JMLR: Workshop and Conference Proceedings vol 40:1–27, 2015

Regularized Linear Regression: A Precise Analysis of the Estimation Error

Christos Thrampoulidis CTHRAMPO@CALTECH.EDU
California Institute of Technology

Samet Oymak OYMAK@EECS.BERKELEY.EDU
University of California, Berkeley

Babak Hassibi HASSIBI@CALTECH.EDU
California Institute of Technology

Abstract

Non-smooth regularized convex optimization procedures have emerged as a powerful tool to recover structured signals (sparse, low-rank, etc.) from (possibly compressed) noisy linear measurements. We focus on the problem of linear regression and consider a general class of optimization methods that minimize a loss function measuring the misfit of the model to the observations with an added structured-inducing regularization term. Celebrated instances include the LASSO, Group-LASSO, Least-Absolute Deviations method, etc.. We develop a quite general framework for how to determine precise prediction performance guaranties (e.g. mean-square-error) of such methods for the case of Gaussian measurement ensemble. The machinery builds upon Gordon's Gaussian min-max theorem under additional convexity assumptions that arise in many practical applications. This theorem associates with a primary optimization (PO) problem a simplified auxiliary optimization (AO) problem from which we can tightly infer properties of the original (PO), such as the optimal cost, the norm of the optimal solution, etc. Our theory applies to general loss functions and regularization and provides guidelines on how to optimally tune the regularizer coefficient when certain structural properties (such as sparsity level, rank, etc.) are known.

Keywords: Linear Regression, mean-square-error , structured signals, sparsity, LASSO, Gaussian min-max Theorem, convexity

1. Introduction

1.1. Linear Regression for structured target vectors

Consider the problem of linear regression with additive noise:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d)^T \in \mathbb{R}^d$ is the "true" parameter, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_d) \in \mathbb{R}^{n\times d}$ is the measurement matrix, $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^T \in \mathbb{R}^n$ are the responses, and, $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n)^T \in \mathbb{R}^n$ is the noise vector. Our task is to learn the target vector $\boldsymbol{\beta}_0$. In order to measure the fit of any vector $\boldsymbol{\beta} \in \mathbb{R}^d$ to the vector of observations $\mathbf{y} \in \mathbb{R}^n$ we introduce a *loss function* $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$, which assigns a penalty $\mathcal{L}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \geq 0$ to the corresponding residual $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$. We are particularly interested in the high-dimensional setup in which the number of observations n is fewer than the dimension d of the ambient space, a scenario which arises in most big-data problems (e.g. high resolution images, gene expression data from a DNA microarray, social network data, etc.). It

is typical in such applications that the properties of the target vector $\boldsymbol{\beta}_0$ lie in some some low-dimensional structure (sparsity, low-rankness, clusters, etc.). With the particular structure of $\boldsymbol{\beta}_0$ we associate a properly chosen *regularizer* $f : \mathbb{R}^d \to \mathbb{R}$. For example, if $\boldsymbol{\beta}_0$ is a *sparse* vector then f can be the ℓ_1-norm, if $\boldsymbol{\beta}_0$ is a $\sqrt{n} \times \sqrt{n}$ *low-rank* matrix then a popular choice for f is the nuclear norm (e.g. Negahban et al. (2012); Chandrasekaran et al. (2012) for more examples). A natural estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}_0$ is then obtained by solving the following[1] optimization problem which we shall henceforth call the *Regression Optimization* (RO):

$$\hat{\boldsymbol{\beta}} := \arg\min_{\boldsymbol{\beta}} \ \mathcal{L}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda f(\boldsymbol{\beta}). \tag{2}$$

Here, $\lambda > 0$ is a regularizer parameter. If the functions $\mathcal{L}$ and f are both convex, then the optimization program in (2) is convex, so it can be solved efficiently Boyd and Vandenberghe (2009). Specific choices of the loss function $\mathcal{L}$ and the regularizer f give rise to different popular instances:

