{2} \mathbb{E}\|\Psi(x)\|_2^2 \text{ with probability no less then } 1 - \delta/8,$$

and

$$\big(2\|\theta^*\| + \|\hat{\theta} - \theta^*\|\big) \leqslant 3\|\theta^*\| \text{ with probability no less then } 1 - \delta/8$$

$$\|\hat{\mathcal{V}}_{1,\text{cov}}(\theta^*) - \hat{\mathcal{V}}_{1,\text{cov}}(\hat{\theta})\|_* \leqslant m \cdot \frac{9}{2} \mathbb{E}\|\Psi(x)\|_2^2 \cdot \|\theta^*\| \cdot$$

$$\left[\frac{2\tau_{\nabla\Psi}\sqrt{m}}{\sqrt{n}\lambda_{\min}(C)} \sqrt{2 \log \tfrac{12m}{\delta}} + \frac{2\|b\|}{\lambda^2_{\min}(C)} \cdot 2\tau_\Psi^2 \sqrt{m^3 d} \left(\frac{\log \frac{24m^2 d}{\delta}}{n} + \sqrt{\frac{2 \log \frac{24m^2 d}{\delta}}{n}} \right) \right]$$

with probability not less then $1 - \delta/2$. Combining and simplifying obtained inequality and (2.7.2) we obtain final inequality: with probablility not less, then $1 - \delta$:

$$\|\mathcal{V}_{1,\text{cov}}(\theta^*) - \hat{\mathcal{V}}_{1,\text{cov}}(\hat{\theta})\|_* \leqslant \frac{1}{\sqrt{n}} \sqrt{2 \log \frac{48m^2 d}{\delta}} \times$$

$$\left[d\sqrt{d}(195\tau_\eta^2 + 2\tau_\ell^2) + \frac{9m}{2} \mathbb{E}\|\Psi(x)\|_2^2 \cdot \|\theta^*\| \left(\frac{2\tau_{\nabla\Psi}\sqrt{m}}{\lambda_0} + \frac{4\|b\|\tau_\Psi^2 \sqrt{m^3 d}}{\lambda_0} \right) \right] + O\left(\frac{1}{n}\right).$$

□

Chapter 3

Constant step-size Stochastic Gradient Descent for probabilistic modeling

Abstract

Stochastic gradient methods enable learning probabilistic models from large amounts of data. While large step-sizes (learning rates) have shown to be best for least-squares (e.g., Gaussian noise) once combined with parameter averaging, these are not leading to convergent algorithms in general. In this chapter, we consider generalized linear models, that is, conditional models based on exponential families. We propose averaging moment parameters instead of natural parameters for constant-step-size stochastic gradient descent. For finite-dimensional models, we show that this can sometimes (and surprisingly) lead to better predictions than the best linear model. For infinite-dimensional models, we show that it always converges to optimal predictions, while averaging natural parameters never does. We illustrate our findings with simulations on synthetic data and classical benchmarks with many observations.

This chapter is based on the conference paper published as a Uncertainty in Artificial Intelligence (UAI) 2018 paper, which was accepted as an oral presentation [Babichev and Bach, 2018a].

3.1 Introduction

Faced with large amounts of data, efficient parameter estimation remains one of the key bottlenecks in the application of probabilistic models. Once cast as an optimization problem, for example through the maximum likelihood principle, difficulties may arise from the size of the model, the number of observations, or the potential non-convexity of the objective functions, and often all three [Koller and Friedman, 2009, Murphy, 2012].

In this chapter we focus primarily on situations where the number of observations is large; in this context, stochastic gradient descent (SGD) methods which look at one sample at a time are usually favored for their cheap iteration cost. However, finding the correct step-size (sometimes referred to as the learning rate) remains a practical and theoretical challenge, for probabilistic modeling but also in most other situations beyond maximum likelihood [Bottou et al., 2016].

In order to preserve convergence, the step size γ_n at the n-th iteration typically has to decay with the number of gradient steps (here equal to the number of data points which are processed), typically as C/n^α for $\alpha \in [1/2, 1]$ [see, e.g., Bach and Moulines, 2011, Bottou et al., 2016]. However, these often leads to slow convergence and the choice of α and C is difficult in practice. More recently, constant step-sizes have been advocated for their fast convergence towards a neighborhood of the optimal solution [Bach and Moulines, 2013], while it is a standard practice in many areas [Goodfellow et al., 2016]. However, it is not convergent in general and thus small step-sizes are still needed to converge to a decent estimator.

Constant step-sizes can however be made to converge in one situation. When the functions to optimize are quadratic, like for least-squares regression, using a constant step-size combined with an averaging of all estimators along the algorithm can be shown to converge to the global solution with the optimal convergence rates [Bach and Moulines, 2013, Dieuleveut and Bach, 2016].

The goal of this chapter is to explore the possibility of such global convergence with a constant step-size in the context of probabilistic modeling with exponential families, e.g., for logistic regression or Poisson regression [McCullagh, 1984]. This would lead to the possibility of using probabilistic models (thus with a principled quantification of uncertainty) with rapidly converging algorithms. Our main novel idea is to replace the averaging of the *natural* parameters of the exponential family by the averaging of the *moment* parameters, which can also be formulated as averaging *predictions* instead of *estimators*. For example, in the context of predicting binary outcomes in $\{0, 1\}$ through a Bernoulli distribution, the moment parameter is the probability $p \in [0, 1]$ that the variable is equal to one, while the natural parameter is the "log odds ratio" $\log \frac{p}{1-p}$, which is unconstrained. This lack of constraint is often seen as a benefit for optimization; it turns out that for stochastic gradient methods, the moment parameter is better suited to averaging. Note that for least-squares, which corresponds to modeling with the Gaussian distribution with fixed variance, moment and natural parameters are equal, so it does not make a difference.

More precisely, our main contributions are:

- For generalized linear models, we propose in Section 3.4 averaging moment parameters instead of natural parameters for constant-step-size stochastic gradient descent.
- For finite-dimensional models, we show in Section 3.5 that this can sometimes (and surprisingly) lead to better predictions than the best linear model.
- For infinite-dimensional models, we show in Section 3.6 that it always converges to optimal predictions, while averaging natural parameters never does.
- We illustrate our findings in Section 3.7 with simulations on synthetic data and classical benchmarks with many observations.

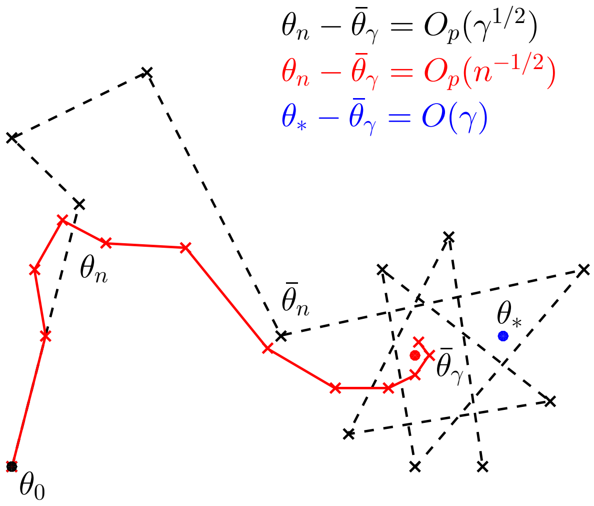
$\theta_n - \bar\theta_\gamma = O_p(\gamma^{1/2})$
$\theta_n - \bar\theta_\gamma = O_p(n^{-1/2})$
$\theta_* - \bar\theta_\gamma = O(\gamma)$
θ_n
$\bar\theta_n$
θ_*
$\bar\theta_\gamma$
θ_0

Figure 3-1 – Convergence of iterates θ_n and averaged iterates $\bar\theta_n$ to the mean $\bar\theta_\gamma$ under the stationary distribution π_γ.

3.2 Constant step size stochastic gradient descent

In this section, we present the main intuitions behind stochastic gradient descent (SGD) with constant step-size. For more details, see Dieuleveut et al. [2017]. We consider a real-valued function F defined on the Euclidean space $\mathbb{R}^d$ (this can be generalized to any Hilbert space, as done in Section 3.6 when considering Gaussian processes and positive-definite kernels), and a sequence of random functions $(f_n)_{n\geqslant 1}$ which are independent and identically distributed and such that $\mathbb{E} f_n(\theta) = F(\theta)$ for all $\theta \in \mathbb{R}^d$. Typically, F will the expected negative log-likelihood on unseen data, while f_n will be the negative log-likelihood for a single observation. Since we require *independent* random functions, we assume that we make single pass over the data, and thus the number of iterations is equal to the number of observations.

Starting from an initial $\theta_0 \in \mathbb{R}^d$, then SGD will perform the following recursion, from $n = 1$ to the total number of observations:

$$\theta_n = \theta_{n-1} - \gamma_n \nabla f_n(\theta_{n-1}). \tag{3.2.1}$$

Since the functions f_n are independent, the iterates $(\theta_n)_n$ form a Markov chain. When the step-size γ_n is constant (equal to γ) and the functions f_n are identically distributed, the Markov chain is *homogeneous*. Thus, under additional assumptions [see, e.g., Dieuleveut et al., 2017, Meyn and Tweedie, 1993], it converges in distribution to a stationary distribution, which we refer to as π_γ. These additional assumptions include that γ is not too large (otherwise the algorithm diverges) and in the traditional analysis of step-sizes for gradient descent techniques, we analyze the situation of small γ's (and thus perform asymptotic expansions around $\gamma = 0$).

The distribution π_γ is in general not equal to a Dirac mass, and thus, constant-step-size SGD is *not* convergent. However, averaging along the path of the Markov chain has some interesting properties. Indeed, several versions of the "ergodic theorem" [see, e.g., Meyn and Tweedie, 1993] show that for functions g from $\mathbb{R}^d$ to any vector space, then the empirical average $\frac{1}{n}\sum_{i=1}^{n} g(\theta_i)$ converges in probability to the expectation $\int g(\theta)d\pi_\gamma(\theta)$ of g under the stationary distribution π_γ. This convergence can also be quantified by a central limit theorem with an error whichs tends to a normal distribution with variance proportional equal to a constant times $1/n$.

Thus, if denote $\bar{\theta}_n = \frac{1}{n+1}\sum_{i=0}^{n}\theta_i$, applying the previous result to the identity function g, we immediately obtain that $\bar{\theta}_n$ converges to $\bar{\theta}_\gamma = \int \theta d\pi_\gamma(\theta)$, with a squared error converging in $O(1/n)$. The key question is the relationship between $\bar{\theta}_\gamma$ and the global optimizer θ_* of F, as this characterizes the performance of the algorithm with an infinite number of observations.

By taking expectations in Eq. (3.2.1), and taking a limit with n tending to infinity we obtain that

$$\int \nabla F(\theta)d\pi_\gamma(\theta) = 0, \tag{3.2.2}$$

that is, under the stationary distribution π_γ, the average gradient is zero. When the gradient is a linear function (like for a quadratic objective F), this leads to

$$\nabla F\left(\int \theta d\pi_\gamma(\theta)\right) = \nabla F(\bar{\theta}_\gamma) = 0,$$

and thus $\bar{\theta}_\gamma$ is a stationary point of F (and hence the global minimizer if F is convex). However this is not true in general.

As shown by Dieuleveut et al. [2017], the deviation $\bar{\theta}_\gamma - \theta_*$ is of order γ, which is an improvement on the non-averaged recursion, which is at average distance $O(\gamma^{1/2})$ (see an illustration in Figure 3-1); thus, small or decaying step-sizes are needed. In this chapter, we explore alternatives which are not averaging the estimators $\theta_1, \dots, \theta_n$, and rely instead on the specific structure of our cost functions, namely negative log-likelihoods.

3.3 Warm-up: exponential families

In order to highlight the benefits of averaging moment parameters, we first consider unconditional exponential families. We thus consider the standard exponential family

$$q(x|\theta) = h(x)\exp(\theta^\top T(x) - A(\theta)),$$

where $h(x)$ is the base measure, $T(x) \in \mathbb{R}^d$ is the sufficient statistics and A the log-partition function. The function A is always convex [see, e.g., Koller and Friedman, 2009, Murphy, 2012]. Note that we do not assume that the data distribution $p(x)$ comes from this exponential family. The expected (with respect to the input distribution $p(x)$) negative log-likelihood is equal to

$$F(\theta) \quad = \quad -\mathbb{E}_{p(x)} \log q(x|\theta)$$

$$= \quad A(\theta) - \theta^\top \mathbb{E}_{p(x)} T(x) - \mathbb{E}_{p(x)} \log h(x).$$

It is known to be minimized by θ_* such that $\nabla A(\theta_*) = \mathbb{E}_{p(x)} T(x)$. Given i.i.d. data $(x_n)_{n \geqslant 1}$ sampled from $p(x)$, then the SGD recursion from Eq. (3.2.1) becomes:

$$\theta_n = \theta_{n-1} - \gamma \big[\nabla A(\theta_{n-1}) - T(x_n) \big],$$

while the stationarity equation in Eq. (3.2.2) becomes

$$\int [\nabla A(\theta) - \mathbb{E}_{p(x)} T(x)] d\pi_\gamma(\theta) = 0,$$

which leads to

$$\int \nabla A(\theta) d\pi_\gamma(\theta) = \mathbb{E}_{p(x)} T(x) = \nabla A(\theta_*).$$

Thus, averaging $\nabla A(\theta_n)$ will converge to $\nabla A(\theta_*)$, while averaging θ_n will *not* converge to θ_*. This simple observation is the basis of our work.

Note that in this context of unconditional models, a simpler estimator exists, that is, we can simply compute the empirical average $\frac{1}{n} \sum_{i=1}^n T(x_i)$ that will converge to $\nabla A(\theta_*)$. Nevertheless, this shows that averaging moment parameters $\nabla A(\theta)$ rather than natural parameters θ can bring convergence benefits. We now turn to conditional models, for which no closed-form solutions exist.

3.4 Conditional exponential families

Now we consider the conditional exponential family

$$q(y|x, \theta) = h(y) \exp \big(y \cdot \eta_\theta(x) - a(\eta_\theta(x)) \big).$$

For simplicity we consider only one-dimensional families where $y \in \mathbb{R}$ — but our framework would also extend to more complex models such as conditional random fields [Lafferty et al., 2001]. We will also assume that $h(y) = 1$ for all y to avoid carrying constant terms in log-likelihoods. We consider functions of the form $\eta_\theta(x) = \theta^\top \Phi(x)$, which are linear in a feature vector $\Phi(x)$, where $\Phi : \mathcal{X} \to \mathbb{R}^d$ can be defined on an arbitrary input set $\mathcal{X}$. Calculating the negative log-likelihood, one obtains:

$$f_n(\theta) = -\log q(y_n|x_n\theta) = -y_n \Phi(x_n)^\top \theta + a\big(\Phi(x_n)\theta\big),$$

and, for any distribution $p(x, y)$, for which $p(y|x)$ may not be a member of the conditional exponential family,

$$\begin{aligned} F(\theta) &= \mathbb{E}_{p(x_n, y_n)} f_n(\theta) \\ &= \mathbb{E}_{p(x_n, y_n)} \Big[- y_n \Phi(x_n)^\top \theta + a\big(\Phi(x_n)\theta\big) \Big]. \end{aligned}$$

The goal of estimation in such generalized linear models is to find an unknown parameter θ given n observations $(x_i, y_i)_{i=1,\dots,n}$:

$$\theta_* = \arg\min_{\theta \in \mathbb{R}^d} F(\theta). \tag{3.4.1}$$

3.4.1 From estimators to prediction functions

Another point of view is to consider that an estimator $\theta \in \mathbb{R}^d$ in fact defines a function $\eta : \mathcal{X} \to \mathbb{R}$, with value a natural parameter for the exponential family $q(y) = \exp(\eta y - a(\eta))$. This particular choice of function η_θ is linear in $\Phi(x)$, and we have, by decomposing the joint probability $p(x_n, y_n)$ in two (and dropping the dependence on n since we have assumed i.i.d. data):

$$\begin{aligned} F(\theta) &= \mathbb{E}_{p(x)}\Big(\mathbb{E}_{p(y|x)}\big[-y\Phi(x)^\top\theta + a(\Phi(x)^\top\theta)\big]\Big) \\ &= \mathbb{E}_{p(x)}\Big(-\mathbb{E}_{p(y|x)}y\Phi(x)^\top\theta + a(\Phi(x)^\top\theta)\Big) \\ &= \mathcal{F}(\eta_\theta), \end{aligned}$$

with $\mathcal{F}(\eta) = \mathbb{E}_{p(x)}\big(-\mathbb{E}_{p(y|x)}y \cdot \eta(x) + a(\eta(x))\big)$ is the performance measure defined for a *function* $\eta : \mathcal{X} \to \mathbb{R}$. By definition $F(\theta) = \mathcal{F}(\eta_\theta) = \mathcal{F}(\theta^\top\Phi(\cdot))$.

However, the global minimizer of $\mathcal{F}(\eta)$ over all functions $\eta : \mathcal{X} \to \mathbb{R}$ may not be attained at a linear function in $\Phi(x)$ (this can only be the case if the linear model is well-specified or if the feature vector $\Phi(x)$ is flexible enough). Indeed, the global minimizer of $\mathcal{F}$ is the function $\eta_{**} : x \mapsto (a')^{-1}(\mathbb{E}_{p(y|x)}y)$ (starting from $\mathcal{F}(\eta) = \int \big[a(\eta(x)) - \mathbb{E}_{p(x|y)}y \cdot \eta(x)\big]p(x)dx$ and writing down the Euler - Lagrange equation: $\frac{\partial \mathcal{F}}{\partial \eta} - \frac{d}{dx}\frac{\partial F}{\partial \eta'} = 0 \Leftrightarrow \big[a'(\eta) - \mathbb{E}_{p(x|y)}y\big]p(x) = 0$ and finally $\eta \mapsto (a')^{-1}(\mathbb{E}_{p(x|y)}y))$ and is typically not a linear function in $\Phi(x)$ (note here that $p(y|x)$ is the conditional data-generating distribution).

The function η corresponds to the *natural* parameter of the exponential family, and it is often more intuitive to consider the *moment* parameter, that is defining functions $\mu : \mathcal{X} \to \mathbb{R}$ that now correspond to moments of outputs y; we will refer to them as *prediction functions*. Going from natural to moment parameter is known to be done through the gradient of the log-partition function, and we thus consider for η a function from $\mathcal{X}$ to $\mathbb{R}$, $\mu(\cdot) = a'(\eta(\cdot))$, and this leads to the performance measure

$$\mathcal{G}(\mu) = \mathcal{F}((a')^{-1}(\mu(\cdot))).$$

Note now, that the global minimum of $\mathcal{G}$ is reached at

$$\mu_{**}(x) = \mathbb{E}_{p(y|x)}y.$$

We introduce also the prediction function $\mu_*(x)$ corresponding to the best η which is linear in $\Phi(x)$, that is:

$$\mu_*(x) = a'\big(\theta_*^\top\Phi(x)\big).$$

We say that the model is well-specified when $\mu_* = \mu_{**}$, and for these models,

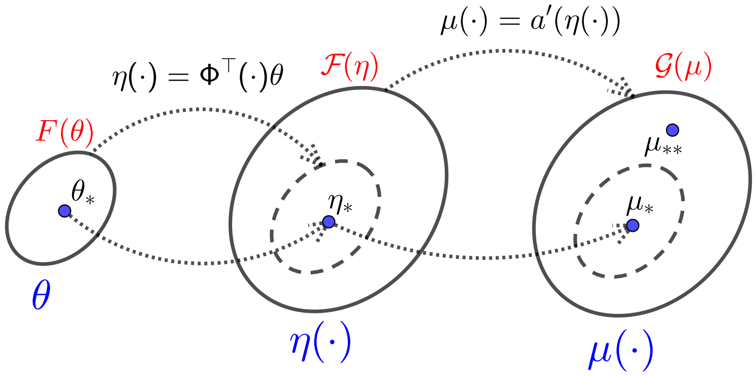

Figure 3-2 – Graphical representation of reparametrization: firstly we expand the class of functions, replacing parameter θ with function $\eta(\cdot) = \Phi^\top(\cdot)\theta$ and then we do one more reparametrization: $\mu(\cdot) = a'(\eta(\cdot))$. Best linear prediction μ_* is constructed using θ_* and the global minimizer of $\mathcal{G}$ is μ_{**}. Model is well-specified if and only if $\mu_* = \mu_{**}$.

$\inf_\theta F(\theta) = \inf_\mu \mathcal{G}(\mu)$. However, in general, we only have $\inf_\theta F(\theta) > \inf_\mu \mathcal{G}(\mu)$ and (very often) the inequality is strict (see examples in our simulations).

To make the further developments more concrete, we now present two classical examples: logistic regression and Poisson regression.

Logistic regression. The special case of conditional family is logistic regression, where $y \in \{0,1\}$, $a(t) = \log(1+e^{-t})$ and $a'(t) = \sigma(t) = \frac{1}{1+e^{-t}}$ is the sigmoid function and the probability mass function is given by $p(y|\eta) = \exp(\eta y - \log(1+e^\eta))$.

Poisson regression. One more special case is Poisson regression with $y \in \mathbb{N}$, $a(t) = \exp(t)$ and the response variable y has a Poisson distribution. The probability mass function is given by $p(y|\eta) = \exp(\eta y - e^\eta - \log(y!))$. Poisson regression may be appropriate when the dependent variable is a count, for example in genomics, network packet analysis, crime rate analysis, fluorescence microscopy, etc. [Hilbe, 2011].

3.4.2 Averaging predictions

Recall from Section 3.2 that π_γ is the stationary distribution of θ. Taking expectation of both parts of Eq. (3.2.1), we get, by using the fact that π_γ is the limiting distribution of θ_n and θ_{n-1}:

$$\begin{aligned}&\mathbb{E}_{\pi_\gamma(\theta_n)}\theta_n\\ =\ &\mathbb{E}_{\pi_\gamma(\theta_{n-1})}\theta_{n-1} - \gamma\mathbb{E}_{\pi_\gamma(\theta_{n-1})}\mathbb{E}_{p(x_n,y_n)}f_n'(\theta_{n-1}),\end{aligned}$$

which leads to $\mathbb{E}_{\pi_\gamma(\theta)}\mathbb{E}_{p(x_n,y_n)}\nabla f_n(\theta) = 0$, that is, now removing the dependence on n (data (x,y) are i.i.d.):

$$\mathbb{E}_{\pi_\gamma(\theta)}\mathbb{E}_{p(x,y)}\Big[-y\Phi(x) + a'\big(\Phi(x)^\top\theta\big)\Phi(x)\Big] = 0,$$

which finally leads to

$$\mathbb{E}_{p(x)}\Big[\mathbb{E}_{\pi_\gamma(\theta)}a'\big(\Phi(x)^\top\theta\big) - \mu_{**}(x)\Big]\Phi(x) = 0. \tag{3.4.2}$$

This is the core equation our method relies on. It does not imply that $b(x) = \mathbb{E}_{\pi_\gamma(\theta)}a'\big(\Phi(x)^\top\theta\big) - \mu_{**}(x)$ is uniformly equal to zero (which we want), but only that $\mathbb{E}_{p(x)}\Phi(x)b(x) = 0$, i.e., $b(x)$ is uncorrelated with $\Phi(x)$.

If the feature vector $\Phi(x)$ is "large enough" then this is equivalent to $b = 0$. [1] For example, when $\Phi(x)$ is composed of an orthonormal basis of the space of integrable functions (like for kernels in Section 3.6), then this is exactly true. Thus, in this situation,

$$\mu_{**}(x) = \mathbb{E}_{\pi_\gamma(\theta)}a'\big(\Phi(x)^\top\theta\big), \tag{3.4.3}$$

and averaging predictions $a'\big(\Phi(x)^\top\theta_n\big)$, along the path (θ_n) of the Markov chain should exactly converge to the optimal prediction.

This exact convergence is weaker (requires high-dimensional fatures) than for the unconditional family in Section 3.3 but it can still bring surprising benefits even when Φ is not large enough, as we present in Section 3.5 and Section 3.6.

3.4.3 Two types of averaging

Now we can introduce two possible ways to estimate the prediction function $\mu(x)$.

Averaging estimators. The first one is the usual way: we first estimate parameter θ, using Ruppert-Polyak averaging [Polyak and Juditsky, 1992]: $\bar\theta_n = \frac{1}{n+1}\sum_{i=0}^n \theta_i$ and then we denote

$$\bar\mu_n(x) = a'(\Phi(x)^\top\bar\theta_n) = a'\Big(\Phi(x)^\top \tfrac{1}{n+1}\textstyle\sum_{i=0}^n \theta_i\Big)$$

the corresponding prediction. As discussed in Section 3.2 it converges to $\bar\mu_\gamma : x \mapsto a'(\Phi(x)^\top\bar\theta_\gamma)$, which is *not* equal to in general to $a'(\Phi(x)^\top\theta_*)$, where θ_* is the optimal parameter in $\mathbb{R}^d$. Since, as presented at the end of Section 3.2, $\bar\theta_\gamma - \theta_*$ is of order $O(\gamma)$, $F(\bar\theta_\gamma) - F(\theta_*)$ is of order $O(\gamma^2)$ (because $\nabla F(\theta_*) = 0$), and thus an error of $O(\gamma^2)$ is added to the usual convergence rates in $O(1/n)$.

Note that we are limited here to prediction functions which corresponds to *linear functions* in $\Phi(x)$ in the natural parameterization, and thus $F(\theta_*) \geqslant \mathcal{G}(\mu_{**})$, and the

[1] 1. Let $\Phi(x) = (\phi_1(x), \ldots, \phi_n(x))^\top$ be an orthogonal basis and $b(x) = \sum_{i=1}^n c_i\phi_i(x) + \varepsilon(x)$, where $\varepsilon(x)$ is small if the basis is big enough. Then $\mathbb{E}_{p(x)}\Phi(x)b(x) = 0 \Leftrightarrow \mathbb{E}\phi_i(x)\big[\sum_{i=1}^n c_i\phi_i(x) + \varepsilon(x)\big] = 0$ for every i, and due to the orthogonality of the basis and the smallness of $\varepsilon(x)$: $c_i \cdot \mathbb{E}_{p(x)}\phi^2(x) \approx 0$ and hence $c_i \approx 0$ and thus $b(x) \approx 0$.

inequality is often strict.

Averaging predictions. We propose a new estimator

$$\bar{\bar{\mu}}_n(x) = \frac{1}{n+1} \sum_{i=0}^{n} a'(\theta_i^\top \Phi(x)).$$

In general $\mathcal{G}(\bar{\bar{\mu}}_n) - \mathcal{G}(\mu_{**})$ does not converge to zero either (unless the feature vector Φ is large enough and Eq. (3.4.3) is satisfied). Thus, on top of the usual convergence in $O(1/n)$ with respect to the number of iterations, we have an extra term that depends only on γ, which we will study in Section 3.5 and Section 3.6.

We denote by $\bar{\bar{\mu}}_\gamma(x)$ the limit when $n \to \infty$, that is, using properties of converging Markov chains, $\bar{\bar{\mu}}_\gamma(x) = \mathbb{E}_{\pi_\gamma(\theta)} a'\big(\Phi(x)^\top \theta\big)$.

Rewriting Eq. (3.4.2) using our new notations, we get:

$$\mathbb{E}\big[(\mu_{**}(x) - \bar{\bar{\mu}}_\gamma(x))\Phi(x_n)\big] = 0.$$

When $\Phi : \mathbb{R} \to \mathbb{R}^d$ is high-dimensional, this leads to $\mu_{**} = \bar{\bar{\mu}}_\gamma$ and in contrast to $\bar{\mu}_\gamma$, averaging predictions potentially converge to the optimal prediction.

Graphical representation. We propose a schematic graphical representation of averaging estimators and averaging predictions in the Figure 3-3.

Computational complexity. Usual averaging of estimators [Polyak and Juditsky, 1992] to compute

$$\bar{\mu}_n(x) = a'(\Phi(x)^\top \bar{\theta}_n)$$

is simple to implement as we can simply update the average $\bar{\theta}_n$ with essentially no extra cost on top the complexity $O(nd)$ of the SGD recursion. Given the number n of training data points and the number m of testing data points, the overall complexity is $O(d(n+m))$.

Averaging prediction functions is more challenging as we have to store all iterates θ_i, $i=1,\dots,n$, and for each testing point x, compute

$$\bar{\bar{\mu}}_n(x) = \frac{1}{n+1} \sum_{i=0}^{n} a'(\theta_i^\top \Phi(x)).$$

Thus the overall complexity is $O(dn + mnd)$, which could be too costly with many test points (i.e., m large).

There are several ways to alleviate this extra cost: (a) using sketching techniques [Woodruff, 2014], (b) using summary statistics like done in applications of MCMC [Gilks et al., 1995], or (c) leveraging the fact that all iterates θ_i will end up being close to $\bar{\theta}_\gamma$ and use a Taylor expansion of $a'\big(\theta^\top \Phi(x)\big)$ around $\bar{\theta}_\gamma$. This expansion is equal to:

$$a'\big(\Phi(x)^\top \overline{\theta}_\gamma\big) + (\theta - \overline{\theta}_\gamma)^\top \Phi(x) \cdot a''\big(\Phi(x)^\top \overline{\theta}_\gamma\big) +$$

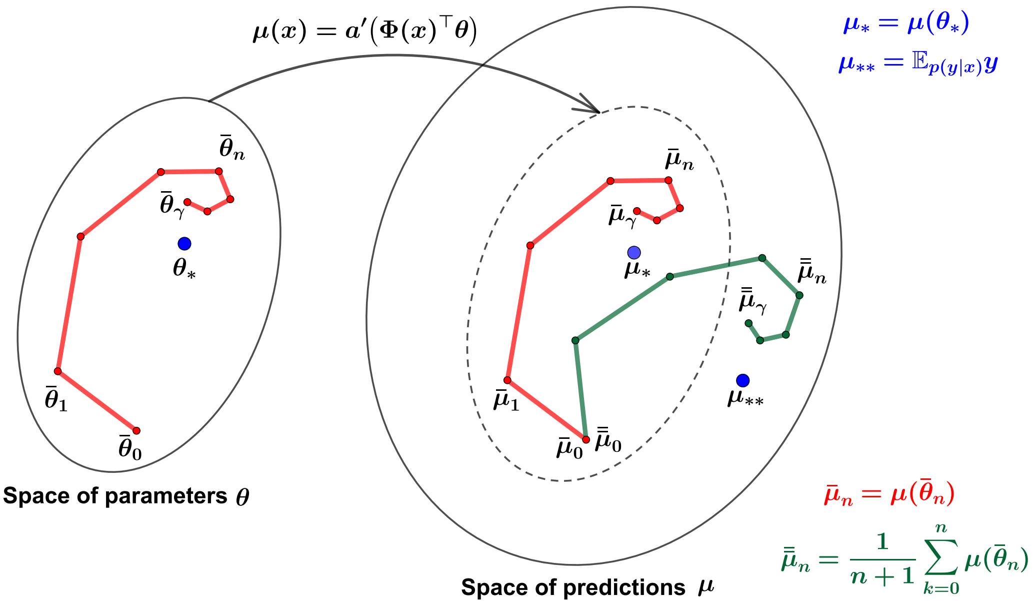
$\mu(x) = a'(\Phi(x)^\top\theta)$
$\mu_* = \mu(\theta_*)$
$\mu_{**} = \mathbb{E}_{p(y|x)} y$
$\bar\theta_n$
$\bar\theta_\gamma$
θ_*
$\bar\theta_1$
$\bar\theta_0$
Space of parameters θ
$\bar\mu_n$
$\bar\mu_\gamma$
μ_*
$\bar{\bar\mu}_n$
$\bar{\bar\mu}_\gamma$
μ_{**}
$\bar\mu_1$
$\bar\mu_0$
$\bar{\bar\mu}_0$
$\bar\mu_n = \mu(\bar\theta_n)$
$\bar{\bar\mu}_n = \frac{1}{n+1}\sum_{k=0}^{n} \mu(\bar\theta_n)$
Space of predictions μ

Figure 3-3 – Visualisation of the space of parameters and its transformation to the space of predictions. Averaging estimators in red vs averaging predictions in green. μ_* is optimal linear predictor and μ_{**} is the global optimum.

$$+\frac{1}{2}\big((\theta - \bar\theta_\gamma)^\top \Phi(x)\big)^2 \cdot a'''\big(\Phi(x)^\top \bar\theta_\gamma\big) + O\big(\|\theta - \bar\theta_\gamma\|^3\big).$$

Taking expectation in both sides above leads to:

$$\bar{\bar\mu}_\gamma(x) \approx \bar\mu_\gamma(x) + \frac{1}{2}\Phi(x)^\top \operatorname{cov}(\theta) \cdot \Phi(x) \cdot a'''\big(\bar\theta_\gamma^\top \Phi(x)\big),$$

where $\operatorname{cov}(\theta)$ is the covariance matrix of θ under π_γ. This provides a simple correction to $\bar\mu_\gamma$, and leads to an approximation of $\bar{\bar\mu}_n(x)$ as

$$\bar\mu_n(x) + \frac{1}{2}\ \Phi(x)^\top \operatorname{cov}_n(\theta)\ \Phi(x) \cdot a'''\big(\bar\theta_n^\top \Phi(x)\big),$$

where $\operatorname{cov}_n(\theta)$ is the empirical covariance matrix of the iterates (θ_i).

The computational complexity now becomes $O(nd^2 + md^2)$, which is an improvement when the number of testing points m is large. In all of our experiments, we used this approximation.

