Università degli Studi di Padova

Dipartimento di Ingegneria dell'Informazione

Master Thesis in Control Systems Engineering

High-Dimensional Analysis of f-divergence Distributionally Regularized M-estimation

Candidate

Riccardo Cescon

Student ID 2026247

Supervisor

Prof. Augusto Ferrante

University of Padova

Academic Year
2021/2022

Abstract

In recent years Distributionally Robust Optimization (DRO) has raised to the status of one of the most promising tools for robust estimation. This because it shares some nice properties such as good out-of-sample performances and well-understood regularization effects. The estimator we obtain within this framework is computed by minimizing the worst-case expected loss under all distributions that are close, in a f-divergence sense, to the empirical distribution which relies just on historical data. In this thesis we propose a new approach to compute an unknown parameter vector using data coming from linear and noisy measurements. In doing so, we will use a slight modification of DRO which amounts to a distributional regularization. The ultimate goal will be to characterize the estimation error which is in general a challenging task but yet very important. Our analysis is performed under the modern assumption of high-dimensional regime in which both the number of measurements and parameters are very large, keeping a fixed proportion while going to infinity which encodes the under/over-parametrization of the problem. Our contribution can be summarized as follows. We show that the estimation error can be recovered solving a scalar minmax convex-concave problem which consists of just four variables. This enables a fast computational method to find the best regularization parameter λ in the bias-variance tradeoff.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis is mainly focused on the problem of estimating an unknown parameter $\theta_0 \in \Theta \subseteq \mathbb{R}^d$ which describes the relationship between two random variables X, Y via the linear model given by $Y = X^T\theta_0 + Z$ where X takes values in $\mathbb{R}^d$, Y takes values in $\mathbb{R}$ and they are distributed accordingly to a nominal and generally unknown distribution, say $\mathbb{P}_\star \in \mathcal{P}(\mathbb{R}^{d+1})$ while Z represents instead some measurements noise distributed accordingly to the unknown distribution $\mathbb{P}_Z$. This problem of estimating parameters by linear functions of measurements has been studied in depth ever since Gauss introduced the theory of least squares.

A possible and simple approach to find an estimate of the true parameter θ_0 is to solve the stochastic optimization problem given as

$$\min_{\theta\in\Theta} \ \mathbb{E}_{\mathbb{P}_\star}[\ell(Y - X^T\theta)] \tag{1.1}$$

where ℓ is a loss function and $\mathbb{P}_\star$ is the true distribution. The most common losses adopted in statistical learning are, among the others, squared loss function, i.e. $\ell(\cdot) = (\cdot)^2$ and absolute loss function, i.e. $\ell(\cdot) = |\cdot|$. For the readers wishing to deepen their knowledge a comprehensive survey of loss functions can be found in [30].

However, as already pointed out, we usually do not know the underlying true distribution $\mathbb{P}_\star$. In almost every data-driven application we have a sample composed by n i.i.d noisy datapoints $\{(x_i, y_i)\}_{i=1}^n$ from $\mathbb{P}_\star$. This is a well-studied problem in which we would like to estimate an unknown vector θ_0 from n noisy linear measurements of the type $y_i = x_i^T\theta_0 + z_i$ where x_i is the vector of known measurements while z_i a noise realization. To solve this problem we can consider an approximation of the true distribution $\mathbb{P}_\star$ constructed giving equal weight to all the datapoints, i.e. $\mathbb{P}_\star \simeq \hat{\mathbb{P}}_n := n^{-1}\sum_{i=1}^n \delta(x_i, y_i)$. Doing this, (1.1) becomes the so-called *M-estimation* problem given by

$$\min_{\theta\in\Theta} \ \frac{1}{n}\sum_{i=0}^{n} \ell(y_i - x_i^T\theta). \tag{1.2}$$

Notice that this is an instance of the *Empirical Risk Minimization* (ERM) [29], whose idea is to minimize the number of training errors (empirical risk). In practice it chooses the parameter θ considering the one that when plugged into the model returns the least discrepancy or error compared to the observed measurements. We refer to the optimal value of (1.2) as $\hat{\theta}_{ERM}$.

In the context of statistical inference we can define two quantities describing how good an estimator is. The concept of "*bias*" of an estimator or model represents its ability to approximate well the true parameter θ_0. It can be formally seen as the difference between θ_0 and the expected value of the estimator. The concept of "*variance*" or "*estimation error*" instead represents the second order moment of the estimator and it appears because the empirical risk is only an estimate of the true error.

In practice, in a high variance scenario we are "*overfitting*" and this is mainly a consequence of the use of a complex model or equivalently of the fact that we are relying solely on the available measurements which are just a sample of the true distribution. Instead, in a high bias setup we are "*underfitting*" meaning that our estimator or model cannot describe well the measurements observed because on average it does not retrieve the true θ_0.

For example, it can be shown that the estimator we obtain from (1.2) in the case of squared loss is unbiased but since it finds the best estimator considering just n i.i.d. samples it might lead to high variance. Therefore, a general rule already adopted in the literature to lower the variance is to add an additional term other than the one accounting for the estimation loss, called regularization term.

The idea of regularization was first introduced by Tikhonov in the context of solving integral equations numerically [28], but it is nowadays used widely for solving ill-posed problems (for example the least square problem when X^TX, with X matrix of measurements, is not invertible) and, as pointed out before, to prevent overfitting in machine learning applications. The regularized M-estimator that we get assumes then the following form

$$\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=0}^{n} \ell(y_i - x_i^T \theta) + \lambda f(\theta) \tag{1.3}$$

where the parameter $\lambda > 0$ constitutes a trade-off between the standard ERM seen previously and the prior structural knowledge brought by f. The optimal solution of (1.3) is denoted by $\hat{\theta}_{REG}$. This trade-off is also known as "*Bias-Variance*" trade-off where the more we increase the value of λ the more we have a biased estimator but with lowered variance while if λ is taken small we obtain an unbiased estimator at the price of having higher variance.

Famous and well-established examples of regularizers are given by Ridge regularization in which the function f is the standard Euclidean norm and the Lasso regularization which instead considers the ℓ_1-norm. However, despite the known ability of these functions to promote particular known-a-priori structures of the parameter to estimate, specifically sparsity for Lasso and small norm in the case of Ridge, they are not able to optimally trade between bias and variance.

Therefore, one can consider the following optimization problem which automatically trades between the two quantities

$$\min_{\theta \in \Theta} \ \mathbb{E}_{\hat{\mathbb{P}}_n}[\ell(Y - X^T\theta)] + \lambda \sqrt{\frac{Var_{\hat{\mathbb{P}}_n}(\ell(Y - X^T\theta))}{n}}. \tag{1.4}$$

In this minimization we can identify the bias with the empirical risk $\mathbb{E}_{\hat{\mathbb{P}}_n}[\ell(Y - X^T\theta)]$ while the variance arises from the second term in the equation, [19]. Notice that this is an instance of the problem in (1.3) where now the regularization term is function of the estimator's variance.

The latter is an important problem which appealed the statistic community because the estimator minimizing the variance regularized risk is optimal in the variance-bias trade off. The main disadvantage is that it is usually not a convex problem even when the loss is convex, thus very difficult to solve.

Moving forward, in nowadays applications the focus has shifted towards the *high dimensional regime*, where both the number of measurements n and the dimension d of the parameter θ_0 are very large, i.e. $d, n \to \infty$, with $n/d = \delta \in (0, +\infty)$. In practice, δ represents the under-parametrization or over-parametrization of the problem, respectively if $\delta \in (1, \infty)$ or $\delta \in (0, 1]$. This is something new compared with standard statistical problems. In general the parameter's space has a relatively low and fixed dimension while we require the number of samples n to be large in order to have possible consistency guarantees, i.e. the estimator converges in probability to the true parameter θ_0. In this work instead we will consider the high-dimensional regime where also the number d of variables to estimate is large. Of particular interest is also the *compressed measurements* scenario (or over-parametrization case), where $d > n$. This case has many challenges because as long as the parameter θ_0 is not structurally constrained the problem in (1.2) is generally ill-posed because there are more variables to estimate than the number of measurements available. This is another reason to adopt the regularized formulation in (1.3) since the structural properties of θ_0 can be encoded in f or one can optimize the variance of the estimator using the formulation in (1.4).

Examples of high-dimensional statistics are ubiquitous throughout science: astronomical projects such as the Large Synoptic Survey Telescope produce terabytes of data in a single evening; each sample is a high-resolution image, with several hundred megapixels, so that $d \gg 10^8$. Financial data are also of high-dimensional nature with lot of financial instruments tracked at a fine time interval for high frequency trading, [20]. Other examples are Magnetic Resonance Imaging (MRI) in medicine which is an essential medical imaging tool inherently slow in the acquiring data process. In here compressed sensing is applied because it offers significant scan time reductions, with benefits for patients and health care economics, [18]. Finally, also hyper-spectral imaging in ecology [8] leads to high-dimensional data sets.

Merging this high dimensional scenario with the non-convexity of problem (1.4) yields to computationally intractable problems, which has limited the applicability of procedures that minimize the variance-corrected empirical risk. On top of this, as pointed out in [27] it is still unknown how to optimally tune the regularization parameter λ. Indeed, it is impossible to perform some sort of cross validation to find the best parameter λ given the huge number of parameters involved in the high-dimensions scenario.

The work [19] has served has inspiration for our procedure because they showed that solving a particular instance of f-divergence DRO we obtain an estimator that near optimally trades between bias and variance. Moreover, they proved that this problem is now convex if the loss is convex in θ which makes the solution easier to find using standard learning algorithms.

The general idea of this line of work called *Distributionally Robust Optimization* (DRO) is to create a "ball" of radius $\epsilon > 0$ in the probability space around the aforementioned empirical distribution $\hat{\mathbb{P}}_n$. Then, the problem solves a minimization problem choosing the parameter θ which best performs under the least favorable distribution in this ball. This intuitive reasoning can be formalized mathematically by the following minmax formulation

$$\min_{\theta \in \Theta} \sup_{\mathbb{Q} \in \mathbb{B}_\epsilon(\hat{\mathbb{P}}_n)} \mathbb{E}_{\mathbb{Q}}[\ell(Y - X^T\theta)]. \tag{1.5}$$

An optimal solution of (1.5) is denoted in the thesis as $\hat{\theta}_{DRO}$. Robustness against perturbations is brought by the inner maximization while $\mathbb{B}_\epsilon(\hat{\mathbb{P}}_n)$ represents the so-called *ambiguity set*. To have a better understanding, this set represents a family of probability distributions which are ϵ-close to the empirical one and may account for the possible adversarial attacks. Instead of solving the standard ERM in (1.2) which considers equal weights for each measurement loss, in (1.5) the expectation is taken accordingly to the worst case distribution present in the ambiguity set. Notice that such set should contain the unknown true distribution with a certain level of confidence and should not be taken too large to avoid the risk of being too conservative.

A question may arise spontaneously, how can we quantify the distance between distributions in order to create the ambiguity set? A possible way, adopted also in [19] to obtain the approximated variance-regularized risk, is to use f-divergences and in particular the χ^2-divergence.

The concept of f-divergence was first introduce in [21] to measure the information that a random variable ξ carries after having observed the event E which is in someway related to ξ. Indeed, the original idea was to quantify the difference between the original (unconditional) distribution of ξ, say $\mathbb{P}$, with the conditional distribution of ξ under the condition that the event E has taken place, say $\mathbb{Q}$, to measure the amount of information contained in the observation of event E. This family of divergences has been also studied by [1] and most importantly by [10].

The f-divergence between two probability measures $\mathbb{P}$ and $\mathbb{Q}$ is defined as

$$D_f(\mathbb{P}\|\mathbb{Q}) \doteq \int f\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{Q} \tag{1.6}$$

where $f(t) : \mathbb{R} \to \overline{\mathbb{R}}_+$ is a convex function for $t \geq 0$ satisfying $f(1) = 0$ and defined as $+\infty$ for $t < 0$. Notice that, f-divergences are not metrics because, in general, they are not symmetric ($D_f(\mathbb{P}\|\mathbb{Q}) \neq D_f(\mathbb{Q}\|\mathbb{P})$) and they do not satisfy the triangle inequality, but they are still useful when quantifying the distance between two probability measures. The ambiguity set constructed adopting a f-divergence takes the form

$$\mathbb{B}_\epsilon(\hat{\mathbb{P}}_n) = \{\mathbb{Q} \ll \hat{\mathbb{P}}_n \mid D_f(\mathbb{Q}\|\hat{\mathbb{P}}_n) \leq \epsilon\}. \tag{1.7}$$

In the literature there are already some tractable reformulations of the DRO problem in (1.5) when the ambiguity set is constructed using f-divergences as distance both in continuous and discrete case, ([4], [12], [24]).

However, in this work we will consider a slightly modification of the DRO problem which at the same time maintains most of the intuitions given so far about the problem and it is also relatively simpler to analyze. What we consider is the *Distributionally Regularized Optimization* problem which can be defined mathematically as follows

$$\min_{\theta\in\Theta} \sup_{\mathbb{Q}\ll\hat{\mathbb{P}}_n} \mathbb{E}_{\mathbb{Q}}[\ell(Y - X^T\theta)] - \lambda D_f(\mathbb{Q}\|\hat{\mathbb{P}}_n). \tag{1.8}$$

We will refer to an optimal solution of (1.8) as $\hat{\theta}_{DRE}$.

As last main ingredient of this thesis we introduce what is called *Convex Gaussian Minmax Theorem* (CGMT), directly related to the Gaussian inequality proposed in [15]. Taking inspiration from [27] and its line of work that uses CGMT to solve convex-regularized M-estimators we will use this tool to show that $\|\hat{\theta}_{DRE} - \theta_0\|^2/d$ can be retrieved solving a deterministic program involving only scalar variables. This is relevant because in this way we are able to find the estimation error and optimally tune the regularization parameter λ by simply solving a convex-concave problem with low dimensionality despite the multidimensional and stochastic nature of the variables involved in the original DRE problem.

To summarize, our contributions brought by this work are:

- Development of a strong dual reformulation of the problem (1.8) and characterization of the estimation error in terms of a scalar deterministic program providing all the steps and proofs to reach such goal.
- Numerical simulations to guarantee the theoretical result.

The rest of this thesis is organized as follows: Chapter 2 is entirely devoted to the description of the CGMT tool and the f-divergence family. In here we present also the f-divergence distributionally regularized problem and its dual formulation. Chapter 3 instead contains the step-by-step solution of the problem that leads to the scalar optimization program returning the estimation error. Chapter 4 instead is devoted to numerical examples.

Chapter 2

Background Material

This chapter is devoted to a self contained and short survey of the basic concepts that are at the core of this work. In particular, we divide it into three sections. The first one provide a brief introduction to the concept of CGMT which is the main theorem upon which we leverage to obtain some theoretical guarantees of our work. In the second section we introduce the reader to the family of distances called f-divergences. Finally, in the third section we present the Distributionally Regularized problem using f-divergences and also a dual reformulation of it. We mostly omit the proofs of the result we present because they are outside of the scope of this thesis, nevertheless for the most interested readers we will provide detailed references.

2.1 Notation

Throughout this thesis we denote sets with upper case calligraphic letters, e.g. $\mathcal{X}$. Probability distributions will be denoted as $\mathbb{P}$ and $\mathbb{Q}$. $\|\cdot\|$ will denote the standard Euclidean (ℓ_2) norm. For a convex function $f : \mathbb{R} \to \mathbb{R}$ we denote with $\partial f(x)$ the subdifferential of f at the point x and with $f'_+(x)$ the quantity $\sup_{s \in \partial f(x)} |s|$. We use the notation $X_n \xrightarrow{P} c$ to denote that the sequence of random variables $\{X_n\}_{n \in \mathbb{N}}$ converges in probability to the constant c. The expectation of a random variable $\xi \sim \mathbb{P}$ will be denoted with $\mathbb{E}_{\mathbb{P}}[\xi]$ while the variance with $Var(\xi)$. We finally denote with $\mathbb{R}_+$ the set of non-negative real numbers while $\overline{\mathbb{R}}_+$ denotes the same set union with $\{+\infty\}$.

2.2 CGMT

The main technical result that we will adopt in this thesis is the asymptotic version of the Convex Gaussian Minmax Theorem (CGMT) which is presented in Prop. 2.2.1. An asymptotic variant is necessary because we will deal with the high-dimensional regime, namely when both the number of measurements n and the parameter's dimension d are very large. In the following we are going to present a self-contained exposition of this theorem.

The theorem can be seen as a tight version of the Gaussian Minmax Theorem (GMT) proved by [15]. Consider indeed the task of analyzing this problem called *Primary Optimization* (PO) problem

$$\Phi(\mathbf{X}) \doteq \min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}} \max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}} \mathbf{u}^T\mathbf{X}\mathbf{w} + \psi(\mathbf{w},\mathbf{u}) \tag{2.1}$$

where $\mathbf{X} \in \mathbb{R}^{n\times d}$ is the known random measurement matrix having entries i.i.d. standard normal, $\mathcal{S}_{\mathbf{w}} \subseteq \mathbb{R}^d, \mathcal{S}_{\mathbf{u}} \subseteq \mathbb{R}^n$ are compact sets and $\psi(\mathbf{w},\mathbf{u})$ is a continuous function defined on $\mathcal{S}_{\mathbf{w}} \times \mathcal{S}_{\mathbf{u}}$. We refer to an optimal solution of (2.1) as $\mathbf{w}_\Phi(\mathbf{X})$.