- Ordinary Least-Squares (LS) ($\mathcal{L}(\cdot) = (1/2)\|\cdot\|_2^2$, $f(\cdot) = 0$).
- Ridge regression ($\mathcal{L}(\cdot) = (1/2)\|\cdot\|_2^2$, $f(\cdot) = \|\cdot\|_2^2$).
- LASSO ($\mathcal{L}(\cdot) = (1/2)\|\cdot\|_2^2$, $f(\cdot) = \|\cdot\|_1$). Popular sparse recovery algorithm. The acronym was introduced in Tibshirani (1996). To distinguish from the ℓ_2-LASSO defined below, we often refer to this version as the ℓ_2^2-LASSO. The "least-squares" nature of the loss function corresponds to a maximum likelihood estimator for the case when $\boldsymbol{\epsilon}$ is gaussian.
- ℓ_2- (or, Square-root) LASSO, ($\mathcal{L}(\cdot) = \|\cdot\|_2$). A sparse-recovery algorithm similar in nature to the LASSO but there exists differences among them, e.g. tuning of the regularizer parameter of the ℓ_2-LASSO does not require knowledge of the standard deviation of the noise Belloni et al. (2011); Oymak et al. (2013).
- Generalized-LASSO, ($\mathcal{L}(\cdot) = (1/2)\|\cdot\|_2^2$ or $\mathcal{L}(\cdot) = \|\cdot\|_2$). A natural generalization of the LASSO to arbitrary convex (and, typically non-smooth) regularizers f, e.g. nuclear norm, $\ell_{1,2}$ norm (Group-LASSO, Yuan and Lin (2006)) and discrete total variation.
- Regularized LAD ($\mathcal{L}(\cdot) = \|\cdot\|_1$). Least Absolute Deviation algorithms are known to have robust properties in linear regression models (e.g. Rao and Toutenburg (1995)). Also, they perform particularly well in the presence of heavy-tailed errors Wang (2013), and, of *sparse* noise Wright and Ma (2010); Foygel and Mackey (2014); Thrampoulidis and Hassibi (2014).
- Support Vector Machines regression, ($\mathcal{L}(\cdot) = \|\cdot\|_\epsilon$, $f(\cdot) = \|\cdot\|_2^2$) Here, $\|\boldsymbol{\beta}\|_\epsilon = \sum_i |\boldsymbol{\beta}_i|_\epsilon$, where $|x|_\epsilon = |x| - \epsilon$ if $|x| \geq \epsilon$ and 0, otherwise, is the Vapniks epsilon-insensitive norm; ϵ can be though of as the resolution at which we want to look at the data Evgeniou et al. (2000)

The list above is not exhaustive. For instance, in a scenario where noise is known to be bounded it might be preferable to choose the ℓ_∞-norm as the loss function.

1. the minimizer of (2) need not be unique. Using a slight abuse of notation, let the operator $\arg\min$ return any one of those optimal values.

1.2. Precise Estimation Performance Analysis

A prevalent problem is characterizing the parameter estimation accuracy of (2): How accurate is $\hat{\boldsymbol{\beta}}$ when compared to the target vector $\boldsymbol{\beta}_0$ in a certain norm? The focus of this work is on the *normalized squared error*[2] $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2/\|\boldsymbol{\epsilon}\|_2^2$, which quantifies robustness of the estimator. Understanding the behavior of this quantity in terms of the choice of the measurement matrix $\mathbf{X}$, the number of measurements m, the convex regularizer f, the value of the regularizer parameter λ and the unknown signal $\boldsymbol{\beta}_0$ itself, is both of theoretical and practical interest. As an example, knowledge of the dependence on λ can provide valuable insights for the challenging task of optimally tuning (2).

Inevitably, the theoretical analysis of (RO) problems as in (2) has attracted enormous attention over the last twenty years or so. In particular, the advances in the study of *noiseless* underdetermined problems, under the prism of "compressive sampling" Candès et al. (2006); Donoho (2006) have resulted in a significant progress on our understanding regarding the performance of (2) in the presence of noise. Sparse linear regression has been the most active area, e.g. Candès and Tao (2007); Bickel et al. (2009); Belloni et al. (2011); Raskutti et al. (2010); Banerjee et al. (2014) and many others. There have also been contributions which characterize general classes of algorithms like (2), e.g. Negahban et al. (2012). The theory holds under standard incoherence or restricted eigenvalue conditions on the measurement matrix $\mathbf{X}$[3]. Although remarkable, those results characterize the normalized squared error only up to unknown absolute constants (order-wise analysis), which yields our understanding (even for the classical Gaussian measurement ensemble) not comparable to more traditional topics in statistical learning theory, such as performance of LS.