3.5 Finite-dimensional models

In this section we study the behavior of $\bar{\bar{A}}(\gamma) = \mathcal{G}(\bar{\bar\mu}_\gamma) - \mathcal{G}(\mu_*)$ for finite-dimensional models, for which it is usually not equal to zero. We know that our estimators $\bar{\bar\mu}_n$ will converge to $\bar{\bar{\mu}}_\gamma$, and our goal is to compare it to $\bar{A}(\gamma) = \mathcal{G}(\bar{\mu}_\gamma) - \mathcal{G}(\mu_*) = F(\bar{\theta}_\gamma) - F(\theta_*)$ which is what averaging estimators tends to. We also consider for completeness the non-averaged performance $A(\gamma) = \mathbb{E}_{\pi_\gamma(\theta)}\big[F(\theta) - F(\theta_*)\big]$.

Note that we must have $A(\gamma)$ and $\bar{A}(\gamma)$ non-negative, because we compare the negative log-likelihood performances to the one of of the best linear prediction (in the natural parameter), while $\bar{\bar{A}}(\gamma)$ could potentially be negative (it will in certain situations), because the the corresponding natural parameters may not be linear in $\Phi(x)$.

We consider the same standard assumptions as Dieuleveut et al. [2017], namely smoothness of the negative log-likelihoods $f_n(\theta)$ and strong convexity of the expected negative log-likelihood $F(\theta)$. We first recall the results from Dieuleveut et al. [2017]. See detailed explicit formulas in the supplementary material.

3.5.1 Earlier work

Without averaging. We have that $A(\gamma) = \gamma B + O(\gamma^{3/2})$, that is γ is *linear* in γ, with B non-negative.

Averaging estimators. We have that $\bar{A}(\gamma) = \gamma^2 \bar{B} + O(\gamma^{5/2})$, that is $\bar{A}$ is *quadratic* in γ, with $\bar{B}$ non-negative. Averaging does provably bring some benefits because the order in γ is higher (we assume γ small).

3.5.2 Averaging predictions

We are now ready to analyze the behavior of our new framework of averaging predictions. The following result is shown in the supplementary material.

Proposition 3.1. *Under the assumptions on the negative loglikelihoods f_n of each observation from Dieuleveut et al. [2017]:*

- *In the case of well-specified data, that is, there exists θ_* such that for all (x, y), $p(y|x) = q(y|x, \theta_*)$, then $\bar{\bar{A}} \sim \gamma^2 \bar{\bar{B}}^{\mathrm{well}}$, where $\bar{\bar{B}}^{\mathrm{well}}$ is a positive constant.*
- *In the general case of potentially mis-specified data, $\bar{\bar{A}} = \gamma \bar{\bar{B}}^{\mathrm{mis}} + O(\gamma^2)$, where $\bar{\bar{B}}^{\mathrm{mis}}$ is constant which may be positive or negative.*

Note, that in contrast to averaging parameters, the constant $\bar{\bar{B}}^{\mathrm{mis}}$ can be negative. It means, that we obtain the estimator better than the optimal linear estimator, which is the limit of capacity for averaging parameters. In our simulations, we show examples for which $\bar{\bar{B}}^{\mathrm{mis}}$ is positive, and examples for which it is negative. Thus, in general, for low-dimensional models, averaging predictions can be worse or better than averaging parameters. However, as we show in the next section, for infinite dimensional models, we always get convergence.

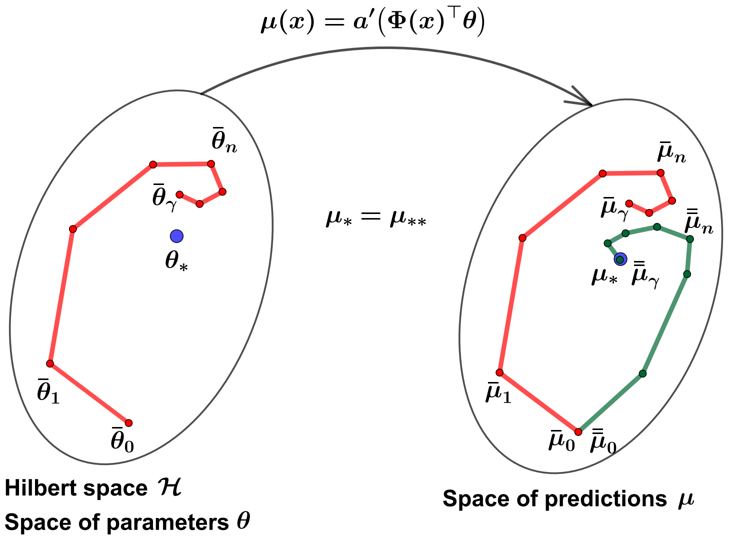

Figure 3-4 – Visualisation of the space of parameters and its transformation to the space of predictions. Averaging estimators in red vs averaging predictions in green. Global optimizer coincides with the best linear: $\mu_* = \mu_{**}$.

3.6 Infinite-dimensional models

Recall, that we have the following objective function to minimize:

$$F(\theta) = \mathbb{E}_{x,y}\Big[-y \cdot \eta_\theta(x) + a\big(\eta_\theta(x)\big)\Big], \tag{3.6.1}$$

where till this moment we consider unknown functions $\eta_\theta(x)$ which were linear in $\Phi(x)$ with $\Phi(x) \in \mathbb{R}^d$, leading to a complexity in $O(dn)$.

We now consider infinite-dimensional features, by considering that $\Phi(x) \in \mathcal{H}$, where $\mathcal{H}$ is a Hilbert space. Note that this corresponds to modeling the function η_θ as a Gaussian process [Rasmussen and Williams, 2006].

This is computationally feasible through the usual "kernel trick" [Scholkopf and Smola, 2001, Shawe-Taylor and Cristianini, 2004], where we assume that the kernel function $k(x, y) = \langle \Phi(x), \Phi(y) \rangle$ is easy to compute. Indeed, following Bordes et al. [2005] and Dieuleveut and Bach [2016], starting from θ_0, each iterate of constant-step-size SGD is of the form $\theta_n = \sum_{t=1}^n \alpha_t \Phi(x_t)$, and the gradient descent recursion $\theta_n = \theta_{n-1} - \gamma[a'(\eta_{\theta_{n-1}}(x_n)) - y_n]\Phi(x_n)$ leads to the following recursion on α_t's:

$$\begin{aligned} \alpha_n &= -\gamma\big[a'\big(\textstyle\sum_{t=1}^{n-1} \alpha_t \langle \Phi(x_t), \Phi(x_n) \rangle\big) - y_n\big] \\ &= -\gamma\big[a'\big(\textstyle\sum_{t=1}^{n-1} \alpha_t k(x_t, x_n)\big) - y_n\big]. \end{aligned}$$

This leads to $\eta_{\theta_n}(x) = \langle \Phi(x), \theta_n \rangle$ and $\mu_{\theta_n}(x) = a'\big(\eta_{\theta_n}(x)\big)$ with

$$\eta_{\theta_n}(x) = \sum_{t=1}^{n} \alpha_t \langle \Phi(x), \Phi(x_t) \rangle = \sum_{t=1}^{n} \alpha_t k(x, x_t),$$

and finally we can express $\bar{\bar{\mu}}_n(x)$ in kernel form as:

$$\bar{\bar{\mu}}_n(x) = \frac{1}{n+1} \sum_{t=0}^{n} a' \Big[\sum_{l=1}^{t} \alpha_l \cdot k(x, x_l) \Big].$$

There is also a straightforward estimator for averaging parameters, i.e., $\bar{\mu}_n(x) = a'\Big(\frac{1}{n+1} \sum_{t=0}^{n} \sum_{l=1}^{t} \alpha_l k(x, x_l)\Big)$. If we assume that the kernel function is *universal*, that is, $\mathcal{H}$ is dense in the space of squared integrable functions, then it is known that if $\mathbb{E}_x b(x)\Phi(x) = 0$, then $b = 0$ [Sriperumbudur et al., 2008]. This implies that we must have $\bar{\bar{\mu}}_\gamma = 0$ and thus averaging predictions does always converge to the global optimum (note that in this setting, we must have a well-specified model because we are in a non-parametric setting).

Graphical representation. We propose a schematic graphical representation of averaging estimators and averaging predictions for Hilbert space setup in the Figure 3-4.

Column sampling. Because of the usual high running-time complexity of kernel method in $O(n^2)$, we consider a "column-sampling approximation" [Williams and Seeger, 2001]. We thus choose a small subset $I = (x_1, \ldots, x_m)$ of samples and construct a new finite m-dimensional feature map $\bar{\Phi}(x) = K(I, I)^{-1/2} K(I, x) \in \mathbb{R}^m$, where $K(I, I)$ is the $m \times m$ kernel matrix of the m points and $K(I, x)$ the vector composed of kernel evaluations $k(x_i, x)$. This allows a running-time complexity in $O(m^2 n)$. In theory and practice, m can be chosen small [Bach, 2013, Rudi et al., 2017].

Regularized learning with kernels. Although we can use an unregularized recursion with good convergence properties [Dieuleveut and Bach, 2016], adding a regularisation by the squared Hilbertian norm is easier to analyze and more stable with limited amounts of data. We thus consider the recursion (in Hilbert space), with λ small:

$$\begin{aligned} \theta_n &= \theta_{n-1} - \gamma \big[f_n'(\theta_{n-1}) + \lambda \theta_{n-1} \big] \\ &= \theta_{n-1} + \gamma (y_n - a'(\langle \Phi(x_n), \theta \rangle)) \Phi(x_n) - \gamma \lambda \theta_{n-1}. \end{aligned}$$

This recursion can also be computed efficiently as above using the kernel trick and column sampling approximations.

In terms of convergence, the best we can hope for is to converge to the minimizer

$\theta_{*,\lambda}$ of the regularized expected negative log-likelihood $F(\theta)+\frac{\lambda}{2}\|\theta\|^2$ (which we assume to exist). When λ tends to zero, then $\theta_{*,\lambda}$ converges to θ_*.

Averaging *parameters* will tend to a limit $\bar{\theta}_{\gamma,\lambda}$ which is $O(\gamma)$-close to $\theta_{*,\lambda}$, thus leading to predictions which deviate from the optimal predictions for two reasons: because of regularization and because of the constant step-size. Since λ should decrease as we get more data, the first effect will vanish, while the second will not.

When averaging *predictions*, the two effects will vanish as λ tends to zero. Indeed, by taking limits of the gradient equation, and denoting by $\bar{\bar{\mu}}_{\gamma,\lambda}$ the limit of $\bar{\bar{\mu}}_n$, we have

$$\mathbb{E}\big[(\mu_{**}(x)-\bar{\bar{\mu}}_{\gamma,\lambda}(x))\Phi(x)\big]=\lambda\bar{\theta}_{\gamma,\lambda}. \tag{3.6.2}$$

Given that $\bar{\theta}_{\gamma,\lambda}$ is $O(\gamma)$-away from θ_*, if we assume that θ_* corresponds to a sufficiently regular [2] element of the Hilbert space $\mathcal{H}$, then the L_2-norm of the deviation satisfies $\|\mu_{**}(x)-\bar{\bar{\mu}}_{\gamma,\lambda}\|=O(\lambda)$ and thus as the regularization parameter λ tends to zero, our predictions tend to the optimal one.

3.7 Experiments

In this section, we compare the two types of averaging (estimators and predictions) on a variety of problems, both on synthetic data and on standard benchmarks. When averaging predictions, we always consider the Taylor expansion approximation presented at the end of Section 3.4.3.

3.7.1 Synthetic data

Finite-dimensional models. we consider the following logistic regression model:

$$q(y|x,\theta)=\exp\big(y\cdot\eta_\theta(x)-a(\eta_\theta(x))\big),$$

where we consider a linear model $\eta_\theta(x)=\theta^\top x$ in x (i.e., $\Phi(x)=x$), the link function $a(t)=\log(1+e^t)$ and $a'(t)=\sigma(t)$ is the sigmoid function. Let x be distributed as a standard normal random variable in dimension $d=2$, $y\in\{0,1\}$ and $\mathbb{P}(y=1|x)=\mu_{**}(x)=\sigma\big(\eta_{**}(x)\big)$, where we consider two different settings:

— Model 1: $\eta_{**}(x)=\sin x_1+\sin x_2$,

— Model 2: $\eta_{**}(x)=x_1^3+x_2^3$.

The global minimum $\mathcal{F}_{**}$ of the corresponding optimization problem can be found as

$$\mathcal{F}_{**}=\mathbb{E}_{p(x)}\big[-\mu_{**}(x)\cdot\eta_{**}(x)+a(\eta_{**}(x))\big].$$

We also introduce the performance measure $\mathcal{F}(\eta)$

$$\mathcal{F}(\eta)=\mathbb{E}_{p(x)}\big[-\mu_{**}(x)\cdot\eta(x)+a(\eta(x))\big], \tag{3.7.1}$$

[2] While our reasoning is informal here, it can be made more precise by considering so-called "source conditions" commonly used in the analysis of kernel methods [Caponnetto and De Vito, 2007], but this is out of the scope of this chapter.

which can be evaluated directly in the case of synthetic data. Note that in our situation, the model is misspecified because $\eta_{**}(x)$ is not linear in $\Phi(x) = x$, and thus, $\inf_\theta F(\theta) > \mathcal{F}_{**}$, and thus our performance measures $\mathcal{F}(\mu_n) - \mathcal{F}_{**}$ for various estimators μ_n will not converge to zero.

The results of averaging 10 replications are shown in Fig. 3-5 and Fig. 3-6. We first observed that constant step-size SGD without averaging leads to a bad performance.

Moreover, we can see, that in the first case (Fig. 3-5) averaging predictions beats averaging parameters, and moreover beats the best linear model: if we use the best linear error $\mathcal{F}_*$ instead of $\mathcal{F}_{**}$, at some moment $\mathcal{F}(\eta_n) - \mathcal{F}_*$ becomes negative. However in the second case (Fig. 3-6), averaging predictions is not superior to averaging parameters. Moreover, by looking at the final differences between performances with different values of γ, we can see the dependency of the final performance in γ for averaging predictions, instead of γ^2 for averaging parameters (as suggested by our theoretical results in Section 3.5). In particular in Fig. 3-5, we can observe the surprising behavior of a larger step-size leading to a better performance (note that we cannot increase too much otherwise the algorithm would diverge).

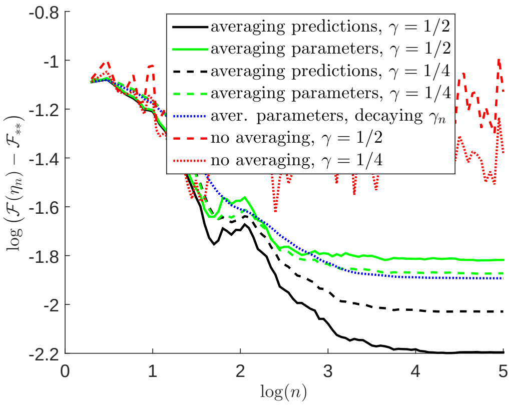

Figure 3-5 – Synthetic data for linear model $\eta_\theta(x) = \theta^\top x$ and $\eta_{**}(x) = \sin x_1 + \sin x_2$. Excess prediction performance vs. number of iterations (both in log-scale).

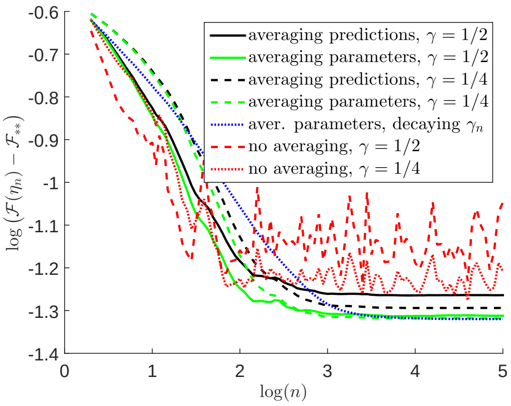

Figure 3-6 – Synthetic data for linear model $\eta_\theta(x) = \theta^\top x$ and $\eta_{**}(x) = x_1^3 + x_2^3$. Excess prediction performance vs. number of iterations (both in log-scale).

Infinite-dimensional models Here we consider the kernel setup described in Section 3.6. We consider Laplacian kernels $k(s,t) = \exp\left(\frac{\|s-t\|_1}{\sigma}\right)$ with $\sigma = 50$, dimension $d = 5$ and generating log odds ratio $\eta_{**}(x) = \frac{5}{5+x^\top x}$. We also use a squared norm regularization with several values of λ and column sampling with $m = 100$ points. We use the exact value of $\mathcal{F}_{**}$ which we can compute directly for synthetic data. The results are shown in Fig. 3-7, where averaging predictions leads to a better performance than averaging estimators.

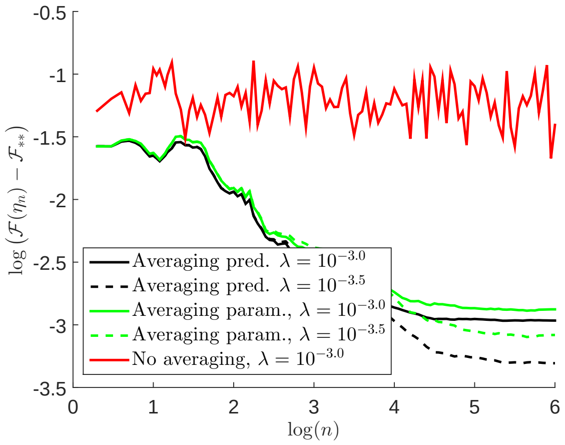

Figure 3-7 – Synthetic data for Laplacian kernel for $\eta_{**}(x) = \frac{5}{5+x^\top x}$. Excess prediction performance vs. number of iterations (both in log-scale).

3.7.2 Real data

Note, that in the case of real data, one does not have access to unknown $\mu_{**}(x)$ and computing the performance measure in Eq. (3.7.1) is inapplicable. Instead of it we use its sampled version on held out data:

$$\hat{\mathcal{F}}(\eta) = -\sum_{i:y_i=1} \log\left(\mu(x_i)\right) - \sum_{j:y_j=0} \log\left(1 - \mu(x_i)\right).$$

We use two datasets, with d not too large, and n large, from [Lichman, 2013]: the "MiniBooNE particle identification" dataset ($d = 50$, $n = 130\ 064$), the "Covertype" dataset ($d = 54$, $n = 581\ 012$).

We use two different approaches for each of them: a linear model $\eta_\theta(x) = \theta^\top x$ and a kernel approach with Laplacian kernel $k(s,t) = \exp\left(\frac{\|s-t\|_1}{\sigma}\right)$, where $\sigma = d$. The results are shown in Figures 3-8 to 3-11. Note, that for linear models we use $\hat{\mathcal{F}}_*$–the estimator of the best performance among linear models (learned on the test set, and hence not reachable from learning on the training data), and for kernels we use $\hat{\mathcal{F}}_{**}$ (same definition as $\hat{\mathcal{F}}_*$ but with the kernelized model), that is why graphs are not comparable (but, as shown below, the value of $\hat{\mathcal{F}}_{**}$ is lower than the value of $\hat{\mathcal{F}}_*$ because using kernels correspond to a larger feature space).

For the "MiniBooNE particle identification" dataset $\hat{\mathcal{F}}_* = 0.35$ and $\hat{\mathcal{F}}_{**} = 0.21$; for the"Covertype" dataset $\hat{\mathcal{F}}_* = 0.46$ and $\hat{\mathcal{F}}_{**} = 0.39$. We can see from the four plots that, especially in the kernel setting, averaging predictions also shows better performance than averaging parameters.

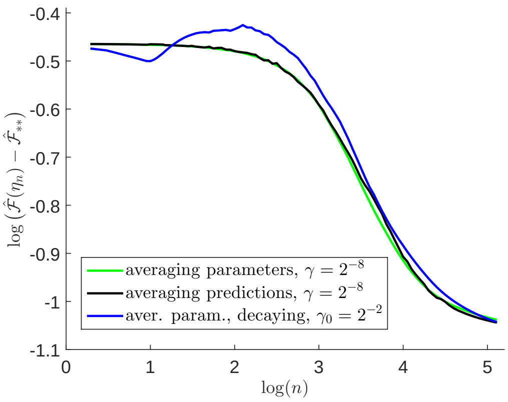

Figure 3-8 – MiniBooNE dataset, dimension $d = 50$, linear model. Excess prediction performance vs. number of iterations (both in log-scale).

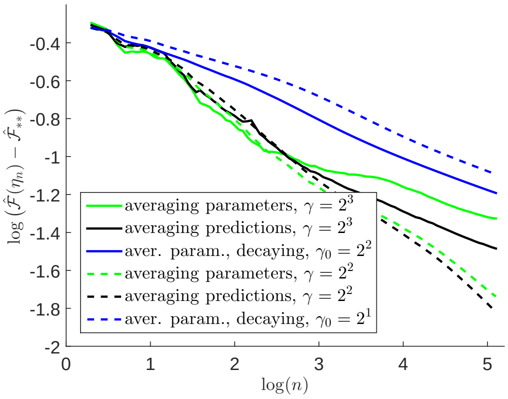

Figure 3-9 – MiniBooNE dataset, dimension $d = 50$, kernel approach, column sampling $m = 200$. Excess prediction performance vs. number of iterations (both in log-scale).

3.8 Conclusion

In this chapter, we have explored how averaging procedures in stochastic gradient descent, which are crucial for fast convergence, could be improved by looking at the specifics of probabilistic modeling. Namely, averaging in the moment parameterization can have better properties than averaging in the natural parameterization.

While we have provided some theoretical arguments (asymptotic expansion in the finite-dimensional case, convergence to optimal predictions in the infinite-dimensional case), a detailed theoretical analysis with explicit convergence rates would provide a better understanding of the benefits of averaging predictions.

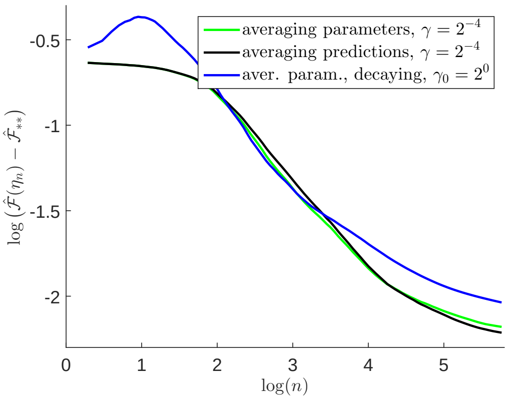

Figure 3-10 – CoverType dataset, dimension $d = 54$, linear model. Excess prediction performance vs. number of iterations (both in log-scale).

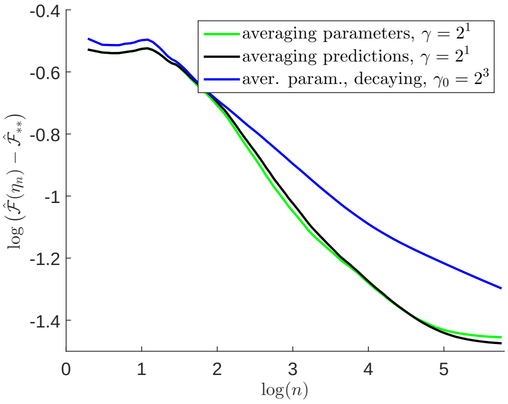

Figure 3-11 – CoverType dataset, dimension $d = 54$, kernel approach, column sampling $m = 200$. Excess prediction performance vs. number of iterations (both in log-scale).

3.9 Appendix. Explicit form of $B, \bar{B}, \bar{\bar{B}}^w$ and $\bar{\bar{B}}^m$

In this appendix we provide explicit expressions for the asymptotic expansions from the chapter. All assumptions from Dieuleveut et al. [2017] are reused, namely, beyond the usual sampling assumptions, smoothness of the cost functions.

We have, even for mis-specified models:

$$\mathcal{F}(\eta) - \mathcal{F}(\eta_*) = \mathcal{G}(\mu) - \mathcal{G}(\mu_*) =$$

$$= \mathbb{E}\big[- \mathbb{E}_{p(y_n|x_n)} y_n \eta(x_n) + a(\eta(x_n)) + \mathbb{E}_{p(y_n|x_n)} y_n \eta_*(x_n) - a(\eta_*(x_n))\big] =$$

$$= \mathbb{E}\Big[a(\eta(x_n)) - a(\eta_*(x_n)) - a'(\eta_*(x_n))(\eta(x_n) - \eta_*(x_n))\Big] +$$

$$+ \mathbb{E}\Big[\big(a'(\eta_*(x_n)) - \mathbb{E}_{p(y_n|x_n)} y_n\big) \cdot \big(\eta(x_n) - \eta_*(x_n)\big)\Big] =$$

$$\mathbb{E}\Big[D_a\big(\eta(x_n)\big|\eta_*(x_n)\big)\Big] + \mathbb{E}\Big[\big(\mu_*(x) - \mu_{**}(x)\big) \cdot \big(\eta(x_n) - \eta_*(x_n)\big)\Big] =$$

$$\mathbb{E}\Big[D_{a^*}\big(\mu(x_n)\big|\mu_*(x_n)\big)\Big] + \mathbb{E}\Big[\big(\mu_*(x) - \mu_{**}(x)\big) \cdot \eta(x_n)\Big] \tag{3.9.1}$$

for D_a the Bregman divergence associated to a, and D_{a^*} the one associated to a^*. We also use the optimality condition for the predictor $\eta_*(x)$: $\mathbb{E}\eta_*(x)[a'(\eta_*(x)) - \mathbb{E}_{p(x|y)} y] = 0$ in the last step.

When the model is well-specified, we have $a'(\eta_*(x_n)) = \mathbb{E}(y_n|x_n)$ and thus $F(\eta) - F(\eta_*) = \mathbb{E}\big[D_{a^*}(\mu_*(x_n)||\mu(x_n))\big]$. If η is linear in $\Phi(x)$, and even if the model is mis-specified, then we also have $F(\eta) - F(\eta_*) = \mathbb{E}\big[D_{a^*}(\mu_*(x_n)||\mu(x_n))\big]$.

Using asymptotic expansions of moments of the averaged SGD iterate with zero-mean statistically independent noise $f_n(\theta) = F(\theta) + \varepsilon_n(\theta)$ from Dieuleveut and Bach [2016], Theorem 2 one obtains:

$$\overline{\theta}_\gamma = \mathbb{E}_{\pi_\gamma}(\theta) = \theta_* + \gamma\Delta + O(\gamma^2), \tag{3.9.2}$$

$$\mathbb{E}_{\pi_\gamma}(\theta - \theta_*)(\theta - \theta_*)^\top = \gamma C + O(\gamma^2), \tag{3.9.3}$$

where

$$C = \big[F''(\theta_*) \otimes I + I \otimes F''(\theta_*)\big]^{-1}\Sigma.$$

and $\Sigma = \int_{\mathbb{R}^d} \varepsilon(\theta)^{\otimes 2} \pi_\gamma(d\theta) \in \mathbb{R}^{d\times d}$.

The "drift" $\overline{\theta}_\gamma - \theta_*$ is linear in γ and can be interpreted as an additional error due to the function is not being quadratic and step sizes are not decaying to zero.

Connection between Δ and C can be easily obtained using $\theta_n = \theta_{n-1} - \gamma\big[F'(\theta_{n-1}) + \varepsilon_n\big]$. Taking expectation of both parts and using Tailor expansion up to the second order:

$$F''(\theta_*)(\overline{\theta}_\gamma - \theta_*) = -\frac{1}{2}F'''(\theta_*)\mathbb{E}_{\pi_\gamma}(\theta - \theta_*)^{\otimes 2} \Rightarrow$$

$$F''(\theta_*)\Delta = -\frac{1}{2}F'''(\theta_*)C. \tag{3.9.4}$$

3.9.1 Estimation without averaging

We start with the simplest estimator of the prediction function: $\mu_0(x) = a'(\Phi^\top \theta_n)$, where we do not use any averaging:

$$\mathcal{G}(\mu_n) - \mathcal{G}(\mu_*) = f(\theta_n) - f(\theta_*) = f'(\theta_*)(\theta_n - \theta_*) + \frac{1}{2} f''(\theta_*)(\theta_n - \theta_*)^{\otimes 2} +$$

$$+ \frac{1}{6} f'''(\theta_*)(\theta_n - \theta_*)^{\otimes 3} + O(\gamma^{3/2})$$

Taking expectation of both sides, when $n \to \infty$ and using Eq. (3.9.3) one obtains:

$$A(\gamma) = \mathbb{E}_{\pi_\gamma} f(\theta_n) - f(\theta_*) = \frac{1}{2} \mathrm{tr} f''(\theta_*) \gamma C + O(\gamma^{3/2}).$$

So, we have linear dependence of γ and $B = \frac{1}{2} \mathrm{tr} f''(\theta_*) C$.

3.9.2 Estimation with averaging parameters

Now, let us estimate $\overline{A}(\gamma)$:

$$\mathcal{G}(\bar{\mu}_n) - \mathcal{G}(\mu_*) = f(\bar{\theta}_n) - f(\theta_*) = f'(\theta_*)(\bar{\theta}_n - \theta_*) + \frac{1}{2}(\bar{\theta}_n - \theta_*) f''(\theta_*)(\bar{\theta}_n - \theta_*) + O(\gamma^3).$$

Taking expectation of both sides, when $n \to \infty$:

$$\mathcal{G}(\bar{\mu}_\gamma) - \mathcal{G}(\mu_*) = f(\bar{\theta}_\gamma) - f(\theta_*) = \frac{1}{2} \mathrm{tr} f''(\theta_*)(\bar{\theta}_\gamma - \theta_*)^{\otimes 2} + O\big(\gamma^3\big) =$$

$$= \frac{1}{2} \mathrm{tr} f''(\theta_*) \gamma^2 \Delta^{\otimes 2} + O(\gamma^3).$$

Finally we have a quadratic dependence of γ:

$$\bar{A}(\gamma) = \frac{1}{2} \mathrm{tr} f''(\theta_*) \gamma^2 \Delta^{\otimes 2} + O(\gamma^3).$$

And the coefficient $\bar{B} = \frac{1}{2} \mathrm{tr} f''(\theta_*) \Delta^{\otimes 2}$.

3.9.3 Estimation with averaging predictions

Recall, that by definition, $\bar{\bar{A}}(\gamma) = \mathcal{G}(\bar{\bar{\mu}}_\gamma) - \mathcal{G}(\mu_*)$, where $\bar{\bar{\mu}}_\gamma(x) = \mathbb{E}_{\pi_\gamma} a'\big(\theta^\top \Phi(x)\big)$. We again use Tailor expansion for $a'\big(\theta^\top \Phi(x)\big)$ at θ^*:

$$a'\big(\theta^\top \Phi(x)\big) = a'\big(\theta_*^\top \Phi(x)\big) + a''\big(\theta_*^\top \Phi(x)\big)(\theta - \theta_*)^\top \Phi(x) +$$

$$+ \frac{1}{2} a'''\big(\theta_*^\top \Phi(x)\big) \cdot \Big((\theta - \theta_*)^\top \Phi(x)\Big)^2 + O(\gamma^{3/2}).$$

Taking expectation of both parts:

$$\bar{\bar{\mu}}_\gamma(x) = \mu_*(x) + a''\big(\theta_*^\top \Phi(x)\big)\big(\bar{\theta}_\gamma - \theta_*\big)^\top \Phi(x) +$$

$$+\frac{1}{2}a'''\big(\theta_*^\top \Phi(x)\big)\mathrm{tr}\big[\Phi(x)\Phi(x)^\top \mathbb{E}(\theta - \theta_*)^{\otimes 2}\big] + O(\gamma^{3/2}) =$$

$$= \mu_*(x) + a''\big(\theta_*^\top \Phi(x)\big)\gamma\Delta^\top \Phi(x) + \frac{1}{2}a'''\big(\theta_*^\top \Phi(x)\big)\mathrm{tr}\big[\Phi(x)\Phi(x)^\top \gamma C\big] + O(\gamma^{3/2}).$$

Finally, we showed, that:

$$\bar{\bar{\mu}}_\gamma(x) - \mu_*(x) = O(\gamma^{3/2}) + \gamma\Big[a''\big(\eta_*(x)\big)\Delta^\top \Phi(x) + \frac{1}{2}a'''\big(\eta_*(x)\big)\mathrm{tr}\big[\Phi(x)^{\otimes 2}C\big]\Big]$$

Now we use Bregram divergence notation Eq. (3.9.1):

$$\bar{\bar{A}}(\gamma) = \mathcal{G}(\bar{\bar{\mu}}_\gamma) - \mathcal{G}(\mu_*) = \mathcal{G}_1 + \mathcal{G}_2,$$

As mentioned above, the term $\mathcal{G}_2$ vanishes if model is well-specified or η is linear in $\Phi(x)$. Note, that for the case $\bar{A}(\gamma)$ indeed linear in $\Phi(x)$.

Estimation of $\mathcal{G}_1$.

By definition $D_{a^*}(\mu_*(x)||\mu(x)) = \dfrac{1}{2}\big(\mu_*(x) - \mu(x)\big)(a^*)''\big(\mu_*(x)\big)\big(\mu_*(x) - \mu(x)\big)$ and

$$\mathcal{G}_1 = \frac{1}{2}\mathbb{E}\frac{\big(\mu_*(x) - \bar{\bar{\mu}}_\gamma(x)\big)^2}{a''\big(\theta_*^\top \Phi(x)\big)} = \frac{\gamma^2}{2}\mathbb{E}\Big[\frac{(a''\big(\eta_*(x)\big)\Delta^\top \Phi(x) + \frac{1}{2}a'''\big(\eta_*(x)\big)\mathrm{tr}\big[\Phi(x)^{\otimes 2}C\big])^2}{a''\big(\theta_*^\top \Phi(x)\big)}\Big].$$

Since

$$\mathbb{E}_x\Big[a''(\big(\theta_*^\top \Phi(x)\big)\big(\Delta^\top \Phi(x)\big)^2\Big] = \Delta^\top f''(\theta_*)\Delta$$

and

$$\mathbb{E}_x\Big[\Delta^\top a'''\big(\theta_*^\top \Phi(x)\big)\Phi(x)^{\otimes 3}C\Big] = \Delta^\top f'''(\theta_*)C = -2\Delta^\top f''(\theta_*)\Delta,$$

$$\mathcal{G}_1 = \gamma^2\Bigg[-\frac{1}{2}\Delta^\top f''(\theta_*)\Delta + \frac{1}{8}\mathbb{E}\Big[\frac{a'''(\eta_*(x))^2}{a''(\eta_*(x))}\cdot\big(\mathrm{tr}\big[\Phi(x)^{\otimes 2}C\big]\big)^2\Big]\Bigg] + O(\gamma^3)$$

And the coefficient $\bar{\bar{B}}^w = -\frac{1}{2}\Delta^\top f''(\theta_*)\Delta + \frac{1}{8}\mathbb{E}\Big[\frac{a'''(\eta_*(x))^2}{a''(\eta_*(x))}\cdot\big(\mathrm{tr}\big[\Phi(x)^{\otimes 2}C\big]\big)^2\Big]$.