Since it is difficult to analyze the (PO) problem due to the bilinear form involving the random matrix $\mathbf{X}$, [15] developed a simpler problem called *Auxiliary Optimization* (AO) and used it to infer properties of the much challenging (PO) problem based on some Gaussian inequalities. The (AO) problem is actually simpler to analyze because the bilinear term $\mathbf{u}^T\mathbf{X}\mathbf{w}$ is split into two separate terms, $\|\mathbf{w}\|\mathbf{g}^T\mathbf{u}$ and $\|\mathbf{u}\|\mathbf{h}^T\mathbf{w}$ where $\mathbf{h} \in \mathbb{R}^d$ and $\mathbf{g} \in \mathbb{R}^n$ are random vectors whose entries are again i.i.d. standard Gaussian. Therefore the (AO) problem is given by

$$\phi(\mathbf{g},\mathbf{h}) \doteq \min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}} \max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}} \|\mathbf{w}\|\mathbf{g}^T\mathbf{u} + \|\mathbf{u}\|\mathbf{h}^T\mathbf{w} + \psi(\mathbf{w},\mathbf{u}). \tag{2.2}$$

We denote by $\mathbf{w}_\phi(\mathbf{g},\mathbf{h})$ an optimal solution of (2.2).

CGMT builds on top of [15] and extends its results proving the following inequality between optimal values of (PO) and (AO)

$$\mathbb{P}(\Phi(\mathbf{X}) < c) \leq 2\mathbb{P}(\phi(\mathbf{g},\mathbf{h}) \leq c) \quad \forall c \in \mathbb{R}. \tag{2.3}$$

This expression has to be intended as follows. Whenever $\mathbb{P}(\phi(\mathbf{g},\mathbf{h}) \leq c)$ is close to zero, namely c is an high-probability lower bound for $\phi(\mathbf{g},\mathbf{h})$ because the probability of $\phi(\mathbf{g},\mathbf{h})$ being greater than c is high, then we can assert the same for $\Phi(\mathbf{X})$. This result is not something new because, as already said, it is a simple extension of the work proposed by [15]. The novelty brought by the CGMT is in a tighter bound. Indeed, up so far (2.3) says that $\Phi(\mathbf{X})$ has high probability to be greater than c but does not say anything on how much greater $\Phi(\mathbf{X})$ is allowed to be. We can potentially have an unbounded by above cost.

To get better the idea let us think about the following example. Let us assume that $\phi(\mathbf{g},\mathbf{h})$ is concentrated around some constant μ, namely for every t > 0 the events

$$\{\phi(\mathbf{g},\mathbf{h}) \leq \mu - t\} \quad \text{and} \quad \{\phi(\mathbf{g},\mathbf{h}) \geq \mu + t\}$$

each occur with small probability. Up to now, the theorem ensures that $\mu - t$ is also a high-probability lower bound on $\Phi(\mathbf{X})$, i.e. the event $\{\Phi(\mathbf{X}) \leq \mu - t\}$ also occurs with small probability, but we do not know nothing about the probability of the event $\{\Phi(\mathbf{X}) \geq \mu+t\}$. It might be that $\Phi(\mathbf{X})$ is arbitrarily greater than μ with high probability.

In order to have the desired tight bound we need to add other requirements not present in the original GMT. These are:

(i) Sets $\mathcal{S}_{\mathbf{w}}, \mathcal{S}_{\mathbf{u}}$ are now convex and compact.

(ii) The function ψ is now assumed to be convex-concave on $\mathcal{S}_{\mathbf{w}} \times \mathcal{S}_{\mathbf{u}}$.

Given these new hypotheses we can formulate the high-probability upper bound for $\Phi(\mathbf{X})$.

$$\mathbb{P}(\Phi(\mathbf{X}) > c) \leq 2\mathbb{P}(\phi(\mathbf{g}, \mathbf{h}) \geq c), \quad \forall c \in \mathbb{R}. \tag{2.4}$$

Again, this has to be interpreted as whenever c is a high-probability upper bound for $\phi(\mathbf{g}, \mathbf{h})$ (the probability of $\phi(\mathbf{g}, \mathbf{h})$ being greater than c is small) then the same holds true for $\Phi(\mathbf{X})$.

Now, if we some up results (2.3) and (2.4) we obtain the tight inequality of CGMT which states, under the additional convexity assumption, that the (AO) problem tightly bounds the optimal value of the (PO) problem, meaning that for every $\mu \in \mathbb{R}$ and $t > 0$

$$\mathbb{P}(|\Phi(\mathbf{X}) - \mu| > t) \leq 2\mathbb{P}(|\phi(\mathbf{g}, \mathbf{h}) - \mu| \geq t). \tag{2.5}$$

This relationship has been widely used in [27] to derive the asymptotic properties of the optimal (PO) problem solution based uniquely on the (AO) problem.

In the following we will explain how these concentration inequalities may be helpful in our analysis. Remember that our goal is to quantify the error norm $\|\hat{\theta}_{DRE} - \theta_0\|^2/d$, therefore it is natural to make the following change of variables

$$\mathbf{w} \doteq \frac{\theta - \theta_0}{\sqrt{d}}. \tag{2.6}$$

We would like to show that the norm of the DRE problem optimal error, i.e. $\|\hat{\mathbf{w}}_{DRE}\|$, converges in probability, as $d, n \to +\infty$, to a value $\alpha_\star$ which can be retrieved by a convex-concave deterministic program involving few scalar variables. CGMT is essential in order to show this convergence as explained in the following.

First we will show that we can bring the DRE problem in (1.8) into a (PO) problem where the random matrix $\mathbf{X}$ is built using the measurement vectors x_i. Then, the step from (PO) to (AO) is quite straightforward. Assume now that we are able to constraint both variables $\mathbf{w}$ and $\mathbf{u}$ to compact sets (we will show this later). In order to arrive at the scalar optimization which determines $\alpha_\star$ we need to pass through the so-called Modified (AO) problem given by

$$\phi(\mathbf{g}, \mathbf{h})' \doteq \max_{0 \leq \beta \leq K_\beta} \min_{\mathbf{w} \in \mathcal{S}_{\mathbf{w}}} \max_{\|\mathbf{u}\| = \beta} \|\mathbf{w}\|\mathbf{g}^T\mathbf{u} + \|\mathbf{u}\|\mathbf{h}^T\mathbf{w} + \psi(\mathbf{w}, \mathbf{u}). \tag{2.7}$$

Notice that this problem differs slightly from (2.2) in the order in which we compute the optimization and one can obtain one problem from the other exchanging the minimization over $\mathbf{w}$ and the maximization over β.

However, in general the two problems are not the same and we cannot do the aforementioned swap because Sion's minmax theorem [25] is not applicable. Indeed, differently from the (PO) objective function the (AO) one is not anymore convex-concave because for different realizations of the random vectors $\mathbf{g}$ and $\mathbf{h}$ the terms $\|\mathbf{w}\|\mathbf{g}^T\mathbf{u}$ and $\|\mathbf{u}\|\mathbf{h}^T\mathbf{w}$ may be convex or concave. For example, if $\mathbf{g}^T\mathbf{u}$ is negative and $\mathbf{h}^T\mathbf{w}$ is positive the term $\|\mathbf{w}\|\mathbf{g}^T\mathbf{u}$ is concave in $\mathbf{w}$ while $\|\mathbf{u}\|\mathbf{h}^T\mathbf{w}$ is convex in $\mathbf{u}$ prohibiting the use of Sion's minimax theorem to interchange the order of minimization and maximization in (2.7).

Nevertheless, CGMT with the previous inequalities helps us because the tight relation between (PO) and (AO) along with convexity of the (PO) problem can be translated to asymptotically (when $d, n \to +\infty$) infer properties of the (PO) optimal solution based on the modified (AO) which we stress again is not equal to the standard (AO) problem.

Since we are focused on the high dimensions regime we present now the asymptotic version of the CGMT that will be actually employed in this work noticing that, as we said, we obtain a characterization of the optimal (PO) problem solution investigating the (AO) problem.

Proposition 2.2.1. *(Asymptotic CGMT). Let $\mathcal{S}$ be an arbitrary open subset of $\mathcal{S}_{\mathbf{w}}$ and $\mathcal{S}^{\mathbf{c}} = \mathcal{S}_{\mathbf{w}}/\mathcal{S}$. Denote with $\phi'_{\mathcal{S}^{\mathbf{c}}}(\mathbf{g}, \mathbf{h})$ the optimal cost of the optimization in (2.7), when the minimization over $\mathbf{w}$ is now constrained over $\mathbf{w} \in \mathcal{S}^{\mathbf{c}}$. Let $\mathbf{w}_{\Phi}(\mathbf{X})$ be any optimal minimizer of (2.1). Under the same convexity conditions of CGMT, suppose there exist constants $\overline{\phi'} < \overline{\phi'}_{\mathcal{S}^{\mathbf{c}}}$ such that $\phi'(\mathbf{g}, \mathbf{h}) \xrightarrow{P} \overline{\phi'}$ and $\phi'_{\mathcal{S}^{\mathbf{c}}}(\mathbf{g}, \mathbf{h}) \xrightarrow{P} \overline{\phi'}_{\mathcal{S}^{\mathbf{c}}}$. Then,*

$$\lim_{n\to\infty} \mathbb{P}(\mathbf{w}_{\Phi}(\mathbf{X}) \in \mathcal{S}) = 1$$

2.3 f-divergences

The notion of divergence to quantify the discrepancy between two probability distribution has its foundations in information theory, specifically in communication theory where the main goal back in the days was to understand how much one can compress data and what is the ultimate transmission rate of communication. These questions are deeply related to probability distributions that underlie the communication and in particular to the concept of entropy of a random variable and mutual information. For example the capacity of a communication channel with input X and output Y is the maximum mutual information.

The entropy of a random variable X with probability density $p(x)$ having support in $\mathcal{X}$ is a measure for the average uncertainty of the random variable and mathematically is defined as

$$H(X) = -\sum_{x\in\mathcal{X}} p(x) \log_2 p(x). \tag{2.8}$$

To get more intuition of the concept, let us think about the toss of a fair coin. The random variable representing this event has entropy 1 because it requires just one bit to describe the outcome of the toss which is simply head or tail (i.e. 0 or 1).

Moving further, entropy is a concept related to a single random variable. When we want to describe the uncertainty of a random variable conditioned on another random variable we introduce the concept of conditional entropy. Therefore, the reduction in uncertainty due to another variable is called mutual information which in formula is

$$I(X;Y) = H(X) - H(X|Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}. \tag{2.9}$$

This small detour to say that mutual information turns out to be a special instance of what is called relative entropy which is a way to quantify the "distance" between two distributions p and q. It is defined as

$$D(p\|q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \tag{2.10}$$

which is mostly known as Kullback-Leibler divergence, [9].

This type of divergence is in turn a special case of distance belonging to the wider family called f-divergences first introduced in [21] and further developed by the works [1] and most importantly [10]. The f-divergence between two measures $\mathbb{P}$ and $\mathbb{Q}$ over $\mathcal{X}$ is defined as

$$D_f(\mathbb{P}\|\mathbb{Q}) \doteq \int_{\mathcal{X}} f\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{Q} \tag{2.11}$$

where $f(t) : \mathbb{R} \to \overline{\mathbb{R}}_+$ is a convex function for $t \geq 0$ satisfying $f(1) = 0$ and defined as $+\infty$ for $t < 0$. Notice that when dealing with probability measures the case $t < 0$ never occurs. Because $f(1) = 0$ we can easily see that when the two measures are the same, i.e. $\mathbb{P} = \mathbb{Q}$ the divergence is identically zero as we would expect. Moreover, if we assume that f is strictly convex then this is actually the only situation in which the divergence is zero, namely $D_f(\mathbb{P}\|\mathbb{Q}) = 0$ iff $\mathbb{P} = \mathbb{Q}$.

If we are able to find a measure μ on the space $\mathcal{X}$ for which $\mathbb{P}, \mathbb{Q}$ are absolutely continuous with respect to μ then we can use Radon-Nikodym theorem and consider the probability densities $p = \frac{d\mathbb{P}}{d\mu}$, $q = \frac{d\mathbb{Q}}{d\mu}$. In this case the f-divergence can be rewritten as

$$D_f(\mathbb{P}\|\mathbb{Q}) = \int_{\mathcal{X}} f\left(\frac{p(x)}{q(x)}\right) q(x) d\mu(x). \tag{2.12}$$

Remark 2.3.1. *Up to now we have considered continuous measures, but there is an analogue definition for discrete ones in which appears a summation instead of the integral. The discrete version is actually what we will consider because we are interested in data-driven applications.*

Notice also that there is another thing to pay attention to. Specifically, if $\mathbb{Q}(x) = 0$ and $\mathbb{P}(x) = a$ we have an non-admissible division by zero in the argument of the function

f. A first way to tackle this issue is by defining

$$f\left(\frac{a}{0}\right) \doteq \begin{cases} a \lim_{x\to 0^+} x f(\frac{a}{x}) & a > 0 \\ 0 & a = 0 \end{cases}$$

However, since the family of divergences is wide and different we might have different limiting behaviour for specific instances of f-divergence. To avoid this, one can restrict the attention to the case where $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$, namely when the two measures $\mathbb{P}$ and $\mathbb{Q}$ share the same support. In this scenario whenever $\mathbb{Q} = 0$ also $\mathbb{P}$ must be equal to zero, thus excluding the above case.

Below we summarize some of the most important properties of f-divergences.

(i) Linearity: $D_{\alpha_1 f_1 + \alpha_2 f_2}(\mathbb{P}\|\mathbb{Q}) = \alpha_1 D_{f_1}(\mathbb{P}\|\mathbb{Q}) + \alpha_2 D_{f_2}(\mathbb{P}\|\mathbb{Q})$, for every non-negative α_1, α_2. This holds because the integral operator is a linear functional.

(ii) Non-negativity: $D_f(\mathbb{P}\|\mathbb{Q}) \geq 0$ with the equality if $\mathbb{P} = \mathbb{Q}$. This is a consequence of Jensen's inequality, indeed
$$D_f(\mathbb{P}\|\mathbb{Q}) = \int f\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{Q} \geq f\left(\int \frac{d\mathbb{P}}{d\mathbb{Q}} d\mathbb{Q}\right) = f(1) = 0.$$

(iii) Joint convexity: for $t \in [0,1]$ and two pairs of measures $(\mathbb{P}_1, \mathbb{Q}_1)$ and $(\mathbb{P}_2, \mathbb{Q}_2)$ it is verified the condition
$$D_f(t\mathbb{P}_1 + (1-t)\mathbb{P}_2 \| t\mathbb{Q}_1 + (1-t)\mathbb{Q}_2) \leq tD_f(\mathbb{P}_1\|\mathbb{Q}_1) + (1-t)D_f(\mathbb{P}_2\|\mathbb{Q}_2)$$
which follows from the convexity of the mapping $(p,q) \mapsto qf(p/q)$ in $\mathbb{R}^2_+$.

(iv) Let $g(x) = f(x) + c(x-1)$, for every $c \in \mathbb{R}$ and $x \geq 0$, then $D_f(\mathbb{P}\|\mathbb{Q}) = D_g(\mathbb{P}\|\mathbb{Q})$. In particular we can always assume that $f \geq 0$ and (if f is differentiable at 1) that $f'(1) = 0$.

This property follows using linearity and noticing that $D_{c(x-1)}(\mathbb{P}\|\mathbb{Q}) = 0$. Moreover, we can reduce to $f \geq 0$ considering $c = -f'(1)$ (or any subdifferential at $x = 1$ if f is not differentiable) and applying the property of convex functions $f(x) \geq f(y) + f'(y)(x-y)$, for every $x,\ y = 1$.

Finally, we grouped some of the most common f-divergences along with their conjugate functions in Tab. 2.1, [4].

2.4 f-divergence distributionally regularized optimization

Now we have all ingredients to formulate the problem from which we will develop our analysis. In particular our problem is an instance of the *Distributionally Regularized Optimization* (DRE) framework where the distance between distributions is measured using f-divergences. This type of optimization differs from the standard *Empirical Risk Minimization* (ERM) because in this latter and more widely adopted framework the goal is to find models that achieve uniformly good performance on almost all possible instances of the input. However, this way of proceeding may lack of performance on *hard* instances of the problem. To give an example, in speech recognition automated algorithms are inaccurate for people with minority accent. Also in other applications such

Table 2.1: Some $f-$divergence examples

Divergence	$f(t),\ t \geq 0$	$f^*(s)$
Kullback-Leibler	$t \log t - t + 1$	$e^s - 1$
Burg entropy	$-\log t + t - 1$	$-\log(1-s), \quad s < 1$
J divergence	$(t-1)\log t$	No closed form
χ^2-distance	$\frac{1}{t}(t-1)^2$	$2 - 2\sqrt{1-s}, \quad s < 1$
Modified χ^2-distance	$\frac{1}{2}(t-1)^2$	$\begin{cases} \frac{s^2}{2} + s & \text{if } s \geq -1 \\ -\frac{1}{2} & \text{otherwise} \end{cases}$
Hellinger distance	$(\sqrt{t} - 1)^2$	$\frac{s}{1-s}, \quad s < -1$
Cressie-Read	$\frac{1-\theta+\theta t - t^\theta}{\theta(1-\theta)}, \ \theta \neq 0, 1$	$\frac{1}{\theta}(1 - s(1-\theta))^{\theta/(\theta-1)} - \frac{1}{\theta}, \, s < \frac{1}{1-\theta}$

as facial recognition, language identification, automatic video captioning performances may vary significantly over different demographic groups such as gender, age or race ([6], [16], [23], [26]). On the contrary, our work explicity optimizes performance on "bad" events that suffer high loss.