It is only very recently that *precise* characterizations of the estimation performance have appeared in the literature. The price paid is that the measurement matrix $\mathbf{X}$ is restricted to have entries i.i.d. Gaussian[4]. Donoho et al. (2011b); Bayati and Montanari (2012) were the first to perform an *asymptotically exact* characterization of the performance of the ℓ_2^2-LASSO algorithm. Stojnic (2013a) derived precise such results for the constrained version of the LASSO, but most significantly, was the first to introduce the idea of analyzing the prediction performance via Gaussian comparison inequalities. In particular, he cleverly combines the Gaussian min-max theorem (GMT), a comparison inequality proved by Gordon (1988), with a duality trick. Our work is motivated by this recent line of work, Stojnic (2013b,d,c).

1.3. Our Contribution

We describe a quite general and unifying theory for how to determine precise performance guaranties (minimum number of measurements, normalized squared-error, etc.) for the (RO) in (2), when the measurement matrix belongs to the Gaussian ensemble. The framework provides guar-

2. similarly defined measures of performance are considered in the literature under the term of *noise sensitivity*, e.g. Wu and Verdú (2012); Donoho et al. (2011a). Also, see for example Zhang et al. (2009) for other typical measures of performance such as prediction accuracy and feature selection accuracy.
3. Such conditions have been shown to be satisfied by a wide class of randomly designed measurement matrices, e.g. Candès and Tao (2007); Raskutti et al. (2010); Adamczak et al. (2011), etc.. Please also refer to the recent line of work by Mendelson (2014); Lecué and Mendelson (2014) where similar (order-wise) bounds are obtained under weaker assumptions on the randomness properties of $\mathbf{X}$.
4. Although restrictive, this assumption is generic in the sense that many of the results derived for the Gaussian ensemble are known/observed to enjoy a *universality* property, i.e. to hold true for fairly broad family of probability ensembles, thus, is typical in the random matrix theory community. In particular, it is a common practice in the literature of compressive sensing (please refer to the tutorials Vershynin (2014); Candès (2014)).

antees for the large-system limit in which the problem dimensions n and d grow to infinity at proportional rates[5]. In principle, the framework can be applied to any instance of (2), for convex $\mathcal{L}$ and f. The proposed methodology builds upon our main Theorem 3, which is a stronger version of the classical Gaussian Min-max Theorem due to Gordon (1988), in the presence of additional convexity assumptions. We expect the theorem to find applications even beyond the error analysis of (RO) problems.

1.4. Overview of the Framework

The Gaussian min-max Theorem (GMT) of Gordon (1988), essentially provides probabilistic lower bounds on the optimal cost of (RO) via a simpler auxiliary optimization (AO). Motivated by recent work of M. Stojnic, we show that under *convexity assumptions* the (AO) problem allows one to *tightly* upper and lower bound both the optimal cost and the norm of the optimal solution of the (RO). We introduce the core ideas here and elaborate in Sections 2–3.

Theorem 1 (GMT Gordon (1988)) [6]. *Let $\mathbf{G} \in \mathbb{R}^{n\times d}$, $g \in \mathbb{R}$, $\mathbf{g} \in \mathbb{R}^n$ and $\mathbf{h} \in \mathbb{R}^d$ have entries i.i.d. $\mathcal{N}(0,1)$, $\mathcal{S}_{\mathbf{w}} \subset \mathbb{R}^d$, $\mathcal{S}_{\mathbf{u}} \subset \mathbb{R}^n$ be compact sets and $\psi : \mathcal{S}_{\mathbf{w}} \times \mathcal{S}_{\mathbf{u}} \to \mathbb{R}$ be continuous. Define,*