Estimation of $\mathcal{G}_2$.

$$\mathcal{G}_2 = \mathbb{E}\Big[\big(\mu_*(x) - \mu_{**}(x)\big)\cdot\big(\bar{\bar{\eta}}(x_n) - \eta_*(x_n)\big)\Big],$$

using properties of conjugated functions,

$$\mathcal{G}_2 = \mathbb{E}\Big[\big((a^*)'(\bar{\bar{\mu}}(x)) - (a^*)'(\mu_*(x))\big)\cdot\big(\mu_*(x) - \mu_{**}(x)\big)\Big] =$$

$$\mathbb{E}\Big[(a^*)''\big(\bar{\bar{\mu}}_*(x))(\bar{\bar{\mu}}(x) - \mu_*(x)\big) \cdot \big(\mu_*(x) - \mu_{**}(x)\big) + O(\gamma^2) =$$

$$= \mathbb{E}\ \frac{\bar{\bar{\mu}}(x) - \mu_*(x)}{a''(\eta_*(x))} \cdot \big(\mu_*(x) - \mu_{**}(x)\big) + O(\gamma^2) =$$

$$= \gamma \cdot \mathbb{E}\Bigg[\bigg(\Delta^\top \Phi(x) + \frac{a'''(\eta_*(x))}{2a''(\eta_*(x))}\mathrm{tr}\big[\Phi(x)^{\otimes 2}C\big]\bigg) \cdot \bigg(\mu_*(x) - \mu_{**}(x)\bigg)\Bigg] + O(\gamma^2).$$

And the coefficient $\bar{\bar{B}}^m = \mathbb{E}\Big(\Delta^\top \Phi(x) + \frac{a'''(\eta_*(x))}{2a''(\eta_*(x))}\mathrm{tr}\big[\Phi(x)^{\otimes 2}C\big]\Big) \cdot \Big(\mu_*(x) - \mu_{**}(x)\Big)$.

Chapter 4

Sublinear Primal-Dual Algorithms for Large-Scale Multiclass Classification

Abstract

In this chapter, we study multiclass classification problems with potentially large number of classes k, number of features d, and sample size n. Our goal is to develop efficient algorithms to train linear classifiers with losses of a certain type, allowing to address, in particular, multiclass support vector machines (SVM) and softmax regression. The developed algorithms are based on first-order proximal primal-dual schemes – Mirror Descent and Mirror Prox. In particular, forthe multiclass SVM with elementwise ℓ_1-regularization we propose a sublinear algorithm with numerical complexity of one iteration being only $O(d+n+k)$. This result relies on the combination of three ideas: (i) passing to the saddle-point problem with a quasi-bilinear objective; (ii) ad-hoc variance reduction technique based on a non-uniform sampling of matrix multiplication; (iii) the proper choice of the proximal setup in Mirror Descent type schemes leading to the balance between the stochastic and deterministic error terms.

The material presented in this chapter is a joint work with Dmitrii M. Ostrovskii and Francis Bach. It is submitted to the ICML 2019 conference.

4.1 Introduction

We focus on efficient training of multiclass linear classifiers, potentially with very large numbers of classes and dimensionality of the feature space, and with losses of a certain type. Formally, consider a dataset comprised of n pairs (x_i, y_i), $i = 1, ..., n$, where $x_i \in \mathbb{R}^d$ is the feature vector of the i-th training example, and $y_i \in \{e_1, ..., e_k\}$ is the label vector encoding one of the k possible classes, $e_1, ..., e_k \subseteq \mathbb{R}^k$ being the standard basis vectors (we assume there are no ambiguities in the class assignment). Given such data, we would like to find a linear classifier that minimizes the regularized empirical risk. This can be formalized as the following minimization problem:

$$\min_{U \in \mathbb{R}^{d\times k}} \frac{1}{n}\sum_{i=1}^{n} \ell(U^\top x_i, y_i) + \lambda \|U\|_{\mathscr{U}}. \tag{4.1.1}$$

Here, $U \in \mathbb{R}^{d\times k}$ is the matrix whose columns specify the parameter vectors for each of the k classes; $\ell(U^\top x, y)$, with $\ell : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}$, is the *loss* corresponding to the margins $U^\top x \in \mathbb{R}^k$ assigned to x when its true class happens to be y; finally, $\lambda\|U\|_{\mathscr{U}}$, with $\lambda > 0$, is the regularization term corresponding to some norm $\|\cdot\|_{\mathscr{U}}$ on $\mathbb{R}^{d\times k}$ which will be specified later on (cf. Section 4.2). For now, we only assume that $\|\cdot\|_{\mathscr{U}}$ is "simple" – intuitively, one can think of $\|U\|_{\mathscr{U}}$ as being quasi-separable in the elements of U, for example, belong to the class of row-wise mixed $\ell_p \times \ell_q$-norms, cf. Section 4.2.

Our focus is on the so-called *Fenchel-Young* losses [Blondel et al., 2018] which can be expressed in the form

$$\ell(U^\top x, y) = \max_{v \in \Delta_k} \left\{ -\mathfrak{f}(v, y) + (v - y)^\top U^\top x \right\}, \tag{4.1.2}$$

where Δ_k is the probability simplex in $\mathbb{R}^k$, and the function $\mathfrak{f} : \Delta_k \to \mathbb{R}$ is convex and "simple" (i.e., quasi-separable in v) implying that the maximization in (4.1.2) can be performed in running time $O(k)$. In particular, this allows to encompass two most commonly used multiclass losses:

— the multiclass *logistic (or softmax) loss*

$$\ell(U^\top x, y) = \log\left(\sum_{j=1}^{k} \exp(U_j^\top x)\right) - y^\top U^\top x, \tag{4.1.3}$$

where U_j is the j-th column of U so that $U_j^\top x$ is the j-th element of $U^\top x$; this loss corresponds to (4.1.2) with the negative entropy $\mathfrak{f}(v, y) = \sum_{i=1}^k v_i \log v_i$;

— the multiclass *hinge loss*, given by

$$\ell(U^\top x, y) = \max_{j \in \{1,\dots,k\}} \left\{ \mathbb{1}[e_j \neq y] + U_j^\top x \right\} - y^\top U^\top x, \tag{4.1.4}$$

and used in multiclass support vector machines (SVM), reduces to (4.1.2) by setting $\mathfrak{f}(v, y) = v^\top y - 1$ (see Appendix 4.7.1 for details).

Additional examples of Fenchel-Young losses are described by Blondel et al. [2018].

Arranging the feature vectors into the design matrix $X \in \mathbb{R}^{n\times d}$, and the label vectors in the matrix $Y \in \mathbb{R}^{n\times k}$, and using the Fenchel dual representation (4.1.2) of the loss, we recast the initial minimization problem (4.1.1) as a convex-concave saddle-point problem of the following form:

$$\min_{U \in \mathbb{R}^{d\times k}} \max_{V \in \mathcal{V}} -\mathcal{F}(V, Y) + \frac{1}{n}\mathrm{tr}\left[(V - Y)^\top X U\right] + \lambda \|U\|_{\mathscr{U}}. \tag{4.1.5}$$

Here we denote

$$\mathcal{F}(V,Y) := \frac{1}{n}\sum_{i=1}^{n} \mathfrak{f}(v_i, y_i) \tag{4.1.6}$$

where $v_i, y_i \in \Delta^k$ are the i-th rows of V and Y, whereas

$$\mathcal{V} := [\Delta_k^{\otimes n}]^\top \subset \mathbb{R}^{n\times k} \tag{4.1.7}$$

is the Cartesian product of probability simplices comprising all matrices in $\mathbb{R}^{n\times k}$ with rows belonging to Δ_k. Taking into account the availability of the Fenchel dual representation (4.1.2), recasting the convex minimization problem (4.1.1) in the form (4.1.5) is quite natural. Indeed, while the objective in (4.1.1) can be non-smooth, the "non-trivial" part of the objective in (4.1.5), given by

$$\Phi(U, V-Y) := \frac{1}{n}\mathrm{tr}\left[(V-Y)^\top XU\right], \tag{4.1.8}$$

is necessarily *bilinear*. On the other hand, the presence of the dual constraints, as given by (4.1.7), also does not seem problematic since $\mathcal{V}$ allows for a computationally cheap projection oracle. Finally, the primal-dual formulation (4.1.5) allows one to control the duality gap, thus providing *online accuracy certificates* for the initial problem (see Nemirovski and Onn [2010]).

In this chapter, we develop efficient algorithms for solving (4.1.1) via the associated saddle-point problem (4.1.5), building upon the two basic schemes for saddle-point optimization: Mirror Descent and Mirror Prox (see Juditsky and Nemirovski [2011a,b], Nesterov and Nemirovski [2013] and references therein). When applied to the class of general convex-concave saddle-point problems with available first-order information given by the partial (sub-)gradients of the objective, these basic schemes are well suited for obtaining *medium-accuracy* solutions – which is not a limitation in the context of empirical risk minimization, where the ultimate goal is to minimize the *true (expected)* risk.

Moreover, these basic schemes must be able, at least in principle, to capture the specific structure of quasi-bilinear saddle-point problems of the form (4.1.5). First, they rely on general Bregman divergences, rather than the standard Euclidean distance, to measure proximity of the points, which allows one to adjust to the particular non-Euclidean geometry associated to the set $\mathcal{V}$ and the norm $\|\cdot\|_{\mathscr{U}}$. Second, Mirror Descent and Mirror Prox retain their favorable convergence guarantees when the partial gradients are replaced with their unbiased estimates. To see why this circumstance is so important, recall that the computation of the partial gradients $\nabla_U[\Phi(U, V-Y)]$ and $\nabla_V[\Phi(U, V-Y)]$, cf. (4.1.8), is reduced to the computation of the matrix products XU and $X^\top(V-Y)$, and thus requires $O(dnk)$ operations. Given the partial gradients, the proximal steps associated to U and V variables, assuming the "simplicity" of $\|\cdot\|_{\mathscr{U}}$ in the aforementioned sense, can be computed in $O(dk+nk)$. Thus, in the regime interesting to us, where both d and k can be very large, the computation of these matrix products becomes the main bottleneck. On the other hand, unbiased estimates of these matrix products, with reduced complexity of computation, can be be obtained by *randomized subsampling* of the rows and columns of U, V, Y, and X. While this approach has been extensively studied by Juditsky and Nemirovski [2011b] in the case of bilinear saddle-point problems with vector-valued variables arising in sparse signal recovery, its extension to problems of the type (4.1.5) is non-trivial. In fact, for a sampling scheme to be deemed "good", it clearly has to satisfy two concurrent requirements:

(a) On one hand, one must control the stochastic variability of the estimates in the chosen sampling scheme. Ideally, the variance – or its suitable analogue in the case of a non-Euclidean geometry – should be comparable to the squared norm of X whose choice depends on the proximal setup.

(b) On the other hand, the estimates must be cheap to compute, much cheaper than $O(dnk)$ required for the full gradient computation. While the apparent goal is $O(dk+nk)$ per iteration (the cost of the primal-dual proximal mapping for the full gradients), one could even want to go beyond that, to $O(d+n+k)$ per iteration, possibly paying a larger price once, during pre/post-processing.

Devising a sampling scheme satisfying these requirements is a delicate task. To solve it, one should exploit the geometric structure of (4.1.5) associated to $\mathcal{V}$ and $\|\cdot\|_{\mathscr{U}}$.

Contributions and outline. We identify "good" sampling schemes satisfying the above requirements, rigorously analyze their statistical properties, carefuly implement the basic algorithms equipped with these sampling schemes, and analyze the total computational complexity of the resulting algorithms. Namely, we consider the situation where $\|\cdot\|_{\mathscr{U}}$ is the entrywise ℓ_1-norm, which corresponds to the implicit assumption that both the classes and the features are sparse. Moreover, we choose the proximal setup adapted to the geometry of the saddle-point problem (4.1.5), thus achieving favorable accuracy guarantees for both of the basic algorithms *without subsampling* (cf. Sections 4.2–4.2.1). In this setup, we propose two sampling schemes with various levels of "aggressiveness", and study their statistical properties, showing that both of the basic primal-dual algorithms preserve their favorable accuracy bounds when combined with the proposed sampling schemes. At the same time, equipping the algorithms with these sampling schemes drastically reduces the total *numerical* complexity (cf. (b)), with improvement depending on the sampling scheme:

— The *Partial Sampling Scheme*, described in Section 4.3.1, is applicable to any problem of the form (4.1.5). Here the main idea is to sample only one row of U and V at a time, with probabilities nearly minimizing the variance proxy of the current estimates of U and V. After careful implementation of the primal-dual proximal step, this leads to the cost $O(dk+nk+dn)$ of one iteration.

— In the more aggressive *Full Sampling Scheme*, described in Section 4.3.2, sampling of the rows of U and V is augmented with column sampling. Applying this scheme to multiclass hinge loss (4.1.4), we are able to carry out the updates in only $O(d+k+n)$ arithmetic operations, with additional cost

$$O(dn+(d+n)\cdot\min(k,T))$$

of pre- and post-processing, T being the number of iterations. Note that these numerical complexity estimates can be justified from the viewpoint of statistical learning theory: according to it, $T = O(n)$ iterations of a stochastic first-order algorithm must be sufficient to extract most of statistical information contained in the sample, since at each iteration we effectively select a single training example by subsampling a row of V. We see that in the high-dimensional regime $d, k \gg n$, this criterion is satisfied since the price of pre/post-processing becomes $O(dn)$.

We conclude the chapter with numerical experiments on synthetic data presented in Section 4.5.

4.1.1 Related work

While we have passed to the saddle-point problem (4.1.5), there is still an option to solve the original problem (4.1.1). The complexity of deterministic approach for it is $O(dnk)$ and the complexity of stochastic approach is $O(dk)$, and because of the structure of the problem, the ideas of Full Sampling Scheme cannot be used. Due to the easy access to noisy gradients, the classical approach is stochastic gradient descent (SGD). Algorithms such as SVRG [Johnson and Zhang, 2013], SDCA [Shalev-Shwartz and Zhang, 2013], SAGA [Defazio et al., 2014], SAG [Schmidt et al., 2017], large-scale linearly convergent algorithms [Palaniappan and Bach, 2016], Breg-SVRG [Shi et al., 2017] use variance-reduced techniques to obtain accelerated convergence rates, but all have $O(dk)$ or $O(dk + nk)$ complexity. Note, however, that our variance reduction technique is different from all these methods, and none of these approaches is sublinear.

Sublinear algorithms. Formally, we call an algorithm *sublinear* if the computational complexity of one iteration has a linear (or linear with a logarithmic factor) dependence on natural dimensions of data – such as sample size, number of features and number of classes. For the biclass setting several results can be found in the literature. Probably, the first result for bilinear problems was considered by Grigoriadis and Khachiyan [1995], which was also considered by Juditsky and Nemirovski [2011b], Xiao et al. [2017]. A sublinear algorithm for SVM was considered by Hazan et al. [2011], for semidefinite programs was considered by Garber and Hazan [2011, 2016] and more general results were given by Clarkson et al. [2012]. However, none of these approaches can be easily extended to the multiclass setting without an extra $O(k)$ factor appearing in the numerical complexity bounds.

4.2 Choice of geometry and basic routines

Preliminaries. We focus on the convex-concave saddle-point problem of the form:

$$\min_{U \in \mathbb{R}^{d\times k}} \max_{V \in (\Delta_k^\top)^{\otimes n}} -\mathcal{F}(V, Y) + \Phi(U, V - Y) + \lambda \|U\|_{\mathscr{U}}, \qquad (4.2.1)$$

with the bilinear part $\Phi(U, V - Y)$ of the objective being given by (4.1.8), and the composite term $\mathcal{F}(V, Y)$ by (4.1.6). From now on, we make the following mild assumption:

Assumption 1. *The $\|\cdot\|_{\mathscr{U}}$-radius of some optimal solution U^* to (4.2.1) is known:*

$$\|U^*\|_{\mathscr{U}} = R_{\mathcal{U}}.$$

Remark 1. *All the algorithms presented later on preserve their accuracy bounds when $R_{\mathcal{U}}$ is an upper bound on $\|U^*\|_{\mathscr{U}}$ instead of being its exact value. However in this case they become suboptimal by factors proportional to the looseness of this bound, hence we are interested in this upper bound to be a tight as possible. Note that since $U = 0$ is a feasible solution, in the case of positive losses $R_{\mathcal{U}} \leqslant 1/\lambda$, but this bound is usually very loose, since λ must decrease with the number of observations as dictated by statistical learning theory. A better approach is to solve a series of problems, starting with a small radius, and increasing it exponentially until the obtained solution leaves the boundary of the feasible set. This strategy works when the solution set is bounded. Since the complexity bounds that we obtain grow linearly with $R_{\mathcal{U}}$, this method enjoys the same overall complexity bound, up to a constant factor, as in the case when Assumption 1 is precisely satisfied.*

With Assumption 1, we can recast the problem (4.2.1) in the constrained form:

$$\min_{U \in \mathcal{U}} \max_{V \in \mathcal{V}} -\mathcal{F}(V, Y) + \Phi(U, V - Y) + \lambda \|U\|_{\mathscr{U}}, \tag{4.2.2}$$

where $\mathcal{U} := \{U \in \mathbb{R}^{d \times k} : \|U\|_{\mathscr{U}} \le R_{\mathcal{U}}\}$ is a $\|\cdot\|_{\mathscr{U}}$-norm ball, and $\mathcal{V} := [\Delta_k^{\otimes n}]^\top$ is the direct product of n probability simplices in $\mathbb{R}^k$. This problem has a very particular structure associated to the sets $\mathcal{U}$ and $\mathcal{V}$, and ideally, this structure should be exploited by the optimization algorithm in order to obtain efficiency estimates scaling with the "correct" norm of the optimal primal-dual solution, see, e.g., Juditsky and Nemirovski [2011a], Nesterov and Nemirovski [2013], Beck and Teboulle [2003], Shi et al. [2017]. This issue can be addressed in the framework of *proximal* algorithms, in which one starts by choosing the norms that capture the inherent geometry of the problem, and then replaces the usual (projected) gradient step with the *proximal step*, i.e., uses the general Bregman divergence corresponding to some *potential* (also called *distance-generating function*) instead of the squared Euclidean distance. The potential must be *compatible* with the chosen norm in the sense of Juditsky and Nemirovski [2011a] (essentially, grow as the squared norm while being 1-strongly convex with respect to the norm, see Section 4.2.2), and at the same time allow for an efficient implementation of the proximal step. We will discuss the question of choosing the potentials in the next section, when describing Mirror Descent [see e.g. Nemirovsky and Yudin, 1983], the simplest general proximal algorithm applicable to (4.2.2). Before that, let us specify the choice of the norms themselves.

Choice of the norms. A natural strategy for choosing the norm in each variable is by ensuring, whenever possible, that its ball of a certain radius is simply the convex hull of the symmetrization of the feasible set (note that scaling is not important since first-order algorithms are invariant with respect to it), see, e.g., Juditsky and Nemirovski [2011a]. Following this strategy, we must choose $\|\cdot\|_{\mathscr{U}}$ itself for U, whatever is the norm $\|\cdot\|_{\mathscr{U}}$. On the other hand, the norm $\|\cdot\|_{\mathscr{U}}$ has not yet been defined at this point; correspondingly, the "implicit" geometry of the problem in U has not been specified. For some reasons explained in Section 4.4, we choose the elementwise ℓ_1-norm for U:

$$\|U\|_{\mathscr{U}} = \|U\|_{1\times 1} = \sum_{i=1}^{d}\sum_{j=1}^{k} |U(i,j)|, \tag{4.2.3}$$

where we use the "Matlab" indexing notation, and define the mixed $\ell_p \times \ell_q$ norms by

$$\|A\|_{p\times q} := \left(\sum_{i} \|A(i,:)\|_q^p\right)^{1/p}, \quad p, q \geq 1, \tag{4.2.4}$$

i.e., the ℓ_p-norm of the vector of ℓ_q-norms of the individual rows of the matrix. Note that this choice of $\|\cdot\|_{\mathscr{U}}$ favors solutions that are sparse in terms of both features and classes.

As for the norm on $\mathcal{V}$, one could choose the ℓ_1-norm in the case $n = 1$, and $\ell_\infty \times \ell_1$-norm in the general case (recall that $\mathcal{V}$ is the direct product of simplices). However, it is known (see Juditsky and Nemirovski [2011a]) that for the ℓ_∞-norm there is no compatible potential in the sense of Juditsky and Nemirovski [2011a] (the existence of such a potential would cotradict the known worst-case complexity lower bounds [Nemirovsky and Yudin, 1983]. The remedy, leading to near-optimal accuracy estimates, is to replace the ℓ_∞-norm with the ℓ_2-norm, leading us to the $\ell_2 \times \ell_1$-norm for V:

$$\|V\|_{\mathscr{V}} = \|V\|_{2\times 1} = \left(\sum_{i=1}^{n} \|V(i,:)\|_1^2\right)^{1/2}. \tag{4.2.5}$$

Alternative choices for both norms will be briefly discussed in Section 4.4. However, we will see that this choice of norms, when equipped with the properly chosen potentials, is the only one in a broad class of those using the mixed $\ell_p \times \ell_q$ norms, for which we can achieve the goal stated in Section 4.1, that is, to combine near-optimal complexity estimates with a numerically efficient implementation of the proximal step.

4.2.1 Basic schemes: Mirror Descent and Mirror Prox

Basic schemes in minimization problems. The classical *Mirror Descent* scheme was introduced by Nemirovsky and Yudin [1983], and generalizes the standard Projected Subgradient Descent to the case of non-Euclidean distance measures. The general presentation of Mirror Descent is described by Beck and Teboulle [2003]; here we only provide a brief exposition. When minimizing a convex function $f(u)$ on a domain $\mathcal{U}$, the Mirror Descent scheme amounts to choosing the potential $\phi_{\mathcal{U}}(u)$ that generalizes the squared Euclidean distance $\frac{1}{2}\|u\|_2^2$, and the sequence of stepsizes γ_t, starting with the initial point $u^0 = \min_{u\in\mathcal{U}} \phi_{\mathcal{U}}(u)$ called the *prox-center*, and then performing iterations of the form

$$u^{t+1} = \operatorname*{arg\,min}_{u\in\mathcal{U}} \left\{ \langle \nabla f(u^t), u - u^t\rangle + \frac{1}{\gamma_t} D_{\phi_{\mathcal{U}}}(u, u^t) \right\},$$

where

$$D_\phi(u, u^t) = \phi(u) - \phi(u^{t-1}) - \langle \nabla\phi(u^t), u - u^t\rangle$$

is the Bregman divergence between the candidate point u and the previous point u^t that corresponds to the chosen potential $\phi(\cdot) = \phi_{\mathcal{U}}(\cdot)$ and replaces the squared Euclidean distance $\frac{1}{2}\|u - u^t\|_2^2$. After computing T iterates, the scheme outputs the averaged point $\bar{u}^T$ as the candidate solution; one can choose the uniform averaging $\bar{u}^T = \frac{1}{T}\sum_{t=0}^{T-1} u^t$ or more complex averaging schedules (for example, with the weights proportional to the stepsizes); for simplicity, we will focus on the uniform averaging.

The counterpart of the Mirror Descent scheme, called *Mirror Prox*, was introduced by Nemirovski [2004], and combines the use of Bregman divergences with an extrapolation step, first proposed in the Euclidean setting by Korpelevich [1977]. In the context of minimization problems, the difference is that Mirror Prox iterates as

$$\begin{aligned} u^{t+1/2} &= \operatorname*{arg\,min}_{u\in\mathcal{U}} \left\{ \langle \nabla f(u^t), u - u^t\rangle + \frac{1}{\gamma_t} D_{\phi_{\mathcal{U}}}(u, u^t) \right\}, \\ u^{t+1} &= \operatorname*{arg\,min}_{u\in\mathcal{U}} \left\{ \langle \nabla f(u^{t+1/2}), u - u^{t+1/2}\rangle + \frac{1}{\gamma_t} D_{\phi_{\mathcal{U}}}(u, u^t) \right\}; \end{aligned}$$

in other words, one first performs the proximal step from the current point u^t to obtain the auxilliary point $u^{t+1/2}$, and then perfoms the *extragradient* step from u^t, i.e., the proximal step in which the gradient at u^t is replaced with that at $u^{t+1/2}$.

Basic schemes for composite saddle-point problems. Both approaches can be extended to composite minimization and, most importantly in our context, to *composite saddle-point problems* such as (4.2.2). In particular, introducing the combined variable $W = (U, V) \in \mathcal{W} \quad [:= \mathcal{U} \times \mathcal{V}]$, the *(Composite) Mirror Descent* scheme for the saddle-point problem (4.2.2) first constructs the joint potential $\phi_{\mathcal{W}}(W)$ from the intial potentials $\phi_{\mathcal{U}}$ and $\phi_{\mathcal{V}}$, in the way specified later on, and then performs iterations

$$\begin{aligned} W^0 &= \min_{W\in\mathcal{W}} \phi_{\mathcal{W}}(W); \\ W^{t+1} &= \operatorname*{arg\,min}_{W\in\mathcal{W}} \left\{ h(W) + \langle G(W^t), W\rangle + \frac{1}{\gamma_t} D_{\phi_{\mathcal{W}}}(W, W^t) \right\}, \ t \ge 1 \end{aligned} \tag{4.2.6}$$

where $h(W) = \mathcal{F}(V, Y) + \lambda\|U\|_{\mathcal{U}}$ is the combined composite term, cf. (4.1.6), and $G(W)$ is the vector field of the partial gradients of $\Phi(U, V - Y)$, cf. (4.1.8):

$$G(W) := (\nabla_U \Phi(U, V - Y), -\nabla_V \Phi(U, V - Y)) = (X^\top(V - Y), -XU). \tag{4.2.7}$$

Accordingly, the *(Composite) Mirror Prox* scheme for (4.2.2) performs iterations

$$\begin{aligned} W^0 &= \min_{W \in \mathcal{W}} \phi_{\mathcal{W}}(W); \\ W^{t+1/2} &= \operatorname*{arg\,min}_{W \in \mathcal{W}} \left\{ h(W) + \langle G(W^t), W \rangle + \frac{1}{\gamma_t} D_{\phi_{\mathcal{W}}}(W, W^t) \right\}, \\ W^{t+1} &= \operatorname*{arg\,min}_{W \in \mathcal{W}} \left\{ h(W) + \langle G(W^{t+1/2}), W \rangle + \frac{1}{\gamma_t} D_{\phi_{\mathcal{W}}}(W, W^t) \right\}, \quad t \geqslant 1. \end{aligned} \tag{4.2.8}$$

Note that the joint minimization in (4.2.6) and (4.2.8) can be performed separately in U and V as long as the joint potential is a linear combination of $\phi_{\mathcal{U}}$ and $\phi_{\mathcal{V}}$. To specify the accuracy guarantees for both schemes, we need to introduce two objects: *the potential differences* $\Omega_{\mathcal{U}}$ and $\Omega_{\mathcal{V}}$, and the *"cross" Lipschitz constant* $\mathcal{L}_{\mathscr{U},\mathscr{V}}$. The choice of the joint potential will be given once we give the definitions of these objects.

Differences of potentials. The potential differences (called "omega-radii" in the literature co-authored by A. Nemirovski), are defined by

$$\Omega_{\mathcal{U}} = \max_{U \in \mathcal{U}} \phi_{\mathcal{U}}(U) - \min_{U \in \mathcal{U}} \phi_{\mathcal{U}}(U), \quad \Omega_{\mathcal{V}} = \max_{V \in \mathcal{V}} \phi_{\mathcal{V}}(V) - \min_{V \in \mathcal{V}} \phi_{\mathcal{V}}(V); \tag{4.2.9}$$

note that all maxima and minima are attained when the potentials are continuous, and $\mathcal{U}, \mathcal{V}$ are compact. When the potentials $\phi_{\mathcal{U}}, \phi_{\mathcal{V}}$ are compatible with the corresponding norms $\|\cdot\|_{\mathcal{U}}, \|\cdot\|_{\mathcal{V}}$ in the sense of Juditsky and Nemirovski [2011a] (see Section 4.2.2), the potential differences grow as the squared radii of $\mathcal{U}$ and $\mathcal{V}$ in the corresponding norms, up to factors logarithmic in the dimensions of $\mathcal{U}, \mathcal{V}$, see, e.g., Nemirovsky and Yudin [1983], Shapiro et al. [2009]; in particular, this holds for the choice of $\phi_{\mathcal{U}}$ and $\phi_{\mathcal{V}}$ discussed later on.

"Cross" Lipschitz constant. Given a smooth convex-concave function $\widetilde{\Phi}(U, V) = \Phi(U, V - Y)$, the "cross" Lipschitz constant $\mathcal{L}_{\mathscr{U},\mathscr{V}}$ of the field

$$G(W) := (\nabla_U \Psi(U, V), -\nabla_V \Psi(U, V))$$

with respect to the pair of norms $\|\cdot\|_{\mathscr{U}}, \|\cdot\|_{\mathscr{V}}$ is defined as $\mathcal{L}_{\mathscr{U},\mathscr{V}} = \max(\mathcal{L}_{\mathscr{U}\to\mathscr{V}}, \mathcal{L}_{\mathscr{V}\to\mathscr{U}})$, where $\mathcal{L}_{\mathscr{U}\to\mathscr{V}}$, $\mathcal{L}_{\mathscr{V}\to\mathscr{U}}$ deliver tight inequalities of the form

$$\begin{aligned} \|\nabla_U \Psi(U', V) - \nabla_U \Psi(U, V)\|_{\mathscr{V}^*} &\leqslant \mathcal{L}_{\mathscr{U}\to\mathscr{V}} \|U' - U\|_{\mathscr{U}}, \\ \|\nabla_V \Psi(U, V') - \nabla_V \Psi(U, V)\|_{\mathscr{U}^*} &\leqslant \mathcal{L}_{\mathscr{V}\to\mathscr{U}} \|V' - V\|_{\mathscr{V}}, \end{aligned}$$

uniformly over $\mathcal{U} \times \mathcal{V}$, where $\|\cdot\|_{\mathscr{U}^*}, \|\cdot\|_{\mathscr{V}^*}$ are the dual norms to $\|\cdot\|_{\mathscr{U}}, \|\cdot\|_{\mathscr{V}}$. For the bilinear function $\Phi(U, V - Y)$ given by (4.1.8), $\mathcal{L}_{\mathscr{U}\to\mathscr{V}}$ and $\mathcal{L}_{\mathscr{V}\to\mathscr{U}}$ are equal, and we can express $\mathcal{L}_{\mathscr{U},\mathscr{V}}$ as a norm of the linear operator acting on $\mathbb{R}^{d\times k} \to \mathbb{R}^{n\times k}$ as $U \mapsto \frac{1}{n} XU$:

$$\mathcal{L}_{\mathscr{U},\mathscr{V}} = \frac{1}{n} \left[\|X\|_{\mathscr{U}\to\mathscr{V}} := \sup_{\|U\|_{\mathscr{U}} \leq 1} \|XU\|_{\mathscr{V}^*} \right]. \tag{4.2.10}$$

Furthermore, for the chosen norms $\|\cdot\|_{\mathscr{U}} = \|\cdot\|_{1\times 1}$ and $\|\cdot\|_{\mathscr{V}} = \|\cdot\|_{2\times 1}$ we can express $\mathcal{L}_{\mathscr{U},\mathscr{V}}$ as a certain mixed norm (cf. (4.2.3)–(4.2.5)) as stated by the following lemma proved in Appendix:

Proposition 4.1. *The "cross" Lipschitz constant can be expressed as*

$$\mathcal{L}_{\mathscr{U},\mathscr{V}} = \frac{\|X^\top\|_{\infty\times 2}}{n}. \tag{4.2.11}$$

Constructing the joint potential. Given the partial potentials $\phi_{\mathcal{U}}(\cdot)$ and $\phi_{\mathcal{V}}(\cdot)$, the natural way to construct the joint potential $\phi_{\mathcal{W}}$ is by taking a weighted average of $\phi_{\mathcal{U}}(\cdot)$ and $\phi_{\mathcal{V}}(\cdot)$, since this leads to $\phi_{\mathcal{W}}$ being separable in U and V, and allows to exploit the structure of the partial potentials to obtain accuracy guarantees. In fact, one can show that the simplistic choice $\phi_{\mathcal{W}}(W) = (\phi_{\mathcal{U}}(U) + \phi_{\mathcal{V}}(V))/2$ leads to the accuracy guarantee proportional to $\mathcal{L}_{\mathscr{U},\mathscr{V}}(\Omega_{\mathcal{U}}+\Omega_{\mathcal{V}})$ where $\mathcal{L}_{\mathscr{U},\mathscr{V}}$ is the "cross" Lipschitz constant, and $\Omega_{\mathcal{U}}, \Omega_{\mathcal{V}}$ are the potential differences defined above. On the other hand, if $\Omega_{\mathcal{U}}$ and $\Omega_{\mathcal{V}}$ are *known*, one can consider, following Juditsky and Nemirovski [2011b] and Ostrovskii and Harchaoui [2018], the "balanced" joint potential

$$\phi_{\mathcal{W}}(W) = \frac{\phi_{\mathcal{U}}(U)}{2\Omega_{\mathcal{U}}} + \frac{\phi_{\mathcal{V}}(V)}{2\Omega_{\mathcal{V}}}, \tag{4.2.12}$$

for which one can achieve the better accuracy guarantee proportional to $\mathcal{L}_{\mathscr{U},\mathscr{V}}\sqrt{\Omega_{\mathcal{U}}\Omega_{\mathcal{V}}}$, cf. Theorem 4.3 in Appendix. Moreover, this construction is, in a sense, "robust": if the ratio of the weights in (4.2.12) is multiplied with a constant factor, the accuracy estimate is preserved up to a constant factor. Besides, note that for the choice of $\phi_{\mathcal{V}}$ considered below $\Omega_{\mathcal{V}}$ is known; for the choice of $\phi_{\mathcal{U}}$ considered below $\Omega_{\mathcal{V}}$ is known when Assumption 1 is satisfied, and is known up to a constant factor when one allows $R_{\mathcal{U}}$ to be an upper bound for $\|U^*\|_{\mathscr{U}}$ as explained in Remark 1. For all these reasons, we choose the joint potential according to (4.2.12).