In the introduction we presented the existing approaches to deal with DRO in the high-dimensions regime presenting the general formulation of our own problem. We will go now deeper in the technicalities related to our problem, in particular let us consider again the minmax program in (1.8) that we report here for convenience

$$\min_{\theta \in \Theta} \sup_{\mathbb{Q} \ll \hat{\mathbb{P}}_n} \mathbb{E}_{\mathbb{Q}}[\ell(Y - X^T\theta)] - \lambda D_f(\mathbb{Q} \| \hat{\mathbb{P}}_n).$$

Recall that ℓ is the convex in θ loss function, D_f is the f-divergence and we would like to stress again that we made the assumption of $\mathbb{Q}$ absolutely continuous with respect to $\hat{\mathbb{P}}_n$. Notice also that this problem is not far from the DRO problem in (1.5). Indeed, it can be thought as a sort of Lagrangian relaxation of the DRO: instead of solving the harder problem of minimizing the expected loss under the constraint $D_f(\mathbb{Q} \| \hat{\mathbb{P}}_n) \leq \epsilon$ we can bring this inequality in the objective function with the Lagrange multiplier λ.

This has served also as an assist to discuss how should be chosen λ. In [11], [19] they select a radius that shrinks as the number of measurements grows, i.e. ϵ/n, and with this choice they obtain also the much desired variance regularization approximation. This is somehow reasonable because when the number of measurements is high we would like to consider a probability measure $\mathbb{Q}$ very close to the empirical one because we are sufficiently confident with the data. This translates to a small admissible difference in f-divergence between $\mathbb{Q}$ and $\hat{\mathbb{P}}_n$. One might think that an identical reasoning holds in our case but with the role of λ flipped compared with the aforementioned radius. In this case, having $\lambda = \lambda_0 d$ forces to choose $\mathbb{Q}$ close to $\hat{\mathbb{P}}_n$ when the number of available data is sufficiently large because a tiny change in the f-divergence will cause the objective function to be $-\infty$ for which any θ is a minimizer. However, in our case we decided to pick λ constant and verify if it works in simulations. Our take on this is that we are dealing with a high-dimensions regime where also d grows to infinity and therefore this might translate into a constant regularization parameter. A thorough study to investigate better this is left as future work.

Back to the problem, we make also some rather mild assumptions to have a tractable formulation. The first one is something very natural and massively used in statistics and machine learning, we indeed assume that the loss function $\ell : \mathbb{R} \mapsto \overline{\mathbb{R}}_+$ is convex in θ and proper. This implies that the problem we would solve is eventually convex enabling us to use established optimization techniques such as Gradient Descent (GD) and Stochastic Gradient Descent (SGD) to find the minimum which is also global. Note that without losing generality we can assume that the loss function takes only non-negative values and $\min_x \ell(x) = 0$ for $x = 0$.

The second assumption which we have made is $\mathbb{Q}$ absolutely continuous with respect to $\hat{\mathbb{P}}_n$. This is needed, as already observed in the previous section, to avoid the argument of f to be an undefined operation. This implies that the possible distributions $\mathbb{Q}$ under which we compute the expectation have the same support as the empirical distribution. Strictly speaking, if $\hat{\mathbb{P}}_n$ assigns zero probability to an event the same thing has to happen with $\mathbb{Q}$, making impossible the emergence of unseen scenarios. Notice that the contrary is still possible, if $\hat{\mathbb{P}}_n$ assigns positive mass to an event, $\mathbb{Q}$ can still give zero mass to the same event. This has also the major implication that we cannot be robust against events that we did not capture with the data. Unfortunately, this is a price we need to pay when working with f-divergences, but on the other hand this very peculiarity of this family of distances makes the problem computationally tractable which is an appreciated property that the same problem with the Wasserstein metric usually does not enjoy.

We continue our discussion presenting a dual reformulation of our minmax problem. This is something not completely new because there are already similar results in the literature presenting the dual problem of the DRO in (1.5). However our instance is slightly different as already pointed out, therefore we present also a proof of how we can obtain the dual problem which we think could be helpful to understand thoroughly it and have some insights about the technicalities. The dual form consists of a single convex minimization problem which can be more easily solved using standard algorithms. Our result leverages on [4] because we are focused on data-driven applications where the reference distribution is discrete. For sake of completeness a similar result holds in the case of continuous distributions, [24].

Proposition 2.4.1. *Let $\hat{\mathbb{P}}_n$ be the empirical measure on $(\mathcal{X}, \mathcal{A})$ which gives to any point the same value $1/n$, $f^*(s) := \sup_{u \in dom(f)}\{u^T s - f(u)\}$ be the usual Fenchel conjugate of*

f then problem (1.8) admits the following strong dual reformulation

$$\inf_{\theta\in\Theta,\eta\in\mathbb{R}} \left\{ \mathbb{E}_{\hat{\mathbb{P}}_n}\left[\lambda f^*\left(\frac{\ell(Y-X^T\theta)-\eta}{\lambda}\right)\right]+\eta\right\}. \tag{2.13}$$

Remark 2.4.2. *(Convexity of the dual reformulation). For convex losses $\theta \mapsto \ell(\theta; X)$ the dual form in Proposition 2.4.1 is jointly convex in (θ,η).*

For illustration purposes, we conclude this section with an example bridging the gap between the distributionally regularized M-estimator with f-divergence and the CGMT tool. In the example we show that the DRE problem can be brought into the formulation required by the CGMT.

Example 2.4.3. *((DRE)→(PO)). Consider the dual reformulation presented in Prop. 2.4.1. For sake of completeness we report it also here*

$$\inf_{\theta\in\Theta,\eta\in\mathbb{R}} \frac{1}{n}\sum_{i=0}^{n}\lambda f^*\left(\frac{\ell(y_i-x_i^T\theta)-\eta}{\lambda}\right)+\eta. \tag{2.14}$$

Remember that we collect measurements of a linear model of the type $y_i = x_i^T\theta_0 + z_i$. Therefore, recalling the change of variables $\mathbf{w} = (\theta-\theta_0)/\sqrt{d}$ we can rewrite the problem as

$$\inf_{\mathbf{w}\in\mathbb{R}^d,\eta\in\mathbb{R}} \frac{1}{n}\sum_{i=0}^{n}\lambda f^*\left(\frac{\ell(z_i-\sqrt{d}x_i^T\mathbf{w})-\eta}{\lambda}\right)+\eta. \tag{2.15}$$

Now, consider the following change of variables $v_i = z_i - \sqrt{d}x_i^T\mathbf{w}$ which lead to this new formulation

$$\begin{aligned}&\inf_{\substack{\mathbf{w}\in\mathbb{R}^d,\eta\in\mathbb{R}\\ \mathbf{v}\in\mathbb{R}^n}} \frac{1}{n}\sum_{i=0}^{n}\lambda f^*\left(\frac{\ell(v_i)-\eta}{\lambda}\right)+\eta\\ &\quad s.t.\ \mathbf{v}=\mathbf{z}-\sqrt{d}\mathbf{X}\mathbf{w}.\end{aligned} \tag{2.16}$$

Let us introduce the Lagrange multiplier $\mathbf{u}\in\mathbb{R}^n$ to bring the constraint into the objective function

$$\inf_{\substack{\mathbf{w}\in\mathbb{R}^d,\eta\in\mathbb{R}\\ \mathbf{v}\in\mathbb{R}^n}} \max_{\mathbf{u}\in\mathbb{R}^n} \frac{1}{\sqrt{d}}\mathbf{u}^T(\sqrt{d}\mathbf{X})\mathbf{w}+\frac{1}{\sqrt{d}}\mathbf{u}^T\mathbf{z}-\frac{1}{\sqrt{d}}\mathbf{u}^T\mathbf{v}+\frac{1}{n}\sum_{i=0}^{n}\lambda f^*\left(\frac{\ell(v_i)-\eta}{\lambda}\right)+\eta. \tag{2.17}$$

Finally it takes not much effort to check that the following objective function is in the format required in CGMT (for now assume that both $\mathbf{w}$ and $\mathbf{u}$ live in compact sets): there is the bilinear form $\mathbf{u}^T\mathbf{X}\mathbf{w}$ where matrix X has entries i.i.d. standard normal while the remaining part is jointly convex-concave in $\mathbf{w},\mathbf{u}$.

Chapter 3

Main Results

This chapter contains the bulk of the results we obtained when solving the high-dimensional regime f-divergence DRE problem. In particular, starting our analysis from the strong dual reformulation presented in the previous chapter we will present all the steps necessary to bring the problem into a convex-concave deterministic formulation of it. We are going to do this using the particular instance of distance called χ^2-divergence.

3.1 Problem statement

In the introduction and in the previous chapter we reviewed some existing approaches to deal with DRO and we also presented our DRE problem along with its strong dual reformulation whose objective function is a function of the estimation error $\mathbf{w}$ as we can see from

$$M(\mathbf{w}) \doteq \inf_{\eta\in\mathbb{R}} \frac{1}{n}\sum_{i=0}^{n} \lambda f^{*}\left(\frac{\ell(z_i - \sqrt{d}x_i^T\mathbf{w}) - \eta}{\lambda}\right) + \eta \tag{3.1}$$

We will now go deeper into the technicalities in order to reach our final goal of finding in a relative easy way the estimation error.

We begin by explicitly state the first main assumptions that will be used in this work.

Assumption 3.1.1.

(i) Isotropic Gaussian features: vectors x_i are all i.i.d. $\mathcal{N}(0, d^{-1}I_d)$, $\forall i$.

(ii) The true parameter θ_0, the measurement noise Z and the feature vectors x_i, $i \in \{1, \ldots, n\}$ are independent random quantities.

(iii) The number of measurements n and the number of variables d go to infinity at a fixed ratio $n/d = \delta \in (0, +\infty)$.

We need to make a simplification that will in some sense facilitate the problem because it restricts ourselves to a particular case of f-divergence. The reason for this decision will become clear later in the analysis of the problem. Also the choice of what particular instance of f to select is not easy, one needs to consider one function for which there exists an expression of its conjugate in closed form and at the same time it should lead to tractable formulations. Our decision is toward χ^2-divergence because it has a relative nice expression for the conjugate function and at the same time it has already been used in the literature as previously mentioned.

The work we will present relies on the result of [27] for the high-dimensional error analysis. In order to make use of these results we need to set up the following assumptions on the loss function ℓ and on the noise Z.

Assumption 3.1.2.

(i) The noise distribution is such that $\mathbb{E}Z^2 < \infty$.

(ii) The loss function ℓ *is proper, lower semicontinuous and convex.*

(iii) The loss function ℓ *satisfies the relation* $\ell(x) \leq K|x|$, *for some* $K > 0$.

Remark 3.1.3. *(Assumption's details). The first assumption is tailored to ensure that the noise has at least finite second order moment. The second one ensures that the problem is well-posed and tractable (it is a standard assumption in the literature). The final one is related to the growth rate of the loss function. We are requiring that* ℓ *has growth rate at most linear.*

Example 3.1.4. *(Examples of noise and loss combination satisfying the assumption). We make here some examples of possible combinations of loss and noise distribution which are admitted by Assumption 3.1.2. Examples of loss functions can be Least Absolute Deviation (LAD) which considers the absolute value as error or Huber loss which has a quadratic growth below than a certain threshold and after that is equivalent to the absolute value. Possible examples of noise distributions with finite second moment are Gaussian and exponential noise.*

Remark 3.1.5. *(Implication of finite second moment). The first assumption of finite second order moment implies that*

$$\|\mathbf{z}\| \leq C_1\sqrt{d}, \quad \textit{for some } C_1 > 0$$

holds with high probability for large d *using* $d = n/\delta$ *by the WLLN. Indeed, as* $n \to +\infty$, $\frac{1}{n}\|\mathbf{z}\|^2 \xrightarrow{P} \mathbb{E}(Z^2) < \infty$

We present now the the main result on the high-dimensional error analysis. However, before stating the main theorem we introduce some notation which is used in its statement. We first introduce the function $k(s,\eta) = \frac{1}{2\lambda}(\ell(s)-\eta)^2+\ell(s)$. Associated with this function we can define its Moreau envelope as presented in Appendix A, namely

$$e_k(c,\tau) = \inf_w \left\{ k(w,\eta) + \frac{1}{2\tau}|w-c|^2 \right\} \tag{3.2}$$

and consequently its *expected Moreau envelope* $\mathcal{K} : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ as

$$\mathcal{K}(c, \tau; \eta) \doteq \mathbb{E}_{\substack{G \sim \mathcal{N}(0,1), \\ Z \sim \mathbb{P}_Z}} \left[e_k \left(cG + Z, \tau \right) \right]. \tag{3.3}$$

Theorem 3.1.6. *(Performance of $\hat{\theta}_{DRE}$). Let Assumptions 3.1.1 and 3.1.2 be satisfied then it holds w.p.a. 1*

$$\lim_{d \to +\infty} \frac{\|\hat{\theta}_{DRE} - \theta_0\|^2}{d} = \alpha_\star^2 \tag{3.4}$$

where $\alpha_\star^2$ is the unique minimizer of the following convex-concave minimax scalar program

$$\inf_{\substack{\alpha \geq 0, \tau > 0, \\ \eta \in \mathbb{R}}} \max_{\beta \geq 0} \frac{\beta \tau}{2} - \alpha \beta + \mathcal{K}\left(\alpha, \frac{\tau}{\beta \delta}; \eta \right) \tag{3.5}$$

where $\mathcal{K}(\cdot, \cdot; \eta)$ is defined in (3.3).

To better have the intuition about the statement of the theorem we can argue as follows. The theorem says that, given $\hat{\mathbf{w}}_{DRE} = (\hat{\theta}_{DRE} - \theta_0)/\sqrt{d}$, we have the convergence in probability $\|\hat{\mathbf{w}}_{DRE}\| \xrightarrow{P} \alpha_\star$ as $n, d \to +\infty$ where $\alpha_\star$ is the deterministic solution of a convex-concave scalar program and this is actually what we should prove. The same convergence in probability is equivalent to say that w.p.a. 1 the optimal solution $\hat{\mathbf{w}}_{DRE}$ belongs to the set

$$\mathcal{S}_\rho = \{ \mathbf{w} \ : \ |\|\mathbf{w}\| - \alpha_\star| < \rho \} \tag{3.6}$$

for every arbitrary $\rho > 0$.

A question may arise: how can we show that the optimal solution of a stochastic program lies in such set? One idea is to consider the complement set $\mathcal{S}_\rho^c$ and instead of looking at the optimal solution we better might check the corresponding optimal value, in particular if

$$M(\hat{\mathbf{w}}) < \inf_{\mathbf{w} \in \mathcal{S}_\rho^c} M(\mathbf{w}) \tag{3.7}$$

holds w.p.a. 1 then we can conclude that $\hat{\mathbf{w}}$ belongs to the set $\mathcal{S}_\rho$.

However, it is generally challenging to verify the previous inequality. Indeed, we are comparing stochastic processes which can hide non-trivial issues (for example it might be that the inequality is not satisfied for all realizations of the random variables involved). One might first consider the convergence in probability of these two random quantities and then compare the deterministic values obtained which is instead a straightforward operation. Indeed, if

$$M(\hat{\mathbf{w}}) \xrightarrow{P} \overline{M} \quad \text{and} \quad \inf_{\mathbf{w} \in \mathcal{S}_\rho^c} M(\mathbf{w}) \xrightarrow{P} \overline{M}_{\mathcal{S}_\rho^c} \tag{3.8}$$

then (3.6) holds as long as the following deterministic comparison is true

$$\overline{M} < \overline{M}_{\mathcal{S}_\rho^c}. \tag{3.9}$$

This approach is what we are going to follow, but there are still some challenges. For example, it is in general not straightforward to work directly with the objective function M and determine the two convergences in probability of the stochastic programs. But here is where CGMT comes to help because we will derive these results working with an auxiliary objective function.

Since the proof of the theorem is definitely long and intricate we briefly present here, in order to get the scheme, what are the main required steps to obtain the desired convergence in probability to a deterministic program. In the next section we are going to develop better all the steps while all the proofs are deferred to the appendix.

- Derivation of the (PO) problem from (2.13) and identification of the corresponding (AO) problem proving convexity-concavity and compactness of the sets $\mathcal{S}_{\mathbf{w}}$ and $\mathcal{S}_{\mathbf{u}}$.
- Scalarization of the (AO) problem.
- Convergence analysis.

3.2 Steps of the solution

In the section we will expand the steps presented before. We decided to divide the proof in steps instead of a continuum to give the reader a sort of methodology to follow explaining how to bring the problem into the desired formulation needed for proving the theorem.

Step 1: We bring the formulation in (1.8) into a (PO) problem. The proof is built on top of the dual formulation presented in Prop. 2.4.1. Indeed after the change of variable $\mathbf{w} = (\theta - \theta_0)/\sqrt{d}$ we can rewrite the dual formulation as

$$\inf_{\mathbf{w}\in\mathbb{R}^d, \eta\in\mathbb{R}} \frac{1}{n}\sum_{i=0}^{n} \lambda f^* \left(\frac{\ell(z_i - \sqrt{d}x_i^T \mathbf{w}) - \eta}{\lambda} \right) + \eta. \tag{3.10}$$

where we have also used the fact that $y_i = x_i^T\theta_0 + z_i$, namely that our measurements come from a linear and noisy model.