$$\Phi(\mathbf{G}, g) := \min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}} \max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}} \mathbf{u}^T\mathbf{G}\mathbf{w} + g\|\mathbf{w}\|_2\|\mathbf{u}\|_2 + \psi(\mathbf{w},\mathbf{u}) \tag{3}$$

$$\phi(\mathbf{g}, \mathbf{h}) := \min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}} \max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}} \|\mathbf{w}\|_2\mathbf{g}^T\mathbf{u} + \|\mathbf{u}\|_2\mathbf{h}^T\mathbf{w} + \psi(\mathbf{w},\mathbf{u}). \tag{4}$$

Then, for any $c \in \mathbb{R}$: $\quad \mathbb{P}(\Phi(\mathbf{G}, g) < c) \le \mathbb{P}\left(\phi(\mathbf{g}, \mathbf{h}) \le c\right).$

Henceforth, we refer to the optimization in (4) as the *Auxiliary Optimization* (AO). Theorem 1 asserts that the lower tail probability of $\Phi(\mathbf{G}, g)$ is upper bounded by that of $\phi(\mathbf{g}, \mathbf{h})$: if c is a high probability *lower* bound on $\phi(\mathbf{g}, \mathbf{h})$ (in the sense that $\mathbb{P}\left(\phi(\mathbf{g}, \mathbf{h}) \le c\right)$ is close to zero), so it is for $\Phi(\mathbf{G}, g)$. At this point it is still unclear how the result relates to the analysis of the (RO) in (2). This is shown in two steps. First, we bring the minimization in (2) in the format of (3). Second, we strengthen the conclusions of Theorem 1.

Let $\mathcal{L}^*$ be the Fenchel conjugate of $\mathcal{L}$; from convexity of $\mathcal{L}$, $\mathcal{L}(\mathbf{v}) = \sup_{\mathbf{u}} \mathbf{u}^T\mathbf{v} - \mathcal{L}^*(\mathbf{u})$. Also let $\mathbf{w} = \boldsymbol{\beta} - \boldsymbol{\beta}_0$ denote the error vector and recall (1). With these, the (RO) in (2) becomes:

$$\min_{\mathbf{w}} \sup_{\mathbf{u}} \mathbf{u}^T\mathbf{X}\mathbf{w} - \mathbf{u}^T\boldsymbol{\epsilon} - \mathcal{L}^*(\mathbf{u}) + \lambda f(\boldsymbol{\beta}_0 + \mathbf{w}). \tag{5}$$

Identifying $\psi(\mathbf{w},\mathbf{u}) := -\mathbf{u}^T\boldsymbol{\epsilon} - \mathcal{L}^*(\mathbf{u}) + \lambda f(\boldsymbol{\beta}_0 + \mathbf{w})$, we see that (5) is almost in the format of (3). The only term missing is "$g\|\mathbf{w}\|_2\|\mathbf{u}\|_2$", but this can be accounted for in Theorem 1 with a simple symmetrization trick. In particular, Theorem 3 shows that slightly changing (3) to the following optimization problem, which we shall henceforth refer to as *Primary Optimization* (PO),

$$\Phi(\mathbf{G}) := \min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}} \max_{\mathbf{u}\in\mathcal{S}_{\mathbf{y}}} \mathbf{u}^T\mathbf{G}\mathbf{w} + \psi(\mathbf{w},\mathbf{u}), \tag{6}$$

5. Numerical simulations suggest that the results hold for matrices with i.i.d. entries from wider class of distributions. Also, Thrampoulidis and Hassibi (2015) leverages the framework to obtain results for the Haar ensemble. Also, simulation results show the predictions to be accurate for problem dimensions ranging over a few hundreds.
6. Thm. 1 is a slight modification of the original result (Gordon, 1988, Lem. 3.1). In contrast to Thm. 1, the latter assumes $\mathcal{S}_{\mathbf{w}}$ to be arbitrary (not necessarily compact) set, $\mathcal{S}_{\mathbf{y}}$ is restricted to be the unit sphere and $\psi(\cdot,\cdot)$ is only a function of $\mathbf{w}$. For completeness, we include some background and a proof of the theorem in Appendix A.