4.2.2 Choice of the partial potentials

We now discuss construction of the partial potentials for the chosen geometry as given by the norms $\|\cdot\|_{\mathscr{U}}$, $\|\cdot\|_{\mathscr{V}}$, cf. (4.2.3), (4.2.5). The choice of potentials for the alternative geometries is discussed in Section 4.4.

Potential for the dual variable. Since $V \in (\Delta_k^\top)^{\otimes n}$ is the direct product of probability simplices, the natural choice for the dual potential $\phi_{\mathcal{V}}(\cdot)$ is the sum of negative entropies [Beck and Teboulle, 2003]: denoting $\log(\cdot)$ the natural logarithm,

$$\phi_{\mathcal{V}}(V) = \sum_{i=1}^{n}\sum_{j=1}^{k} V_{ij}\log(V_{ij}). \tag{4.2.13}$$

This choice reflects the fact that $V \in \mathcal{V}$ corresponds to the marginals of an n-fold product distribution for which the entropy is the sum of the marginal entropies. By

Pinsker's inequality Kemperman [1969], $\phi_{\mathcal{V}}(V)$ is 1-strongly convex on $\mathcal{V}$. On the other hand,

$$\Omega_{\mathcal{V}} = n \log k, \tag{4.2.14}$$

whereas the squared $\|\cdot\|_{\mathscr{V}}$-norm of any feasible solution V to (4.2.2) is precisely n, cf. (4.2.5). Finally, $\phi_{\mathcal{V}}(V)$ is clearly continuously differentiable in the interior of $\mathcal{U}$. As such, we see that the potential (4.2.13) is *compatible* with $\mathcal{U}$ and $\|\cdot\|_{\mathscr{U}}$, i.e., the triple $(\mathcal{U}, \|\cdot\|_{\mathscr{U}}, \phi_{V}(\cdot))$ comprises a valid *proximal setup* in the sense of Juditsky and Nemirovski [2011a], and grows nearly as the squared $\|\cdot\|_{\mathscr{V}}$-radius of the feasible set $\mathcal{V}$. Note that the Bregman divergence corresponding to (4.2.13) is the sum of the Kullback-Leibler divergences between the individual rows:

$$D_{\phi_{\mathcal{V}}}(V, V^t) = \sum_{i=1}^{n} D_{\text{KL}}(V(i,:) \| V^t(i,:)), \tag{4.2.15}$$

where $D_{D_{\text{KL}}}(p,q) := \sum_j p_j \log(p_j/q_j)$ for two discrete measures p, q (not necessarily normalized to one).

Final reduction and the primal potential. Recall that we chose the elementwise norm $\|\cdot\|_{\mathscr{U}} = \|U\|_{1\times 1}$, and the primal feasible set of the problem (4.2.2) corresponds to the ball with radius $R_{\mathcal{U}}$ in this norm. In this situation, the standard option is the power potential $\phi(U) = C_{d,k}\|U\|_p^2$, where $p = 1 + 1/\log(dk)$, and the constant $C_{d,k}$ can be chosen, depending on d and k, so that $\phi(\cdot)$ is 1-strongly convex on the whole space $\mathbb{R}^{d\times k}$, see Nesterov and Nemirovski [2013]. This leads to the compatible proximal setup in the sense defined above, and the correct scaling $\Omega_{\mathcal{U}} = O(\log(dk) R_{\mathcal{U}}^2)$.

However, it turns out that the specific algebraic structure of this potential does not allow for "sublinear" numerical complexity in the full sampling scheme which we will present later on in Section 4.3.2. Hence, we advocate an alternative approach: first transform the ℓ_1-constraint into the "solid" simplex constraint (borrowing the idea from Juditsky and Nemirovski [2011b]), and then construct a compatible potential for the new problem based on the *unnormalized* negative entropy, using the fact that $R_{\mathcal{U}}$ is known, cf. Assumption 1. To this end, let $\widehat{\mathcal{U}}$ be the "solid" simplex in $\mathbb{R}^{2d\times k}$:

$$\widehat{\mathcal{U}} := \left\{ \widehat{U} \in \mathbb{R}^{2d\times k} : \widehat{U}_{ij} \geqslant 0, \;\; \text{tr}[\mathbb{1}_{2d\times k}^{\top} \widehat{U}] \leqslant R_{\mathcal{U}} \right\}, \tag{4.2.16}$$

where $\mathbb{1}_{2d\times k} \in \mathbb{R}^{2d\times k}$ is the matrix of all ones so that $\text{tr}[\mathbb{1}_{2d\times k}^{\top}\widehat{U}] = \sum_{i,j} \widehat{U}_{ij} = \|\widehat{U}\|_{1\times 1}$ whenever $\widehat{U} \in \widehat{\mathcal{U}}$. Consider the following saddle-point problem:

$$\min_{\widehat{U} \in \widehat{\mathcal{U}}} \max_{V \in \mathcal{V}} -\mathcal{F}(V,Y) + \widehat{\Phi}(\widehat{U}, V - Y) + \lambda \, \text{tr}[\mathbb{1}_{2d\times k}^{\top}\widehat{U}], \tag{4.2.17}$$

where $\mathcal{F}(V,Y)$ and $\mathcal{V}$ are the same as before (cf. (4.1.6)–(4.1.7)), and

$$\widehat{\Phi}(\widehat{U}, V-Y) := \frac{1}{n}\text{tr}\left[(V-Y)^{\top}\widehat{X}\widehat{U}\right], \quad \text{where} \;\; \widehat{X} := [X, -X] \in \mathbb{R}^{n\times 2d}. \tag{4.2.18}$$

It can be easily verified that the new saddle-point problem (4.2.17) is equivalent to (4.2.2) in the following sense: any ε-accurate (in the sence of duality gap, cf Section 4.2.4) solution $(\widehat{U}, V)$ to (4.2.17) with $\widehat{U} = [\widehat{U}_1; \widehat{U}_2]$ provides an ε-accurate solution (U, V) to (4.2.2) by taking $U = \widehat{U}_1 - \widehat{U}_2$.[1] Clearly, the quantities $\mathcal{L}_{\mathscr{U},\mathscr{V}}, \Omega_{\mathcal{U}}, \Omega_{\mathcal{V}}$ for the new problem remain the same as their counterparts for the problem (4.2.2); the only difference is a slight change of $\Omega_{\mathcal{U}}$ (see below). On the other hand, the primal feasible set of the new problem (4.2.17) – the "solid" simplex (4.2.16) – admits a compatible entropy-type potential. Indeed, consider first the "unit solid" simplex given by (4.2.16) with $R_{\mathcal{U}} = 1$. On this set, the *unnormalized negative entropy*

$$\mathcal{H}(\widehat{U}) = \sum_{i,j} \widehat{U}_{ij} \log \widehat{U}_{ij} - \widehat{U}_{ij} \tag{4.2.19}$$

is 1-strongly convex (see Yu [2013] for an elementary proof), and clearly is continuously differentiable in its interior; thus, $\mathcal{H}(\cdot)$ is a compatible potential on the "unit solid" simplex in the sense of Juditsky and Nemirovski [2011a]. Finally, using Assumption 1, we can construct a compatible (i.e., 1-strongly convex and continuously differentiable in the interior) potential on the initial "solid" simplex (4.2.16) with arbitrary radius by scaling the argument of $\mathcal{H}(\cdot)$ and then renormalizing it as follows:

$$\begin{aligned} \phi_{\widehat{\mathcal{U}}}(\widehat{U}) &:= R_{\mathcal{U}}^2 \cdot \mathcal{H}(\widehat{U}/R_{\mathcal{U}}) \\ &= R_{\mathcal{U}} \left[\sum_{i=1}^{2d} \sum_{j=1}^{k} \widehat{U}_{ij} \log \left(\frac{\widehat{U}_{ij}}{R_{\mathcal{U}}} \right) - \widehat{U}_{ij} \right]. \end{aligned} \tag{4.2.20}$$

Note that the potential difference in this case is

$$\Omega_{\widehat{\mathcal{U}}} = R_{\mathcal{U}}^2 \log(2dk), \tag{4.2.21}$$

and the corresponding Bregman divergence is given by

$$D_{\phi_{\widehat{\mathcal{U}}}}(\widehat{U}, \widehat{U}^t) = R_{\mathcal{U}}^2 D_{\text{KL}} \left(\frac{\widehat{U}}{R_{\mathcal{U}}} \middle\| \frac{\widehat{U}^t}{R_{\mathcal{U}}} \right) - R_{\mathcal{U}} \operatorname{tr}[\mathbb{1}_{2d\times k}^\top \widehat{U}] + R_{\mathcal{U}} \operatorname{tr}[\mathbb{1}_{2d\times k}^\top \widehat{U}^t], \tag{4.2.22}$$

where the last term does not depend on $\widehat{U}$.

4.2.3 Recap of the deterministic algorithms

We can now formulate the iterations of the Mirror Descent and Mirror Prox schemes applied to the reformulation (4.2.17) of the saddle-point problem (4.2.2)

[1] The idea of this reduction is that for any optimal primal solution $\widehat{U}$ to (4.2.17), only one of the blocks $\widehat{U}_1, \widehat{U}_2$ is non-zero, and these blocks can then be interpreted as the positive and negative parts of the corresponding optimal solution to (4.2.2).

with the specified choice of geometry. Both of them are initialized with

$$V^0 = \frac{\mathbb{1}_{n\times k}}{k}, \quad \widehat{U}^0 = \frac{R_{\mathcal{U}}\mathbb{1}_{2d\times k}}{2dk}, \tag{\textbf{Init}}$$

corresponding to the uniform distributions. **Mirror Descent** then iterates ($t \geq 1$):

$$\boxed{\begin{aligned} V^{t+1} &= \underset{V\in\mathcal{V}}{\arg\min}\left\{\mathcal{F}(V,Y) - \widehat{\Phi}(\widehat{U}^t, V) + \frac{D_{\phi_{\mathcal{V}}}(V,V^t)}{2\gamma_t\Omega_{\mathcal{V}}}\right\}, \\ \widehat{U}^{t+1} &= \underset{\widehat{U}\in\widehat{\mathcal{U}}}{\arg\min}\left\{\lambda\,\mathrm{tr}[\mathbb{1}_{2d\times k}^\top\widehat{U}] + \widehat{\Phi}(\widehat{U}, V^t - Y) + \frac{D_{\phi_{\widehat{\mathcal{U}}}}(\widehat{U},\widehat{U}^t)}{2\gamma_t\Omega_{\widehat{\mathcal{U}}}}\right\}, \end{aligned}} \tag{\textbf{MD}}$$

where $\mathcal{F}(V,Y)$, $\widehat{\Phi}(\cdot,\cdot)$, $\mathcal{V}$, $\widehat{\mathcal{U}}$, $D_{\phi_{\mathcal{V}}}(V,V^t)$, $D_{\phi_{\widehat{\mathcal{U}}}}(\widehat{U},\widehat{U}^t)$, $\Omega_{\mathcal{V}}$, $\Omega_{\widehat{\mathcal{U}}}$ were defined above. On the other hand, **Mirror Prox** performs iterations

$$\boxed{\begin{aligned} V^{t+1/2} &= \underset{V\in\mathcal{V}}{\arg\min}\left\{\mathcal{F}(V,Y) - \widehat{\Phi}(\widehat{U}^t, V) + \frac{D_{\phi_{\mathcal{V}}}(V,V^t)}{2\gamma_t\Omega_{\mathcal{V}}}\right\}, \\ \widehat{U}^{t+1/2} &= \underset{\widehat{U}\in\widehat{\mathcal{U}}}{\arg\min}\left\{\lambda\,\mathrm{tr}[\mathbb{1}_{2d\times k}^\top\widehat{U}] + \widehat{\Phi}(\widehat{U}, V^t - Y) + \frac{D_{\phi_{\widehat{\mathcal{U}}}}(\widehat{U},\widehat{U}^t)}{2\gamma_t\Omega_{\widehat{\mathcal{U}}}}\right\}; \\ V^{t+1} &= \underset{V\in\mathcal{V}}{\arg\min}\left\{\mathcal{F}(V,Y) - \widehat{\Phi}(\widehat{U}^{t+1/2}, V) + \frac{D_{\phi_{\mathcal{V}}}(V,V^t)}{2\gamma_t\Omega_{\mathcal{V}}}\right\}, \\ \widehat{U}^{t+1} &= \underset{\widehat{U}\in\widehat{\mathcal{U}}}{\arg\min}\left\{\lambda\,\mathrm{tr}[\mathbb{1}_{2d\times k}^\top\widehat{U}] + \widehat{\Phi}(\widehat{U}, V^{t+1/2} - Y) + \frac{D_{\phi_{\widehat{\mathcal{U}}}}(\widehat{U},\widehat{U}^t)}{2\gamma_t\Omega_{\widehat{\mathcal{U}}}}\right\}. \end{aligned}} \tag{\textbf{MP}}$$

Complexity of one iteration. Note that the primal updates in both algorithms can be expressed in closed form: using Lemma 4.1 in Appendix 4.7.3 we can verify that the primal update in (**MD**) amounts to

$$\begin{aligned} &\widehat{U}_{ij}^{t+1} = \widehat{U}_{ij}^t \cdot \exp(-2\gamma_t S_{ij}^t R_{\mathcal{U}} \log(2dk)) \cdot \min\left\{\exp(-2\gamma_t \lambda R_{\mathcal{U}} \log(2dk)), \frac{R_{\mathcal{U}}}{M}\right\}, \\ &\text{where } \; S^t = \frac{1}{n}\widehat{X}^\top(V^t - Y), \; \text{and } M = \sum_{i=1}^{2d}\sum_{j=1}^{k} \widehat{U}_{ij}^t \cdot \exp(-2\gamma_t S_{ij}^t R_{\mathcal{U}} \log(2dk)). \end{aligned} \tag{4.2.23}$$

On the other hand, the primal updates depend on the expression for $\mathfrak{f}(v,y)$ in the definition (4.1.2) of Fenchel-Young losses (cf. also (4.1.6)). Crucially, when $\mathfrak{f}(v,y)$ is *separable* in v – which is the case for the Fenchel-Young losses including the multiclass logistic (4.1.3) and hinge (4.1.4) loss – we can perform the update for V as follows:

- pass to the Lagrange dual formulation of the proximal step for V;
- minimize the Lagrangian for the given value of the multiplier by solving $O(nk)$ one-dimensional problems exactly (when possible) or to numerical tolerance;
- on top of that, find the optimal Lagrange multiplier via one-dimensional search.

See Ostrovskii and Harchaoui [2018, Supp. Mat.] for an illustration of this technique. Note also that in the case of multiclass SVM (cf. (4.1.4)), we have an explicit formula:

$$V_{ij}^{t+1} = \frac{V_{ij}^t \exp\left(2\gamma_t Q_{ij}^t \log(k)\right)}{\sum\limits_{l=1}^k V_{il}^t \exp\left(2\gamma_t Q_{il}^t \log(k)\right)}, \text{ where } Q^t = \widehat{X}\widehat{U}^{t-1} - Y. \tag{4.2.24}$$

Overall, we see that the computational bottleneck in the deterministic schemes is the matrix-matrix multiplication which costs $O(dnk)$ arithmetic operations; the remaining part of the combined proximal step takes only $O(dk + nk)$ operations.

We now present accuracy guarantees that follow from the general analysis of the proximal saddle-point optimization schemes (see Appendix 4.7.2) combined with the obtained expressions for $\mathcal{L}_{\mathscr{U},\mathscr{V}}$, $\Omega_{\widehat{\mathcal{U}}}$, and $\Omega_{\mathcal{V}}$.

4.2.4 Accuracy bounds for the deterministic algorithms

Preliminaries. In what follows, we denote by $O(1)$ generic constants. Recall that the accuracy of a candidate solution $(\bar{U}, \bar{V})$ to a saddle-point problem

$$\min_{U \in \mathcal{U}} \max_{V \in \mathcal{V}} f(U, V), \tag{4.2.25}$$

assuming continuity of f and compactness of $\mathcal{U}$ and $\mathcal{V}$, can be quantified via the *duality gap*

$$\Delta_f(\bar{U}, \bar{V}) := \max_{V \in \mathcal{V}} f(\bar{U}, V) - \min_{U \in \mathcal{U}} f(U, \bar{V}). \tag{4.2.26}$$

In particular, under Slater's conditions (which holds for (4.2.17) in particular), the problem (4.2.25) possesses an optimal solution $W^* = (U^*, V^*)$, called a *saddle point*, for which it holds $f(U^*, V^*) = \max_{V \in \mathcal{V}} f(U^*, V) = \min_{U \in \mathcal{U}} f(U, V^*)$ – that is, U^* (resp. V^*) is optimal in the primal problem of minimizing $f_{\text{prim}}(U) := \max_{V \in \mathcal{V}} f(U, V)$ in U (resp. the dual problem of maximizaing $f_{\text{dual}}(V) := \min_{U \in \mathcal{U}} f(U, \bar{V})$ in V). Hence, in this case $\Delta_f(\bar{U}, \bar{V})$ is the sum of the primal and dual accuracies, and bounds from above the primal accuracy, which is of main interest in the initial problem (4.1.1).

We can derive the following accuracy guarantee for the composite saddle-point Mirror Descent (**MD**) applied to the saddle-point problem (4.2.17).[2] To simplify the exposition, we only consider constant stepsize and simple averaging of the iterates; empirically we observe similar accuracy for the sample sizes descreasing as $\propto 1/\sqrt{t}$.

[2] Surprisingly, we could not find accuracy guarantees for the composite-objective variants of Mirror Descent and Mirror Prox when applied to *saddle-point* problems. Note that the results of Duchi et al. [2010] only hold for Mirror Descent in a composite minimization problem, and those of Nesterov and Nemirovski [2013] use a different (in fact, more robust) formulation of the algorithms.

Theorem 4.1. *Let $(\bar{U}^T, \bar{V}^T) = \frac{1}{T}\sum_{t=0}^{T-1}(\widehat{U}^t, V^t)$ be the average of the first T iterates of Mirror Descent* (**MD**) *with initialization* (**Init**) *and constant stepsize*

$$\gamma_t \equiv \frac{1}{\mathcal{L}_{\mathscr{U},\mathscr{V}}\sqrt{5T\Omega_{\widehat{\mathcal{U}}}\Omega_{\mathcal{V}}}}$$

with the values of $\mathcal{L}_{\mathscr{U},\mathscr{V}}$, $\Omega_{\mathcal{V}}$, $\Omega_{\widehat{\mathcal{U}}}$ given by (4.2.11), (4.2.14), (4.2.21)*. Then it holds*

$$\begin{aligned}\Delta_f(\bar{U}^T, \bar{V}^T) &\left[\leqslant \frac{2\sqrt{5}\mathcal{L}_{\mathscr{U},\mathscr{V}}\sqrt{\Omega_{\mathcal{U}}\Omega_{\mathcal{V}}}}{\sqrt{T}} + \frac{F(V^0, Y) - \min_{V\in\mathcal{V}} F(V,Y)}{T}\right] \\ &\leqslant \frac{2\sqrt{5}\|X^\top\|_{\infty\times 2}}{\sqrt{n}} \cdot \frac{\log(2dk)\|U^*\|_{1\times 1}}{\sqrt{T}} + \frac{\boldsymbol{r}}{T},\end{aligned} \tag{4.2.27}$$

where $\boldsymbol{r} = \max_{y\in\Delta_k}\{\mathfrak{f}(\mathbb{1}/k, y) - \min_{v\in\Delta_k}\mathfrak{f}(v,y)\}$. Moreover, for the multiclass hinge loss (4.1.4) *the $O(1/T)$ term vanishes from the brackets, and we can put $\boldsymbol{r} = 0$.*

Proof. The bracketed bound in (4.2.27) follows from the general accuracy bound for the composite saddle-point Mirror Descent scheme (Theorem 4.3 in Appendix), combined with the observation that the primal "simple" term $\lambda \text{tr}[\mathbb{1}_{2d\times k}^\top \widehat{U}]$ of the objective is linear, whence the corresponding error term is not present. In the case of the hinge loss (4.1.4), the same holds for the primal "simple" term $\mathcal{F}(V,Y)$, hence the final remark in the claim. To obtain the right-hand side of (4.2.27), we use (4.2.11), (4.2.14), (4.2.21), and bound $F(V^0,Y) - \min_{V\in\mathcal{V}} F(V,Y)$ from above. □

This result merits some discussion.

Remark 2. *Denoting $X(:,j) \in \mathbb{R}^n$ the j-th column of X for $1 \le j \le d$, we have*

$$\frac{\|X^\top\|_{\infty\times 2}}{\sqrt{n}} = \max_{j\leqslant d}\sqrt{\frac{\|X(:,j)\|_2^2}{n}}. \tag{4.2.28}$$

Note that $X(:,j)$ represents the empirical distribution of the j-th feature φ_j. Hence, for i.i.d. data, $\|X^\top\|_{\infty,2}/\sqrt{n}$ almost surely converges to the largest L_2-norm of an individual feature:

$$\frac{\|X^\top\|_{\infty\times 2}}{\sqrt{n}} \xrightarrow{a.s.} \max_{j\leqslant d}(\mathbf{E}\varphi_j^2)^{1/2}, \tag{4.2.29}$$

In the non-asymptotic regime, (4.2.28) *is the largest empirical L_2-moment of an individual feature, and can be controlled if the features are bounded, or, more generally, if they are sufficiently light-tailed, via standard concentration inequalities. In particular,*

$$\frac{\|X^\top\|_{\infty\times 2}}{\sqrt{n}} \leqslant B$$

if the features are uniformly bounded by B. Also, using the standard χ^2-bound, *see Laurent and Massart [2000, Lemma 1], with probability at least $1-\delta$ it holds*

$$\frac{\|X^\top\|_{\infty\times 2}}{\sqrt{n}} \leqslant O(1)\sigma\left(1+\sqrt{\frac{\log(d/\delta)}{n}}\right) \tag{4.2.30}$$

if the features are zero-mean and Gaussian with variances uniformly bounded by σ^2.

Remark 3. *The accuracy bound in Theorem 4.1 includes the remainder term $\boldsymbol{r}/T$ due to the presence of $\mathcal{F}(V,Y)$. Note that $\boldsymbol{r}$ does not depend on d; moreover, we can expect that $\boldsymbol{r} = O(\log(k))$ in the case when $\mathfrak{f}(v,y)$ corresponds to some notion of entropy on Δ_k as is the case, in particular, for the multiclass logistic loss (4.1.3). Finally, as stated in the theorem, one can set $\boldsymbol{r}=0$ for the multiclass hinge loss (4.1.4). Thus, overall we see that the additional term is relatively small, and can be neglected.*

Remark 4. *Finally, note that the $O(1/\sqrt{T})$ can be improved to the $O(1/T)$ one if instead of Mirror Descent we use Mirror Prox. While here we do not state the precise accuracy bound for the "direct" version of the algorithm (**MP**), such results are known for the more flexible "epigraph" version, in which the "simple" terms are moved into the constraints (which allows to address a more general class of semi-separable problems), see, e.g., He et al. [2015]. Also, in the case of multiclass hinge loss (4.1.4), both "simple" terms are linear, and can be formally absorbed into $\Psi(U,V)=\Phi(U,V-Y)$ without changing $\mathcal{L}_{\mathscr{U},\mathscr{V}}$. This would result in the bound*

$$\Delta_f(\bar{U}^T,\bar{V}^T) \leqslant \frac{O(1)\mathcal{L}_{\mathscr{U},\mathscr{V}}\sqrt{\Omega_{\mathcal{U}}\Omega_{\mathcal{V}}}}{T} \leqslant \frac{\widetilde{O}_{d,k}(1)\|U^*\|_{1\times 1}}{T}\cdot\frac{\|X^\top\|_{\infty\times 2}}{\sqrt{n}}, \tag{4.2.31}$$

where $\widetilde{O}_{d,k}(1)$ is a logarithmic factor in d and k, see Juditsky and Nemirovski [2011b]. While the $O(1/T)$ rate is not preserved for Mirror Prox with stochastic oracle, this result is still useful if we are willing to use the mini-batching technique.

4.3 Sampling schemes

As prevously noted, the main drawback of the deterministic approach is the high numerical complexity $O(dnk)$ of operations due to the cost of matrix multiplications when computing the partial gradients $\widehat{X}\widehat{U}$ and $\widehat{X}^\top(V-Y)$. Following Juditsky and Nemirovski [2011b], the natural approach to accelerate the computation of these matrix products is by sampling the matrices U and $V-Y$. Denoting $\xi_{\widehat{U}}$ and $\eta_{V,Y}$ unbiased estimates of the matrix products $\widehat{X}\widehat{U}$ and $\widehat{X}^\top(V-Y)$ obtained in this way, we arrive at the following version of **Stochastic Mirror Descent** scheme (cf. (**MD**)):

$$\boxed{\begin{aligned} V^{t+1} &= \mathop{\arg\min}_{V\in\mathcal{V}}\left\{\mathcal{F}(V,Y)-\frac{1}{n}\mathrm{tr}\left[\xi_{\widehat{U}^t}^\top V\right]+\frac{D_{\phi_{\mathcal{V}}}(V,V^t)}{2\gamma_t\Omega_{\mathcal{V}}}\right\},\\ \widehat{U}^{t+1} &= \mathop{\arg\min}_{\widehat{U}\in\widehat{\mathcal{U}}}\left\{\lambda\,\mathrm{tr}[\mathbb{1}_{2d\times k}^\top\widehat{U}]+\frac{1}{n}\mathrm{tr}\left[\eta_{V^t,Y}\widehat{U}\right]+\frac{D_{\phi_{\widehat{\mathcal{U}}}}(\widehat{U},\widehat{U}^t)}{2\gamma_t\Omega_{\widehat{\mathcal{U}}}}\right\},\end{aligned}} \tag{S-MD}$$

and its counterpart in the case of Mirror Prox, which can be formulated analogously.

The immediate question is how to obtain $\xi_{\widehat{U}}$ and $\eta_{V,Y}$. For example, we might sample one row of U and $V-Y$ at a time, obtaining unbiased estimates of the true partial gradients that can be computed in just $O(dk+nk)$ operations – comparable with the remaining volume of computations for proximal mapping itself. In fact, one can try to go even further, sampling a *single element* at a time, and using explicit expressions for the updates available in the case of SVM, cf. (4.2.23)–(4.2.24), allowing to reduce the complexity even further – to $O(d+n+k)$ as we show below. Yet, there is a price to pay: the gradient oracle becomes stochastic, and the accuracy bounds should be augmented with an extra term that reflects stochastic variability of the gradients. In fact, the effect of the gradient noise on the accuracy bounds for both schemes is known: we get the extra term

$$O\left(\frac{\sqrt{\Omega_{\widehat{\mathcal{U}}}\sigma^2_{\mathcal{V}}+\Omega_{\mathcal{V}}\sigma^2_{\widehat{\mathcal{U}}}}}{\sqrt{T}}\right), \tag{4.3.1}$$

where $\sigma^2_{\widehat{\mathcal{U}}}$ and $\sigma^2_{\mathcal{V}}$ are "variance proxies" – expected squared dual norms of the noise in the gradients (see Juditsky and Nemirovski [2011b, Sec. 2.5.1]) and Appendix 4.7.2):

$$\sigma^2_{\widehat{\mathcal{U}}}=\frac{1}{n^2}\sup_{\widehat{U}\in\widehat{\mathcal{U}}}\mathbb{E}\left[\left\|\widehat{X}\widehat{U}-\xi_{\widehat{U}}\right\|^2_{\mathscr{V}^*}\right], \quad \sigma^2_{\mathcal{V}}=\frac{1}{n^2}\sup_{(V,Y)\in\mathcal{V}\times\mathcal{V}}\mathbb{E}\left[\left\|\widehat{X}^\top(V-Y)-\eta_{V,Y}\right\|^2_{\mathscr{U}^*}\right], \tag{4.3.2}$$

In this section, we consider two sampling schemes for the estimates $\xi_{\widehat{U}}$ and $\eta_{V,Y}$: the *partial scheme*, in which the estimates are obtained by sampling (non-uniformly) the rows of U and $V-Y$; and the *full scheme*, in which one also samples the columns of these matrices. In both cases, we derive the probabilities that nearly minimize the variance proxies. As it turns out, for the choice of the norms and potentials done in Section 4.2, and under weak assumptions on the feature distribution, these choices of probabilities force the additional term (4.3.1) to be of the same order of magnitude as the accuracy bound in Theorem 4.1 for the deterministic Mirror Descent (cf. Theorems 4.2–1 below). Due to the reduced cost of iterations, this leads to the drastic reductions in the overall numerical complexity.

4.3.1 Partial sampling

In the *partial sampling scheme*, we draw the rows of $\widehat{U}$ and V with non-uniform probabilities. In other words, we choose a pair of distributions $p=(p_1,\dots,p_{2d})\in\Delta_{2d}$ and $q=(q_1,\dots,q_n)\in\Delta_n$, and sample $\xi_{\widehat{\mathcal{U}}}=\xi_{\widehat{U}}(p)$ and $\eta_{\mathcal{V},Y}=\eta_{V,Y}(q)$ according to

$$\xi_{\widehat{U}}(p)=\widehat{X}\frac{e_ie_i^\top}{p_i}\widehat{U}, \quad \eta_{V,Y}(q)=\widehat{X}^\top\frac{e_je_j^\top}{q_j}(V-Y) \tag{Part-SS}$$

where $e_i\in\mathbb{R}^{2d}$ and $e_j\in\mathbb{R}^n$ are the standard basis vectors, and $i\in[2d]:=\{1,...,2d\}$ and $j\in[n]$ have distributions p and q correspondingly (clearly, this guarantees unbiasedness). Thus, in this scheme we sample the features and the training examples, but not the classes. The main challenge is to correctly choose the distributions p, q. This can be done in a data-dependent manner to approximately minimize the variance proxies $\sigma^2_{\widehat{\mathcal{U}}}, \sigma^2_{\mathcal{V}}$ defined in (4.3.2). Note that minimization of the variance proxies can only be done explicitly in the Euclidean case, when $\| \cdot \|_{\mathscr{U}^*}, \| \cdot \|_{\mathscr{V}^*}$ are the Frobenius norms. Instead, we minimize the expected squared norm of the gradient estimate itself (corresponding to the second moment in the Euclidean case), that is, choose $p^* = p^*(\widehat{X}, \widehat{U})$ and $q^* = q^*(\widehat{X}, V, Y)$ according to

$$\begin{aligned} p^*(\widehat{X}, \widehat{U}) &\in \operatorname*{arg\,min}_{p \in \Delta_{2d}} \mathbb{E}\left[\|\xi_{\widehat{U}}(p)\|^2_{\mathscr{V}^*} \right], \\ q^*(\widehat{X}, V, Y) &\in \operatorname*{arg\,min}_{q \in \Delta_n} \mathbb{E}\left[\|\eta_{V,Y}(q)\|^2_{\mathscr{U}^*} \right]. \end{aligned} \tag{4.3.3}$$

As we show next in Proposition 4.2, these minimization problems can indeed be solved explicitly. On the other hand, we can then easily bound the variance proxies for p^* and q^* as well via the triangle inequality. The approach is justified by the observation that these bounds result in the stochastic term (4.3.1) of the same order as the terms already present in the accuracy bound of Theorem 4.1 (cf. Theorem 4.2).

Remark 5. *For the stochastic Mirror Descent it is sufficient to control the second moment proxies directly, and passing to the variance proxies only deteriorates the constants. Nonetheless, we choose to provide bounds for the variance proxies: such bounds are needed in the analysis of Stochastic Mirror Prox.*

The next result gives the optimal sampling distributions (4.3.3) and upper bounds for their variance proxies.