We can now make this other change of variables $\mathbf{v} = \mathbf{z} - \sqrt{d}\mathbf{X}\mathbf{w}$ which leads to the following

$$\begin{gathered} \inf_{\substack{\mathbf{w}\in\mathbb{R}^d, \eta\in\mathbb{R} \\ \mathbf{v}\in\mathbb{R}^n}} \frac{1}{n}\sum_{i=0}^{n} \lambda f^* \left(\frac{\ell(v_i) - \eta}{\lambda} \right) + \eta \\ \text{s.t. } \mathbf{v} = \mathbf{z} - \sqrt{d}\mathbf{X}\mathbf{w} \end{gathered} \tag{3.11}$$

and using Lagrangian duality to bring the equality constraint into the objective function with associated Lagrange multiplier $\mathbf{u} \in \mathbb{R}^n$ we obtain the following problem which has the form of a (PO) problem

$$\inf_{\substack{\mathbf{w}\in\mathbb{R}^d, \eta\in\mathbb{R} \\ \mathbf{v}\in\mathbb{R}^n}} \max_{\mathbf{u}\in\mathbb{R}^n} -\frac{1}{\sqrt{d}}\mathbf{u}^T(\sqrt{d}\mathbf{X})\mathbf{w} + \frac{1}{\sqrt{d}}\mathbf{u}^T\mathbf{z} - \frac{1}{\sqrt{d}}\mathbf{u}^T\mathbf{v} + \frac{1}{n}\sum_{i=0}^{n} \lambda f^* \left(\frac{\ell(v_i) - \eta}{\lambda} \right) + \eta. \tag{3.12}$$

This is partially incorrect, because in order to be a (PO) problem it should have the bilinear form which is present in our case (note that the minus sign in front is not important since the matrix $\mathbf{X}$ has entries that are $\mathcal{N}(0, 1/d)$) and the remaining part must be convex-concave in $(\mathbf{w}, \mathbf{u})$ which instead it has not been verified yet. Moreover, for now we have written that variables $\mathbf{w}, \mathbf{u}$ live in $\mathbb{R}^d$ and $\mathbb{R}^n$ respectively, but CGMT requires those sets to be compact convex and clearly we lack of compactness. The following will be devoted to investigate these requirements.

Convexity-concavity is immediate to verify, because we can notice that we have a linear function in $\mathbf{u}$ which is concave while since $\mathbf{w}$ does not appear except in the bilinear form, convexity is ensured. As remark that will be useful later we would like to point out that the remaining part is jointly convex in $(\mathbf{w}, \mathbf{v}, \eta)$. Indeed, $f^*\left(\frac{\ell(v_i)-\eta}{\lambda}\right)$ is jointly convex in (v_i, η) because from the definition of Fenchel conjugate we have

$$f^*\left(\frac{\ell(v_i)-\eta}{\lambda}\right) = \sup_{z\in\mathbb{R}}\left\{z\left(\frac{\ell(v_i)-\eta}{\lambda}\right) - f(z)\right\}.$$

Again, we can notice that $f(z) = +\infty$ for $z < 0$ implying that the supremum will be never attained for negative values.

Therefore, fixing $z \geq 0$ (which is the effective domain of f) the function $\frac{z}{\lambda}\ell(v_i) - \frac{z}{\lambda}\eta - f(z)$ is jointly convex in (v_i, η) simply because by hypothesis ℓ is convex and η appears linearly. Thus, using Proposition A.0.2 taking the supremum over z we prove joint convexity. Therefore the term

$$\frac{1}{n}\sum_{i=0}^{n}\lambda f^*\left(\frac{\ell(v_i)-\eta}{\lambda}\right) + \eta$$

is finally jointly convex in $(\mathbf{v}, \eta)$ because sum of convex functions with weight $\lambda > 0$ is still convex and we have another term η which appears linearly.

We move forward now to check compactness which will require two separate lemmas. The general idea is to add artificial constraints that do not invalidate the optimization problem but allow to apply the theorem by restricting to compact sets. The first one regards the variable $\mathbf{w}$. We can define the set $\mathcal{S}_\mathbf{w} = \{\mathbf{w} \in \mathbb{R}^d \mid \|\mathbf{w}\| \leq K_\alpha\}$ where $K_\alpha := \alpha_\star + \zeta$ for a constant $\zeta > 0$ and $\alpha_\star$ being the estimation error defined in Theorem 3.1.6. If we restrict the optimization variable $\mathbf{w} \in \mathcal{S}_\mathbf{w}$ in (3.12) we expect that the optimization will not be affected with high probability when d is large enough. This is formally stated in the following Lemma.

Lemma 3.2.1. *Consider the optimization in (3.12) and its "bounded" version when* $\mathbf{w} \in \mathcal{S}_\mathbf{w}$. *Let* $\hat{\mathbf{w}}$ *and* $\hat{\mathbf{w}}_b$ *be optimal solutions of these two problems, respectively. Recall also the definition of* K_α *given before. If* $\|\hat{\mathbf{w}}_b\| \xrightarrow{P} \alpha_\star$, *then it holds also* $\|\hat{\mathbf{w}}\| \xrightarrow{P} \alpha_\star$.

Moving on, variable $\mathbf{u}$ is unconstrained as well, therefore we would like to add a constraint of the same type adopted for $\mathbf{w}$ without modifying the optimization result.

Defining the set $\mathcal{S}_{\mathbf{u}} = \{\mathbf{u} \in \mathbb{R}^n \mid \|\mathbf{u}\| \leq K_\beta\}$ where K_β is taken sufficiently large we will show that if we constraint variable $\mathbf{u}$ to $\mathcal{S}_{\mathbf{u}}$ the optimization result is not affected.

However, before stating the lemma, we are going to consider a specific instance of f as we mentioned before. The main reason to consider a particular instance of function is that it makes the remaining analysis a little bit easier. Secondly, in [19] they obtain the bias-variance approximation from the DRE problem using this specific instance of f. Recall that the χ^2-divergence function is

$$f(t) = \begin{cases} \frac{1}{2}(t-1)^2 & \text{if } t \geq 0 \\ +\infty & \text{otherwise} \end{cases}$$

whose convex conjugate is

$$f^*(s) = \begin{cases} \frac{s^2}{2} + s & \text{if } s \geq -1 \\ -\frac{1}{2} & \text{otherwise.} \end{cases}$$

Clearly this is a convex function but we do not like this two-cases formulation. However, we assume that only the first case is satisfied w.p.a. 1, namely when the dimension grows only the first case happens with high probability. We will then verify if it is a reasonable assumption in simulation.

Therefore substituting this expression into the problem we get

$$\inf_{\substack{\mathbf{w}\in\mathcal{S}_{\mathbf{w}},\\ \mathbf{v}\in\mathbb{R}^n,\eta\in\mathbb{R}}} \max_{\mathbf{u}\in\mathbb{R}^n} \frac{1}{\sqrt{d}}\left\{-\mathbf{u}^T(\sqrt{d}\mathbf{X})\mathbf{w} + \mathbf{u}^T\mathbf{z} - \mathbf{u}^T\mathbf{v}\right\} + \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{2\lambda}\left(\ell(v_i)-\eta\right)^2 + \ell(v_i)\right] \tag{3.13}$$

Before presenting the lemma it is important to have an insight on the value of the optimal η and on the subdifferential of the loss function. This is what the following remarks do.

Remark 3.2.2. *(η expression). We would like to understand what is the expression of η which can be useful in later proofs. In particular what we can do is to solve the minimization over it in closed form and find the expression for its optimal value. From (3.13) we can differentiate over η obtaining*

$$\frac{1}{n}\sum_{i=0}^{n}\frac{1}{\lambda}(\ell(v_i)-\eta)(-1)$$

and since (3.13) is convex in η as previously said, by simply imposing derivative equal to zero we obtain the minimum

$$\eta_\star = \frac{1}{n}\sum_{i=0}^{n}\ell(v_i). \tag{3.14}$$

Notice that since the loss is always non-negative we are ensured that the optimal $\eta_\star$ is *actually non-negative.*

Remark 3.2.3. *(Subdifferential of ℓ). In the case of loss with at most linear growth we have that the subdifferential is always bounded, i.e. $\ell_+'(\cdot) \leq Q$, for some $Q > 0$. This simply implies the natural normalization condition: for every $c > 0$ there exists a constant $P > 0$ s.t. if $\|\mathbf{v}\| \leq c\sqrt{d}$ then $\sup_{\mathbf{s}\in\partial L(\mathbf{v})}\|\mathbf{s}\| \leq P\sqrt{d}$.*

Lemma 3.2.4. *(Growth rate of $\mathbf{u}_\star$ in (3.13)). If Assumption 3.1.2 holds then the optimal $\mathbf{u}_\star$ in (3.13) satisfies $\|\mathbf{u}_\star\| \leq K_\beta$, for some constant $K_\beta > 0$.*

Finally we can identify the corresponding (AO) problem which is given by

$$\inf_{\substack{\mathbf{w}\in\mathcal{S}_\mathbf{w},\\ \mathbf{v}\in\mathbb{R}^n,\eta\in\mathbb{R}}} \max_{\mathbf{u}\in\mathcal{S}_\mathbf{u}} \frac{1}{\sqrt{d}}\left\{\|\mathbf{w}\|\mathbf{g}^T\mathbf{u} - \|\mathbf{u}\|\mathbf{h}^T\mathbf{w} + \mathbf{u}^T\mathbf{z} - \mathbf{u}^T\mathbf{v}\right\} + \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{2\lambda}\left(\ell(v_i) - \eta\right)^2 + \ell(v_i)\right]. \tag{3.15}$$

Step 2: We consider the modified (AO) problem, and show that it can be brought to a scalar optimization problem. When we presented the CGMT in the previous chapter we discussed that in order to arrive at the scalar optimization we cannot work directly with the (AO) problem in (3.15) but we might consider a similar program but yet with an important difference: the order of optimization operations is slightly different. Thus, we refer to the following as Modified (AO) problem

$$\begin{aligned}\max_{0\leq\beta\leq K_\beta} \inf_{\substack{\mathbf{w}\in\mathcal{S}_\mathbf{w},\\ \mathbf{v}\in\mathbb{R}^n,\eta\in\mathbb{R}}} \max_{\|\mathbf{u}\|=\beta} \frac{1}{\sqrt{d}}\left\{\|\mathbf{w}\|\mathbf{g}^T\mathbf{u} - \|\mathbf{u}\|\mathbf{h}^T\mathbf{w} + \mathbf{u}^T\mathbf{z} - \mathbf{u}^T\mathbf{v}\right\} + \\ + \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{2\lambda}\left(\ell(v_i) - \eta\right)^2 + \ell(v_i)\right].\end{aligned} \tag{3.16}$$

We would like to point out again that (AO) and Modified (AO) problems could be derived one from the other if we could interchange the maximization over β and the infimum, but this is not possible due to the non-convex nature that might arise for different realizations of the random vectors involved. However, [27, Lemma 7] provides the theoretical guarantees allowing to consider the modified version of the (AO) problem in place of the "original" one.

Now, observing (3.16) one might notice that variables $\mathbf{w}$ and $\mathbf{u}$ appear in the objective only through either linear terms or their magnitudes. This suggests that one can easily optimize over their direction keeping fixed the magnitudes. Indeed, one can use this general fact when optimizing a linear function over a variable with fixed magnitude

$$\max_{\|\mathbf{x}\|=\epsilon} \mathbf{x}^T\mathbf{y} = \epsilon\|\mathbf{y}\|. \tag{3.17}$$

We can notice that we already have the maximization over $\|\mathbf{u}\| = \beta$ in (3.16), therefore we can apply the previous relation obtaining the following

$$\max_{0\leq\beta\leq K_\beta} \inf_{\substack{\mathbf{w}\in\mathcal{S}_\mathbf{w},\\ \mathbf{v}\in\mathbb{R}^n,\eta\in\mathbb{R}}} \frac{\beta}{\sqrt{d}}\|\|\mathbf{w}\|\mathbf{g}+\mathbf{z}-\mathbf{v}\| - \frac{\beta}{\sqrt{d}}\mathbf{h}^T\mathbf{w} + \frac{1}{n}\sum_{i=1}^n\left[\frac{1}{2\lambda}\left(\ell(v_i)-\eta\right)^2+\ell(v_i)\right]. \tag{3.18}$$

We can also do the same for the minimization over $\mathbf{w}$ which lives in the compact set $\mathcal{S}_\mathbf{w}$, i.e. $\|\mathbf{w}\|\leq K_\alpha$.

$$\max_{0\leq\beta\leq K_\beta} \inf_{\substack{0\leq\alpha\leq K_\alpha,\\ \mathbf{v}\in\mathbb{R}^n,\eta\in\mathbb{R}}} \inf_{\|\mathbf{w}\|=\alpha} \frac{\beta}{\sqrt{d}}\|\|\mathbf{w}\|\mathbf{g}+\mathbf{z}-\mathbf{v}\| - \frac{\beta}{\sqrt{d}}\mathbf{h}^T\mathbf{w} + \frac{1}{n}\sum_{i=1}^n\left[\frac{1}{2\lambda}\left(\ell(v_i)-\eta\right)^2+\ell(v_i)\right] =$$
$$\max_{0\leq\beta\leq K_\beta} \inf_{\substack{0\leq\alpha\leq K_\alpha,\\ \mathbf{v}\in\mathbb{R}^n,\eta\in\mathbb{R}}} \frac{\beta}{\sqrt{d}}\|\alpha\mathbf{g}+\mathbf{z}-\mathbf{v}\| - \frac{\alpha\beta}{\sqrt{d}}\|\mathbf{h}\| + \frac{1}{n}\sum_{i=1}^n\left[\frac{1}{2\lambda}\left(\ell(v_i)-\eta\right)^2+\ell(v_i)\right]. \tag{3.19}$$

In this way we have been able to transform the problem from an optimization over two vector variables to only two scalar values. We would like to stress that this is necessary once the number of measurements and parameters go to infinity otherwise the optimization would be performed over variables that are infinite dimensional vectors.

Next, we wish to simplify the minimization over the remaining vector variable $\mathbf{v}\in\mathbb{R}^n$ using the same technique but unfortunately this is not possible because this time the variable appears in the objective also as argument of a function and not only linearly or through its magnitude. The new idea that we will use here is the so-called "square-root trick" which in formula can be written as

$$\chi = \inf_{\tau>0}\left\{\frac{\tau}{2}+\frac{\chi^2}{2\tau}\right\}.$$

We apply this trick to the term $\frac{1}{\sqrt{d}}\|\alpha\mathbf{g}+\mathbf{z}-\mathbf{v}\|$ obtaining the following reformulation of (3.19)

$$\max_{0\leq\beta\leq K_\beta} \inf_{\substack{0\leq\alpha\leq K_\alpha,\\ \eta\in\mathbb{R},\mathbf{v}\in\mathbb{R}^n}} \inf_{\tau>0} \beta\left\{\frac{\tau}{2}+\frac{\|\alpha\mathbf{g}+\mathbf{z}-\mathbf{v}\|^2}{2d\tau}\right\} - \frac{\alpha\beta}{\sqrt{d}}\|\mathbf{h}\| + \frac{1}{n}\sum_{i=1}^n\left[\frac{1}{2\lambda}\left(\ell(v_i)-\eta\right)^2+\ell(v_i)\right] \tag{3.20}$$

And now, rearranging some terms we can write

$$\max_{0\leq\beta\leq K_\beta} \inf_{\substack{0\leq\alpha\leq K_\alpha,\\ \eta\in\mathbb{R},\tau>0}} \frac{\beta\tau}{2} - \frac{\alpha\beta\|\mathbf{h}\|}{\sqrt{d}} + \frac{1}{n}\sum_{i=1}^n \inf_{v_i\in\mathbb{R}}\left\{\frac{1}{2\lambda}\left(\ell(v_i)-\eta\right)^2+\ell(v_i)+\frac{\beta\delta}{2\tau}(\alpha g_i+z_i-v_i)^2\right\}. \tag{3.21}$$

At this point, if we use the function $k(v_i,\eta) := \frac{1}{2\lambda}\left(\ell(v_i)-\eta\right)^2+\ell(v_i)$ it can be readily seen that the minimization over v_i gives rise to the Moreau envelope function of k evaluated at αg_i+z_i with index $\tau/\beta\delta$, i.e.

$$\inf_{v_i\in\mathbb{R}}\left\{\frac{1}{2\lambda}\left(\ell(v_i)-\eta\right)^2+\ell(v_i)+\frac{\beta\delta}{2\tau}(\alpha g_i+z_i-v_i)^2\right\} := e_k\left(\alpha g_i+z_i,\frac{\tau}{\beta\delta}\right)$$

which substituted in the previous expression leads to

$$\max_{0\leq\beta\leq K_\beta}\ \inf_{\substack{0\leq\alpha\leq K_\alpha,\\ \eta\in\mathbb{R},\tau>0}} \frac{\beta\tau}{2}-\frac{\alpha\beta\|\mathbf{h}\|}{\sqrt{d}}+\frac{1}{n}\sum_{i=1}^{n}e_k\left(\alpha g_i+z_i,\frac{\tau}{\beta\delta}\right). \tag{3.22}$$

We refer to the objective function in (3.22) as $\mathcal{O}_d(\alpha,\eta,\tau,\beta)$ where the pedix d stands for the dimension of the problem.