Proposition 4.2. *Consider the choice of norms $\| \cdot \|_{\mathscr{U}} = \| \cdot \|_{1 \times 1}$, $\| \cdot \|_{\mathscr{V}} = \| \cdot \|_{2 \times 1}$. Then, optimal solutions $p^* = p^*(\widehat{X}, \widehat{U})$ and $q^* = q^*(\widehat{X}, V, Y)$ to (4.3.3) are given by*

$$p^*_i = \frac{\|\widehat{X}(:, i)\|_2 \cdot \|\widehat{U}(i, :)\|_\infty}{\sum_{\imath=1}^{2d} \|\widehat{X}(:, \imath)\|_2 \cdot \|\widehat{U}(\imath, :)\|_\infty}, \quad q^*_j = \frac{\|\widehat{X}(j, :)\|_\infty \cdot \|V(j, :) - Y(j, :)\|_\infty}{\sum_{\jmath=1}^{n} \|\widehat{X}(\jmath, :)\|_\infty \cdot \|V(\jmath, :) - Y(\jmath, :)\|_\infty}. \tag{4.3.4}$$

Moreover, their respective variance proxies satisfy the bounds

$$\sigma^2_{\widehat{\mathcal{U}}}(p^*) \leqslant \frac{4R^2_{\mathcal{U}} \|X^\top\|^2_{\infty \times 2}}{n^2}, \quad \sigma^2_{\mathcal{V}}(q^*) \leqslant \frac{8\|X^\top\|^2_{\infty \times 2}}{n} + \frac{8\|X\|^2_{1 \times \infty}}{n^2}. \tag{4.3.5}$$

Proof. See Appendix 4.7.5 for the proof of a generalized result with mixed norms. □

Combining Proposition 4.2 with the general accuracy bound for Stochastic Mirror Descent (see Theorem 4.4 in Appendix 4.7.2), we arrive at the following result:

Theorem 4.2. *Let $(\bar{U}^T, \bar{V}^T) = \frac{1}{T} \sum_{t=0}^{T-1} (\widehat{U}^t, V^t)$ be the average of the first T iterates of Stochastic Mirror Descent (**S-MD**) with initialization (**Init**), equipped with* *sampling scheme* (**Part-SS**) *with distributions given by* (4.3.4)*, and constant stepsize*

$$\gamma_t \equiv \frac{1}{\sqrt{T}} \min\left\{ \frac{1}{\sqrt{10}\mathcal{L}_{\mathscr{U},\mathscr{V}}\sqrt{\Omega_{\widehat{\mathcal{U}}}\Omega_{\mathcal{V}}}}, \frac{1}{\sqrt{2}\sqrt{\Omega_{\widehat{\mathcal{U}}}\bar{\sigma}_{\mathcal{V}}^2 + \Omega_{\mathcal{V}}\bar{\sigma}_{\widehat{\mathcal{U}}}^2}} \right\}, \tag{4.3.6}$$

with the values of $\mathcal{L}_{\mathscr{U},\mathscr{V}}$*,* $\Omega_{\mathcal{V}}$*,* $\Omega_{\widehat{\mathcal{U}}}$ *given by* (4.2.11)*,* (4.2.14)*,* (4.2.21)*, and the upper bounds* $\bar{\sigma}_{\widehat{\mathcal{U}}}^2, \bar{\sigma}_{\mathcal{V}}^2$ *on the variance proxies given by* (4.3.5)*. Then it holds*

$$\begin{aligned}
&\mathbb{E}[\Delta_f(\bar{U}^T, \bar{V}^T)] \\
&\quad \left[\leqslant \frac{2\sqrt{10}\mathcal{L}_{\mathscr{U},\mathscr{V}}\sqrt{\Omega_{\widehat{\mathcal{U}}}\Omega_{\mathcal{V}}}}{\sqrt{T}} + \frac{2\sqrt{2}\sqrt{\Omega_{\mathcal{U}}\bar{\sigma}_{\mathcal{V}}^2 + \Omega_{\mathcal{V}}\bar{\sigma}_{\mathcal{U}}^2}}{\sqrt{T}} + \frac{F(V^0, Y) - \min_{V\in\mathcal{V}} F(V,Y)}{T}\right] \\
&\quad \leqslant \left(\frac{(4\sqrt{6} + 2\sqrt{10})\|X^\top\|_{\infty\times 2}}{\sqrt{n}} + \frac{8\|X\|_{1\times\infty}}{n} \right) \cdot \frac{\log(2dk)\|U^*\|_{1\times 1}}{\sqrt{T}} + \frac{\boldsymbol{r}}{T},
\end{aligned} \tag{4.3.7}$$

where $\mathbb{E}[\cdot]$ *is the expectation over the randomness of the algorithm, and* $\boldsymbol{r}$ *is the same as in Theorem* 4.1*.*

Discussion. The main message of Theorem 4.2 (cf. Theorem 4.1) is as follows: for the chosen geometry of $\|\cdot\|_{\mathscr{U}}$ and potentials, partial sampling scheme (**Part-SS**) does not result in *any* growth of computational complexity, in terms of the number of iterations to guarantee a given value of the duality gap, as long as $\|X\|_{1\times\infty}/n$ does not dominate $\|X^\top\|_{\infty\times 2}/\sqrt{n}$. This is a reasonable assumption as soon as the data has a *light-tailed* distribution. Indeed, recall that $\|X^\top\|_{\infty\times 2}/\sqrt{n}$ is the largest L_2-norm of an individual feature φ_j (cf. Remark 2); on the other hand, $\|X\|_{1\times\infty}/n$ is the sample average of $\max_{j\in[d]}|\varphi_j|$, and thus has the finite limit $\mathbf{E}\max_{j\leqslant d}|\varphi_j|$ when $n\to\infty$. If φ_j are subgaussian, $\mathbf{E}\max_{j\leqslant d}|\varphi_j| \leqslant \widetilde{O}_d(1)\max_{j\leqslant d}\mathbf{E}|\varphi_j| \leqslant \widetilde{O}_d(1)\max_{j\leqslant d}(\mathbf{E}\varphi_j^2)^{1/2}$, whence (cf. (4.2.29)):

$$\lim_{n\to\infty}\frac{\|X\|_{1\times\infty}}{n} \leqslant \widetilde{O}_d(1)\lim_{n\to\infty}\frac{\|X^\top\|_{\infty\times 2}}{\sqrt{n}}.$$

Similar observations can be made in the finite-sample regime. In particular, both terms admit the same bound in terms of the uniform a.s. bound on the features. Also, if $\varphi_j \sim \mathcal{N}(0,\sigma_j^2)$ with $\sigma_j \le \sigma$ for any $j\in[d]$, then with probability at least $1-\delta$,

$$\frac{\|X\|_{1\times\infty}}{n} \leqslant \sigma\sqrt{\log(dn/\delta)}$$

with a similar bound for $\|X^\top\|_{\infty\times 2}/\sqrt{n}$, cf. (4.2.30).

Computational complexity. Computation of $\xi_{\widehat{U}}(p^*)$ and $\eta_{V,Y}(q^*)$ requires $O(dn + dk + nk)$ operations; note that this includes computing the optimal sampling probabilities (4.3.4). The combined complexity of the proximal steps, given these estimates, is $O(dk + nk)$; in particular, in the case of SVM we have explicit formulae both for the primal and dual updates, and in the general case the dual updates are reduced to the root search on top of an explicit $O(nk)$ computation – cf. (4.2.23)–(4.2.24) and the accompanying discussion.

4.3.2 Full sampling

In the *full sampling scheme*, the sampling of the rows of $\widehat{U}$ and $V - Y$ is augmented with the subsequent column sampling. This can be formalized (see Figure 4-1) as

$$\xi_{\widehat{U}}(p, P) = \widehat{X}\frac{e_i e_i^\top}{p_i}\widehat{U}\frac{e_l e_l^\top}{P_{il}}, \quad \eta_{V,Y}(q, Q) = \widehat{X}^\top \frac{e_j e_j^\top}{q_j}(V - Y)\frac{e_l e_l^\top}{Q_{jl}}, \qquad \textbf{(Full-SS)}$$

where the indices $i \in [2d]$ and $j \in [n]$ are sampled with distributions $p \in \Delta_{2d}$ and $q \in \Delta_n$ as before, and the rows of the stochastic matrices $P \in (\Delta_k^\top)^{\otimes 2d}$ and $Q \in (\Delta_k^\top)^{\otimes n}$ specify the conditional sampling distribution of the class $l \in [k]$, provided that we drew the i-th feature (in the primal) and j-th example (in the dual). Clearly, this gives us unbiased estimates of the matrix products.

Next we derive the optimal sampling distributions and bound their variance proxies (see Appendix 4.7.6 for the proof).

Proposition 4.3. *Consider the choice of norms $\|\cdot\|_{\mathscr{U}} = \|\cdot\|_{1\times 1}$, $\|\cdot\|_{\mathscr{V}} = \|\cdot\|_{2\times 1}$. Optimal solutions $(p*, P^*)$, (q^*, Q^*) to the optimization problems*

$$\min_{p\in\Delta_{2d}, P\in(\Delta_k^\top)^{\otimes 2d}} \mathbb{E}\|\xi_{\widehat{U}}(p, P)\|^2_{\mathscr{V}^*}, \qquad \min_{q\in\Delta_n, Q\in(\Delta_k^\top)^{\otimes n}} \mathbb{E}\|\eta_{V,Y}(q, Q)\|^2_{\mathscr{U}^*}$$

are given by

$$\begin{aligned} p_i^* &= \frac{\|\widehat{X}(:, i)\|_2 \cdot \|\widehat{U}(i, :)\|_1}{\sum_{\imath=1}^{2d} \|\widehat{X}(:, \imath)\|_2 \cdot \|\widehat{U}(\imath, :)\|_1}, & P_{il}^* &= \frac{|\widehat{U}_{il}|}{\|\widehat{U}(i, :)\|_1}, \\ q_j^* &= \frac{\|\widehat{X}(j, :)\|_\infty \cdot \|V(j, :) - Y(j, :)\|_1}{\sum_{\jmath=1}^{n} \|\widehat{X}(\jmath, :)\|_\infty \cdot \|V(\jmath, :) - Y(\jmath, :)\|_1}, & Q_{jl}^* &= \frac{|V_{jl} - Y_{jl}|}{\|V(j, :) - Y(j, :)\|_1}. \end{aligned} \qquad (4.3.8)$$

The variance proxies $\sigma^2_{\widehat{\mathcal{U}}}(p^*, P^*)$, $\sigma^2_{\mathcal{V}}(q^*, Q^*)$ admit the same upper bounds as in (4.3.5).

Combining Proposition 4.3 with the general bound used in Theorem 4.2, we obtain

Corollary 1. *The accuracy bound (4.3.7) of Theorem 4.2 remains true when we replace the sampling scheme* (**Part-SS**) *with the scheme* (**Full-SS**) *with parameters chosen according to (4.3.8).*

Computational complexity. In the next section, we describe efficient implementation for the multiclass hinge loss (4.1.4) which has $O(d + k + n)$ complexity of one

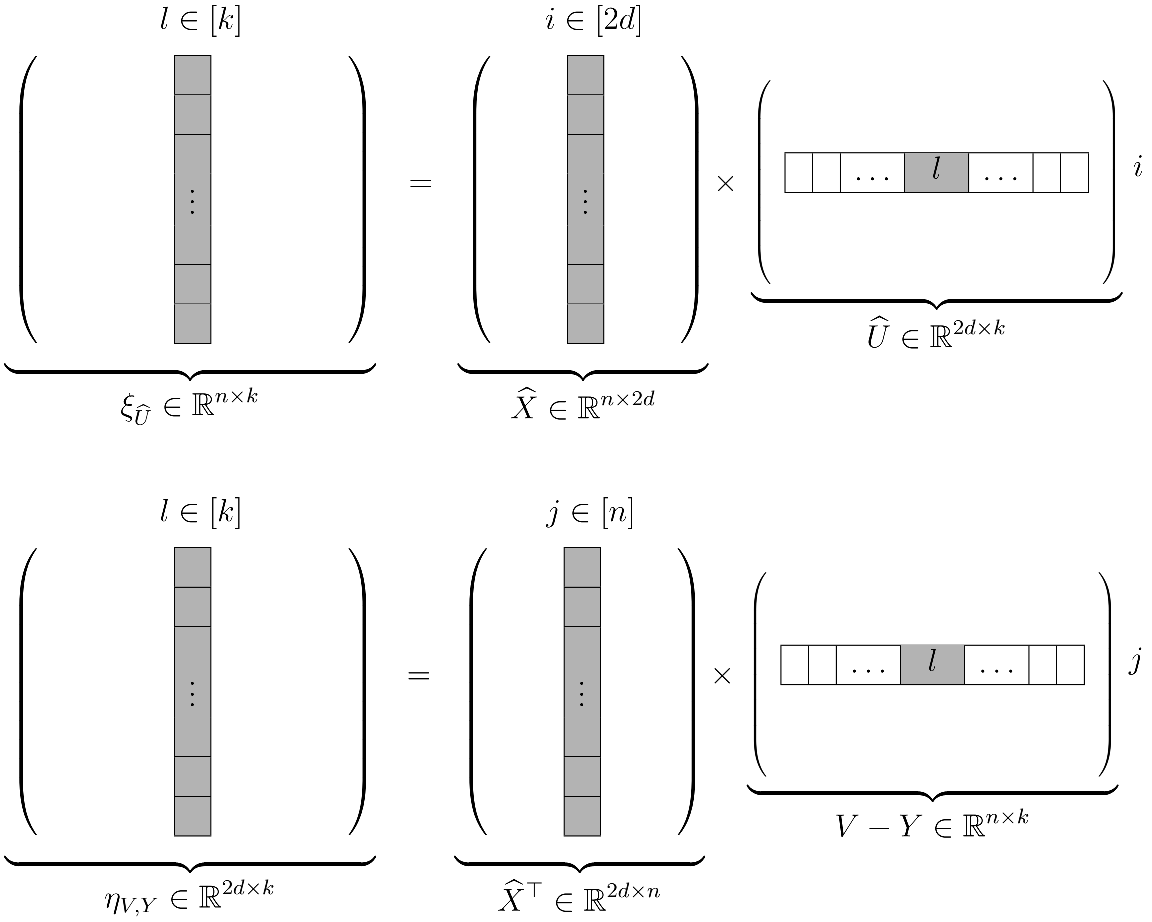
$l \in [k]$
$i \in [2d]$
$=$
$\times$
l
i
$\widehat{U} \in \mathbb{R}^{2d \times k}$
$\xi_{\widehat{U}} \in \mathbb{R}^{n \times k}$
$\widehat{X} \in \mathbb{R}^{n \times 2d}$
$l \in [k]$
$j \in [n]$
$=$
$\times$
l
j
$V - Y \in \mathbb{R}^{n \times k}$
$\eta_{V,Y} \in \mathbb{R}^{2d \times k}$
$\widehat{X}^\top \in \mathbb{R}^{2d \times n}$

Figure 4-1 – Depiction of the Full Sampling Scheme (**Full-SS**).

iteration, plus pre- and postprocessing of

$$O(dn + (d+n) \cdot \min(k,T))$$

arithmetic operations (a.o.'s). While the detailed analysis is deferred to the next section, let us briefly present the main ideas on which it relies.

- For the optimal sampling distributions p^* and q^*, we need to compute the norms $\sigma_i = \|\widehat{X}(:,i)\|_2$ and $\tau_j = \|\widehat{X}(j,:)\|_\infty$ for all $i \in [2d]$ and $j \in [n]$. This requires $O(dn)$ iterations but only once, during preprocessing.
- We also need to calculate $O(d)$ quantities $\pi_i = \|\widehat{U}(i,:)\|_1$ and $O(n)$ quantities $\rho_j = \|V(j,:) - Y(j,:)\|_1$. While the direct calculation of each of them takes $O(k)$ a.o.'s, we can update them dynamically in $O(1)$ in the case of the hinge loss. Examining the explicit formulae for the updates in the case of hinge loss (cf. (4.2.23)–(4.2.24)), we notice that in the case of full sampling, but one elements of each row of $\widehat{U}$ and V are simply scaled by the same coefficient, and thus one can maintain each of the quantities $\|\widehat{U}(i,:)\|_1$ and $\|V(j,:) - Y(j,:)\|_1$ in $O(1)$ a.o.'s, leading to $O(d+n)$ complexity of this step. This result is fragile, requiring the specific combination of hinge loss (for which $\mathfrak{f}(v,y)$ is linear) and entropic potentials.
- We do not need to compute the full matrices P, Q. Instead, we can compute only one row of each of them, corresponding to the drawn pair i, j, which requires $O(k)$ a.o.'s. Hence we can implement sampling in $O(d+n+k)$ a.o.'s
- Note that updates of the matrices U and V are *not* dense even in the case of the full sampling scheme, so implementing them naively would result in $O(dk+nk)$ a.o.'s per iteration. Instead, we employ "lazy updates": for each $i \in [2d]$ and $j \in [n]$, instead of $\widehat{U}(i,:)$ and $V(j,:)$ we maintain the pairs $(\widetilde{U}(i,:), \alpha_i)$, $(V(j,:), \beta_j)$ such that
$$\widehat{U}(i,:) = \alpha_i \widetilde{U}(i,:), \quad V(j,:) = \beta_j \widetilde{V}(j,:).$$
Recalling that at each iteration all but one elements of $\widehat{U}(i,:)$ are scaled by the same coefficient, we can update any such pair in $O(1)$, with the total complexity of $O(d+n)$ a.o.'s.
- It can be shown that given the pairs $(\widetilde{U}(i,:), \alpha_i)$ and $(V(j,:), \beta_j)$, the computation of the averages $\bar{U}^T, \bar{V}^T$ after T iterations requires $O(d+n)$ a.o.'s per iteration plus the post-processing of $O((d+n) \cdot \min(k,T))$, resulting in the outlined complexity estimate. Note that the computation of the duality gap (which can be used as the online stopping criterion, see, e.g., Ostrovskii and Harchaoui [2018]) also has the same complexity. Hence, in practice it is reasonable to employ the "doubling" technique, computing the averages and the duality gap at iterations $T_m = 2^m$ with increasing values of m. This will result in the same overall complexity while allowing for an online stopping criterion.

Remark 6. *The way to avoid averaging of iterations is to consider averaging of stochastic gradients as done by Juditsky and Nemirovski [2011b]. Stochastic gradients $\xi_{\widehat{U}}$ and η_V have only $O(n)$ and $O(d)$ non-zero elements respectively, and their running averages can be computed in $O(d+n)$ a.o.'s. However, in this case we lose* *the duality gap guarantee (and thus an online stopping criterion), and only have a guarantee on the primal accuracy.*[3]

Remark 7. *Unfortunately, we were not able to achieve the $O(d+n+k)$ complexity for the multiclass logistic loss (4.1.3). The reason is that in this case the composite term $\mathcal{F}(V,Y)$ is the sum of negative entropies (same as the potential (4.2.13)), and the corresponding proximal updates are not reduced to rescalings of rows (modulo $O(1)$ elements). However, due to the sparsity of stochastic gradients, the updates have a special form, and we believe that $O(d+n+k)$ complexity is possible to achieve in this case as well. We are planning to investigate it in the future.*

4.3.3 Efficient implementation of SVM with Full Sampling

In this section we consider updates for fully stochastic case and show that one iteration complexity is indeed $O(d+n+k)$. We will show, that it is possible to the special form of "lazy" updates for scaling coefficients. Recall, that one iteration of the fully stochastic mirror descent is written as **S-MD**:

$$\boxed{\begin{aligned} V^{t+1} &= \underset{V\in\mathcal{V}}{\arg\min}\left\{\mathcal{F}(V,Y) - \frac{1}{n}\mathrm{tr}\left[\xi_{\widehat{U}^t}^\top V\right] + \frac{D_{\phi_{\mathcal{V}}}(V,V^t)}{2\gamma_t\Omega_{\mathcal{V}}}\right\}, \\ \widehat{U}^{t+1} &= \underset{\widehat{U}\in\widehat{\mathcal{U}}}{\arg\min}\left\{\lambda\,\mathrm{tr}[\mathbb{1}_{2d\times k}^\top\widehat{U}] + \frac{1}{n}\mathrm{tr}\left[\eta_{V^t,Y}\widehat{U}\right] + \frac{D_{\phi_{\widehat{\mathcal{U}}}}(\widehat{U},\widehat{U}^t)}{2\gamma_t\Omega_{\widehat{\mathcal{U}}}}\right\}, \end{aligned}} \qquad \textbf{(S-MD)}$$

where $\xi_{\widehat{U}^t}$ and $\eta_{V^t,Y}$ are stochastic gradients.

Lazy Updates.

Note that although the estimates $\xi_{\widehat{U}}, \eta_{V,Y}$ produced in (**Full-SS**) are sparse (each contains a single non-zero column), the updates in (**S-MD**), which can be expressed as (4.2.23)–(4.2.24) with $\xi_{\widehat{U}}, \eta_{V,Y}$ instead of the corresponding matrix products, are *dense*, and implementing them naively costs $O(dk+nk)$ a.o.'s. Fortunately, these updates have a special form: all elements in each row of $\widehat{U}^t$ and V^t are simply rescaled with the same factor – except for at most two elements corresponding to a single non-zero element of $\eta_{V^t,Y}$ and at most two non-zero elements of $\xi_{\widehat{U}^t} - Y$ in this row. To exploit this fact, we perform "lazy" updates: instead of explicitly computing the actual iterates $(\widehat{U}^t, V^t)$, we maintain the quadruple $(\widetilde{U}, \alpha, \widetilde{V}, \beta)$, where $\widetilde{U}, \widetilde{V}$ have the same dimensions as U, V, while $\alpha \in \mathbb{R}^{2d}$ and $\beta \in \mathbb{R}^n$ are the "scaling vectors", so that at any iteration t it holds

$$\widehat{U}^t(i,:) = \widetilde{U}(i,:)\cdot\alpha(i), \quad V^t(j,:) = \widetilde{V}(j,:)\cdot\beta(j) \tag{4.3.9}$$

for any row of $\widehat{U}^t$ and V^t. Initializing with $(\widetilde{U},\widetilde{V}) = (\widehat{U},V)$, $\alpha = \mathbb{1}_{2d}$, $\beta = \mathbb{1}_n$, we can update the whole quadruple, while maintaining (4.3.9), by updating at most two

[3] This situation corresponds to the "general case" in the terminology [Juditsky and Nemirovski, 2011b, Sec. 2.5.2.1 and Prop. 2.6(ii)].

elements in each row of $\widetilde{U}$ and $\widetilde{V}$, and encapsulating the overall scaling of rows in α and β. Clearly, this update requires only $O(d+n)$ operations once $\xi_{\widehat{U}^t}, \eta_{V^t,Y}$ have been drawn.

Sampling.

Computing the distributions p^*, q^* from (4.3.8) requires the knowledge of $\|\widehat{X}(:,i)\|_2$ and $\|\widehat{X}(j,:)\|_\infty$ which can be precomputed in $O(dn)$ a.o.'s, and maintaining $O(d+n)$ norms π_i, ρ_j of the rows of $\widehat{U}^t$ and $V^t - Y$ that can maintained in $O(1)$ a.o.'s each using (4.3.9). Thus, p^* and q^* can be updated in $O(d+n)$. Once it is done, we can sample $i^t \sim p^*$ and $j^t \sim q^*$, and then sample the class from P^* and Q^*, cf. (4.3.8), by computing only the i^t-th row of P^* and the j^t-th row of Q^*, both in $O(k)$ a.o.'s. Thus, the total complexity of producing $\xi_{U^t}, \eta_{V^t,Y}$ is $O(d+n+k)$.

Tracking the Averages.

Similar "lazy" updates can be performed for the running averages of the iterates. Omitting the details, this requires $O(d+n)$ a.o.'s per iteration, plus post-processing of $O(dk+nk)$ a.o.'s.

The above ideas are implemented in Algorithm 1 whose correctness is formally shown in Appendix, Sec. 4.7.7 (see also Sec. 4.7.8 for an additional discussion). Its close inspection shows the iteration cost of $O(d+n+k)$ a.o.'s, plus $O(dn+dk+nk)$ a.o.'s for pre/post-processing, and the memory complexiy of $O(dn+dk+nk)$. Moreover, the term $O(dk)$, which dominates in high-dimensional and highly multiclass problems, can be removed if one exploits sparsity of the corresponding primal solution to the ℓ_1-constrained problem (4.2.2), and outputs it directly, bypassing the explicit storage of $\widetilde{U}$ (see Appendix Sec. 4.7.7 for details). Note that when $n = O(\min(d,k))$, the resulting algorithm enters the sublinear regime after as few as $O(n)$ iterations.

4.4 Discussion of Alternative Geometries

Here we consider alternative choices of the proximal geometry in mirror descent applied to the saddle-point formulation of the CCSPP (4.1.1), possibly with other choices of regularization than the entrywise ℓ_1-norm. The goal is to show that the geometry chosen in Sec. 4.2 is the only one for which we can obtain favorable accuracy guarantees for stochastic mirror descent (**S-MD**).

Given the structure of the primal and dual feasible sets, it is reasonable to consider general mixed norms of the type (4.2.4):

$$\|\cdot\|_{\mathscr{U}} = \|\cdot\|_{p_U^1 \times p_U^2}, \quad \|\cdot\|_{\mathscr{V}} = \|\cdot\|_{p_V^1 \times p_V^2},$$

where $p_U^{1,2}, p_V^{1,2} \geq 1$ (in the case of $\|\cdot\|_{\mathscr{U}}$, we also assume the same norm for regularization). Note that their dual norms can be easily computed: the dual norm of $\|\cdot\|_{p^1 \times p^2}$ is $\|\cdot\|_{q^1 \times q^2}$, where $q^{1,2}$ are the corresponding conjugates to $p^{1,2}$, i.e., $1/p^i + 1/q^i = 1$ (see, e.g., Lemma 3 in Sra [2012]). Moreover, it makes sense to fix $p_V^1 = 2$ for

Algorithm 1 Sublinear Multiclass ℓ_1-Regularized SVM

```
Require: $X \in \mathbb{R}^{n\times d}$, $y \in [k]^{\otimes n}$, $\lambda$, $R_1$, $T \geqslant 1$, $\{\gamma_t\}_{t=0}^{T-1}$
Obtain $Y \in \Delta_k^{\otimes n}$ from the labels $y$;   $\widehat{X} \equiv [X, -X]$
$\alpha \leftarrow \mathbb{1}_{2d}$;   $\widetilde{U} \leftarrow \frac{R_* \mathbb{1}_{2d\times k}}{2dk}$;   $\beta \leftarrow \mathbb{1}_n$;   $\widetilde{V} \leftarrow \frac{\mathbb{1}_{n\times k}}{k}$
for $\imath = 1$ to $2d$ do
    $\sigma(\imath) \equiv \|\widehat{X}(:,\imath)\|_2$;   $\pi(\imath) \leftarrow \|\widetilde{U}(\imath,:)\|_1$
for $\jmath = 1$ to $n$ do
    $\tau(\jmath) \equiv \|\widehat{X}(\jmath,:)\|_\infty$;   $\rho(\jmath) \leftarrow \|\widetilde{V}(\jmath,:) - Y(\jmath,:)\|_1$
# Initialize machinery to track the cumulative sums
$U_\Sigma \leftarrow 0_{2d\times k}$;   $V_\Sigma \leftarrow 0_{n\times k}$                                # Cumulative sums
$A \leftarrow 0_{2d}$;   $B \leftarrow 0_n$;   $A_{\mathrm{pr}} \leftarrow 0_{2d\times k}$;   $B_{\mathrm{pr}} \leftarrow 0_{n\times k}$
for $t = 0$ to $T-1$ do                                                # (S-MD) iterations
    Draw $\boldsymbol{j} \sim \tau \circ \rho$                                # $\circ$ is the elementwise product
    Draw $\boldsymbol{l} \sim |\widetilde{V}(\boldsymbol{j},:) \cdot \beta_{\boldsymbol{j}} - Y(\boldsymbol{j},:)|$
    $[U_\Sigma, A_{\mathrm{pr}}, A] \leftarrow \textsc{TrackPrimal}(\widetilde{U}, U_\Sigma, A_{\mathrm{pr}}, A, \alpha, \boldsymbol{l})$
# The only non-zero column of $\eta_{V^t,Y}$, cf. (Full-SS):
    $\eta \leftarrow \widehat{X}(\boldsymbol{j},:) \cdot \frac{\sum_{\jmath=1}^{n} \tau(\jmath) \cdot \rho(\jmath) \cdot \mathrm{sgn}[\beta_\jmath \cdot V(\jmath,\boldsymbol{l}) - Y(\jmath,\boldsymbol{l})]}{\tau(\boldsymbol{j})}$
    $[\widetilde{U}, \alpha, \pi] \leftarrow \textsc{UpdatePrimal}(\widetilde{U}, \alpha, \pi, \eta, \boldsymbol{l}, \gamma_t, \lambda, R_*)$
    Draw $\boldsymbol{i} \sim \sigma \circ \pi$
    Draw $\boldsymbol{\ell} \sim \widetilde{U}(\boldsymbol{i},:)$
    $[V_\Sigma, B_{\mathrm{pr}}, B] \leftarrow \textsc{TrackDual}(\widetilde{V}, V_\Sigma, B_{\mathrm{pr}}, B, \beta, \boldsymbol{\ell}, y)$
# The only non-zero column of $\xi_{U^t}$, cf. (Full-SS):
    $\xi \leftarrow \widehat{X}(:,\boldsymbol{i}) \cdot \frac{\sum_{\imath=1}^{2d} \sigma(\imath) \cdot \pi(\imath)}{\sigma(\boldsymbol{i})}$
    $[\widetilde{V}, \beta, \rho] \leftarrow \textsc{UpdateDual}(\widetilde{V}, Y, \beta, \rho, \xi, \boldsymbol{\ell}, y, \gamma_t)$
for $l = 1$ to $k$ do                                      # Postprocessing of cumulative sums
    $U_\Sigma(:,l) \leftarrow U_\Sigma(:,l) + \widetilde{U}(:,l) \circ (\alpha + A - A_{\mathrm{pr}}(:,l))$
    $V_\Sigma(:,l) \leftarrow V_\Sigma(:,l) + \widetilde{V}(:,l) \circ (\beta + B - B_{\mathrm{pr}}(:,l))$
Ensure: $\frac{1}{T+1} U_\Sigma, \frac{1}{T+1} V_\Sigma$                                # Averages $(\bar{U}^{T+1}, \bar{V}^{T+1})$
```

Procedure 1 UpdatePrimal

```
Require: $\widetilde{U} \in \mathbb{R}^{2d \times k}$, $\alpha, \pi, \eta \in \mathbb{R}^{2d}$, $l \in [k]$, $\gamma$, $\lambda$, $R_1$
$L \equiv \log(2dk)$
for $i = 1$ to $2d$ do
    $\mu_i = \pi_i - \alpha_i \cdot \widetilde{U}(i,l) \cdot (1 - e^{-2\gamma L R_* \eta_i/n})$
$M = \sum_{i=1}^{2d} \mu_i$
$\nu = \min\{e^{-2\gamma L R_* \lambda}, R_*/M\}$
for $i = 1$ to $2d$ do
    $\widetilde{U}(i,l) \leftarrow \widetilde{U}(i,l) \cdot e^{-2\gamma L R_* \eta_i/n}$
    $\alpha_i^+ = \nu \cdot \alpha_i$
    $\pi_i^+ = \nu \cdot \mu_i$
Ensure: $\widetilde{U}, \alpha^+, \pi^+$
```

Procedure 2 UpdateDual

```
Require: $\widetilde{V}, Y \in \mathbb{R}^{n \times k}$, $\beta, \rho, \xi \in \mathbb{R}^n$, $\ell \in [k]$, $y \in [k]^{\otimes n}$, $\gamma$
$\theta = e^{-2\gamma \log(k)}$
for $j = 1$ to $n$ do
    $\omega_j = e^{2\gamma \log(k) \xi_j}$
    $\varepsilon_j = e^{-2\gamma \log(k) Y(j,\ell)}$
    $\chi_j = 1 - \beta_j \cdot \widetilde{V}(j,\ell) \cdot (1 - \omega_j \cdot \varepsilon_j)$
    if $\ell \neq y_j$ then                                # not the actual class of $j$ drawn
        $\chi_j \leftarrow \chi_j - \beta_j \cdot \widetilde{V}(j, y_j) \cdot (1 - \theta)$
    $\beta_j^+ = \beta_j / \chi_j$
    $\widetilde{V}(j,\ell) \leftarrow \widetilde{V}(j,\ell) \cdot \omega_j \cdot \varepsilon_j$
    $\widetilde{V}(j,y_j) \leftarrow \widetilde{V}(j,y_j) \cdot \omega_j \cdot \theta$
    $\rho_j^+ = 2 - 2\beta_j^+ \cdot \widetilde{V}(j, y_j)$
Ensure: $\widetilde{V}, \beta^+, \rho^+$
```

Procedure 3 TrackPrimal

```
Require: $\widetilde{U}, U_\Sigma, A_{\mathrm{pr}} \in \mathbb{R}^{2d \times k}$, $A, \alpha \in \mathbb{R}^{2d}$, $l \in [k]$
for $i = 1$ to $2d$ do
    $U_\Sigma(i,l) \leftarrow U_\Sigma(i,l) + \widetilde{U}(i,l) \cdot (A_i + \alpha_i - A_{\mathrm{pr}}(i,l))$
    $A_{\mathrm{pr}}(i,l) \leftarrow A_i + \alpha_i$
    $A_i \leftarrow A_i + \alpha_i$
Ensure: $U_\Sigma$, $A_{\mathrm{pr}}$, $A$
```

Procedure 4 TrackDual

```
Require: $\widetilde{V}, V_\Sigma, B_{\mathrm{pr}} \in \mathbb{R}^{n\times k}$, $B, \beta \in \mathbb{R}^n$, $\ell \in [k]$, $y \in [k]^{\otimes n}$
for $j = 1$ to $n$ do
    for $l \in \{\ell, y_j\}$ do                    # $\{\ell, y_j\}$ has 1 or 2 elements
        $V_\Sigma(j,l) \leftarrow V_\Sigma(j,l) + \widetilde{V}(j,l) \cdot (B_j + \beta_j - B_{\mathrm{pr}}(j,l))$
        $B_{\mathrm{pr}}(j,l) \leftarrow B_j + \beta_j$
    $B_j \leftarrow B_j + \beta_j$
Ensure: $V_\Sigma$, $B_{\mathrm{pr}}$, $B$
```

the reasons discussed in Section 4.2. This leaves us with the obvious choices $p_V^2 \in \{1,2\}$, $p_U^2 \in \{1,2\}$ which corresponds to the sparsity-inducing or the standard Euclidean geometry of the *classes* in the dual/primal; $p_U^1 \in \{1,2\}$ which corresponds to the sparsity-inducing or Euclidean geometry of the *features*. Finally, the choice $p_U^1 = 2$ (i.e., the Euclidean geometry in the features) can also be excluded: its combination with $p_V^1 = 2$ is known to lead to the large variance term in the *biclass* case.[4] This leaves us with the possibilities

$$p_U^2, p_V^2 \in \{1,2\} \times \{1,2\}. \tag{4.4.1}$$

In all these cases, the quantity $\mathcal{L}_{\mathcal{U},\mathcal{V}}$ defined in (4.2.10) can be controlled by extending Proposition 4.1:

Proposition 4.4. *For any $\alpha \geqslant 1$ and $\beta \geqslant 1$ such that $\beta \geqslant \alpha$ it holds:*

$$\mathcal{L}_{\mathcal{U},\mathcal{V}} := \frac{\|X\|_{1\times\alpha,\,2\times\beta}}{n} = \frac{\|X^\top\|_{\infty\times 2}}{n}.$$

The proof of this proposition follows the steps in the proof of Proposition 4.1, and is omitted.

Finally, the corresponding partial potentials could be constructed by combining the Euclidean and an entropy-type potential in a way similar to the one described in Sec. 4.2 for the dual variable; alternatively, one could use the power potential of Nesterov and Nemirovski [2013] that results in the same rates up to a constant factor.

Using Proposition 4.4, we can also compute the potential differences for the four remaining setups (4.4.1). The results are shown in Table 4.1. Up to logarithmic factors, we have equivalent results for all four geometries, with the radius R_* evaluated in the corresponding norm $\|\cdot\|_{1\times 2}$ or $\|\cdot\|_{1\times 1} = \|\cdot\|_1$.

As a result, for the deterministic Mirror Descent (with balanced potentials) we

[4] Note that in the biclass case, our variance estimate for the partial sampling scheme (cf. Theorem 4.2) reduces to those in [Juditsky and Nemirovski, 2011b, Section 2.5.2.3]. They consider the cases of ℓ_1/ℓ_1 and ℓ_1/ℓ_2 geometries for the primal/dual, and omit the case of ℓ_2/ℓ_2-geometry, in which the sampling variance "explodes".

		Norm for $V \in \mathbb{R}^{n\times k}$	
		2×1	2×2
Norm for $U \in \mathbb{R}^{d\times k}$	1×2	$\Omega_{\mathcal{U}} = \|U^*\|_{1\times 2}^2 \log d$ $\Omega_{\mathcal{V}} = n\log k$	$\Omega_{\mathcal{U}} = \|U^*\|_{1\times 2}^2 \log d$ $\Omega_{\mathcal{V}} = n$
	1×1	$\Omega_{\mathcal{U}} = \|U^*\|_1^2 \log(dk)$ $\Omega_{\mathcal{V}} = n\log k$	$\Omega_{\mathcal{U}} = \|U^*\|_1^2 \log(dk)$ $\Omega_{\mathcal{V}} = n$

Table 4.1 – Comparison of the potential differences for the norms corresponding to (4.4.1).

obtain the accuracy bound (cf. (4.2.27)):

$$\Delta_f(\bar{U}^T, \bar{V}^T) \leqslant \frac{O(1)\mathcal{L}_{\mathcal{U},\mathcal{V}}\sqrt{\Omega_{\mathcal{U}}\Omega_{\mathcal{V}}}}{\sqrt{T}} \leqslant \frac{\widetilde{O}_{d,k}(1)R_*}{\sqrt{T}} \frac{\|X^\top\|_{\infty\times 2}}{\sqrt{n}} + \frac{\mathtt{r}}{T}.$$

in all four cases, where $\widetilde{O}_{d,k}(1)$ is a logarithmic factor in d and k, and $R_* = \|U^*\|_{1\times 2}$ or $R_* = \|U^*\|_1$ depending on $p_U^2 \in \{1, 2\}$. In other words, the deterministic accuracy bound of Theorem 4.1 is essentially preserved for all four geometries in (4.4.1). On the other hand, using Proposition 4.5, we obtain that in the case of (**Part-SS**), the extra part of the accuracy bound due to sampling (cf. (4.3.7)) is also essentially preserved:

$$\mathbb{E}[\Delta_f(\bar{U}^T, \bar{V}^T)] \leqslant \frac{\widetilde{O}_{d,k}(1)R_*}{\sqrt{T}} \left(\frac{\|X^\top\|_{\infty\times 2}}{\sqrt{n}} + \frac{\|X\|_{1\times\infty}}{n} \right) + \frac{\mathtt{r}}{T}.$$

However, if we consider *full sampling*, the situation changes: in the case $p_U^2 = 2$ the variance bound that holds for (**Part-SS**) is not preserved for (**Full-SS**). This is because our argument to control the variance of the full sampling scheme always requires that $p_U^2 \leqslant 1$ (see the proof of Proposition 4.3 in Appendix 4.7.6 for details; note that for q_V^2 we do not have such a restriction since the variance proxy $\sigma_{\mathcal{V}}^2$ is controlled on the *set* $\mathcal{V}$ given by (4.1.7) that has $\ell_\infty \times \ell_1$-type geometry regardless of the norm $\|\cdot\|_{\mathscr{V}}$. This leaves us with the final choice between the $\|\cdot\|_{2\times 1}$ and $\|\cdot\|_{2\times 2}$ norm in the dual, as we have to use the elementwise $\|\cdot\|_1$-norm in the primal. Both choices result in essentially the same accuracy bound (note that this choice only influences the algorithm but not the saddle-point problem itself). We have focused on the $\|\cdot\|_{2\times 1}$ norm because of the algorithmic considerations: with this norm, we have multiplicative updates in the case of the multiclass hinge loss, which allows for a sublinear algorithm presented in Section 4.3.3.

4.5 Experiments

The natural way to estimate the performance measure for saddle-point problems is the so-called *duality gap*:

$$\Delta_f(\widetilde{U}, \widetilde{V}) = \max_{V \in \mathcal{V}} f(\widetilde{U}, V) - \min_{U \in \mathcal{U}} f(U, \widetilde{V}).$$

In the case of the multi-class SVM formulation:

$$\min_{\substack{\|\widehat{U}\|_1 \leqslant \mathcal{R}_1 \\ \widehat{U} \in \mathbb{R}_+^{2d \times k}}} \max_{V \in (\Delta_k^\top)^{\otimes n}} -\frac{1}{n}\mathrm{tr}[V^\top Y] + \frac{1}{n}\mathrm{tr}\big[(V-Y)^\top \widehat{X}\widehat{U}\big] + \lambda \|\widehat{U}\|_1,$$

the solution can be found in closed form:

$$\Delta_{dual}(\widetilde{U}, \widetilde{V}) = -\frac{1}{n}\mathrm{tr}[Y^\top \widehat{X}\widetilde{U}] + \lambda\|\widetilde{U}\|_1 + \frac{1}{n}\sum_{i=1}^n \max_j \Big[(\widehat{X}\widetilde{U} - Y)(i,j)\Big] +$$

$$+\frac{1}{n}\mathrm{tr}[\widetilde{V}^\top Y] + \left[\max_{i,j}\Big[\Big(-\frac{1}{n}\widehat{X}^\top(\widetilde{V} - Y) - \lambda I\Big)(i,j)\Big] \cdot \mathcal{R}_1\right]_+$$

Sublinear Runtime.

To illustrate the sublinear iteration cost of Algorithm 1, we consider the following experiment. Fixing $n = d = k$, we generate X with i.i.d. standard Gaussian entries, take U^o to be the identity matrix (thus very sparse), and generate the labels by $\arg\max_{l \in [k]} x_j U_l^o + \frac{1}{\sqrt{d}}\mathcal{N}(0, I_d)$, where x_j's are the rows of X, and U_l^o's are the columns of U^o. This is repeated 10 times with $n = d = k$ increasing by a constant factor κ; each time we run Algorithm 1 for a fixed (large) number of iterations to dominate the cost of pre/post-processing, with $R_* = \|U^o\|_1$ and $\lambda = 10^{-3}$, and measure its runtime. We observe (see Tab. 4.2) that the runtime is proportional to κ, as expected.

$n = d = k$	200	400	800	1600	3200	6400
$T = 10^4$	0.80	1.17	2.07	4.27	7.55	15.56
$T = 2 \cdot 10^4$	1.63	2.47	4.27	8.74	14.65	30.77

Table 4.2 – Runtime (in seconds) of Algorithm 1 on synthetic data.

Synthetic Data Experiment.

We compare Algorithm 1 with two competitors: $\|\cdot\|_1$-composite stochastic subgradient method (SSM) for the primal problem (4.1.1), in which one uniformly samples one training example at a time Shalev-Shwartz et al. [2011], leading to $O(dk)$ iteration cost; deterministic saddle-point Mirror Prox (MP) with geometry chosen as in Algorithm 1, for which we have $O(dnk)$ cost of iterations but $O(1/T)$ convergence in terms of the number of iterations. We generate data as in the previous experiment, fixing $n = d = k = 10^3$. The randomized algorithms are run 10 times for $T \in \{10^{m/2}, m = 1, ..., 12\}$ iterations with constant stepsize (we use stepsize (4.3.6) in Algorithm 1, choose the one recommended in Theorem 4.1 for MP, and use the theoretical stepsize for SSM, explicitly computing the variance of subgradients and the Lipschitz constant). Each time we compute the duality gap and the primal accuracy, and measure the runtime (see Fig. 4-2). We see that Algorithm 1 outperforms SSM, which might be the combined effect of sublinearity and our choice of geometry. It also outmatches MP up to high accuracy due to the sublinear effect (MP eventually "wins" because of its $O(1/T)$ rate).[5]

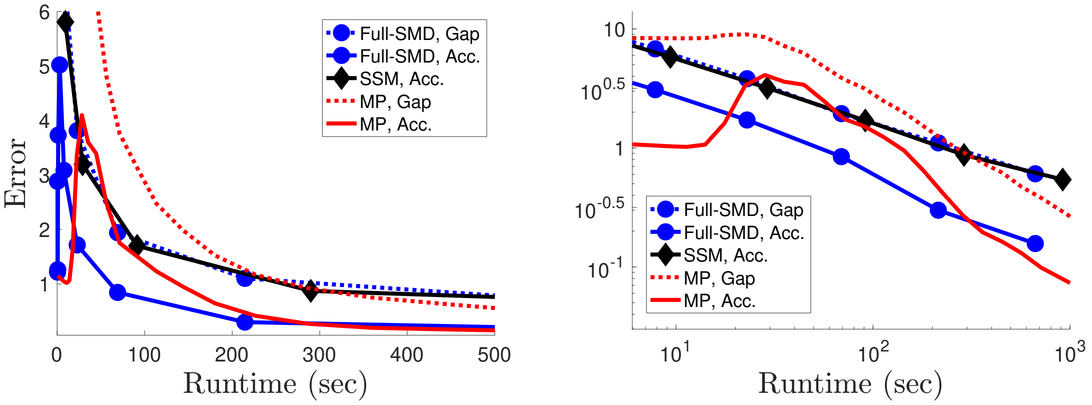

Figure 4-2 – Primal accuracy and duality gap (when available) for Algorithm 1, stochastic subgradient method (SSM), and Mirror Prox (MP) with exact gradients, on a synthetic data benchmark.

Practical issue: "explosion" of rescaling coefficients.

When running the fully stochastic algorithm, we observed one practical issue: the scaling coefficients α and β tend to rapidly increase or decrease, causing occasional arithmetic over/underflows. We suggest two solutions for this problem:

- The immediate solution is to use arbitrary-precision arithmetic. In theory, this results in the $O(d+n+k)$ cost of an iteration being inflated by the allowed limit of digits. The resulting complexity can still be beneficial, vis-à-vis the partial sampling scheme, when k is large enough. This solution is not practical, since one has to use external libraries or implement arbitrary-precision arithmetic.
- A better solution, actually used in our codes, is to perform "urgent" rescales in the case of an over/underflow. Hopefully, it will be required not too often, and the more seldom the closer we are to the optimal solution. For example, in the above experiment the "urgent" rescaling was required in about 10 iterations.

[5] We provide the Matlab codes of our experiments in Supp. Mat.

4.6 Conclusion and perspectives

In this chapter we considered sublinear algorithms for the saddle-points problems constructed for certain type Fenchel-Young losses. The main contribution of this work is the algorighm for the ℓ_1-regularized multiclass SVM with numerical cost $O(d+n+k)$ of one iteration, where d is the number of features, n the sample size, and k the number of classes. This was possible due to the right choice of the proximal setup, and ad-hoc sampling techniques for matrix multiplication.

We envision the following directions for future research:

- Extension to the multi-class logistic regression (softmax) model, which is widely used in Natural Language Processing (NLP) problems, where d, n and k are on order of millions or even billions [Chelba et al., 2013, Partalas et al., 2015].
- Provide more experiments with larger dimensions to highlight the advantages of the algorithm; use real data as well.
- Implement more flexible stepsizes, including the online stepsize search in the vein of Juditsky and Nemirovski [2011a].

4.7 Appendix

4.7.1 Motivation for the multiclass hinge loss

We justify the multiclass extension (4.1.4) of the hinge loss due to Shalev-Shwartz and Ben-David [2014]. In the binary case, the soft-margin SVM objective is

$$\frac{1}{n}\sum_{i=1}^{n}\left[\max(0, 1 - \widetilde{y}_i u^\top x_i)\right] + \lambda\|u\|,$$

where $u \in \mathbb{R}^d$, $\|\cdot\|$ is some norm, and $\widetilde{y}_i \in \{-1, 1\}$. Introducing $y = e_{\widetilde{y}} \in \{e_{-1}, e_1\}$ where e_j is the j-th standard basis vector (the dimension of space are symbolically indexed in $\{-1, 1\}$), and putting $u_1 = -u_{-1} = \frac{u}{2}$, we can rewrite the loss as

$$\max(0, 1 - \widetilde{y}u^\top x) = \max_{k\in\{1,-1\}} \left\{\mathbb{1}\{e_k \neq y\} + u_k^\top x - u_{\widetilde{y}}^\top x\right\}.$$

The advantage of this reformulation is that we can naturally pass to the multiclass case, by replacing the set $\{-1, 1\}$ with $\{1, ..., K\}$ and introducing $u_1, ..., u_K \in \mathbb{R}^d$ without any restrictions:

$$\begin{aligned}\max_{k\in\{1,\dots,K\}} \left\{\mathbb{1}\{e_k \neq y\} + u_k^\top x - u_{\widetilde{y}}^\top x\right\} &= \max_{v\in\{e_1,\dots,e_K\}} \left\{\mathbb{1}\{v \neq y\} + \sum_{j=1}^{K}(v[j] - y[j])u_j^\top x\right\} \\ &= \max_{v\in\{e_1,\dots,e_K\}} \left\{\mathbb{1}\{v \neq y\} + \sum_{j=1}^{K}(v - y)^\top U^\top x\right\} =: \ell(U, (x, y)),\end{aligned}$$

where $\cdot[j]$ denotes the j-th element of a vector, and $U \in \mathbb{R}^{d\times K}$ has u_j as its j-th column. Finally, we can rewrite $\ell(U, (x, y))$ as follows:

$$\ell(U, (x, y)) = \max_{v\in\Delta_K} \left\{1 - v^\top y + \sum_{j=1}^{K}(v - y)^\top U^\top x\right\}.$$

This is because we maximize an affine function of v, and $1 - v^\top y = \mathbb{1}\{v \neq y\}$ at the vertices. Thus, we arrive at the announced saddle-point problem

$$\min_{U\in\mathbb{R}^{d\times k}} \max_{V\in(\Delta_k^\top)^{\otimes n}} 1 - \frac{1}{n}\mathrm{tr}[V^\top Y] + \frac{1}{n}\mathrm{tr}\left[(V - Y)^\top XU\right] + \lambda\|U\|_{\mathscr{U}}.$$

4.7.2 General accuracy bounds for the composite saddle-point Mirror Descent

Deterministic case. Here we provide general accuracy bounds which are instantiated in Theorems 4.1–1. Below we outline the general setting that encompasses, in particular, the case of (4.2.17) solved via (**MD**) or (**MP**) with initialization (**Init**).

— We consider a convex-concave saddle-point problem

$$\min_{U \in \mathcal{U}} \max_{V \in \mathcal{V}} f(U,V)$$

with a composite objective,

$$f(U,V) = \Phi(U, V - Y) + \Upsilon(U) - \mathcal{F}(V),$$

where

$$\Phi(U,V) = \frac{1}{n} V^\top X U$$

is a bilinear function, and $\Upsilon(U), \mathcal{F}(V)$ are convex "simple" terms. Moreover, we assume that the primal feasible set $\mathcal{U}$ belongs to the $\|\cdot\|_{\mathscr{U}}$-norm ball with radius $R_{\mathcal{U}}$, the dual constraint set $\mathcal{V}$ belongs to the $\|\cdot\|_{\mathscr{V}}$-norm ball with radius $R_{\mathcal{V}}$, and $\|Y\|_{\mathscr{V}} \le R_{\mathcal{V}}$.[6] To simplify the results, we make the mild assumption (satisfied in all known to us situations):

$$\Omega_{\mathcal{U}} \geqslant R_{\mathcal{U}}^2, \quad \Omega_{\mathcal{V}} \geqslant R_{\mathcal{V}}^2. \tag{4.7.1}$$

— Recall that the vector field of partial gradients of $\Psi(U,V) := \Phi(U, V-Y)$ is

$$\begin{aligned} G(W) :&= (\nabla_U \Psi(U,V), -\nabla_V \Psi(U,V)) \\ &= \frac{1}{n}(X^\top (V-Y), -XU) \end{aligned} \tag{4.7.2}$$

— Given the partial proximal setups $(\|\cdot\|_{\mathscr{U}}, \phi_{\mathcal{U}}(\cdot))$ and $(\|\cdot\|_{\mathscr{V}}, \phi_{\mathcal{V}}(\cdot))$, we run Composite Mirror Descent (4.2.6) or Mirror Prox (4.2.8) on the vector field $G(W)$ with the joint penalty term

$$h(W) = \Upsilon(U) + \mathcal{F}(V),$$

the "balanced" joint potential given by (4.2.12), and stepsizes γ_t.

We now provide the convergence analysis of the Mirror Descent scheme, extending the argument of [Duchi et al., 2010, Lemma 1] to composite saddle-point optimization.

Theorem 4.3. *In the above setting, let $(\bar{U}^T, \bar{V}^T) = \frac{1}{T}\sum_{t=0}^{T-1}(U^t, V^t)$ be the average of the first T iterates of the composite Mirror Descent (4.2.6) with constant stepsize*

$$\gamma_t \equiv \frac{1}{\mathcal{L}_{\mathscr{U},\mathscr{V}} \sqrt{5T\Omega_{\mathcal{U}}\Omega_{\mathcal{V}}}},$$

where

$$\mathcal{L}_{\mathscr{U},\mathscr{V}} := \frac{1}{n} \sup_{\|U\|_{\mathscr{U}} \le 1} \|XU\|_{\mathscr{V}^*}.$$

[6] Note that the linear term $\frac{1}{n}Y^\top XU$ can be absorbed into the simple term $\Upsilon(U)$, which will slightly improve the bound in Theorem 4.3. However, this improvement is impossible in the stochastic version of the algorithm, in which we sample the linear form $Y^\top X$ but not the gradient of $\Upsilon(U)$.

Then we have the following guarantee for the duality gap:

$$\Delta_f(\bar{U}^T, \bar{V}^T) \leqslant \frac{2\sqrt{5}\mathcal{L}_{\mathscr{U},\mathscr{V}}\sqrt{\Omega_{\mathcal{U}}\Omega_{\mathcal{V}}}}{\sqrt{T}} + \frac{\Upsilon(U^0) - \min_{U\in\mathcal{U}}\Upsilon(U)}{T} + \frac{\mathcal{F}(V^0) - \min_{V\in\mathcal{V}}\mathcal{F}(V)}{T}.$$

Moreover, if one of the functions $\Upsilon(U)$, $\mathcal{F}(V)$ is affine, the corresponding $O(1/T)$ error term vanishes from the bound.

Proof. **1^o**. We begin by introducing the norm for $W = (U, V)$:

$$\|W\|_{\mathscr{W}} = \sqrt{\frac{\|U\|_{\mathscr{U}}^2}{2\Omega_{\mathcal{U}}} + \frac{\|V\|_{\mathscr{V}}^2}{2\Omega_{\mathcal{V}}}} \tag{4.7.3}$$

and its dual norm defined for $G = (G_U, G_V)$ with $G_U \in \mathbb{R}^{d\times k}$ and $G_V \in \mathbb{R}^{n\times k}$:

$$\|G\|_{\mathscr{W}^*} = \sqrt{2\Omega_{\mathcal{U}}\|G_U\|_{\mathscr{U}^*}^2 + 2\Omega_{\mathcal{V}}\|G_V\|_{\mathscr{V}^*}^2}, \tag{4.7.4}$$

where $\|\cdot\|_{\mathscr{U}^*}$ and $\|\cdot\|_{\mathscr{V}^*}$ are the dual norms for $\|\cdot\|_{\mathscr{U}}$ and $\|\cdot\|_{\mathscr{V}}$ correspondingly. We now make a few observations. First, the joint potential $\phi_{\mathcal{W}}(W)$ given by (4.2.12) is 1-strongly convex with respect to the norm $\|\cdot\|_{\mathscr{W}}$. Second, we can compute the potential difference corresponding to $\phi_{\mathcal{W}}$:

$$\Omega_{\mathcal{W}} := \max_{W\in\mathcal{W}} \phi_{\mathcal{W}}(W) - \min_{W\in\mathcal{W}} \phi_{\mathcal{W}}(W) = 1 \tag{4.7.5}$$

Finally, by (4.7.2) and (4.2.10) we have

$$\max_{W\in\mathcal{W}} \|G_U(W)\|_{\mathscr{U}^*} \leqslant 2\mathcal{L}_{\mathscr{U},\mathscr{V}} R_{\mathcal{V}}, \quad \max_{W\in\mathcal{W}} \|G_V(W)\|_{\mathscr{V}^*} \leqslant \mathcal{L}_{\mathscr{U},\mathscr{V}} R_{\mathcal{U}},$$

combining which with (4.7.1) we bound the $\|\cdot\|_{\mathscr{W}^*}$-norm of $G(W)$ on $\mathcal{W}$:

$$\max_{W\in\mathcal{W}} \|G(W)\|_{\mathscr{W}^*} \leqslant \sqrt{10}\mathcal{L}_{\mathscr{U},\mathscr{V}}\sqrt{\Omega_{\mathcal{U}}\Omega_{\mathcal{V}}}. \tag{4.7.6}$$

2^o. We now follow the classical convergence analysis of composite Mirror Descent, see Duchi et al. [2010], extending it for convex-concave objectives. By the convexity properties of $\Psi(U, V) = \Phi(U, V - Y)$, for any $(\bar{U}, \bar{V}) \in \mathcal{W}$ and $(U, V) \in \mathcal{W}$ it holds

$$\begin{aligned}\Psi(\bar{U}, V) - \Psi(U, \bar{V}) &= \Psi(\bar{U}, V) - \Psi(\bar{U}, \bar{V}) + \Psi(\bar{U}, \bar{V}) - \Psi(U, \bar{V})\\ &\leqslant \left\langle \nabla_U \Psi(\bar{U}, \bar{V}), \bar{U} - U \right\rangle - \left\langle \nabla_V \Psi(\bar{U}, \bar{V}), \bar{V} - V \right\rangle\\ &= \left\langle G(\bar{W}), \bar{W} - W \right\rangle,\end{aligned}$$

Let $W^t = (U^t, V^t)$ be the t-th iterate of (4.2.6) for $t \geq 1$. By convexity of $\Upsilon(U)$ and $\mathcal{F}(V)$, and denoting $h(W) = \Upsilon(U) + \mathcal{F}(V)$, we have, for any $W = [U, V]$, that

$$\begin{aligned}&\Psi(U^{t-1}, V) - \Psi(U, V^{t-1}) + h(W^t) - h(W)\\ &\qquad\leqslant \left\langle G(W^{t-1}), W^{t-1} - W \right\rangle + \left\langle \partial h(W^t), W^t - W \right\rangle.\end{aligned} \tag{4.7.7}$$

Let us now bound the right-hand side. Note that the first-order optimality condition for (4.2.6) (denoting $\phi(\cdot) := \phi_{\mathcal{W}}(\cdot)$ the joint potential) writes [7]

$$\left\langle \gamma_t[G(W^{t-1}) + \partial h(W^t)] + \nabla\phi(W^t) - \nabla\phi(W^{t-1}), W^t - W \right\rangle \leqslant 0. \tag{4.7.8}$$

Combining this with (4.7.7), we get

$$\begin{aligned}\gamma_t[\Psi(U^{t-1}, V) - \Psi(U, V^{t-1}) + h(W^t) - h(W)] \leqslant \left\langle \nabla\phi(W^{t-1}) - \nabla\phi(W^t), W^t - W \right\rangle \\ + \gamma_t \left\langle G(W^{t-1}), W^{t-1} - W^t \right\rangle.\end{aligned} \tag{4.7.9}$$

By the well-known identity,

$$\left\langle \nabla\phi(W^{t-1}) - \nabla\phi(W^t), W^t - W \right\rangle = D_\phi(W, W^{t-1}) - D_\phi(W, W^t) - D_\phi(W^t, W^{t-1}), \tag{4.7.10}$$

see, e.g., [Beck and Teboulle, 2003, Lemma 4.1]. On the other hand, by the Fenchel-Young inequality we have

$$\begin{aligned}\gamma_t \left\langle G(W^{t-1}), W^{t-1} - W^t \right\rangle &\leqslant \frac{\gamma_t^2 \|G(W^{t-1})\|_{\mathscr{W}^*}^2}{2} + \frac{\|W^{t-1} - W^t\|_{\mathscr{W}}^2}{2} \\ &\leqslant 5\gamma_t^2 \mathcal{L}_{\mathscr{U},\mathscr{V}}^2 \Omega_{\mathcal{U}}\Omega_{\mathcal{V}} + D_\phi(W^t, W^{t-1}),\end{aligned} \tag{4.7.11}$$

where we used (4.7.6) and 1-strong convexity of $\phi(\cdot)$ with respect to $\|\cdot\|_{\mathscr{W}}$. Thus, we obtain

$$\begin{aligned}\gamma_t[\Psi(U^{t-1}, V) - \Psi(U, V^{t-1}) + h(W^t) - h(W)] \leq & D_\phi(W, W^{t-1}) - D_\phi(W, W^t) \\ & + 5\gamma_t^2 \mathcal{L}_{\mathscr{U},\mathscr{V}}^2 \Omega_{\mathcal{U}}\Omega_{\mathcal{V}}.\end{aligned} \tag{4.7.12}$$

3^o. Now, assuming the constant stepsize, by the convexity properties of $\Psi(\cdot,\cdot)$ and $h(\cdot)$ we obtain

$$\begin{aligned}f(\bar{U}^T, V) - f(U, \bar{V}^T) &= \Psi(\bar{U}^T, V) - \Psi(U, \bar{V}^T) + h(\bar{W}^T) - h(W) \\ &\leqslant \frac{1}{T}\sum_{t=1}^{T} \Psi(U^{t-1}, V) - \Psi(U, V^{t-1}) + h(W^{t-1}) - h(W) \\ &\leqslant \frac{1}{T}\left(h(W^0) - h(W^T) + \frac{D_\phi(W, W^0)}{\gamma} + 5T\gamma \mathcal{L}_{\mathscr{U},\mathscr{V}}^2 \Omega_{\mathcal{U}}\Omega_{\mathcal{V}} \right) \\ &\leqslant \frac{1}{T}\left(h(W^0) - \min_{W \in \mathcal{W}} h(W) + \frac{1}{\gamma} + 5T\gamma \mathcal{L}_{\mathscr{U},\mathscr{V}}^2 \Omega_{\mathcal{U}}\Omega_{\mathcal{V}} \right).\end{aligned} \tag{4.7.13}$$

where for the third line we substituted (4.7.12), simplified the telescoping sum, and used that $D(W, W^T) \geqslant 0$, and in the last line we used $D(W, W^0) \leqslant \Omega_{\mathcal{W}} \leqslant 1$, cf. (4.7.5). The choice

$$\gamma = \frac{1}{\mathcal{L}_{\mathscr{U},\mathscr{V}}\sqrt{5T\Omega_{\mathcal{U}}\Omega_{\mathcal{V}}}},$$

results in the accuracy bound from the premise of the theorem:

$$\Delta_f(\bar{U}^T, \bar{V}^T) \leqslant \frac{2\sqrt{5}\mathcal{L}_{\mathscr{U},\mathscr{V}}\sqrt{\Omega_{\mathcal{U}}\Omega_{\mathcal{V}}}}{\sqrt{T}} + \frac{h(W^0) - \min_{W\in\mathcal{W}} h(W)}{T}.$$

Finally, assume that one of the terms $\Upsilon(U)$, $\mathcal{F}(V)$ is affine – w.l.o.g. let it be $\Upsilon(U)$. Then, since $\nabla\Upsilon(U)$ is constant, $\partial h(W^t) = (\nabla\Upsilon(U^t), \partial\mathcal{F}(V^t))$ in (4.7.8) can be replaced with $(\nabla\Upsilon(U^{t-1}), \partial\mathcal{F}(V^t))$. Then in (4.7.12) we can replace $h(W^t) - h(W^0)$ with $\Upsilon(U^{t-1}) - \Upsilon(U) + \mathcal{F}(V^t) - \mathcal{F}(V)$, implying that the term $h(W^0) - h(W^T)$ in the right-hand side of (4.7.13) gets replaced with $\mathcal{F}(V^0) - \mathcal{F}(V^t)$. The claim is proved. □

Stochastic Mirror Descent. We now consider the stochastic setting that allows to encompass (**S-MD**). Stochastic Mirror Descent is given by

$$\begin{aligned} W^0 &= \min_{W\in\mathcal{W}} \phi_{\mathcal{W}}(W); \\ W^t &= \operatorname*{arg\,min}_{W\in\mathcal{W}} \left\{ h(W) + \langle \Xi(W^{t-1}), W\rangle + \frac{1}{\gamma_t} D_{\phi_{\mathcal{W}}}(W, W^{t-1}) \right\}, \; t \geq 1, \end{aligned} \tag{4.7.14}$$

where

$$\Xi(W) := \frac{1}{n}(\eta_{V,Y}, -\xi_U)$$

is the unbiased estimate of the first-order oracle $G(W) = \frac{1}{n}(X^\top(V - Y), -XU)$. Let us introduce the corresponding variance proxies (refer to the preamble of Section 4.3 for the discussion):

$$\sigma_{\mathcal{U}}^2 = \frac{1}{n^2} \sup_{U\in\mathcal{U}} \mathbb{E}\left[\|XU - \xi_U\|_{\mathscr{V}^*}^2\right], \quad \sigma_{\mathcal{V}}^2 = \frac{1}{n^2} \sup_{(V,Y)\in\mathcal{V}\times\mathcal{V}} \mathbb{E}\left[\left\|X^\top(V - Y) - \eta_{V,Y}\right\|_{\mathscr{U}^*}^2\right]. \tag{4.7.15}$$

We assume that the noises $G(W^{t-1}) - \Xi(W^{t-1})$ are independent along the iterations of (4.7.14). In this setting, we prove the following generalization of Theorem 4.3:

Theorem 4.4. *Let $(\bar{U}^T, \bar{V}^T) = \frac{1}{T}\sum_{t=0}^{T-1}(U^t, V^t)$ be the average of the first T iterates of the Stochastic Composite Mirror Descent (4.7.14) with constant stepsize*

$$\gamma_t \equiv \frac{1}{\sqrt{T}} \min\left\{ \frac{1}{\sqrt{10}\mathcal{L}_{\mathscr{U},\mathscr{V}}\sqrt{\Omega_{\mathcal{U}}\Omega_{\mathcal{V}}}}, \frac{1}{\sqrt{2}\sqrt{\Omega_{\mathcal{U}}\bar{\sigma}_{\mathcal{V}}^2 + \Omega_{\mathcal{V}}\bar{\sigma}_{\mathcal{U}}^2}} \right\},$$

where $\mathcal{L}_{\mathscr{U},\mathscr{V}}, \Omega_{\mathcal{U}}, \Omega_{\mathcal{V}}$ are the same as in Theorem 4.4, and $\bar{\sigma}_{\mathcal{U}}^2, \bar{\sigma}_{\mathcal{V}}^2$ are the upper bounds

[7] Note that $\phi(W)$ is continuously differentiable in the interior of $\mathcal{W}$, and $\nabla\phi$ diverges on the border of $\mathcal{W}$, then the iterates are guaranteed to stay in the interior of $\mathcal{W}$ Beck and Teboulle [2003].

for $\sigma_{\mathcal{U}}^2, \sigma_{\mathcal{V}}^2$, cf. (4.7.15)*. Then it holds*

$$\mathbb{E}[\Delta_f(\bar{U}^T,\bar{V}^T)] \leqslant \frac{2\sqrt{10}\mathcal{L}_{\mathscr{U},\mathscr{V}}\sqrt{\Omega_{\mathcal{U}}\Omega_{\mathcal{V}}}}{\sqrt{T}} + \frac{2\sqrt{2}\sqrt{\Omega_{\mathcal{U}}\bar{\sigma}_{\mathcal{V}}^2+\Omega_{\mathcal{V}}\bar{\sigma}_{\mathcal{U}}^2}}{\sqrt{T}} + \frac{\Upsilon(U^0)-\min_{U\in\mathcal{U}}\Upsilon(U)}{T} + \frac{\mathcal{F}(V^0)-\min_{V\in\mathcal{V}}\mathcal{F}(V)}{T},$$

where $\mathbb{E}[\cdot]$ is the expectation over the randomness in (4.7.14)*. Moreover, if one of the functions $\Upsilon(U)$, $\mathcal{F}(V)$ is affine, the corresponding $O(1/T)$ term can be discarded.*

Proof. The proof closely follows that of Theorem 4.3. First, **1°** remains unchanged. Then, in the first-order condition (4.7.8) one must replace $G(W^{t-1})$ with $\Xi(W^{t-1})$, which results in replacing (4.7.9) with

$$\begin{aligned}\gamma_t[\Psi(U^{t-1},V)-\Psi(U,V^{t-1})+h(W^t)-h(W)]\\ \leqslant \left\langle \nabla\phi(W^{t-1})-\nabla\phi(W^t), W^t-W\right\rangle\\ +\gamma_t\left\langle \Xi(W^{t-1}), W^{t-1}-W^t\right\rangle\\ +\gamma_t\left\langle G(W^{t-1})-\Xi(W^{t-1}), W^{t-1}-W\right\rangle,\end{aligned}$$

where the last term has zero mean. The term $\gamma_t\left\langle \Xi(W^{t-1}), W^{t-1}-W^t\right\rangle$ can be bounded using Young's inequality, and 1-strong convexity of $\phi(\cdot)$, cf. (4.7.11):

$$\begin{aligned}&\gamma_t\left\langle \Xi(W^{t-1}), W^{t-1}-W^t\right\rangle\\ &\leqslant \frac{\gamma_t^2\|\Xi(W^{t-1})\|_{\mathscr{W}^*}^2}{2}+\frac{\|W^{t-1}-W^t\|_{\mathscr{W}}^2}{2}\\ &\leqslant \gamma_t^2\|G(W^{t-1})\|_{\mathscr{W}^*}^2+\gamma_t^2\|\Xi(W^{t-1})-G(W^{t-1})\|_{\mathscr{W}^*}^2+D_\phi(W^t,W^{t-1}).\end{aligned}$$

Combining (4.7.6), (4.7.4), and (4.7.15), this implies

$$\gamma_t\left\langle \Xi(W^{t-1}), W^{t-1}-W^t\right\rangle \leqslant 10\gamma_t^2\mathcal{L}_{\mathscr{U},\mathscr{V}}^2\Omega_{\mathcal{U}}\Omega_{\mathcal{V}}+2\gamma_t^2\left(\Omega_{\mathcal{U}}\bar{\sigma}_{\mathcal{V}}^2+\Omega_{\mathcal{V}}\bar{\sigma}_{\mathcal{U}}^2\right)+D_\phi(W^t,W^{t-1}).$$

Using (4.7.10) and (4.7.5), this results in

$$\begin{aligned}\Delta_f(\bar{U}^T,\bar{V}^T) &= \mathbb{E}\left[\max_{(U,V)\in\mathcal{W}}\left\{f(\bar{U}^T,V)-f(U,\bar{V}^T)\right\}\right]\\ &\leqslant \frac{1}{T}\left[h(W^0)-\min_{W\in\mathcal{W}}h(W)+\frac{1}{\gamma}+2T\gamma\left(5\mathcal{L}_{\mathscr{U},\mathscr{V}}^2\Omega_{\mathcal{U}}\Omega_{\mathcal{V}}+\Omega_{\mathcal{U}}\bar{\sigma}_{\mathcal{V}}^2+\Omega_{\mathcal{V}}\bar{\sigma}_{\mathcal{U}}^2\right)\right];\end{aligned}$$

note that maximization on the left is *under* the expectation (and not vice versa) because the right hand side is independent from $W=(U,V)$. Choosing γ to balance the terms, we arrive at the desired bound. Finally, improvement in the case of affine $\Upsilon(U)$, $\mathcal{F}(V)$ is obtained in the same way as in Theorem 4.4. □

4.7.3 Auxiliary lemmas

Lemma 4.1. *Let* $X^0 \in \mathbb{R}^n_+$ *and* $X^1 = \operatorname*{arg\,min}_{\substack{\|X\|_1 \leqslant R \\ X \in \mathbb{R}^n_+}} \left\{ C_1\|X\|_1 + \langle S, X\rangle + C_2 \sum_{i=1}^n X_i \log \frac{X_i}{X_i^0} \right\}$.