Step 3: We show that $\mathcal{O}_d$ is jointly convex in (α,τ,η) and concave in β and it is also continuous on its domain. First of all, in order to understand convexity-concavity for $\mathcal{O}_d$ we need to study it for the Moreau envelope. In particular, we show that the Moreau envelope is jointly convex in (α,τ,η) while concave in β. In doing so we will make use of the following technical lemma

Lemma 3.2.5. *The function $h(\alpha,\tau,v_i):=\frac{1}{2\tau}(\alpha g_i+z_i-v_i)^2$ is jointly convex in (α,τ,v_i).*

Let us first consider the function inside the minimization in the definition of Moreau envelope, namely

$$\frac{1}{2\lambda}\left(\ell(v_i)-\eta\right)^2+\ell(v_i)+\frac{\beta\delta}{2\tau}(\alpha g_i+z_i-v_i)^2.$$

Concavity is simple to verify since β appears only linearly. As for the convexity, we already saw that $\frac{1}{2\lambda}\left(\ell(v_i)-\eta\right)^2+\ell(v_i)$ is jointly convex in (v_i,η). Now, from Lemma 3.2.5 we have joint convexity in (α,v_i,τ) for $\frac{1}{2\tau}(\alpha g_i+z_i-v_i)^2$ therefore we can say that the function inside the minimization is jointly convex in (α,v_i,η,τ) and concave in β. Thus, applying Proposition A.0.3 we can conclude that the Moreau envelope is jointly convex in (α,η,τ) and concave in β.

Now that we have shown joint convexity in (α,τ,η) and concavity in β for the Moreau envelope we can conclude convexity-concavity for the stochastic function $\mathcal{O}_d$. This because the sum of Moreau envelopes is still convex-concave while the remaining terms are simply linear in α,τ and β thus implying convexity-concavity. As a consequence, convexity-concavity of $\mathcal{O}_d$ allows us to apply Sion's minmax theorem to flip the order between inf and max obtaining

$$\inf_{\substack{0\leq\alpha\leq K_\alpha,\\ \eta\in\mathbb{R},\tau>0}}\ \max_{0\leq\beta\leq K_\beta} \frac{\beta\tau}{2}-\frac{\alpha\beta\|\mathbf{h}\|}{\sqrt{d}}+\frac{1}{n}\sum_{i=1}^{n}e_k\left(\alpha g_i+z_i,\frac{\tau}{\beta\delta}\right). \tag{3.23}$$

We are left with the continuity of $\mathcal{O}_d$. Notice that this is nothing more than proving the continuity for the Moreau envelope $e_k(\alpha g+z,\tau/\beta\delta)$ because all the other terms in the objective function are trivially continuous. As for the Moreau envelope, if the function $k(\cdot\ ,\cdot;\eta)$ is lower semicontinuous and convex we can apply [22, Theorem 2.26(b)] to conclude the continuity. Convexity has already been shown while lower semicontinuity follows because of the assumption 3.1.2(ii) on the loss $\ell(\cdot)$.

Step 4: We study the convergence in probability of $\mathcal{O}_d$. We have now arrived at the optimization over scalar variables in (3.23) which has the same optimal value as the modified (AO) problem in (3.16). Following the reasoning presented at the beginning of this chapter we are interested in computing the asymptotic behaviour (in probability) of this problem for easier comparisons. To do so, we start with the convergence in probability of its objective function $\mathcal{O}_d$.

Notice that the first term in (3.22) is already a deterministic quantity that does not depend on the dimensions. The second one instead contains the norm of the stochastic vector $\mathbf{h} \in \mathbb{R}^d$ whose convergence is simple to compute using the WLLN (Theorem B.0.1). Indeed, we know that given $h \sim \mathcal{N}(0,1)$

$$\frac{1}{d}\sum_{i=0}^{d} h_i^2 \xrightarrow{P} \mathbb{E}h^2 = 1 \tag{3.24}$$

thus, taking the square root both sides and applying this result to the second term we obtain the desired result

$$\alpha\beta\frac{\|\mathbf{h}\|}{\sqrt{d}} \xrightarrow{P} \alpha\beta. \tag{3.25}$$

Finally, we need to show the convergence in probability for the sum of Moreau envelopes. If the Moreau envelope shows absolute integrability, namely its absolute value is finite when taking the expectation, then we can apply again WLLN to conclude the following

$$\frac{1}{n}\sum_{i=0}^{n} e_k\left(cg_i + z_i, \tau\right) \xrightarrow{P} \mathbb{E}_{\substack{G\sim\mathcal{N}(0,1),\\ Z\sim\mathbb{P}_Z}}\left[e_k\left(cG+Z,\tau\right)\right] = \mathcal{K}(c,\tau;\eta) \tag{3.26}$$

As for the integrability of the Moreau envelope we can argue as follows

$$\begin{aligned}\left|e_k\left(\alpha G + Z, \frac{\tau}{\beta\delta}\right)\right| &= \inf_{v\in\mathbb{R}} \frac{\beta\delta}{2\tau}(\alpha G + Z - v)^2 + \ell(v) + \frac{1}{2\lambda}(\ell(v)-\eta)^2 \\ &\leq \frac{\beta\delta}{2\tau}(\alpha G+Z)^2 + \ell(0) + \frac{1}{2\lambda}(\ell(0)-\eta)^2 = \frac{\beta\delta}{2\tau}(\alpha G+Z)^2 + \frac{\eta^2}{2\lambda}\end{aligned} \tag{3.27}$$

which is integrable due to the fact that both G and Z have finite second moment and the quantity $\eta^2/2\lambda$ is finite as well. In the previous equation we have also used $\ell(0) = 0$.

Wrapping up the results we discussed we obtain the convergence in probability of $\mathcal{O}_d$ to the following deterministic objective function

$$\mathcal{O}(\alpha,\tau,\eta,\beta) \doteq \frac{\beta\tau}{2} - \alpha\beta + \mathcal{K}\left(\alpha, \frac{\tau}{\beta\delta};\eta\right). \tag{3.28}$$

As last remark we would like to point out that the function $\mathcal{O}$ is jointly convex in (α,τ,η) and concave in β because it is obtained by pointwise limit (in probability, for each (α,τ,η,β)) of the sequence of convex-concave functions $\mathcal{O}_d$ and convexity is preserved by pointwise limits.

Step 5: We explain how the uniqueness of $\alpha_\star$ can be used to conclude

the proof of the theorem. At the beginning of this chapter we briefly explained the idea behind Theorem 3.1.6 and a possible direction to prove the statement. What we are going to do now is a step forward to the final solution. In particular we make evident how uniqueness of $\alpha_\star$ is essential in the convergence analysis of Prop. 2.2.1. In particular, recall the set presented in (3.6) and knowing that $\|\mathbf{w}\| = \alpha$ we can consider this new variant for that set more suited for the scalar problem

$$\mathcal{S}_\rho = \{\alpha \geq 0 \;:\; |\alpha - \alpha_\star| < \rho\} \tag{3.29}$$

with $\alpha_\star$ unique solution of the optimization in (3.23). If we consider now the set $\mathcal{S}_\rho^c$ it is clear that uniqueness of $\alpha_\star$ is sufficient to guarantee that

$$\inf_{\substack{0\leq\alpha\leq K_\alpha,\\ \eta\in\mathbb{R},\tau>0}} \max_{0\leq\beta\leq K_\beta} \mathcal{O}(\alpha,\tau,\eta,\beta) < \inf_{\substack{\alpha\in\mathcal{S}_\rho^c,\\ \eta\in\mathbb{R},\tau>0}} \max_{0\leq\beta\leq K_\beta} \mathcal{O}(\alpha,\tau,\eta,\beta) \tag{3.30}$$

which is nothing more than the deterministic comparison (3.9) presented in the theorem's discussion of the previous section, but contextualized in our scenario.

Now, due to (3.30), if we prove that the optimal value of the scalar version of the modified (AO) problem in (3.23) satisfies

$$\inf_{\substack{0\leq\alpha\leq K_\alpha,\\ \eta\in\mathbb{R},\tau>0}} \max_{0\leq\beta\leq K_\beta} \mathcal{O}_d(\alpha,\tau,\eta,\beta) \xrightarrow{P} \inf_{\substack{0\leq\alpha\leq K_\alpha,\\ \eta\in\mathbb{R},\tau>0}} \max_{0\leq\beta\leq K_\beta} \mathcal{O}(\alpha,\tau,\eta,\beta) \tag{3.31}$$

and that the same one when additionally restricted to $\alpha \in \mathcal{S}_\rho^c$, for any $\rho > 0$, satisfies

$$\inf_{\substack{\alpha\in\mathcal{S}_\rho^c,\\ \eta\in\mathbb{R},\tau>0}} \max_{0\leq\beta\leq K_\beta} \mathcal{O}_d(\alpha,\tau,\eta,\beta) \xrightarrow{P} \inf_{\substack{\alpha\in\mathcal{S}_\rho^c,\\ \eta\in\mathbb{R},\tau>0}} \max_{0\leq\beta\leq K_\beta} \mathcal{O}(\alpha,\tau,\eta,\beta) \tag{3.32}$$

then we can conclude the desired result of Theorem 3.1.6 by simply applying Prop. 2.2.1.

Therefore, the next part will be devoted to proving such convergences in probability. Notice that it will be easier to prove the following version of convergence in probability in which we modified the order of optimizations by a simple application of Sion's minmax theorem (we stress again that this is possible because the objective function is convex-concave and variable β lives in a compact set)

$$\min_{\alpha\geq0,\eta\in\mathbb{R}} \max_{0\leq\beta\leq K_\beta} \inf_{\tau>0} \mathcal{O}(\alpha,\tau,\eta,\beta) < \min_{\alpha\geq0,\eta\in\mathbb{R}} \max_{0\leq\beta\leq K_\beta} \inf_{\tau>0} \mathcal{O}(\alpha,\tau,\eta,\beta). \tag{3.33}$$

Step 6: We prove the two convegences in probability and we conclude the proof. The main ingredient that we will adopt to prove these convergences is [27, Lemma 10] which allows us to prove the convergence of the infimum of a sequence of convex converging stochastic processes. Indeed, if we know that a sequence of stochastic processes converges pointwise in probability to a deterministic function, it suffices to show that this function is level-bounded to extend the convergence also to the respective infima. Level-boundedness condition for a convex function G is equivalent to say that exists $z > 0$ such that $G(x) > \inf_{x>0} G(x)$ for all $x \geq z$.

First, for fixed $(\alpha \geq 0, \eta \in \mathbb{R}, \beta > 0)$, $\{\mathcal{O}_d(\alpha, \cdot, \eta, \beta)\}_{d\in\mathbb{N}}$ is a family of real-valued convex stochastic functions defined on the open interval $(0, +\infty)$ converging pointwise (in probability, for every $\tau > 0$) to the deterministic function $\mathcal{O}(\alpha, \cdot, \eta, \beta)$. Thus, if we prove that $\mathcal{O}(\alpha, \tau, \eta, \beta)$ is level-bounded in $\tau > 0$ we can apply the lemma to conclude that

$$\inf_{\tau>0} \ \mathcal{O}_d(\alpha, \tau, \eta, \beta) \xrightarrow{P} \inf_{\tau>0} \ \mathcal{O}(\alpha, \tau, \eta, \beta). \tag{3.34}$$

The condition $\mathcal{O}(\alpha, \tau, \eta, \beta)$ being level-bounded is equivalent, by [27, Lemma 11], to proving the following limit

$$\lim_{\tau\to+\infty} \mathcal{O}(\alpha, \tau, \eta, \beta) = +\infty \tag{3.35}$$

or equivalently

$$\lim_{\tau\to+\infty} \frac{\beta}{2} + \frac{1}{\tau}\mathcal{K}(\alpha, \tau/\beta\delta; \eta) > 0. \tag{3.36}$$

But this is immediate noticing that $\beta/2 > 0$ and that the Moreau envelope is always non-negative which in turn implies that the expected Moreau envelope is itself non-negative.

Let us define $\mathcal{O}_d^\tau(\alpha, \eta, \beta) \doteq \inf_{\tau>0} \mathcal{O}_d(\alpha, \tau, \eta, \beta)$ and $\mathcal{O}^\tau(\alpha, \eta, \beta) \doteq \inf_{\tau>0} \mathcal{O}(\alpha, \tau, \eta, \beta)$. Consider for now the case $\beta > 0$ and until further notice restrict to the case $\alpha > 0$. For fixed $(\alpha > 0, \eta \in \mathbb{R})$, $\{\mathcal{O}_d^\tau(\alpha, \eta, \cdot)\}_{d\in\mathbb{N}}$ is a sequence of real-valued stochastic concave functions (minimization over $\tau > 0$ of a concave in β function) defined in (0,$+\infty$) converging pointwise (in probability, for every β) to the deterministic function $\mathcal{O}^\tau(\alpha, \eta, \cdot)$ by (3.34).

Thus, if we prove $\lim_{\beta\to+\infty} \mathcal{O}^\tau(\alpha, \eta, \beta) = -\infty$ we can use again [27, Lemma 10] and conclude

$$\sup_{\beta>0} \ \mathcal{O}_d^\tau(\alpha, \eta, \beta) \xrightarrow{P} \sup_{\beta>0} \ \mathcal{O}^\tau(\alpha, \eta, \beta). \tag{3.37}$$

First notice that $-\alpha\beta \to -\infty$ as $\beta \to +\infty$. Then, let us consider the sequence $\{\tau\}_j \to 0^+$. Along this sequence the remaining terms $\beta\tau/2 + \mathcal{K}(\alpha, \tau/\beta\delta; \eta)$ converge to $\mathbb{E}[\lim_{\tau\to0^+} e_k(\alpha G + Z, \tau/\beta\delta)]$ where we have used monotone convergence theorem to flip limit and expectation. This is possible because, calling $n = 1/\tau$, we have that if $n_1 < n_2$ $(\tau_1 > \tau_2)$ then $e_k(\cdot\ , 1/n_1) \leq e_k(\cdot\ , 1/n_2)$ since the Moreau envelope is monotone non-increasing in τ.

Indeed, the derivative with respect to τ of the Moreau envelope is always negative $\forall\tau$

$$\frac{\partial e_k(\chi, \tau)}{\partial\tau} = -\frac{1}{2\tau^2}(\chi - prox_k(\chi; \tau))^2 \leq 0 \tag{3.38}$$

where $prox_k(\chi; \tau)$ is the proximal mapping defined in A.0.6. Recall also that by [22, Theorem 1.25]

$$\lim_{\tau\to0^+} e_k(\alpha G + Z, \tau/\beta\delta) = \ell(\alpha G + Z) + \frac{1}{2\lambda}(\ell(\alpha G + Z) - \eta)^2. \tag{3.39}$$

Therefore, if we can prove that the previous quantity is finite when taking the expectation we conclude because

$$\inf_{\tau>0} \mathcal{O}(\alpha,\tau,\eta,\beta) \leq \lim_{\tau\to 0^+} \mathcal{O}(\alpha,\tau,\eta,\beta) = -\alpha\beta + \mathbb{E}\left[\lim_{\tau\to 0^+} e_k(\alpha G+Z,\tau/\beta\delta)\right] \tag{3.40}$$

and then taking the limit as $\beta \to +\infty$ both sides we have that the RHS goes to $-\infty$ and thus we get the desired result $\lim_{\beta\to+\infty} \mathcal{O}^\tau(\alpha,\eta,\beta) = -\infty$.

But proving that $\mathbb{E}[\ell(\alpha G+Z)+\frac{1}{2\lambda}(\ell(\alpha G+Z)-\eta)^2] < +\infty$ is quite straightforward using Assumption 3.1.2(iii) and the hypothesis that both G and Z have finite second moments which in turn implies finite first moments.

Finally, consider the case $\beta = 0$ to conclude

$$\sup_{\beta\geq 0} \mathcal{O}_d^\tau(\alpha,\eta,\beta) \xrightarrow{P} \sup_{\beta\geq 0} \mathcal{O}^\tau(\alpha,\eta,\beta). \tag{3.41}$$

In this case $\mathcal{O}_d^\tau(\alpha,\eta,0) = \frac{1}{n}\sum_{i=0}^n \min_{v_i\in\mathbb{R}}\{\frac{1}{2\lambda}(\ell(v_i)-\eta)^2+\ell(v_i)\}$ which by WLLN converges in probability to $\mathcal{O}^\tau(\alpha,\eta,0) = \mathbb{E}[\min_x\{\frac{1}{2\lambda}(\ell(x)-\eta)^2+\ell(x)\}]$ since $\min_x\{\frac{1}{2\lambda}(\ell(x)-\eta)^2+\ell(x)\}$ is trivially absolutely integrable (use again $\min_x \ell(x) = \ell(0) = 0$). Notice that the expectation is another time computed over the joint distribution $\mathcal{N}(0,1)\otimes\mathbb{P}_Z$.

Let us now define $\mathcal{O}_d^{\tau,\beta}(\alpha,\eta) \doteq \sup_{\beta\geq 0}\mathcal{O}_d^\tau(\alpha,\eta,\beta)$ and $\mathcal{O}^{\tau,\beta}(\alpha,\eta) \doteq \sup_{\beta\geq 0}\mathcal{O}(\alpha,\eta,\beta)$. Recall now Remark 3.2.2 which says that the optimal $\eta_\star$ is actually positive. This enable us to restrict the minimization over $\eta \geq 0$.

Consider for now the case $\eta > 0$. For fixed $\alpha > 0$, $\{\mathcal{O}_d^{\tau,\beta}(\alpha,\cdot)\}_{d\in\mathbb{N}}$ is a sequence of real-valued stochastic convex functions (since they were obtained by first minimizing over τ a jointly convex function in (α,τ,η), and then maximizing over β a jointly convex function in (α,η)) defined in (0,+∞) converging pointwise (in probability, for every η) to the deterministic function $\mathcal{O}^{\tau,\beta}(\alpha,\cdot)$ by (3.41).

Following the same line as before, if we prove that $\lim_{\eta\to+\infty}\mathcal{O}^{\tau,\beta}(\alpha,\eta) = +\infty$ we can use [27, Lemma 10] to conclude

$$\inf_{\eta>0} \mathcal{O}_d^{\tau,\beta}(\alpha,\eta) \xrightarrow{P} \inf_{\eta>0} \mathcal{O}^{\tau,\beta}(\alpha,\eta). \tag{3.42}$$

To prove the previous limit we can argue as follows

$$\begin{aligned}\mathcal{O}^{\tau,\beta}(\alpha,\eta) &= \sup_{\beta\geq 0}\inf_{\tau>0} \frac{\beta\tau}{2} - \alpha\beta + \mathcal{K}(\alpha,\tau/\beta\delta;\eta) \geq \inf_{\tau>0} \mathcal{K}(\alpha,\tau/0;\eta) = \\ &= \mathbb{E}\left[\min_x\left\{\ell(x)+\frac{1}{2\lambda}(\ell(x)-\eta)^2\right\}\right] = \mathbb{E}\left[\min_x\left\{\ell(x)+\frac{1}{2\lambda}(\ell(x)^2-2\eta\ell(x)+\eta^2)\right\}\right] \geq \\ &\geq \mathbb{E}\left[\min_x\left\{-\frac{\eta}{\lambda}\ell(x)\right\}\right] + \frac{\eta^2}{\lambda} = \frac{\eta^2}{\lambda} \to +\infty, \text{ as } \eta\to+\infty\end{aligned} \tag{3.43}$$

where the first inequality holds because the supremum over β is for sure greater than the same function evaluated in $\beta = 0$ (specifically the function we are evaluating is $\mathcal{O}^{\tau}(\alpha, \eta, \beta)$). After that, we can discard the infimum over τ because there is no more dependency on this variable. Finally, last inequality follows because if $f(x) \geq g(x)\ \forall x$, then the same relation holds for the minimums. Therefore, since $\ell(x) + \frac{1}{2\lambda}(\ell(x)^2 - 2\eta\ell(x) + \eta^2) \geq -\frac{\eta}{\lambda}\ell(x)$ the same inequality holds when considering the minima. Thus, we can conclude that also $\lim_{\eta \to +\infty} \mathcal{O}^{\tau,\beta}(\alpha, \eta) = +\infty$.

Consider now the case $\eta = 0$ where

$$\mathcal{O}^{\tau,\beta}(\alpha, 0) = \sup_{\beta \geq 0} \inf_{\tau > 0} \frac{\beta\tau}{2} - \alpha\beta + \mathbb{E}\left[\min_{x}\left\{\ell(x) + \frac{1}{2\lambda}\ell(x)^2 + \frac{\beta\delta}{2\tau}(\alpha G + Z - x)^2\right\}\right]. \tag{3.44}$$

We know that for every $\eta \in \mathbb{R}$ we have the convergence in probability in (3.41). Therefore, the same holds when $\eta = 0$ concluding the following convergence

$$\inf_{\eta \geq 0} \mathcal{O}_d^{\tau,\beta}(\alpha, \eta) \xrightarrow{P} \inf_{\eta \geq 0} \mathcal{O}^{\tau,\beta}(\alpha, \eta). \tag{3.45}$$

Finally, define $\mathcal{O}_d^{\tau,\beta,\eta}(\alpha) \doteq \inf_{\eta \geq 0} \mathcal{O}_d^{\tau,\beta}(\alpha, \eta)$ and $\mathcal{O}^{\tau,\beta,\eta}(\alpha) \doteq \inf_{\eta \geq 0} \mathcal{O}(\alpha, \eta)$. These functions are convex in α because they were obtained by minimizing jointly convex in (α, η) functions. Moreover, by (3.45) we know that $\mathcal{O}_d^{\tau,\beta,\eta}(\alpha) \xrightarrow{P} \mathcal{O}^{\tau,\beta,\eta}(\alpha)$ pointwise for every $\alpha \geq 0$. Thus, since we have assumed that $\alpha_\star$ is the unique minimizer, level-boundedness is satisfied and we can ensure that

$$\min_{\alpha > 0} \mathcal{O}_d^{\tau,\beta,\eta}(\alpha) \xrightarrow{P} \min_{\alpha > 0} \mathcal{O}^{\tau,\beta,\eta}(\alpha). \tag{3.46}$$

This is equivalent to the convergence in probability in (3.31) (still the case $\alpha = 0$ to be done).

A similar reasoning holds also for the convergence in probability in (3.32) where the last convergence of the minimization over $\alpha \in \mathcal{S}_\eta^c$ is obtained applying a generalization of [27, Lemma 10] provided by [2, Lemma A.3]. With this we conclude the section.

3.3 Choice of λ.

In this section we are going to investigate how the choice of λ affects the problem. In particular, up to now we have considered λ as a constant, or better as if its value is proportional to the ratio n/d. However, in the literature, [11] and [19] obtained some results with χ^2-divergence among which there is also the variance regularization approximation using a radius that shrinks as the number of measurements increases. As already explained in Chapter 2, this is reasonable but for our best understanding it holds when considering a problem with fixed dimension d. In our case also d grows to infinity thus preventing us to use $\lambda = \lambda_0 d$.

In the next we will show that actually considering $\lambda = \lambda_0 d$ leads us to a final scalar convex-concave optimization that does not encode anymore the parameter λ_0 and instead it is the same as solving the problem starting from the simple ERM in (1.2). As side note we would like to point out that in the case of λ growing with d it is actually possible to prove that the convex conjugate of χ^2-divergence in our problem is just $f^*(s) = \frac{s^2}{2} + s$ w.p.a. 1.

Going back to the previous section, before going into the convergence analysis we arrived to an objective function containing the following term

$$\frac{1}{n}\sum_{i=1}^{n} \inf_{v_i \in \mathbb{R}} \left\{ \frac{1}{2\lambda} \left(\ell(v_i) - \eta\right)^2 + \ell(v_i) + \frac{\beta\delta}{2\tau}(\alpha g_i + z_i - v_i)^2 \right\}. \tag{3.47}$$

Now, if we pick λ not as a constant but $\lambda = \lambda_0 d$ we can look at the minimization over v_i and immediately see that the first squared term is weighted by $1/d$ while the remaining terms not. This implies that when d is large, despite the choice of v_i, the first term is close to zero (this is true because ℓ is a proper function therefore there exist values of v_i for which the first term is not identically $+\infty$) and hence it is not relevant when considering the minimum over v_i. Therefore, we can equivalently consider the following when $d \to +\infty$

$$\frac{1}{n}\sum_{i=1}^{n} \inf_{v_i \in \mathbb{R}} \left\{ \ell(v_i) + \frac{\beta\delta}{2\tau}(\alpha g_i + z_i - v_i)^2 \right\}. \tag{3.48}$$

However, one can notice that in the previous expression there is no more the regularization parameter and actually this is the same thing we can derive from the ERM in (1.2) as we will do in the next chapter.

Chapter 4

Numerical simulations

In this chapter we will validate our findings with some experiments involving the solution of the scalar minmax problem in (3.5) compared with its ERM counterpart. We are going to perform numerical simulations involving different levels of noise and also in different scenarios of under and over-parametrization.

Before presenting our results, we want to determine the scalar problem resulting from the ERM in (1.2) in order to make a comparison with the one coming from χ^2-divergence DRE. Without going trough all the details which we have discussed previously, we will show the derivation of the scalar problem following a similar analysis of the previous chapter.

We start from (1.2) and with the same change of variable $\mathbf{w} = (\theta - \theta_0)/\sqrt{d}$ we obtain

$$\inf_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}} \frac{1}{n}\sum_{i=0}^{n} \ell(z_i - \sqrt{d}x_i^T\mathbf{w}). \tag{4.1}$$

Next, we can perform another change of variable $\mathbf{v} = \mathbf{z} - \sqrt{d}\mathbf{X}\mathbf{w}$ which leads to

$$\begin{gathered}\inf_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}},\mathbf{v}\in\mathbb{R}^n} \frac{1}{n}\sum_{i=0}^{n} \ell(v_i) \\ \text{s.t. } \mathbf{v} = \mathbf{z} - \sqrt{d}\mathbf{X}\mathbf{w}\end{gathered} \tag{4.2}$$

and using Lagrangian duality to bring the equality constraint into the objective function with associated Lagrange multiplier $\mathbf{u} \in \mathbb{R}^n$ we obtain the following problem which we can think as (PO) problem

$$\inf_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}},\mathbf{v}\in\mathbb{R}^n} \max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}} -\frac{1}{\sqrt{d}}\mathbf{u}^T(\sqrt{d}\mathbf{X})\mathbf{w} + \frac{1}{\sqrt{d}}\mathbf{u}^T\mathbf{z} - \frac{1}{\sqrt{d}}\mathbf{u}^T\mathbf{v} + \frac{1}{n}\sum_{i=0}^{n}\ell(v_i). \tag{4.3}$$

The next step is to derive the (AO) problem and its associated modified version which takes the form

$$\max_{0\le\beta\le K_\beta} \inf_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}},\mathbf{v}\in\mathbb{R}^n} \max_{\|\mathbf{u}\|=\beta} \frac{1}{\sqrt{d}}\left\{\|\mathbf{w}\|\mathbf{g}^T\mathbf{u} - \|\mathbf{u}\|\mathbf{h}^T\mathbf{w} + \mathbf{u}^T\mathbf{z} - \mathbf{u}^T\mathbf{v}\right\} + \frac{1}{n}\sum_{i=1}^{n}\ell(v_i) \tag{4.4}$$

and using (3.17) we get

$$\max_{0\leq\beta\leq K_\beta} \min_{\mathbf{w}\in\mathcal{S}_\mathbf{w},\mathbf{v}\in\mathbb{R}^n} \frac{\beta}{\sqrt{d}}\|\|\mathbf{w}\|\mathbf{g}+\mathbf{z}-\mathbf{v}\| - \frac{\beta}{\sqrt{d}}\mathbf{h}^T\mathbf{w} + \frac{1}{n}\sum_{i=0}^{n}\ell(v_i). \tag{4.5}$$

Doing the same over $\mathbf{w}$, we can write

$$\max_{0\leq\beta\leq K_\beta} \min_{0\leq\alpha\leq K_\alpha,\mathbf{v}\in\mathbb{R}^n} \frac{\beta}{\sqrt{d}}\|\alpha\mathbf{g}+\mathbf{z}-\mathbf{v}\| - \frac{\alpha\beta}{\sqrt{d}}\|\mathbf{h}\| + \frac{1}{n}\sum_{i=0}^{n}\ell(v_i). \tag{4.6}$$

Finally, using again the square-root trick we arrive to

$$\max_{0\leq\beta\leq K_\beta} \min_{0\leq\alpha\leq K_\alpha,\tau>0} \frac{\beta\tau}{2} - \alpha\beta\frac{\|\mathbf{h}\|}{\sqrt{d}} + \frac{1}{n}\sum_{i=0}^{n}\min_{v_i\in\mathbb{R}}\left\{\ell(v_i)+\frac{\beta\delta}{2\tau}(\alpha g_i+z_i-v_i)^2\right\} \tag{4.7}$$

which can be easily shown (simple application of WLLN) to converge in probability to the following optimization

$$\max_{0\leq\beta\leq K_\beta} \min_{0\leq\alpha\leq K_\alpha,\tau>0} \frac{\beta\tau}{2} - \alpha\beta + \mathbb{E}_{\substack{G\sim\mathcal{N}(0,1),\\ Z\sim\mathbb{P}_Z}}\left[\min_{x\in\mathbb{R}}\left\{\ell(x)+\frac{\beta\delta}{2\tau}(\alpha G+Z-x)^2\right\}\right]. \tag{4.8}$$

Example 4.0.1 (Moreau envelope for the LAD estimator)**.** *Consider the LAD estimator where the loss function is the absolute value, i.e. $\ell(\cdot)=|\cdot|$. To recover the expression of the Moreau envelope $e_\ell(c,\tau)$ we will make use of the following relationship between the Moreau envelope of ℓ and its convex conjugate ℓ^*,*

$$e_\ell(c,\tau)+e_{\ell^*}\left(\frac{c}{\tau},\frac{1}{\tau}\right)=\frac{c^2}{2\tau}. \tag{4.9}$$

Now, given the convex conjugate of the absolute value function

$$\ell^*(v)=\sup_{u\in\mathbb{R}} uv-|u| = \begin{cases} 0 & \textit{if} \quad |v|\leq 1 \\ \infty & \textit{otherwise,} \end{cases} \tag{4.10}$$

we can compute its Moreau envelope

$$\begin{aligned} e_{\ell^*}\left(\frac{c}{\tau},\frac{1}{\tau}\right) &= \min_{u\in\mathbb{R}}\ell^*(u)+\frac{\tau}{2}\left(\frac{c}{\tau}-u\right)^2 = \min_{|u|\leq 1}\frac{\tau}{2}\left(\frac{c}{\tau}-u\right)^2 \\ &= \begin{cases} 0 & \textit{if} \quad |c|\leq\tau \\ \tau(c/\tau-\textit{Sign}(c))^2/2 & \textit{otherwise,} \end{cases} \end{aligned} \tag{4.11}$$

from which we can easily derive

$$e_\ell(c,\tau)=\begin{cases} c^2/2\tau & \textit{if} \quad |c|\leq\tau \\ |c|-\tau/2 & \textit{otherwise.} \end{cases} \tag{4.12}$$

For the simulations, we selected the LAD estimator which considers as loss the absolute value function. Since the expected Moreau envelopes, both for ERM and DRE, require the calculation of an integral coming from the expected value over the joint distribution $\mathcal{N}(0,1) \otimes \mathbb{P}_Z$ we decided to approximate this using "*Sample Average Approximation*". This method is quite simple, it consists in drawing a random sample W and approximating the expected value function by the corresponding sample average function. The obtained sample average optimization problem is solved, and the procedure is iterated several times until convergence, [17]. In our case, we took two random samples of size 5000 composed by i.i.d. points from a standard gaussian and from the noise distribution which we assumed to be gaussian with zero mean and different levels of variance. Notice that we can consider two different samples because the two distributions are assumed to be independent.

To simulate the high-dimensional regime we are interested in, we decided to pick $n = 2500$. Since we would like to quantify the estimation error we also selected a value for θ_0, in particular each entry of this vector is i.i.d. from the following distribution $p_x(x) = 0.9\delta(x) + 0.1\phi(x)/\sqrt{0.1}$ with $\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$. In simple words, 90% of the entries is 0 while the remaining 10% is drawn from a gaussian distribution with zero mean and variance 10. As for the noise, we added to each measurement an additive stochastic noise drawn from a Gaussian with zero mean and different values of variance, $[0.01, 0.1]$.

At this point we need to compute the Moreau envelope

$$\inf_{x\in\mathbb{R}} \left\{ \frac{1}{2\lambda}\left(|x| - \eta\right)^2 + |x| + \frac{\beta\delta}{2\tau}(c-x)^2 \right\}. \tag{4.13}$$

To do so we can simply calculate the minimum of that expression and then substitute back the corresponding value. Since the function is convex, we can set the derivative equal to zero to find the minimum. We just need to be careful because the absolute value is not differentiable in the origin.

The value of the minimum that we obtain is therefore

$$x_\star = \begin{cases} |\beta\lambda\delta \cdot c - \lambda\tau + \eta\tau|/(\tau + \beta\lambda\delta) & \textit{if} \quad |c| \geq \tau\frac{\lambda-\eta}{\beta\lambda\delta} \\ 0 & \textit{otherwise.} \end{cases} \tag{4.14}$$

In order to solve the scalar minmax problem both for ERM (4.8) and χ^2-divergence DRE (3.5) we adopted a gradient ascent-descent approach in which at each minimizer iteration we performed 500 maximizer iterations over the variable β. The learning algorithm adopted is *Adam* from *PyTorch*.

In the following we can observe various plots of our simulations. What we would expect from the theory is that for small values of λ our robust approach performs worse or similar to ERM because it is overly conservative. Then, DRE should perform better as λ increases meaning that when there is additive stochastic noise our approach is robust against it while simple risk minimizer not. Lastly, as λ grows toward infinity our approach should converge to ERM because the only admissible distribution in the ambiguity set is the empirical center.

Looking at the figures we can clearly see that this is actually what happens. Indeed, for smaller values of λ the green line is greater or close to the red line of the ERM. Then, as λ grows, we have a decrease in the green line which guarantees that the robust approach performs better while after a certain value it starts to increase again converging finally to the ERM value.

Looking at Fig. 4.1 and Fig. 4.2 it is evident the benefit brought by our robust approach which for a range of λ values performs better compared to ERM, while if we take λ sufficiently large our approach is equivalent to ERM as we expected. If instead we look at Fig. 4.3 and Fig. 4.4 we can see a similar behaviour even though in this case the noise is smaller and hence our robust approach does not really outperform standard ERM.