Then,

$$X_i^1 = \rho \cdot \frac{X_i^0 \cdot \exp(-S_i/C_2)}{M},$$

where $M = \sum_{j=1}^n X_j^0 \cdot \exp(-S_j/C_2)$ *and* $\rho = \min(M \cdot e^{-\frac{C_1+C_2}{C_2}}, R)$.

Proof. Clearly, we have

$$X^1 = \operatorname*{arg\,min}_{r \leqslant R} \min_{\substack{\|X\|_1 = r \\ X \in \mathbb{R}^n_+}} \left\{ C_1 r + \langle S, X\rangle + C_2 \sum_{i=1}^n X_i \log \frac{X_i}{X_i^0} \right\}.$$

Let us first do the internal minimization. By simple algebra, the first-order optimality condition for the Lagrangian dual problem (with constraint $\|X\|_1 = R$) amounts to

$$S_i + C_2 + C_2 \log X_i^1 - C_2 \log X_i^0 + \kappa = 0,$$

where κ is Lagrange multiplier, and $\sum_{i=1}^n X_i^1 = r$. Equivalently,

$$X_i^1 = X_i^0 \cdot \exp\left(- \left(\frac{\kappa + C_2 + S_i}{C_2} \right) \right),$$

that is,

$$X_i^1 = r \cdot \frac{X_i^0 \cdot \exp(-S_i/C_2)}{\sum_j X_j^0 \cdot \exp(-S_j/C_2)}.$$

Denoting $D_j = \exp(-S_j/C_2)$ and $M = \sum_j X_j^0 \cdot D_j$ and substituting for X_1 in the external minimization problem, we arrive at

$$\rho = \operatorname*{arg\,min}_{r \leqslant R} \left\{ C_1 r + r \sum_i \frac{X_i^0 D_i S_i}{M} + C_2 \cdot r \sum_i \frac{X_i^0 D_i}{M} \cdot \log \left[\frac{r \cdot D_i}{M} \right] \right\}.$$

One can easily verify that the counterpart of this minimization problem with $R = \infty$ has a unique stationary point $r^* = M \cdot e^{-\frac{C_1+C_2}{C_2}} > 0$. As the minimized function is convex, the minimum is attained at the point $\rho = \min(r^*, R)$. □

Lemma 4.2. *Given* $X \in \mathbb{R}^{n \times d}$ *and mixed norms (cf. (4.2.4))* $\| \cdot \|_{p_U^1 \times p_U^2}$ *on* $\mathbb{R}^{d \times k}$

and $\|\cdot\|_{q_V^1 \times q_V^2}$ *on* $\mathbb{R}^{n\times k}$ *with* $p_U^2 \leqslant q_V^2$*, one has*

$$\sup_{\|U\|_{p_U^1 \times p_U^2} \leqslant 1} \left\{ \sum_{i=1}^{d} \|X(:,i)\|_{q_V^1} \cdot \|U(i,:)\|_{q_V^2} \right\} = \|X^\top\|_{q_U^1 \times q_V^1},$$

where q_U^1 *is the conjugate of* p_U^1*, i.e.,* $1/p_U^1 + 1/q_U^1 = 1$*.*

Proof. First assume $p_U^2 = q_V^2$. Let $a_i = \|X(:,i)\|_{q_V^1}$, $u_i = \|U(i,:)\|_{q_V^2}$, $1 \leqslant i \leqslant d$. Then,

$$\sup_{\|U\|_{p_U^1 \times p_U^2} \leqslant 1} \left\{ \sum_{i=1}^{d} \|X(:,i)\|_{q_V^1} \cdot \|U(i,:)\|_{q_V^2} \right\} = \sup_{\|u\|_{p_U^1} \leqslant 1} \sum_{i=1}^{d} a_i u_i = \|a\|_{q_U^1} = \|X^\top\|_{q_U^1 \times q_V^1}.$$

Now let $p_U^2 < q_V^2$. Then, for any $i \le d$ one has $\|U(i,:)\|_{q_V^2} < \|U(i,:)\|_{p_U^2}$ unless $U(i,:)$ has a single non-zero element, in which case $\|U(i,:)\|_{q_V^2} = \|U(i,:)\|_{p_U^2}$. Hence, the supremum must be attained on such U, for which the previous argument applies. □

Lemma 4.3. *In the setting of Lemma 4.2, for any* $q_U^1 \geqslant 1$ *and* $q_U^2 \geqslant 1$ *one has:*

$$\sup_{\|V\|_{\infty \times 1} \le 1} \left\{ \sum_{i=1}^{n} \|X^\top(:,i)\|_{q_U^1} \cdot \|V(i,:)\|_{q_U^2} \right\} = \|X\|_{1 \times q_U^1}.$$

Proof. The claim follows by instatiating Lemma 4.2. □

4.7.4 Proof of Proposition 4.1

By (4.2.10), and verifying that the dual norm to $\|\cdot\|_{2\times 1}$ is $\|\cdot\|_{2\times\infty}$, we have

$$\mathcal{L}_{\mathscr{U},\mathscr{V}} = \sup_{\|U\|_{1\times 1} \leqslant 1} \|XU\|_{2\times\infty}.$$

The maximization over the unit ball $\|U\|_{1\times 1} \leqslant 1$ can be replaced with that over its extremal points, which are the matrices U that have zeroes in all positions except for one in which there is 1. Let (i,j) be this position, then for every such U we have:

$$\|XU\|_{2\times\infty} = \sqrt{\sum_{l=1}^{n} \sup_j |X(l,:)U(:,j)|^2} = \sqrt{\sum_{l=1}^{n} |X(l,i)|^2} = \sqrt{\sum_{l=1}^{n} |X^\top(i,l)|^2}.$$

As a result,

$$\sup_{\|U\|_{1\times 1} \leqslant 1} \|XU\|_{2\times\infty} = \sup_{1 \le i \le k} \sqrt{\sum_{l=1}^{n} |X^\top(i,l)|^2} = \|X^\top\|_{\infty \times 2}. \quad \square$$

4.7.5 Proof of Proposition 4.2

We prove an extended result that holds when $\|\cdot\|_{\mathscr{U}}$ and $\|\cdot\|_{\mathscr{V}}$ are more general mixed $(\ell_p \times \ell_q)$-norms, cf. (4.2.4).

Proposition 4.5. *Let $\|\cdot\|_{\mathscr{U}} = \|\cdot\|_{p_U^1 \times p_U^2}$ on $\mathbb{R}^{2d\times k}$, and $\|\cdot\|_{\mathscr{V}} = \|\cdot\|_{p_V^1 \times p_V^2}$ on $\mathbb{R}^{n\times k}$. Then, optimal solutions $p^* = p^*(\widehat{X}, \widehat{U})$ and $q^* = q^*(\widehat{X}, V, Y)$ to (4.3.3) are given by*

$$p_i^* = \frac{\|\widehat{X}(:,i)\|_{q_V^1} \cdot \|\widehat{U}(i,:)\|_{q_V^2}}{\sum_{\imath=1}^{2d} \|\widehat{X}(:,\imath)\|_{q_V^1} \cdot \|\widehat{U}(\imath,:)\|_{q_V^2}}, \quad q_j^* = \frac{\|\widehat{X}(j,:)\|_{q_U^1} \cdot \|V(j,:) - Y(j,:)\|_{q_U^2}}{\sum_{\jmath=1}^{n} \|\widehat{X}(\jmath,:)\|_{q_U^1} \cdot \|V(\jmath,:) - Y(\jmath,:)\|_{q_U^2}}.$$

Moreover, we can bound their respective variance proxies (cf. (4.3.2)): introducing

$$\widehat{\mathcal{L}}^2_{\mathscr{U},\mathscr{V}} = \frac{1}{n^2} \sup_{\|\widehat{U}\|_{\mathscr{U}} \le 1} \|\widehat{X}\widehat{U}\|^2_{\mathscr{V}^*},$$

we have, as long as $p_U^2 \leqslant q_V^2$,

$$\sigma^2_{\widehat{\mathcal{U}}}(p^*) \leqslant 2R_{\mathcal{U}}^2 \widehat{\mathcal{L}}^2_{\mathscr{U},\mathscr{V}} + \frac{2}{n^2} R_{\mathcal{U}}^2 \|\widehat{X}^\top\|^2_{q_U^1 \times q_V^1},$$

and, as long as $p_V^1 \geqslant 2$,

$$\sigma^2_{\mathcal{V}}(q^*) \leqslant 8n\widehat{\mathcal{L}}^2_{\mathscr{U},\mathscr{V}} + \frac{8}{n^2}\|\widehat{X}\|^2_{1\times q_U^1}.$$

Proof. Note that the dual norms to $\|\cdot\|_{p_U^1 \times p_U^2}$ and $\|\cdot\|_{p_V^1 \times p_V^2}$ are given by $\|\cdot\|_{q_U^1 \times q_U^2}$ and $\|\cdot\|_{q_V^1 \times q_V^2}$ correspondingly, see, e.g., Sra [2012].

1^o. For $\mathbb{E}\left[\|\xi_{\widehat{U}}(p)\|^2_{\mathscr{V}^*}\right]$ we have:

$$\begin{aligned}
\mathbb{E}\left[\|\xi_{\widehat{U}}(p)\|^2_{\mathscr{V}^*}\right] = \sum_{i=1}^{2d} p_i \left\| \widehat{X} \frac{e_i e_i^\top}{p_i} \widehat{U} \right\|^2_{q_V^1 \times q_V^2} &= \sum_{i=1}^{2d} \frac{1}{p_i} \left\| \widehat{X}(:,i) \cdot \widehat{U}(i,:) \right\|^2_{q_V^1 \times q_V^2} \\
&= \sum_{i=1}^{2d} \frac{1}{p_i} \|\widehat{X}(:,i)\|^2_{q_V^1} \cdot \|\widehat{U}(i,:)\|^2_{q_V^2},
\end{aligned}$$

where the last transition can be verified directly. The right-hand side can be easily minimized on Δ_{2d} explicitly, which results in

$$p_i^* = \frac{\|\widehat{X}(:,i)\|_{q_V^1} \cdot \|\widehat{U}(i,:)\|_{q_V^2}}{\sum_{\imath=1}^{2d} \|\widehat{X}(:,\imath)\|_{q_V^1} \cdot \|\widehat{U}(\imath,:)\|_{q_V^2}}$$

and

$$\mathbb{E}\left[\|\xi_{\widehat{U}}(p^*)\|^2_{\mathscr{V}^*}\right] = \left[\sum_{i=1}^{2d} \|\widehat{X}(:,i)\|_{q_V^1} \cdot \|\widehat{U}(i,:)\|_{q_V^2}\right]^2.$$

Now we can bound $\sigma^2_{\widehat{\mathcal{U}}}(p^*)$ via the triangle inequality:

$$\begin{aligned}\sigma^2_{\widehat{\mathcal{U}}}(p^*) &\leqslant \frac{2}{n^2}\sup_{\widehat{U}\in\widehat{\mathcal{U}}}\|\widehat{X}\widehat{U}\|^2_{\mathscr{V}^*} + \frac{2}{n^2}\sup_{\widehat{U}\in\widehat{\mathcal{U}}}\mathbb{E}\left[\|\xi_{\widehat{U}}(p^*)\|^2_{\mathscr{V}^*}\right] \\ &= 2R^2_{\mathcal{U}}\widehat{\mathcal{L}}^2_{\mathscr{U},\mathscr{V}} + \frac{2}{n^2}\sup_{\widehat{U}\in\widehat{\mathcal{U}}}\left[\sum_{i=1}^{2d}\|\widehat{X}(:,i)\|_{q^1_V}\cdot\|\widehat{U}(i,:)\|_{q^2_V}\right]^2 \\ &= 2R^2_{\mathcal{U}}\widehat{\mathcal{L}}^2_{\mathscr{U},\mathscr{V}} + \frac{2}{n^2}\sup_{\|\widehat{U}\|_{\mathscr{U}}\le R_{\mathcal{U}}}\left[\sum_{i=1}^{2d}\|\widehat{X}(:,i)\|_{q^1_V}\cdot\|\widehat{U}(i,:)\|_{q^2_V}\right]^2 \\ &= 2R^2_{\mathcal{U}}\widehat{\mathcal{L}}^2_{\mathscr{U},\mathscr{V}} + \frac{2}{n^2}R^2_{\mathcal{U}}\|\widehat{X}^\top\|^2_{q^1_U\times q^1_V},\end{aligned}$$

where we used Lemma 4.2 (see Appendix 4.7.3) in the last transition.

2^o. We now deal with $\mathbb{E}\left[\|\eta_{V,Y}(q)\|^2_{\mathscr{U}^*}\right]$. As previously, we can explicitly compute

$$q^*_j = \frac{\|\widehat{X}(j,:)\|_{q^1_U}\cdot\|V(j,:)-Y(j,:)\|_{q^2_U}}{\sum_{j=1}^n\|\widehat{X}(j,:)\|_{q^1_U}\cdot\|V(j,:)-Y(j,:)\|_{q^2_U}}$$

and

$$\mathbb{E}\left[\|\eta_{V,Y}(q^*)\|^2_{\mathscr{U}^*}\right] = \left[\sum_{j=1}^n\|\widehat{X}(j,:)\|_{q^1_U}\cdot\|V(j,:)-Y(j,:)\|_{q^2_U}\right]^2.$$

Thus, by the triangle inequality,

$$\begin{aligned}\sigma^2_{\mathcal{V}}(q_*) &\leqslant \frac{2}{n^2}\sup_{(V,Y)\in\mathcal{V}\times\mathcal{V}}\|\widehat{X}^\top(V-Y)\|^2_{\mathscr{U}^*} + \frac{2}{n^2}\sup_{(V,Y)\in\mathcal{V}\times\mathcal{V}}\mathbb{E}\left[\|\eta_{V,Y}(q^*)\|^2_{\mathscr{U}^*}\right] \\ &\leqslant \frac{2}{n^2}\sup_{\|V\|_{\infty\times1}\leqslant2}\|\widehat{X}^\top V\|^2_{\mathscr{U}^*} + \frac{2}{n^2}\sup_{\|V\|_{\infty\times1}\leqslant2}\left[\sum_{j=1}^n\|\widehat{X}(j,:)\|_{q^1_U}\cdot\|V(j,:)\|_{q^2_U}\right]^2 \\ &= \frac{8}{n}\sup_{\|V\|_{2\times1}\leqslant1}\|\widehat{X}^\top V\|^2_{\mathscr{U}^*} + \frac{8}{n^2}\|\widehat{X}\|^2_{1\times q^1_U} \\ &\leqslant \frac{8}{n}\sup_{\|V\|_{p^1_V\times p^2_V}\leqslant1}\|\widehat{X}^\top V\|^2_{\mathscr{U}^*} + \frac{8}{n^2}\|\widehat{X}\|^2_{1\times q^1_U} \\ &= 8n\widehat{\mathcal{L}}^2_{\mathscr{U},\mathscr{V}} + \frac{8}{n^2}\|\widehat{X}\|^2_{1\times q^1_U}.\end{aligned}$$

Here in the second line we used that the Minkowski sum $\Delta_k+(-\Delta_k)$ belongs to the ℓ_1-ball with radius 2 (whence $\mathcal{V}+(-\mathcal{V})$ belongs to the $(\ell_\infty\times\ell_1)$-ball with radius 2); in the third line we used Lemma 4.3 (see Appendix 4.7.3) and the relation on $\mathbb{R}^{n\times k}$:

$$\|\cdot\|_{2\times1}\leqslant\sqrt{n}\|\cdot\|_{\infty\times1};$$

lastly, we used that $p^1_V\ge2$ and that $\|\cdot\|_{p^1_V\times p^2_V}$ is non-increasing in $p^1_V, p^2_V\ge1$. □

Proof of Proposition 4.2. We instantiate Proposition 4.5 with $\|\cdot\|_{\mathscr{U}} = \|\cdot\|_{1\times 1}$ and $\|\cdot\|_{\mathscr{V}} = \|\cdot\|_{2\times 1}$, and observe, using Proposition 4.1, that for $\widehat{X} = [X, -X] \in \mathbb{R}^{n\times 2d}$ it holds

$$\widehat{\mathcal{L}}_{\mathscr{U},\mathscr{V}} = \frac{1}{n}\|\widehat{X}^\top\|_{\infty\times 2} = \frac{1}{n}\|X^\top\|_{\infty\times 2}, \quad \|\widehat{X}\|_{1\times\infty} = \|X\|_{1\times\infty}. \quad \square$$

4.7.6 Proof of Proposition 4.3

We have

$$\begin{aligned}\min_{\substack{p\in\Delta_{2d},\\ P\in(\Delta_k^\top)^{\otimes 2d}}} \mathbb{E}\|\xi_{\widehat{U}}(p,P)\|_{2\times\infty}^2 &= \min_{\substack{p\in\Delta_{2d},\\ P\in(\Delta_k^\top)^{\otimes 2d}}} \sum_{i=1}^{2d} \frac{1}{p_i}\|\widehat{X}(:,i)\|_2^2 \cdot \left[\sum_{l=1}^{k} \frac{1}{P_{il}} \cdot |\widehat{U}(i,l)|^2\right] \\ &= \min_{p\in\Delta_{2d}} \sum_{i=1}^{2d} \frac{1}{p_i}\|\widehat{X}(:,i)\|_2^2 \cdot \|\widehat{U}(i,:)\|_1^2,\end{aligned}$$

where we carried out the internal minimization explicitly, obtaining

$$P_{il}^* = \frac{|\widehat{U}_{il}|}{\|\widehat{U}(i,:)\|_1}.$$

Optimization in p gives:

$$p_i^* = \frac{\|\widehat{X}(:,i)\|_2 \cdot \|\widehat{U}(i,:)\|_1}{\sum\limits_{\imath=1}^{2d} \|\widehat{X}(:,\imath)\|_2 \cdot \|\widehat{U}(\imath,:)\|_1}, \qquad \mathbb{E}\|\xi_U(p^*,P^*)\|_{2\times\infty}^2 = \sum_{i=1}^{2d} \|\widehat{X}(:,i)\|_2 \cdot \|\widehat{U}(i,:)\|_1.$$

Defining

$$\widehat{\mathcal{L}}_{\mathscr{U},\mathscr{V}} = \sup_{\|\widehat{U}\|_{\mathscr{U}}\le 1} \|X\widehat{U}\|_{\mathscr{V}^*}$$

and proceeding as in the proof of Proposition 4.5, we get

$$\begin{aligned}\sigma_{\widehat{\mathcal{U}}}^2(p^*,P^*) &\leqslant \frac{2}{n^2}\sup_{\widehat{U}\in\widehat{\mathcal{U}}} \|\widehat{X}\widehat{U}\|_{2\times\infty}^2 + \frac{2}{n^2}\sup_{\widehat{U}\in\widehat{\mathcal{U}}} \mathbb{E}\left[\|\xi_{\widehat{U}}(p^*,P^*)\|_{2\times\infty}^2\right] \\ &= 2R_{\mathcal{U}}^2\widehat{\mathcal{L}}_{\mathscr{U},\mathscr{V}}^2 + \frac{2}{n^2}\sup_{\widehat{U}\in\widehat{\mathcal{U}}}\left[\sum_{i=1}^{2d} \|\widehat{X}(:,i)\|_2 \cdot \|\widehat{U}(i,:)\|_1\right]^2 \\ &= 2R_{\mathcal{U}}^2\widehat{\mathcal{L}}_{\mathscr{U},\mathscr{V}}^2 + \frac{2R_{\mathcal{U}}^2}{n^2}\sup_{\|\widehat{U}\|_{1\times 1}\leqslant 1}\left[\sum_{i=1}^{2d} \|\widehat{X}(:,i)\|_2 \cdot \|\widehat{U}(i,:)\|_1\right]^2 \\ &= 2R_{\mathcal{U}}^2\widehat{\mathcal{L}}_{\mathscr{U},\mathscr{V}}^2 + \frac{2}{n^2}R_{\mathcal{U}}^2\|\widehat{X}^\top\|_{2\times 1}^2 \\ &= \frac{4}{n^2}R_{\mathcal{U}}^2\|X^\top\|_{2\times 1}^2,\end{aligned}$$

where in the last two transitions we used Lemma 4.2 and Proposition 4.1 (note that $\|\widehat{X}^\top\|_{2\times 1} = \|X^\top\|_{2\times 1}$). Note that the last transition requires that $\|\cdot\|_{\mathscr{U}}$ has ℓ_1-geometry in the classes – otherwise, Lemma 4.2 cannot be applied.

To obtain (q^*, Q^*) we proceed in a similar way:

$$
\begin{aligned}
&\min_{\substack{q\in\Delta_n,\\ Q\in(\Delta_k^\top)^{\otimes n}}} \mathbb{E}\|\eta_{V,Y}(q,Q)\|^2_{\infty\times\infty} \\
&= \min_{\substack{q\in\Delta_n,\\ Q\in(\Delta_k^\top)^{\otimes n}}} \sum_{j=1}^n \frac{1}{q_j}\|\widehat{X}(j,:)\|^2_\infty \cdot \left[\sum_{l=1}^k \frac{1}{Q_{jl}} \cdot |V(j,l) - Y(j,l)|^2\right] \\
&= \min_{q\in\Delta_n} \sum_{j=1}^n \frac{1}{q_j}\|\widehat{X}(j,:)\|^2_\infty \cdot \|V(j,:) - Y(j,:)\|_1^2,
\end{aligned}
$$

which results in

$$
q_j^* = \frac{\|\widehat{X}(j,:)\|_\infty \cdot \|V(j,:) - Y(j,:)\|_1}{\sum_{j=1}^n \|\widehat{X}(j,:)\|_\infty \cdot \|V(j,:) - Y(j,:)\|_1} \qquad Q_{jl}^* = \frac{|V_{jl} - Y_{jl}|}{\|V(j,:) - Y(j,:)\|_1}.
$$

The corresponding variance proxy can then be bounded in the same way as in the proof of Proposition 4.5. □

4.7.7 Correctness of subroutines in Algorithm 1

In this section, we recall the subroutines used in Algorithm 1 – those for performing the lazy updates and tracking the running averages – and demonstrate their correctness.

Procedure 1 UpdatePrimal

```
Require: $\widetilde{U} \in \mathbb{R}^{2d\times k}$, $\alpha, \pi, \eta \in \mathbb{R}^{2d}$, $l \in [k]$, $\gamma$, $\lambda$, $R_1$
$L \equiv \log(2dk)$
for $i = 1$ to $2d$ do
    $\mu_i = \pi_i - \alpha_i \cdot \widetilde{U}(i,l) \cdot (1 - e^{-2\gamma L R_* \eta_i/n})$
$M = \sum_{i=1}^{2d} \mu_i$
$\nu = \min\{e^{-2\gamma L R_* \lambda}, R_*/M\}$
for $i = 1$ to $2d$ do
    $\widetilde{U}(i,l) \leftarrow \widetilde{U}(i,l) \cdot e^{-2\gamma L R_* \eta_i/n}$
    $\alpha_i^+ = \nu \cdot \alpha_i$
    $\pi_i^+ = \nu \cdot \mu_i$
Ensure: $\widetilde{U}, \alpha^+, \pi^+$
```

Primal updates (Procedure 1). To demonstrate the correctness of Procedure 1, we prove the following result:

Lemma 4.4. *Suppose that at t-iteration of Algorithm 1, Procedure 1 was fed with $\widetilde{U} = \widetilde{U}^t, \alpha = \alpha^t, \pi = \pi^t, \eta = \eta^t, l = l^t$ for which one had*

$$\widetilde{U}^t(:,l) \circ \alpha^t = U^t(:,l), \quad \forall l \in [k], \tag{4.7.16}$$

*where U^t is the t-th primal iterate of (**S-MD**) equipped with (**Full-SS**) with the optimal sampling distributions (4.3.8), and η^t was the only non-zero column $\eta_{V^t,Y}(:,l^t)$ of $\eta_{V^t,Y}$. Moreover, suppose also that*

$$\pi^t(\imath) = \|U^t(\imath,:)\|_1 = \sum_{l \in [k]} U^t(\imath,l), \quad \imath \in [2d], \tag{4.7.17}$$

were the correct norms at the t-th step. Then Procedure 1 will output $\widetilde{U}^{t+t}, \alpha^{t+1}, \pi^{t+1}$ such that

$$\widetilde{U}^{t+1}(:,l) \circ \alpha^{t+1} = U^{t+1}(:,l), \; \forall l \in [k]$$

and

$$\pi^{t+1}(\imath) = \sum_{l \in [k]} U^{t+1}(\imath,l), \; \imath \in [2d].$$

Proof. Recall that the matrix $\eta^t = \eta_{V^t,Y}$ produced in (**Full-SS**) has a single non-zero column $\eta^t = \eta^t(:,l^t)$, and according to (**S-MD**), the primal update $U^t \to U^{t+1}$ writes as (cf. (4.2.23)):

$$U_{il}^{t+1} = U_{il}^t \cdot e^{-2\gamma_t R_* L \eta_{V^t,Y}(i,l)/n} \cdot \min\left\{e^{-2\gamma_t R_* L \lambda}, R_*/M\right\},$$

where

$$L := \log(2dk), \quad M_t := \sum_{i=1}^{2d} \sum_{l=1}^{k} U_{il}^t \cdot e^{-2\gamma_t R_* L \eta_{V^t,Y}(i,l)/n},$$

and $\eta_{V^t,Y}$ has a single non-zero column $\eta^t = \eta^t_{V^t,Y}(:,l^t)$. This can be rewritten as

$$U_{il}^{t+1} = \begin{cases} U_{il}^t \cdot \nu \cdot q_i^t, & l = l^t, \\ U_{il}^t \cdot \nu, & l \neq l^t, \end{cases} \tag{4.7.18}$$

where

$$\begin{aligned} q_i^t &= e^{-2\gamma_t L R_* \eta_i^t/n}, \\ \nu &= \min\{e^{-2\gamma_t L R_* \lambda}, R_*/M\}, \\ M &= \sum_{i \in [2d]} U_{il^t}^t \cdot q_i^t + \sum_{i \in [2d]} \sum_{l \in [k] \setminus \{l^t\}} U_{il}^t. \end{aligned}$$

Thus, M and ν can be expressed via $\pi^t(i) = \sum_{l \in [k]} U^t(i,l)$, cf. (4.7.17):

$$M = \sum_{i \in [2d]} \pi_i^t - \underbrace{\alpha_i^t \widetilde{U}_{i,l^t}^t}_{U_{i,l^t}^t}(1 - q_i^t), \tag{4.7.19}$$

where we used the premise (4.7.16). Now we can see that lazy updates of $\widetilde{U}$ can be expressed as

$$\begin{aligned} \alpha_i^{t+1} &= \nu \cdot \alpha_i^t, \\ \widetilde{U}_{i,l^t}^{t+1} &= \widetilde{U}_{i,l^t}^t \cdot q_i^t, \end{aligned} \tag{4.7.20}$$

and the updates for the norms π^{t+1} as

$$\pi_i^{t+1} = \nu^t \Big[\pi_i^t + \alpha_i^t \widetilde{U}_{i,l^t}^t (q_i^t - 1) \Big] \tag{4.7.21}$$

One can immediately verify that this is exactly the update produced in the call of Procedure 1 in line 13 of Algorithm 1. □

Procedure 2 UpdateDual

```
Require: $\widetilde{V}, Y \in \mathbb{R}^{n \times k}$, $\beta, \rho, \xi \in \mathbb{R}^n$, $\ell \in [k]$, $y \in [k]^{\otimes n}$, $\gamma$
$\theta = e^{-2\gamma \log(k)}$
for $j = 1$ to $n$ do
    $\omega_j = e^{2\gamma \log(k) \xi_j}$
    $\varepsilon_j = e^{-2\gamma \log(k) Y(j,\ell)}$
    $\chi_j = 1 - \beta_j \cdot \widetilde{V}(j,\ell) \cdot (1 - \omega_j \cdot \varepsilon_j)$
    if $\ell \neq y_j$ then                                  # not the actual class of j drawn
        $\chi_j \leftarrow \chi_j - \beta_j \cdot \widetilde{V}(j, y_j) \cdot (1 - \theta)$
    $\beta_j^+ = \beta_j / \chi_j$
    $\widetilde{V}(j,\ell) \leftarrow \widetilde{V}(j,\ell) \cdot \omega_j \cdot \varepsilon_j$
    $\widetilde{V}(j,y_j) \leftarrow \widetilde{V}(j,y_j) \cdot \omega_j \cdot \theta$
    $\rho_j^+ = 2 - 2\beta_j^+ \cdot \widetilde{V}(j, y_j)$
Ensure: $\widetilde{V}, \beta^+, \rho^+$
```

Dual updates (Procedure 2). To demonstrate the correctness of Procedure 2, we prove the following result:

Lemma 4.5. *Suppose that at t-iteration of Algorithm 1, Procedure 2 was fed with $\widetilde{V} = \widetilde{V}^t, \beta = \beta^t, \rho = \rho^t, \xi = \xi^t, \ell = \ell^t$, for which one had*

$$\widetilde{V}^t(:, l) \circ \beta^t = V^t(:, l), \quad \forall l \in [k], \tag{4.7.22}$$

*where V^t is the t-th dual iterate of (**S-MD**) equipped with (**Full-SS**) with the optimal sampling distributions (4.3.8), and ξ^t was the only non-zero column $\xi_{U^t}(:, \ell^t)$ of ξ_{U^t}. Moreover, suppose also that*

$$\rho^t(\jmath) = \|V^t(\jmath, :) - Y(\jmath, :)\|_1, \quad \jmath \in [n] \tag{4.7.23}$$

were the correct norms at the t-th step. Then Procedure 2 will output $\widetilde{V}^{t+t}, \beta^{t+1}, \rho^{t+1}$ such that

$$\widetilde{V}^{t+1}(:, l) \circ \beta^{t+1} = V^{t+1}(:, l), \ \forall l \in [k]$$

and

$$\rho^{t+1}(\jmath) = \|V^{t+1}(\jmath,:) - Y(\jmath,:)\|_1, \quad \jmath \in [n].$$

Proof. Recall that the random matrix ξ_{U^t} has a single non-zero column $\xi^t := \xi_{U^t}(:, \ell^t)$, and according to (**S-MD**), the update $V^t \to V^{t+1}$ writes as (cf. (4.2.24)):

$$V_{jl}^{t+1} = V_{jl}^t \cdot \frac{\exp[2\gamma_t \log(k) \cdot (\xi_{U^t}(j,l) - Y(j,l))]}{\sum\limits_{\ell=1}^{k} V_{j\ell}^t \cdot \exp[2\gamma_t \log(k) \cdot (\xi_{U^t}(j,\ell) - Y(j,\ell))]}. \tag{4.7.24}$$

Note that all elements of the matrix $\xi_{U^t} - Y$ in each row j have value 1, except for at most two elements in the columns ℓ^t and y_j, where y_j is the actual label of the j-th training example, that is, the only $l \in [k]$ for which $Y(j,l) = 1$. Recall also that $\sum_{\ell \in [k]} V^t(j,\ell) = 1$ for any $j \in [n]$. Thus, introducing

$$\omega_j = e^{2\gamma_t \log(k) \xi_j^t}, \quad \varepsilon_j = e^{-2\gamma_t \log(k) Y(j,l)}$$

as defined in Procedure 2, we can express the denominator in (4.7.24) as

$$\chi_j = \begin{cases} 1 - \underbrace{\beta_j^t \cdot \widetilde{V}_{j,\ell^t}^t}_{V_{j,\ell^t}^t} \cdot (1 - \omega_j \cdot \varepsilon_j), & \text{if } \ell^t = y_j, \\ 1 - \underbrace{\beta_j^t \cdot \widetilde{V}_{j,\ell^t}^t}_{V_{j,\ell^t}^t} \cdot (1 - \omega_j \cdot \varepsilon_j) - \underbrace{\beta_j^t \cdot \widetilde{V}_{j,y_j}^t}_{V_{j,y_j}^t} \cdot (1 - e^{-2\gamma_t \log(k)}), & \text{if } \ell^t \neq y_j, \end{cases} \tag{4.7.25}$$

where we used the premise (4.7.22). One can verify that this corresponds to the value of χ_j produced by line 7 of Procedure 2. Then, examining the numerator in (4.7.24), we can verify that lines 8–10 guarantee that

$$\widetilde{V}_{j,l}^{t+1} \cdot \beta_j^{t+1} = V_{j,l}^{t+1}, \ \forall l \in [k]$$

holds for the updated values. To verify the second invariant, we combining this result with the premise (4.7.23). This gives

$$\rho_j^{t+1} = 2 - V_{j,y_j}^{t+1} = 2 - 2\beta_j^{t+1} \cdot \widetilde{V}_{j,y_j}^{t+1}, \tag{4.7.26}$$

which indeed corresponds to the update in line 11 of the procedure. □

Correctness of tracking the cumulative sums.