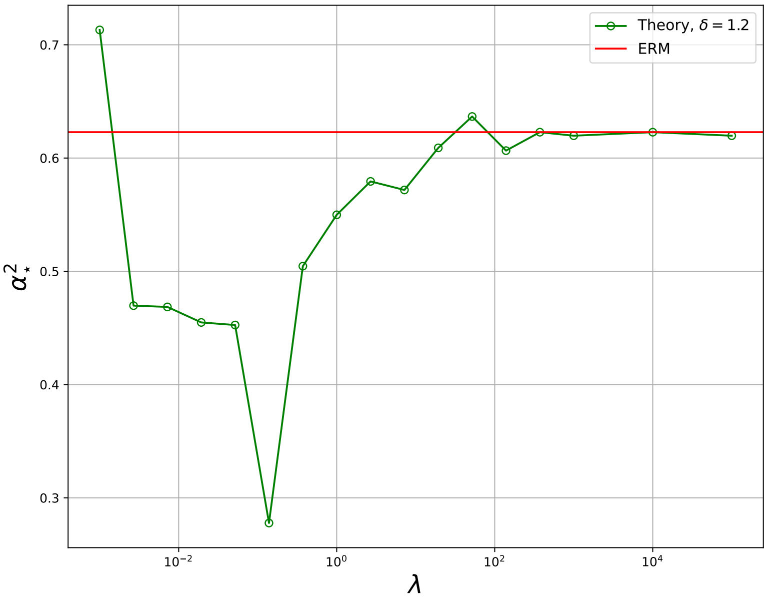

Figure 4.1: Comparison between DRE and ERM. In this plot we have in green the mean squared error for the DRE while in red the same error for the ERM. The noise is drawn from a Gaussian distribution with zero mean and variance 0.1. The ratio $n/d = 1.2$.

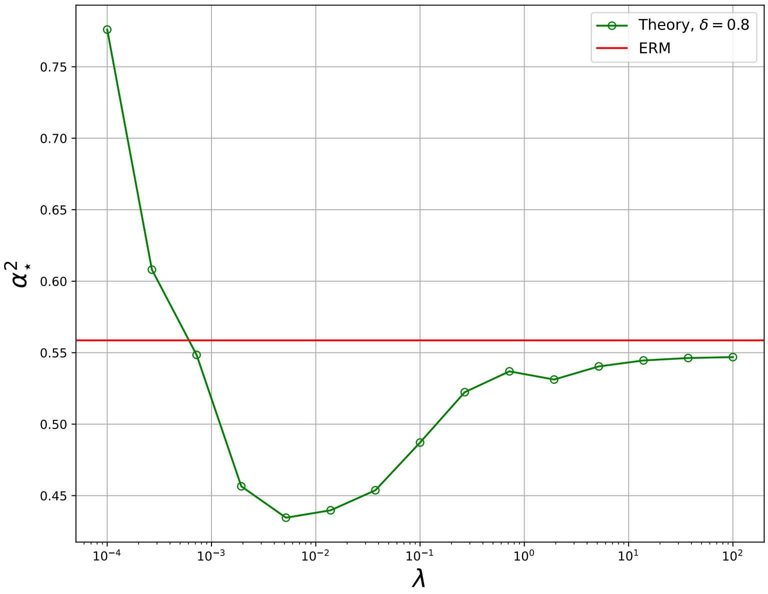

Figure 4.2: Comparison between DRE and ERM. In this plot we have in green the mean squared error for the DRE while in red the same error for the ERM. The noise is drawn from a Gaussian distribution with zero mean and variance 0.1. The ratio $n/d = 0.8$.

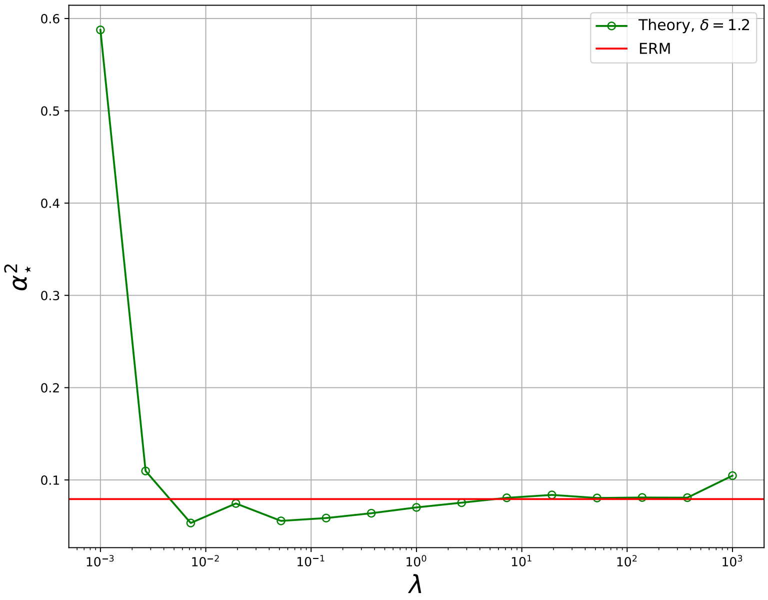

Figure 4.3: Comparison between DRE and ERM. In this plot we have in green the mean squared error for the DRE while in red the same error for the ERM. The noise is drawn from a Gaussian distribution with zero mean and variance 0.01. The ratio $n/d = 1.2$.

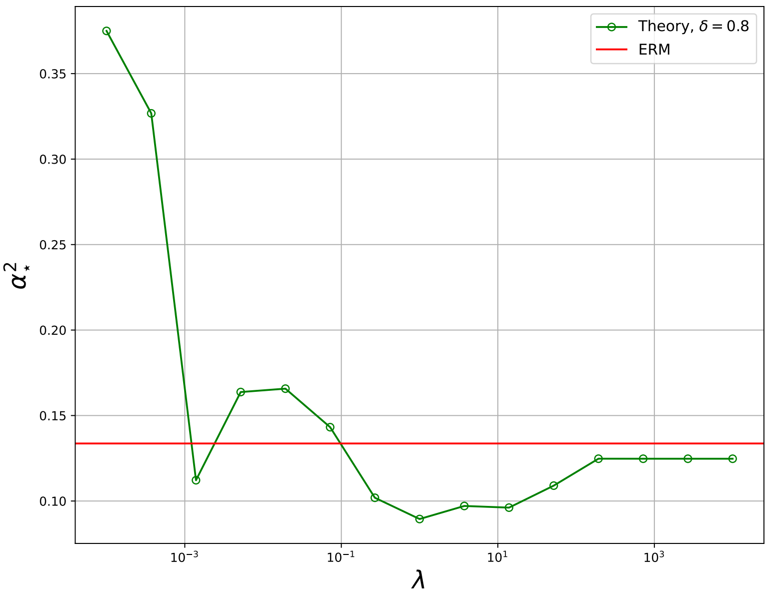

Figure 4.4: Comparison between DRE and ERM. In this plot we have in green the mean squared error for the DRE while in red the same error for the ERM. The noise is drawn from a Gaussian distribution with zero mean and variance 0.01. The ratio $n/d = 0.8$.

Bibliography

[1] S. M. Ali and S. D. Silvey. “A General Class of Coefficients of Divergence of One Distribution from Another”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 28.1 (1966), pp. 131–142. (Visited on 08/08/2022).

[2] Liviu Aolaritei, Soroosh Shafieezadeh-Abadeh, and Florian Dörfler. “The Performance of Wasserstein Distributionally Robust M-Estimators in High Dimensions”. In: *arXiv preprint arXiv:2206.13269* (2022).

[3] Robert Bassett and Julio Deride. “One-Step Estimation with Scaled Proximal Methods”. In: *Mathematics of Operations Research* (2021).

[4] Aharon Ben-Tal et al. “Robust solutions of optimization problems affected by uncertain probabilities”. In: *Management Science* 59.2 (2013), pp. 341–357.

[5] Dimitri Bertsekas. *Convex optimization theory*. Vol. 1. Athena Scientific, 2009.

[6] Su Lin Blodgett, Lisa Green, and Brendan O’Connor. “Demographic dialectal variation in social media: A case study of African-American English”. In: *arXiv preprint arXiv:1608.08868* (2016).

[7] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[8] Gustavo Camps-Valls et al. “Advances in hyperspectral image classification: Earth monitoring with statistical learning methods”. In: *IEEE signal processing magazine* 31.1 (2013), pp. 45–54.

[9] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[10] Imre Csiszár. “Eine informationstheoretische ungleichung und ihre anwendung auf beweis der ergodizitaet von markoffschen ketten”. In: *Magyer Tud. Akad. Mat. Kutato Int. Koezl.* 8 (1964), pp. 85–108.

[11] John Duchi, Peter Glynn, and Hongseok Namkoong. “Statistics of robust optimization: A generalized empirical likelihood approach”. In: *arXiv preprint arXiv:1610.03425* (2016).

[12] John C. Duchi and Hongseok Namkoong. “Learning models with uniform performance via distributionally robust optimization”. In: *The Annals of Statistics* 49.3 (2021), pp. 1378–1406.

[13] Rick Durrett. *Probability: theory and examples.* Vol. 49. Cambridge university press, 2019.

[14] Gerald B Folland. *Real analysis: modern techniques and their applications.* Vol. 40. John Wiley & Sons, 1999.

[15] Yehoram Gordon. "Some inequalities for Gaussian processes and applications". In: *Israel Journal of Mathematics* 50.4 (1985), pp. 265–289.

[16] Dirk Hovy and Anders Søgaard. "Tagging performance correlates with author age". In: *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 2: Short papers).* 2015, pp. 483–488.

[17] Anton J Kleywegt, Alexander Shapiro, and Tito Homem-de-Mello. "The sample average approximation method for stochastic discrete optimization". In: *SIAM Journal on Optimization* 12.2 (2002), pp. 479–502.

[18] Michael Lustig et al. "Compressed sensing MRI". In: *IEEE signal processing magazine* 25.2 (2008), pp. 72–82.

[19] Hongseok Namkoong and John C Duchi. "Variance-based Regularization with Convex Objectives". In: *Advances in Neural Information Processing Systems.* Curran Associates, Inc., 2017.

[20] Sahand N Negahban et al. "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers". In: *Statistical science* 27.4 (2012), pp. 538–557.

[21] Alfréd Rényi et al. "On measures of entropy and information". In: *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability.* Vol. 1. 547-561. Berkeley, California, USA. 1961.

[22] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis.* Vol. 317. Springer Science & Business Media, 2009.

[23] Piotr Sapiezynski, Valentin Kassarnig, and Christo Wilson. "Academic performance prediction in a gender-imbalanced environment". In: *Proceedings of the Eleventh ACM Conference on Recommender Systems.* Vol. 1. 2017, pp. 48–51.

[24] Alexander Shapiro. "Distributionally robust stochastic programming". In: *SIAM Journal on Optimization* 27.4 (2017), pp. 2258–2275.

[25] Maurice Sion. "On general minimax theorems". In: *Pacific Journal of Mathematics* 8 (1958), pp. 171–176.

[26] Rachael Tatman. "Gender and dialect bias in YouTubes automatic captions". In: *Proceedings of the first ACL workshop on ethics in natural language processing.* 2017, pp. 53–59.

[27] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. “Precise error analysis of regularized M-estimators in high dimensions”. In: *IEEE Transactions on Information Theory* 64.8 (2018), pp. 5592–5628.

[28] A. N. Tikhonov. “On the stability of inverse problems”. In: *Proceedings of the USSR Academy of Sciences* 39 (1943), pp. 195–198.

[29] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

[30] Qi Wang et al. “A comprehensive survey of loss functions in machine learning”. In: *Annals of Data Science* 9.2 (2022), pp. 187–212.

Appendix A

Basic Concepts of Convex Analysis

Let us have a brief overview of what are the main results from convex analysis adopted in the thesis. Everything reported in this section can also be found in standard convex analysis books. As main references we will use [5], [7] and [22].

Definition A.0.1. *A set $\mathcal{C} \subset \mathbb{R}^n$ is called convex if*

$$\alpha x + (1-\alpha)y \in \mathcal{C}, \quad \forall \alpha \in [0,1], \quad \forall x, y \in \mathcal{C}.$$

Definition A.0.2. *Let $\mathcal{C}$ be a convex subset of $\mathbb{R}^n$. We say that a function $f : \mathcal{C} \to \mathbb{R}$ is convex if*

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y), \quad \forall x, y \in \mathcal{C}, \quad \forall \alpha \in [0,1].$$

Similarly we say that a function is concave if $-f$ is convex. Examples of convex functions are affine functions of the form $f(x) = ax + b$ where $a \in \mathbb{R}^n$, $b \in \mathbb{R}$ or any norm $\|\cdot\|$. We can introduce variants of the definition of convex function. For example, we can say that a convex function f is *strictly convex* if the inequality in the definition holds strictly for all $x, y \in \mathcal{C}$, $x \neq y$ and for all $\alpha \in (0,1)$.

We want also to describe some operations that preserve convexity because they are used throughout the thesis.

Proposition A.0.1. *[5, Proposition 1.1.5] Let $f_i : \mathbb{R}^n \mapsto (-\infty, +\infty]$, $i = 1, \ldots, m$, be given functions, let $\gamma_1, \ldots, \gamma_m$ be positive scalars, and let $F : \mathbb{R}^n \mapsto (-\infty, +\infty]$ be the function*

$$F(x) = \gamma_1 f_1(x) + \cdots + \gamma_m f_m(x), \quad x \in \mathbb{R}^n.$$

If $f_1, \ldots, f_m$ are convex, then F is also convex.

Proposition A.0.2. *[7] If for each $y \in A$, $f(x,y)$ is convex in x, then the function g, defined as*

$$g(x) = \sup_{y \in A} f(x,y)$$

is convex in x.

Proposition A.0.3. *[5, Proposition 3.3.1] Consider a function $F : \mathbb{R}^{n+m} \mapsto (-\infty, +\infty]$ and the function $f : \mathbb{R}^n \mapsto (-\infty, +\infty]$ defined by*

$$f(x) = \inf_{z \in \mathbb{R}^m} F(x, z).$$

then f is convex if F is jointly convex.

Generally we prefer to deal with convex functions that are real-valued and defined over the entire space $\mathbb{R}^n$ rather than just a convex subset of it because they are simpler to study. In some cases this is not possible (e.g. $f : (0, \infty) \mapsto \mathbb{R}$ defined by $f(x) = 1/x$), but it may be convenient to extend the domain to be $\mathbb{R}^n$ and let the function take infinite values.
When dealing with *extended real-valued* functions the forbidden sum $+\infty - \infty$ can arise when checking convexity. To avoid this kind of problem we can give another characterization for convex functions in term of *epigraphs*, intuitively the set of points above the graph of a function.

Definition A.0.3. *The epigraph of a function $f : \mathcal{X} \mapsto [-\infty, +\infty]$, where $\mathcal{X} \in \mathbb{R}^n$ is defined to be the subset of $\mathbb{R}^{n+1}$ given by*

$$epi(f) = \{(x, w) \mid x \in \mathcal{X}, w \in \mathbb{R}, f(x) \leq w\}.$$

We then say that a function is convex if its epigraph is a convex subset of $\mathbb{R}^{n+1}$.

When dealing with extended real-valued functions we can introduce also other important notions that are not relevant when the function cannot take infinite values, namely effective domain and properness.

The *effective domain* of f is defined to be the set

$$dom(f) = \{x \in \mathcal{X} \mid f(x) < \infty\}.$$

It can be seen that the effective domain is obtainable by projecting the epigraph on $\mathbb{R}^n$. Notice also that if we restrict f only to its effective domain the epigraph remains unaffected.

A function is instead *proper* if $f(x) < +\infty$ for at least one $x \in \mathcal{X}$ and it never coincides with $-\infty$. This property is important to exclude degenerate cases. Indeed, it is not meaningful in optimization to have a function which is always $+\infty$ (true if and only if the epigraph of f is empty), neither having some points in which the function is $-\infty$ (true if and only if the epigraph at those points is a vertical line). In both cases there is nothing to optimize since the minimum can be simply found checking the function expression.

For twice differentiable convex functions, there is another characterization of convexity, given by the following proposition

Proposition A.0.4. *[5, Proposition 1.1.10] Let C be a nonempty subset of $\mathbb{R}^n$ and let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be twice continuously differentiable over an open set that contains C.*

(a) If $\nabla^2 f(x)$ is positive semidefinite for all $x \in C$, then f is convex over C.

(b) If $\nabla^2 f(x)$ is positive definite for all $x \in C$, then f is strictly convex over C.

(c) If C is open and f is convex over C, then $\nabla^2 f(x)$ is positive semidefinite for all $x \in C$.

Another very important notion adopted is the one of *Fenchel conjugate.*

Definition A.0.4. *(Legendre-Fenchel transform). For any function $f : \mathbb{R}^n \mapsto \overline{\mathbb{R}}$, the function $f^* : \mathbb{R}^n \mapsto \overline{\mathbb{R}}$ defined by*

$$f^*(v) := \sup_{x \in dom(f)} \{\langle v, x\rangle - f(x)\}$$

is conjugate to f.

Next, we present an important result that can be used when it is desirable to verify if for an optimization problem strong duality holds. The following proposition is actually a particular case when the constraints are simple linear equalities of more general results that hold when the constraints are both equalities and inequalities, possibly non-linear.

Proposition A.0.5. *(Convex Programming - Linear Equality Constraints)[5, Proposition 5.3.3]. Consider the problem*

$$\begin{aligned} &\min\ f(x) \\ &s.t.\ x \in X,\ Ax = b. \end{aligned} \tag{A.1}$$

Assume that f^* is finite and that there exists $\bar{x} \in ri(X)$ such that $A\bar{x} = b$. Then $f^* = q^*$ and there exists at least one dual optimal solution.

Sometimes it is also useful to deal with function approximations. We therefore present a method of approximating general functions in terms of “envelope functions”. But let us first present the concept of semicontinuity.