We only consider the primal variables (Procedure 3 and line 21 of Algorithm 1); the complimentary case can be treated analogously. Note that due to the previous two lemmas, at any iteration t of Algorithm 1 Procedure 3 is fed with $l = l^t$, $\widetilde{U} = \widetilde{U}^t$, $\alpha = \alpha^t$ for which it holds $\widetilde{U}^t \alpha^t = U^t$. Now, assume that all previous input values $A^\tau, \tau \leq t$, of variable A, and the current inputs $A_{\text{pr}}^t, U_\Sigma^t$ of variables A_{pr}, U_Σ, satisfy the following:

$$A^{\tau} = \sum_{s=0}^{\tau-1} \alpha^s, \quad \forall \tau \le t, \tag{4.7.27}$$

$$A^t_{\mathrm{pr}}(i,l) = A_i^{\tau^t(i,l)}, \tag{4.7.28}$$

$$U^t_{\Sigma}(i,l) = \sum_{s=0}^{\tau^t(i,l)} U^s(i,l), \tag{4.7.29}$$

where $0 \leqslant \tau^t(i,l) \leqslant t-1$ is the latest moment s, **strictly before** t, when the sampled $l^s \in [k]$ coincided with the given l:

$$\tau^t(i,l) = \arg\max_{s \leqslant t-1} \{s : l^s = l\}. \tag{4.7.30}$$

Let us show that this invariant will be preserved aftet the call of Procedure 3 – in other words, that (4.7.27)–(4.7.30) hold for $t+1$, i.e., for the ouput values U^{t+1}_{Σ}, A^{t+1}_{pr}, A^{t+1} (note that the variables U_{Σ}, A_{pr}, A^{t+1} only changed within Procedure 3, so their output values are also the input values at the next iteration).

Proof. Indeed, it is clear that (4.7.27) will be preserved (cf. line 4 of Procedure 3). To verify (4.7.28), note that $A_{\mathrm{pr}}(i,l)$ only gets updated when $l = l^t$ (cf. line 3), and in this case we will have $\tau^{t+1}(i,l) = t$, and otherwise $\tau^{t+1}(i,l) = \tau^t(i,l)$, cf. (4.7.30).

Thus, it only remains to verify the validity of (4.7.29) after the update. To this end, note that by (4.7.30) we know that the value $\widetilde{U}^s(i,l)$ of the variable $\widetilde{U}(i,l)$ remained constant for $\tau^t(i,l) \le s < t$, and it will not change after the call at t-th iteration unless $l^t = l$, that is, unless $\tau^{t+1}(i,l) = t$. This is exactly when line (2) is invoked, and it ensures (4.7.29) for $t+1$. □

Finally, invoking (4.7.27)–(4.7.30) at $t = T$, we see that line 21 results in the correct final value $\sum_{t=0}^{T} U^t$ of the cumulative sum U_{Σ}. Thus, the correctness of Algorithm 1 is verified.

4.7.8 Additional remarks on Algorithm 1

Removing the $O(dk)$ complexity term. In fact, the extra term $O(dk)$ in the runtime and memory complexities of Algorithm 1 can be easily avoided. To see this, recall that when solving the simplex-constrained CCSPP (4.2.17), we are foremost interested in solving the ℓ_1-constrained CCSPP (4.2.2), and an ε-accurate solution $U = [U_1; U_2] \in \mathbb{R}^{2d \times k}$ to (4.2.17) yields an ε-accurate solution $\widehat{U} = U_1 - U_2 \in \mathbb{R}^{d \times k}$ to (4.2.2). Recall that we initialize Algorithm 1 with $\widetilde{U}^0 = \mathbb{1}_{2d \times k}$ and $\alpha^0 = \mathbb{1}_{2d}$, which corresponds to $\widehat{U}^0 = 0_{2d \times k}$. Moreover, at any iteration we change a single entry of $\widetilde{U}^t$, and scale the whole scaling vector $\alpha^0 = \mathbb{1}_{2d}$ by a constant (in fact, all entries of α^t are always equal to each other; we omitted this fact in the main text to simplify the presentation, since the entries of β generally have different values). Hence, the final candidate solution $\widehat{U}^T = [\bar{U}_1^T - \bar{U}_2^T]$ to the ℓ_1-constrained problem will actually have at most $O(dT)$ non-zero entries that correspond to the entries of $\widetilde{U}$ that were changed in the course of the algorithm. To exploit this, we can modify Algorithm 1 as follows:

— Instead of explicitly initializing and storing the whole matrices $\widetilde{U}$, U_Σ, and $A_{\rm pr}$, we can hard-code the "default" value $\widetilde{U}(i,l) = 1$ (cf. line 2 of Algorithm 1), and use a bit mask to flag the entries $\widetilde{U}(i,l)$ that have already been changed at least once. This mask can be stored as a list `Changed` of pairs (i,l), i.e. in a sparse form.

— When post-processing the cumulative sum U_Σ (see line (21) of Algorithm 1), instead of post-processing all entries of U_Σ, we can only process those in the list `Changed`, and ignore the remaining ones, since the corresponding to them entries in $\widehat{U}^T$ (a candidate solution to (4.2.2)) will have zero values. We can then directly output $\widehat{U}^T$ in a sparse form.

It is clear that such modification of Algorithm 1 results in the replacement of the $O(dk)$ term in runtime complexity with $O(dT)$ (which is always an improvement since $O(d)$ a.o.'s are done anyway in each iteration); moreover, the memory complexity changes from $O(dn + nk + dk)$ to $O(dn + nk + d\min(T,k))$.

Infeasibility of the noisy dual iterates. Note that when we generate an estimate of the primal gradient $X^T(V^t - Y)$ according to (**Full-SS**) or (**Part-SS**), we also obtain an unbiased estimate of the dual iterate V^t, and vice versa. In the setup with vector variables, Juditsky and Nemirovski [2011b] propose to average such noisy iterates instead of the acutal iterates (U^t, V^t) as we do in (**S-MD**). Averaging of noisy iterates is easier to implement since they are sparse (one does not need to track the cumulative sums), and one could show similar guarantees for the primal accuracy of their running averages. However, in the case of the dual variable its noisy counterpart is infeasible (see Juditsky and Nemirovski [2011b, Sec. 2.5.1]); as a result, one loses the guarantee for the duality gap. Hence, we prefer to track the averages of the actual iterates of (**S-MD**) as we do in Algorithm 1.

Chapter 5

Conclusion and Future Work

In this thesis we considered several parameter estimation problems: the extension of the moment-matching technique for non-linear regression, a special type of averaging for generalized linear models and sublinear algorithms for multiclass Fenchel-Young losses. We get theoretical guarantees for the corresponding convex and saddle-point minimization problems.

In Chapter 2 we considered a general non-linear regression model and the dependence on an unknown k-dimensional subspace assumption. Our goal was direct estimation of this unknown k-dimensional space, which is often called the effective dimension reduction or e.d.r. space. We proposed new approaches (SADE and SPHD), combining two existing techniques (sliced inverse regression (SIR) and score function-based estimation). We obtained consistent estimation for $k > 1$ only using the first-order score and proposed explicit approaches to learn the score from data.

It would be interesting to extend our sliced extensions to learning neural networks: indeed, our work focused on the subspace spanned by w and cannot identify individual columns, while Janzamin et al. [2015] used tensor approaches to to estimate columns of w in polynomial time. Another direction is to use smarter score matching techniques to estimate score functions from the data: try to learn only relevant directions. Indeed, our current approach for learning score functions uses a parametric assumption, i.e., that the score is the linear combination of m basis functions and this number grows with the number of observations. This disadvantage can be avoided if we somehow manage to learn scores only in relevant directions.

In Chapter 3 we have explored how averaging procedures in stochastic gradient descent, which are crucial for fast convergence, could be improved by looking at the specifics of probabilistic modeling. Namely, averaging in the moment parameterization can have better properties than averaging in the natural parameterization.

While we have provided some theoretical arguments (asymptotic expansion in the finite-dimensional case, convergence to optimal predictions in the infinite-dimensional case), a detailed theoretical analysis with explicit convergence rates would provide a better understanding of the benefits of averaging predictions.

In Chapter 4 we considered the Fenchel-Young losses and the corresponding saddle-point problems for multi-class classification problems. We develop sublinear algorithm, i.e., with computational complexity of one iteration of $O(d+n+k)$, where d is the number of features, n is the sample size and k it the number of classes.

Though our approach has a sublinear rates for the classical SVM setup, it is interesting to investigate the efficiency estimates for more general geometries and losses (which is actually partly done). Multinomial logit loss or softmax loss is widely used in Natural Language Processing (NLP) and recommendation systems, where n, d and k are on order of millions or even billions (Chelba et al. [2013], Partalas et al. [2015]) and developing an $O(n+d+k)$ approach will be an important development in this area. Even though we provided simple experiments in this chapter, bigger experiments with more expressed convergence rates will be a good illustration of our approach. We are planning to put the code online, due to the full implementation of Algorithm 1 is time-consuming. Another direction of research is to develop the approach with flexible step-sizes, which can be learned from the data as well.

Bibliography

H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer, 1998.

G. Arfken. Divergence. In *Mathematical Methods for Physicists*, chapter 1.7, pages 37–42. Academic Press, Orlando, FL, 1985.

A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

D. Babichev and F. Bach. Constant step size stochastic gradient descent for probabilistic modeling. *Proceedings in Uncertainty in Artificial Intelligence*, pages 219–228, 2018a.

D. Babichev and F. Bach. Slice inverse regression with score functions. *Electronic Journal of Statistics*, 12(1):1507–1543, 2018b.

F. Bach. Sharp analysis of low-rank kernel matrix approximations. In *Conference on Learning Theory*, pages 185–209, 2013.

F. Bach and E. Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Adv. NIPS*, 2011.

F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. *Advances in Neural Information Processing Systems (NIPS)*, 2013.

H. H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2016.

A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

S. Ben-David, N. Eiron, and P. M. Long. On the difficulty of approximately maximizing agreements. *Journal of Computer and System Sciences*, 66(3):496–514, 2003.

D. P. Bertsekas. *Nonlinear programming.* Athena scientific Belmont, 1999.

C.M. Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006.

M. Blondel, A. F. T. Martins, and V. Niculae. Learning classifiers with fenchel-young losses: Generalized entropies, margins, and algorithms. *arXiv preprint arXiv:1805.09717*, 2018.

A. Bordes, S. Ertekin, J. Weston, and L. Bottou. Fast kernel classifiers with online and active learning. *Journal of Machine Learning Research*, 6(Sep):1579–1619, 2005.

J. Borwein and A. S. Lewis. *Convex analysis and nonlinear optimization: theory and examples.* Springer Science & Business Media, 2010.

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. Technical Report 1606.04838, arXiv, 2016.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence.* Oxford University Press, 2013.

D. R. Brillinger. A Generalized Linear Model with 'Gaussian' Regressor Variables. In K.A. Doksum P.J. Bickel and J.L. Hodges, editors, *A Festschrift for Erich L. Lehmann.* Woodsworth International Group, Belmont, California, 1982.

S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

S. Cambanis, S Huang, and G Simons. On the Theory of Elliptically Contoured Distributions. *Journal of Multivariate Analysis*, 11(3):368–385, 1981.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

G. Casella and R. L. Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

C. Chelba, T. Mikolov, M. Schuster, Q. Ge, T. Brants, P. Koehn, and T. Robinson. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*, 2013.

K. L. Clarkson, E. Hazan, and D. P. Woodruff. Sublinear optimization for machine learning. *Journal of the ACM (JACM)*, 59(5):23, 2012.

R. D. Cook. Save: a method for dimension reduction and graphics in regression. *Communications in Statistics - Theory and Methods*, 29:2109–2121, 2000.

R. D. Cook and H. Lee. Dimension Reduction in Binary Response Regression. *Journal of the American Statistical Association*, 94:1187–1200, 1999.

R. D. Cook and S. Weisberg. Discussion of 'Sliced Inverse Regression' by K. C. Li. *Journal of the American Statistical Association*, 86:328–332, 1991.

C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

A. S. Dalalyan, A. Juditsky, and V. Spokoiny. A new algorithm for estimating the effective dimension-reduction subspace. *Journal of Machine Learning Research*, 9: 1647–1678, 2008.

A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pages 1646–1654, 2014.

A. Dieuleveut and F. Bach. Nonparametric stochastic approximation with large step-sizes. *Ann. Statist.*, 44(4):1363–1399, 08 2016.

A. Dieuleveut, A. Durmus, and F. Bach. Bridging the gap between constant step size stochastic gradient descent and markov chains. Technical Report 1707.06386, arXiv, 2017.

W.F. Donoghue, Jr. *Monotone Matrix Functions and Analytic Continuation*. Springer, 1974.

N. Duan and K.-C. Li. Slicing regression: a link-free regression method. *The Annals of Statistics*, 19:505–530, 1991.

J. C Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *COLT*, pages 14–26, 2010.

V. Feldman, V. Guruswami, P. Raghavendra, and Y. Wu. Agnostic learning of monomials by halfspaces is hard. *SIAM Journal on Computing*, 41(6):1558–1590, 2012.

K. Fukumizu, F. R. Bach, and M. I. Jordan. Kernel dimension reduction in regression. *The Annals of Statistics*, 37(4):1871–1905, 2009.

D. Garber and E. Hazan. Approximating semidefinite programs in sublinear time. In *Advances in Neural Information Processing Systems*, pages 1080–1088, 2011.

D. Garber and E. Hazan. Sublinear time algorithms for approximate semidefinite programming. *Mathematical Programming*, 158(1-2):329–361, 2016.

W. R. Gilks, S. Richardson, and D. Spiegelhalter. *Markov chain Monte Carlo in practice*. CRC press, 1995.

A.S. Goldberger. Econometric theory. *Econometric theory.*, 1964.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

M. D. Grigoriadis and L. G. Khachiyan. A sublinear-time randomized approximation algorithm for matrix games. *Operations Research Letters*, 18(2):53–58, 1995.

L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A distribution-free theory of nonparametric regression.* Springer series in statistics. Springer, New York, 2002.

L. P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054, 1982.

E. Hazan, T. Koren, and N. Srebro. Beating SGD: Learning SVMs in sublinear time. In *Advances in Neural Information Processing Systems*, pages 1233–1241, 2011.

Niao He, Anatoli Juditsky, and Arkadi Nemirovski. Mirror prox algorithm for multi-term composite minimization and semi-separable problems. *Computational Optimization and Applications*, 61(2):275–319, 2015.

J. M. Hilbe. *Negative binomial regression.* Cambridge University Press, 2011.

J. Hooper. Simultaneous Equations and Canonical Correlation Theory. *Econometrica*, 27:245–256, 1959.

J. L. Horowitz. *Semiparametric methods in econometrics*, volume 131. Springer Science & Business Media, 2012.

M. Hristache, A. Juditsky, and V. Spokoiny. Direct estimation of the index coefficient in a single index model. *The Annals of Statistics*, 29(3):595–623, 2001.

T. Hsing and R. J. Carroll. An asymptotic theory for sliced inverse regression. *The Annals of Statistics*, 20(2):1040–1061, 1992.

A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*, volume 46. John Wiley & Sons, 2004.

M. Janzamin, H. Sedghi, and A. Anandkumar. Score function features for discriminative learning: Matrix and tensor framework. *arXiv preprint arXiv:1412.2863*, 2014.

M. Janzamin, H. Sedghi, and A. Anandkumar. Generalization Bounds for Neural Networks through Tensor Factorization. *arXiv preprint arXiv:1412.2863*, 2015.

R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

A. Juditsky and A. Nemirovski. First-order methods for nonsmooth convex large-scale optimization, I: General purpose methods. *Optimization for Machine Learning*, pages 121–148, 2011a.

A. Juditsky and A. Nemirovski. First order methods for nonsmooth convex large-scale optimization, ii: utilizing problems structure. *Optimization for Machine Learning*, pages 149–183, 2011b.

A. Juditsky, A. Nemirovski, and C. Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

Johannes HB Kemperman. On the optimum rate of transmitting information. In *Probability and information theory*, pages 126–169. Springer, 1969.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009.

G. M. Korpelevich. Extragradient method for finding saddle points and other problems. *Matekon*, 13(4):35–49, 1977.

J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. ICML*, 2001.

G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1-2):365–397, 2012.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.

L. LeCam. On some asymptotic properties of maximum likelihood estimates and related bayes estimates. *Univ. California Pub. Statist.*, 1:277–330, 1953.

E. L. Lehmann and G. Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

K.-C. Li. Sliced Inverse Regression for Dimensional Reduction. *Journal of the American Statistical Association*, 86:316–327, 1991.

K.-C. Li. On Principal Hessian Directions for Data Visualization and Dimension Reduction: Another Application of Stein's Lemma. *Journal of the American Statistical Association*, 87:1025–1039, 1992.

K.-C. Li and N. Duan. Regression analysis under link violation. *The Annals of Statistics*, 17:1009–1052, 1989.

M. Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

Q. Lin, Z. Zhao, and J. S. Liu. On consistency and sparsity for sliced inverse regression in high dimensions. *The Annals of Statistics*, 46(2):580–610, 2018.

P. McCullagh. Generalized linear models. *European Journal of Operational Research*, 16(3):285–292, 1984.

P. McCullagh and J. A. Nelder. *Generalized linear models*, volume 37. CRC press, 1989.

A. M. McDonald, M. Pontil, and S. Stamos. Spectral k-support norm regularization. In *Advances in Neural Information Processing Systems*, 2014.

S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability.* Springer-Verlag Inc, Berlin; New York, 1993.

J.-J. Moreau. Proximité et dualité dans un espace hilbertien. *Bull. Soc. Math. France*, 93(2):273–299, 1965.

K. P. Murphy. *Machine Learning: A Probabilistic Perspective.* The MIT Press, 2012.

A. Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

A. Nemirovski and U. G. Onn, S.and Rothblum. Accuracy certificates for computational problems with convex structure. *Mathematics of Operations Research*, 35(1): 52–78, 2010.

A. Nemirovsky and D. Yudin. *Problem complexity and method efficiency in optimization.* Chichester, 1983.

Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

Y Nesterov. Gradient methods for minimizing composite objective function, 2007.

Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Y. Nesterov and A. Nemirovski. On first-order algorithms for l_1/nuclear norm minimization. *Acta Numerica*, 22:509–575, 2013.

Y. E. Nesterov. A method for solving the convex programming problem with convergence rate o (1/k^ 2). In *Dokl. Akad. Nauk SSSR*, volume 269, pages 543–547, 1983.

D. Ostrovskii and Z. Harchaoui. Efficient first-order algorithms for adaptive signal denoising. In *Proceedings of the 35th ICML conference*, volume 80, pages 3946–3955, 2018.

B. Palaniappan and F. Bach. Stochastic variance reduction methods for saddle-point problems. In *Advances in Neural Information Processing Systems*, pages 1416–1424, 2016.

I. Partalas, A. Kosmopoulos, N. Baskiotis, T. Artieres, G. Paliouras, E. Gaussier, I. Androutsopoulos, M.-R. Amini, and P. Galinari. LSHTC: A benchmark for large-scale text classification. *arXiv preprint arXiv:1503.08581*, 2015.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning.* MIT Press, 2006.

H. Robbins and S. Monro. ªa stochastic approximation method, º annals math. *Statistics*, 22:400–407, 1951.

R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

R. T. Rockafellar. *Convex analysis.* Princeton university press, 2015.

A. Rudi, L. Carratino, and L. Rosasco. Falkon: An optimal large scale kernel method. In *Advances in Neural Information Processing Systems*, pages 3891–3901, 2017.

M. Schmidt, N. Le Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

B. Scholkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and beyond.* MIT press, 2001.

S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms.* Cambridge university press, 2014.

S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(Feb): 567–599, 2013.

S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter. Pegasos: Primal estimated sub-gradient solver for SVM. *Mathematical programming*, 127(1):3–30, 2011.

A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on stochastic programming: modeling and theory.* SIAM, 2009.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis.* Cambridge university press, 2004.

Z. Shi, X. Zhang, and Y. Yu. Bregman divergence for stochastic variance reduction: saddle-point and adversarial prediction. In *Advances in Neural Information Processing Systems*, pages 6031–6041, 2017.

S. Sra. Fast projections onto mixed-norm balls with applications. *Data Mining and Knowledge Discovery*, 25(2):358–377, 2012.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Injective hilbert space embeddings of probability measures. In *Proc. COLT*, 2008.

C.M. Stein. Estimation of the Mean of a Multivariate Normal Distribution. *The Annals of Statistics*, 9:1135–1151, 1981.

G.W. Stewart and J.-G. Sun. Matrix perturbation theory (computer science and scientific computing), 1990.

T.M. Stoker. Consistent estimation of scaled coefficients. *Econometrica*, 54: 1461–1481, 1986.

A. B. Tsybakov. *Introduction to Nonparametric Estimation.* Springer, 2009.

V. Q. Vu and J. Lei. Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics*, 41(6):2905–2947, 2013.

H. Wang and Y. Xia. On directional regression for dimension reduction. In *J. Amer. Statist. Ass.* Citeseer, 2007.

H. Wang and Y. Xia. Sliced regression for dimension reduction. *Journal of the American Statistical Association*, 103(482):811–821, 2008.

C. K. I. Williams and M. Seeger. Using the nyström method to speed up kernel machines. In *Advances in neural information processing systems*, pages 682–688, 2001.

D. P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

Y. Xia, H. Tong, W. K. Li, and L.-X. Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):363–410, 2002a.

Y. Xia, H. Tong, W.K. Li, and L.-X. Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):363–410, 2002b.

L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11(Oct):2543–2596, 2010.

L. Xiao, A. W. Yu, Q. Lin, and W. Chen. Dscovr: Randomized primal-dual block coordinate algorithms for asynchronous distributed optimization. *arXiv preprint arXiv:1710.05080*, 2017.

S. S. Yang. General distribution theory of the concomitants of order statistics. *The Annals of Statistics*, 5:996–1002, 1977.

Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

Y.-L. Yu. The strong convexity of von Neumann's entropy. *Unpublished note*, June 2013. URL `http://www.cs.cmu.edu/~yaoliang/mynotes/sc.pdf`.

M. Yuan. On the identifiability of additive index models. *Statistica Sinica*, 21(4): 1901–1911, 2011.

L.-X. Zhu and K. W. Ng. Asymptotics of sliced inverse regression. *Statistica Sinica*, 5:727–736, 1995.

List of Figures

1-1 Graphical representation of classical loss functions. 10

1-2 Graphical representation of Lipschitz smoothness and strong convexity. 12

1-3 Graphical representation of projected gradient descent: firstly we do gradient step and then project onto $\mathcal{X}$. 13

2-1 Graphical explanation of SADE: firstly we divide y into H slices, let $\hat{p}_h$ be a proportion of y_i in slice I_h. This is an empirical estimator of $\mathbb{P}(y \in I_h)$. Then we evaluate the empirical estimator $(\hat{\mathcal{S}}_1)_h$ of $\mathbb{E}(\mathcal{S}_1(x)|y \in I_h)$. Finally, we evaluate weighted covariance matrix and find the k largest eigenvectors. 33

2-2 Contour lines of Gaussian mixture pdf in 2D case. 42

2-3 Mean and standard deviation of $R^2(\mathcal{E}, \hat{\mathcal{E}})$ divided by 10 for the rational function (2.5.2). 42

2-4 Mean and standard deviation of $R^2(\mathcal{E}, \hat{\mathcal{E}})$ divided by $\sqrt{10}$ for the rational function (2.5.3) with $\sigma = 1/4$. 43

2-5 Mean and standard deviation of $R^2(\mathcal{E}, \hat{\mathcal{E}})$ divided by $\sqrt{10}$ for the the classification problem (2.5.4). 43

2-6 Mean and standard deviation of $R^2(\mathcal{E}, \hat{\mathcal{E}})$ divided by $\sqrt{10}$ for the model (2.5.2) with $\sigma = 1/4$. 44

2-7 Square trace error $R(\mathcal{E}, \hat{\mathcal{E}})$ for different sample sizes. 45

2-8 Quadratic model, $d = 10$, $n = 1000$. 45

2-9 Rational model, $d = 10$, $n = 1000$. 45

2-10 Rational model, $d = 20$, $n = 2000$. 45

2-11 Rational model, $d = 50$, $n = 5000$. 46

2-12 Increasing dimensions, with rational model 2.5.3, $n = 10000$, $k = 10$, $d = 20$ to 150. 46

3-1 Convergence of iterates θ_n and averaged iterates $\bar{\theta}_n$ to the mean $\bar{\theta}_\gamma$ under the stationary distribution π_γ. 59

3-2 Graphical representation of reparametrization: firstly we expand the class of functions, replacing parameter θ with function $\eta(\cdot) = \Phi^\top(\cdot)\theta$ and then we do one more reparametrization: $\mu(\cdot) = a'(\eta(\cdot))$. Best linear prediction μ_* is constructed using θ_* and the global minimizer of $\mathcal{G}$ is μ_{**}. Model is well-specified if and only if $\mu_* = \mu_{**}$. 63

3-3 Visualisation of the space of parameters and its transformation to the space of predictions. Averaging estimators in red vs averaging predictions in green. μ_* is optimal linear predictor and μ_{**} is the global optimum. 66
3-4 Visualisation of the space of parameters and its transformation to the space of predictions. Averaging estimators in red vs averaging predictions in green. Global optimizer coincides with the best linear. 68
3-5 Synthetic data for linear model $\eta_\theta(x) = \theta^\top x$ and $\eta_{**}(x) = \sin x_1 + \sin x_2$. Excess prediction performance vs. number of iterations (both in log-scale). 71
3-6 Synthetic data for linear model $\eta_\theta(x) = \theta^\top x$ and $\eta_{**}(x) = x_1^3 + x_2^3$. Excess prediction performance vs. number of iterations (both in log-scale). 71
3-7 Synthetic data for Laplacian kernel for $\eta_{**}(x) = \frac{5}{5+x^\top x}$. Excess prediction performance vs. number of iterations (both in log-scale). 72
3-8 MiniBooNE dataset, dimension $d = 50$, linear model. Excess prediction performance vs. number of iterations (both in log-scale). 73
3-9 MiniBooNE dataset, dimension $d = 50$, kernel approach, column sampling $m = 200$. Excess prediction performance vs. number of iterations (both in log-scale). 73
3-10 CoverType dataset, dimension $d = 54$, linear model. Excess prediction performance vs. number of iterations (both in log-scale). 74
3-11 CoverType dataset, dimension $d = 54$, kernel approach, column sampling $m = 200$. Excess prediction performance vs. number of iterations (both in log-scale). 74

4-1 Depiction of the Full Sampling Scheme (**Full-SS**). 99
4-2 Primal accuracy and duality gap (when available) for Algorithm 1, stochastic subgradient method (SSM), and Mirror Prox (MP) with exact gradients, on a synthetic data benchmark. 108

List of Tables

2.1 Comparison of different methods using score functions. 31

4.1 Comparison of the potential differences for the norms corresponding to (4.4.1). 106

4.2 Runtime (in seconds) of Algorithm 1 on synthetic data. 107

RÉSUMÉ

Dans cette thèse, nous examinons plusieurs aspects de l'estimation des paramètres pour les statistiques et les techniques d'apprentissage automatique, aussi que les méthodes d'optimisation applicables à ces problèmes. Le but de l'estimation des paramètres est de trouver les paramètres cachés inconnus qui régissent les données, par exemple les paramètres dont la densité de probabilité est inconnue. La construction d'estimateurs par le biais de problèmes d'optimisation n'est qu'une partie du problème, trouver la valeur optimale du paramètre est souvent un problème d'optimisation qui doit être résolu, en utilisant diverses techniques. Ces problèmes d'optimisation sont souvent convexes pour une large classe de problèmes, et nous pouvons exploiter leur structure pour obtenir des taux de convergence rapides. La première contribution principale de la thèse est de développer des techniques d'appariement de moments pour des problèmes de régression non linéaire multi-index. Nous considérons le problème classique de régression non linéaire, qui est irréalisable dans des dimensions élevées en raison de la malédiction de la dimensionnalité. Nous combinons deux techniques existantes : ADE et SIR pour développer la méthode hybride sans certain des aspects faibles de ses parents. Dans la deuxième contribution principale, nous utilisons un type particulier de calcul de la moyenne pour la descente stochastique du gradient. Nous considérons les familles exponentielles conditionnelles (comme la régression logistique), où l'objectif est de trouver la valeur inconnue du paramètre. Nous proposons le calcul de la moyenne des paramètres de moments, que nous appelons fonctions de prédiction. Pour les modèles à dimensions finies, ce type de calcul de la moyenne peut entraîner une erreur négative, c'est-à-dire que cette approche nous fournit un estimateur meilleur que tout estimateur linéaire ne peut jamais le faire. La troisième contribution principale de cette thèse porte sur les pertes de Fenchel-Young. Nous considérons des classificateurs linéaires multi-classes avec les pertes d'un certain type, de sorte que leur double conjugué a un produit direct de simplices comme support. La formulation convexe-concave à point-selle correspondante a une forme spéciale avec un terme de matrice bilinéaire et les approches classiques souffrent de la multiplication des matrices qui prend beaucoup de temps. Nous montrons que pour les pertes SVM multi-classes avec des techniques d'échantillonnage efficaces, notre approche a une complexité d'itération sous-linéaire, c'est-à-dire que nous devons payer seulement trois fois $O(n+d+k)$: pour le nombre de classes k, le nombre de caractéristiques d et le nombre d'échantillons n, alors que toutes les techniques existantes sont plus complexes.

MOTS CLÉS

Estimation des paramètres, méthode des moments, SGD à pas constant, famille exponentielle conditionnelle, fonction objectif du Fenchel-Young, descente en miroir.

ABSTRACT

In this thesis we consider several aspects of parameter estimation for statistics and machine learning and optimization techniques applicable to these problems. The goal of parameter estimation is to find the unknown hidden parameters, which govern the data, for example parameters of an unknown probability density. The construction of estimators through optimization problems is only one side of the coin, finding the optimal value of the parameter often is an optimization problem that needs to be solved, using various optimization techniques. Hopefully these optimization problems are convex for a wide class of problems, and we can exploit their structure to get fast convergence rates. The first main contribution of the thesis is to develop moment-matching techniques for multi-index non-linear regression problems. We consider the classical non-linear regression problem, which is unfeasible in high dimensions due to the curse of dimensionality. We combine two existing techniques: ADE and SIR to develop the hybrid method without some of the weak sides of its parents. In the second main contribution we use a special type of averaging for stochastic gradient descent. We consider conditional exponential families (such as logistic regression), where the goal is to find the unknown value of the parameter. Classical approaches, such as SGD with constant step-size are known to converge only to some neighborhood of the optimal value of the parameter, even with averaging. We propose the averaging of moment parameters, which we call prediction functions. For finite-dimensional models this type of averaging can lead to negative error, i.e., this approach provides us with the estimator better than any linear estimator can ever achieve. The third main contribution of this thesis deals with Fenchel-Young losses. We consider multi-class linear classifiers with the losses of a certain type, such that their dual conjugate has a direct product of simplices as a support. We show, that for multi-class SVM losses with smart matrix-multiplication sampling techniques, our approach has an iteration complexity which is sublinear, i.e., we need to pay only trice $O(n+d+k)$: for number of classes k, number of features d and number of samples n, whereas all existing techniques have higher complexity.

KEYWORDS

Parameter estimation, method of moments, constant step-size SGD, conditional exponential family, Fenchel-Young loss, mirror descent.