Definition A.0.5. *(Lower semicontinuity). The function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is lower semicontinuous (lsc) at $\bar{x}$ if*

$$\liminf_{x \to \bar{x}} f(x) \geq f(\bar{x}), \quad \text{or equivalently} \quad \liminf_{x \to \bar{x}} f(x) = f(\bar{x})$$

and lower semicontinuous on $\mathbb{R}^n$ if this holds for every $x \in \mathbb{R}^n$ where

$$\liminf_{x \to \bar{x}} f(x) \doteq \lim_{\delta \to 0^+} \left[\inf_{x \in \mathbb{B}(\bar{x}, \delta)} f(x) \right].$$

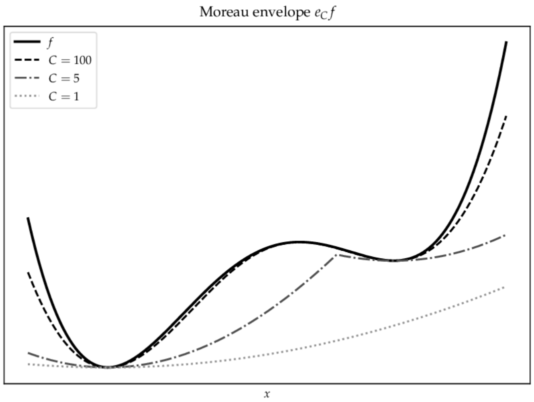

Figure A.1: Example of comparison between a function and its Moreau envelope with different parameter's values, [3] (in the paper the role of C is inverse of our λ).

Definition A.0.6. *(Moreau envelopes and proximal mappings). For a proper, lower semicontinuous function $f : \mathbb{R}^n \mapsto \overline{\mathbb{R}}$ and parameter value $\lambda > 0$, the Moreau envelope function $e_\lambda f$ and proximal mapping $P_\lambda f$ are defined by*

$$e_\lambda f(x) \doteq \inf_w \left\{ f(w) + \frac{1}{2\lambda}|w - x|^2 \right\} \leq f(x). \tag{A.2}$$

$$P_\lambda f(x) \doteq \arg\min_w \left\{ f(w) + \frac{1}{2\lambda}|w - x|^2 \right\}. \tag{A.3}$$

What we can observe is that generally $e_\lambda f$ approximates f from below and one pictorial example of it can be found in Fig. A.1. The idea is that for smaller and smaller values of λ the Moreau envelope approximates f better and better, and indeed, $1/\lambda$ can be interpreted as a penalty parameter for the constraint $w - x = 0$ in the minimization defining $e_\lambda f(x)$.

Appendix B

Basic Concepts of Probability and Real Analysis

In this section we want to give the reader some basic concepts of probability theory that are majorly utilized throughout the thesis. Moreover, we present also some important results related to real analysis which are helpful when dealing with convergence's analysis. Everything that appears in this section can be found in [13] and in [14] which we have used as main references.

Definition B.0.1. *(Convergence in probability). We say that a sequence of random variables $\{Y_n\}_{n\in\mathbb{N}}$ converges in probability to the constant Y if for all $\epsilon > 0$,*

$$\lim_{n\to\infty} Pr(|Y_n - Y| > \epsilon) = 0.$$

In this case we will say that $|Y_n - Y| \leq \epsilon$ holds with probability approaching 1 (abbreviated as w.p.a. 1).

Theorem B.0.1. *(Weak Law of Large Numbers (WLLN)). Let $X_1, X_2, \ldots$ be uncorrelated random variables with $\mathbb{E}X_i = \mu$ and $Var(X_i) < \infty$. Denoting with $S_n = X_1 + \cdots + X_n$, then as $n \to +\infty$ $S_n/n \to \mu$ in probability.*

Theorem B.0.2. *(Monotone Convergence Theorem)[14]. If $\{f\}_n$ is a sequence of measurable non-negative functions such that $f_j \leq f_{j+1}$ for all j, and $f = \lim_{n\to+\infty} f_n$, then*

$$\int f = \lim_{n\to+\infty} \int f_n.$$

Appendix C

Proofs for Chapter 3

Proof of Proposition 2.4.1. In order to prove the proposition we use duality theory to bring the inner maximization in (1.8) to a minimization problem. The reference distribution is empirical therefore we can consider both expectation and f-divergence as sums because the distribution has support on a finite number of points. Notice also that in this discrete setting the supremum over the probability measure $\mathbb{Q}$ in (1.8) becomes the supremum over a n-dimensional vector with non-negative entries which are the weights of the distribution. Therefore, we can rewrite the maximization as follows

$$\sup_{q\in\mathbb{R}^n_+} \sum_{i=0}^{n} q_i \ell(y_i - x_i^T\theta) - \lambda p_i f\left(\frac{q_i}{p_i}\right) \quad \text{s.t.} \sum_{i=0}^{n} q_i = 1. \tag{C.1}$$

Now, using a Lagrangian multiplier we can bring the equality constraint into the objective function. Thus, the Lagrangian of problem (C.1) is

$$\begin{aligned}\mathcal{L}(q,\eta) =& \sum_{i=0}^{n} \left\{ p_i \frac{q_i}{p_i} \ell(y_i - x_i^T\theta) - \lambda p_i f\left(\frac{q_i}{p_i}\right)\right\} + \eta\left(1 - \sum_{i=0}^{n} q_i\right) \\ =& \sum_{i=0}^{n} \left\{ p_i \frac{q_i}{p_i} (\ell(y_i - x_i^T\theta) - \eta) - \lambda p_i f\left(\frac{q_i}{p_i}\right)\right\} + \eta.\end{aligned} \tag{C.2}$$

The Lagrangian dual of problem (C.1) is the problem

$$\inf_{\eta\in\mathbb{R}} \sup_{q\in\mathbb{R}^n_+} \mathcal{L}(q,\eta). \tag{C.3}$$

Next we can define the likelihood ratio $L_i = \frac{q_i}{p_i}$ to reformulate our dual problem as

$$\begin{aligned}&\inf_{\eta\in\mathbb{R}} \sup_{L\in\mathbb{R}^n_+} \sum_{i=0}^{n} p_i \left\{ L_i(\ell(y_i - x_i^T\theta) - \eta) - \lambda f(L_i)\right\} + \eta = \\ &\inf_{\eta\in\mathbb{R}} \sup_{L\in\mathbb{R}^n_+} \sum_{i=0}^{n} \lambda p_i \left\{ L_i \frac{\ell(y_i - x_i^T\theta) - \eta}{\lambda} - f(L_i)\right\} + \eta\end{aligned} \tag{C.4}$$

Finally, we can swap the supremum and the sum because we are maximizing over a vector whose components appear independently in the summation.

$$\inf_{\eta\in\mathbb{R}} \sum_{i=0}^{n} \sup_{L_i\in\mathbb{R}_+} \quad \lambda p_i \left\{ L_i \frac{\ell(y_i - x_i^T\theta) - \eta}{\lambda} - f(L_i) \right\} + \eta \tag{C.5}$$

Notice that the supremum is over the non-negative real variable L_i which does not allow to use the definition of convex conjugate. However, if we extend the domain of f to be the real line and define f to be $+\infty$ in the interval $(-\infty, 0)$ as done in Section 2.3 we can use the convex conjugate given by $f^*(s) = \sup_{L_i\in\mathbb{R}}\{L_i s - f(L_i)\}$. Indeed, if we consider a negative value for L_i this immediately implies that $-f$ is equal to $-\infty$ and the supremum will never take this value. Finally, the dual reformulation of the DRE problem is given by

$$\inf_{\theta\in\Theta,\eta\in\mathbb{R}} \quad \sum_{i=0}^{n} \lambda p_i f^* \left(\frac{\ell(y_i - x_i^T\theta) - \eta}{\lambda} \right) + \eta. \tag{C.6}$$

This reformulation works for any kind of reference distributions with weights p_i, but in our case $p_i = 1/n \ \forall i$, therefore the sum with weights p_i can be seen as empirical average (expectation under $\hat{\mathbb{P}}_n$) as we wanted.

To finish the proof we need to verify that strong duality actually holds, namely the gap between the primal optimal solution and the optimal dual one is zero. With Prop. A.0.5 we can guarantee strong duality. In this case, since $ri(\mathbb{R}^n) = \mathbb{R}^n$ we just need to verify that the problem is feasible, namely there exists at least one feasible point for which the equality constraint is satisfied. By simply choosing $\bar{q}_i = 1/n \ \forall i$ we satisfy the equality thus concluding the proof. □

Appendix D

Proofs for Chapter 4

Proof of Lemma 3.2.1. We prove this lemma by contradiction, in particular we assume that $\|\hat{\mathbf{w}}\|$ does not convergence in probability to $\alpha_\star$. Let us define the set $\mathcal{D} := \{\mathbf{w} \in \mathbb{R}^d : \big|\|\mathbf{w}\| - \alpha_\star\big| \leq \epsilon\}$ given $\epsilon > 0$ such that $\alpha_\star + \epsilon < K_\alpha$. In practice we are defining an interval of radius ϵ around $\alpha_\star$ and $\mathcal{D}$ contains vectors such that their norm is inside this interval. Notice also that we can consider $\mathcal{D} \subset \mathcal{S}_{\mathbf{w}}$ whenever we choose $\epsilon < \zeta$.

By hypothesis, $\|\hat{\mathbf{w}}_b\|$ converges in probability to $\alpha_\star$. This means that by taking sufficiently large d, $\big|\|\hat{\mathbf{w}}_b\| - \alpha_\star\big| \leq \epsilon$ w.p.a. 1 which is equivalent to say that w.p.a. 1 $\hat{\mathbf{w}}_b \in \mathcal{D}$. Denoting with $M(\mathbf{w})$ the value of the objective function in (3.12) we can also state that

$$M(\hat{\mathbf{w}}) \leq M(\hat{\mathbf{w}}_b) \tag{D.1}$$

because the optimal cost obtained with the unbounded solution can only be better than the one obtained with $\mathbf{w}$ constrained in $\mathcal{S}_{\mathbf{w}}$.

Assume now by contradiction that $\|\hat{\mathbf{w}}\|$ does not convergence in probability to $\alpha_\star$, i.e. $\hat{\mathbf{w}} \notin \mathcal{D}$ w.p.a. 1. There are two possible cases, either $\hat{\mathbf{w}} \in \mathcal{S}_{\mathbf{w}}/\mathcal{D}$ or $\hat{\mathbf{w}} \notin \mathcal{S}_{\mathbf{w}}$.

The first case is trivial because we have an optimal solution of the "bounded" problem ($\hat{\mathbf{w}} \in \mathcal{S}_{\mathbf{w}}$) which we know by hypothesis that converges to $\alpha_\star$, thus contradicting the assumption. In the second case instead we can always define $\mathbf{w}_\theta := \theta\hat{\mathbf{w}} + (1-\theta)\hat{\mathbf{w}}_b$ with $\theta \in (0,1)$ such that $\mathbf{w}_\theta \in \mathcal{S}_{\mathbf{w}}/\mathcal{D}$. This because we can always find a point in the segment between $\hat{\mathbf{w}} \notin \mathcal{S}_{\mathbf{w}}$ and $\hat{\mathbf{w}}_b \in \mathcal{D}$ satisfying that property. Notice that $\mathbf{w}_\theta$ is a solution of the "bounded" optimization problem.

Now, using convexity in $\mathbf{w}$ of the objective function in (3.12) and property (D.1) we can formulate the following result

$$\begin{aligned} M(\mathbf{w}_\theta) &= \theta M(\hat{\mathbf{w}}) + (1-\theta)M(\hat{\mathbf{w}}_b) \\ &\leq \theta M(\hat{\mathbf{w}}_b) + (1-\theta)M(\hat{\mathbf{w}}_b) = M(\hat{\mathbf{w}}_b). \end{aligned}$$

But this clearly contradicts the optimality of $\hat{\mathbf{w}}_b$ concluding the proof. □

Proof of Lemma 3.2.4. Consider the optimal values of the optimization in (3.13), $\mathbf{w}_\star, \mathbf{v}_\star, \mathbf{u}_\star, \eta_\star$.

Simple first order optimality conditions lead to

$$\mathbf{v}_\star = \mathbf{z} - \sqrt{d}\mathbf{X}\mathbf{w}_\star, \tag{D.2}$$

$$\mathbf{u}_\star = \frac{1}{\delta\sqrt{d}}\frac{\partial}{\partial \mathbf{v}}\sum_{i=0}^{n}\frac{1}{2\lambda}(\ell(v_{\star,i})-\eta_\star)^2+\ell(v_{\star,i}) = \frac{1}{\delta\sqrt{d}}\begin{bmatrix}\ell'_+(v_{\star,1})+\frac{1}{\lambda}\left(\ell(v_{\star,1})-\eta_\star\right)\ell'_+(v_{\star,1})\\ \vdots \\ \ell'_+(v_{\star,n})+\frac{1}{\lambda}\left(\ell(v_{\star,n})-\eta_\star\right)\ell'_+(v_{\star,n})\end{bmatrix} \tag{D.3}$$

where we make use of the subdifferential of ℓ because it might be that it is not a differentiable function.

We want to verify if $\|\mathbf{u}_\star\|$ is bounded.

$$\|\mathbf{u}_\star\| = \frac{1}{\delta\sqrt{d}}\left\|\begin{matrix}\ell'_+(v_{\star,1})+\frac{1}{\lambda}\ell(v_{\star,1})\ell'_+(v_{\star,1})-\frac{\eta_\star}{\lambda}\ell'_+(v_{\star,1})\\ \vdots \\ \ell'_+(v_{\star,n})+\frac{1}{\lambda}\ell(v_{\star,n})\ell'_+(v_{\star,n})-\frac{\eta_\star}{\lambda}\ell'_+(v_{\star,n})\end{matrix}\right\| \leq$$
$$\frac{1}{\delta\sqrt{d}}\left[\sum_{i=0}^{n}\left(\ell'_+(v_{\star,i})\right)^2\right]^{\frac{1}{2}}+\frac{1}{\lambda\delta\sqrt{d}}\left[\sum_{i=0}^{n}\left(\ell(v_{\star,i})\ell'_+(v_{\star,i})\right)^2\right]^{\frac{1}{2}}+\frac{\eta_\star}{\lambda}\frac{1}{\delta\sqrt{d}}\left[\sum_{i=0}^{n}\left(\ell'_+(v_{\star,i})\right)^2\right]^{\frac{1}{2}} \tag{D.4}$$

which follows using simple triangular inequality.

Now, recall from Lemma 6 proof in [27] that $\|\mathbf{v}_\star\| \leq C\sqrt{d}$ for some $C > 0$ with high probability as $d \to +\infty$. This inequality comes from (D.2) and a standard high probability bound on the spectral norm of Gaussian matrices.

From Remark 3.2.3 the normalization condition ensures that $\frac{1}{\sqrt{d}}\left[\sum_{i=0}^{n}\left(\ell'_+(v_{\star,i})\right)^2\right]^{\frac{1}{2}}$ is upper bounded by a constant $P > 0$. At the same time we also know that $\ell'_+(v_{\star,i}) \leq Q$ $\forall i$, given some $Q > 0$. Moreover, $\eta_\star$ is bounded, indeed developing from (3.14) using Assumption 3.1.2(iii)

$$\eta_\star = \frac{1}{n}\sum_{i=0}^{n}\ell(v_{\star,i}) \leq \frac{K}{n}\sum_{i=0}^{n}|v_{\star,i}| \leq \frac{K\sqrt{n}}{n}\left[\sum_{i=0}^{n}|v_{\star,i}|^2\right]^{\frac{1}{2}} \leq \frac{KC}{\sqrt{\delta}}. \tag{D.5}$$

We can use these results to prove that $\mathbf{u}_\star$ has finite norm. In particular the first and third term in the previous equation are bounded using the normalization condition while also the second one can be bounded as follows

$$\frac{1}{\lambda\delta\sqrt{d}}\left[\sum_{i=0}^{n}\left(\ell(v_{\star,i})\ell'_+(v_{\star,i})\right)^2\right]^{\frac{1}{2}} \leq \frac{Q}{\lambda\delta\sqrt{d}}\left[\sum_{i=0}^{n}(K|v_{\star,i}|)^2\right]^{\frac{1}{2}} \leq \frac{QKC}{\lambda\delta}. \tag{D.6}$$

To conclude we have

$$\|\mathbf{u}_\star\| \leq \frac{P}{\delta}\left(1+\frac{\eta_\star}{\lambda}\right)+\frac{QKC}{\lambda\delta}. \tag{D.7}$$

□

Proof of Lemma 3.2.5. We first consider convexity for $p(\alpha, v_i) := \frac{1}{2}(\alpha g_i - v_i)^2$. If this is convex then its perspective function $\frac{1}{2\tau}(\alpha g_i - v_i)^2$ is jointly convex in all its arguments. Finally, its shifted version $\frac{1}{2\tau}(\alpha g_i + z_i - v_i)^2$ is jointly convex as well.

We are left with proving convexity for the initial function p. In order to do so consider the Hessian matrix of the function p which can be easily found to be

$$\mathcal{H}_p = \begin{bmatrix} (g_i)^2 & -g_i \\ -g_i & 1 \end{bmatrix}.$$

From Prop. A.0.4(a) we know that by checking if the Hessian is positive semidefinite we can guarantee convexity.

Therefore, let us consider $x = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \neq 0$ and verify that $x^T \mathcal{H}_p x \geq 0$. What we obtain is

$$x^T \mathcal{H}_p x = x_1^2 (g_i)^2 - 2 g_i x_1 x_2 + x_2^2 = (g_i x_1 - x_2)^2 \geq 0$$

for every possible realization of g_i. □