only changes the conclusion of the theorem to

$$\mathbb{P}(\Phi(\mathbf{G}) < c) \leq 2\mathbb{P}\left(\phi(\mathbf{g}, \mathbf{h}) \leq c\right). \tag{7}$$

Note that this does not affect the essence of the result of Theorem 1: if $\mathbb{P}(\phi(\mathbf{g}, \mathbf{h}) \leq c)$ is close to zero, then c is still a high probability lower bound on $\Phi(\mathbf{G})$. This result is remarkable, since it relates the (PO) (and, essentially the (RO) thanks to (5)) to a seemingly unrelated, but potentially easier to analyze, (AO) problem as given by (4). Yet, this only establishes a lower bound type of relation regarding the optimal cost of the two optimizations. How could this possibly lead to any conclusion regarding the minimizer of (2)? Theorem 3 provides an answer to this question.

In short, Theorem 3 shows that in the presence of appropriate *convexity assumptions* on the sets $\mathcal{S}_{\mathbf{w}}$, $\mathcal{S}_{\mathbf{u}}$ and on the function ψ the (AO) problem tightly bounds the optimal cost of the (PO) in the sense that for all $\mu \in \mathbb{R}$ and $t > 0$,

$$\mathbb{P}\left(|\Phi(\mathbf{G}) - \mu| > t\right) \leq 2\mathbb{P}\left(|\phi(\mathbf{g}, \mathbf{h}) - \mu| > t\right). \tag{8}$$

In (2), the principal objective is not characterizing the optimal cost of the optimization, but rather, its optimal minimizer $\hat{\boldsymbol{\beta}}$ and concluding about the achieved parameter estimation accuracy $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|$. With this serving as our motivation, we show that, in an asymptotic setting and under proper additional assumptions, the optimal solutions of the problems (AO) and (PO) are also closely related:

$$\|\mathbf{w}_{\Phi}(\mathbf{G})\| \approx \|\mathbf{w}_{\phi}(\mathbf{g}, \mathbf{h})\|, \tag{9}$$

where $\mathbf{w}_{\Phi}(\mathbf{G})$ and $\mathbf{w}_{\phi}(\mathbf{g}, \mathbf{h})$ denote the optimal minimizers in (6) and (4), respectively.

2. The Convex Gaussian Min-max Theorem

We start by fixing some notation and introducing the asymptotic setting under which the analysis holds.

Definition 2 (GMT admissible sequence) *The sequence* $\{\mathbf{G}^{(d)}, \mathbf{g}^{(d)}, \mathbf{h}^{(d)}, \mathcal{S}_{\mathbf{w}}^{(d)}, \mathcal{S}_{\mathbf{u}}^{(d)}, \psi^{(d)}\}_{d\in\mathbb{N}}$ *indexed by* d*, with* $\mathbf{G}^{(d)} \in \mathbb{R}^{n\times d}, \mathbf{h}^{(d)} \in \mathbb{R}^{d}, \mathbf{g}^{(d)} \in \mathbb{R}^{n}$*,* $\mathcal{S}_{\mathbf{w}}^{(d)} \subset \mathbb{R}^{d}$*,* $\mathcal{S}_{\mathbf{u}}^{(d)} \subset \mathbb{R}^{n}$*,* $\psi^{(d)} : \mathcal{S}_{\mathbf{w}}^{(d)} \times \mathcal{S}_{\mathbf{u}}^{(d)} \to \mathbb{R}$ *and* $n = n(d)$*, is said to be* admissible *if, for each* $d \in \mathbb{N}$*,* $\mathcal{S}_{\mathbf{w}}^{(d)}$ *and* $\mathcal{S}_{\mathbf{u}}^{(d)}$ *are compact sets and* $\psi^{(d)}$ *is is continuous on its domain. Onwards, we will drop the superscript* (d) *from* $\mathbf{G}^{(d)}$,$\mathbf{g}^{(d)}$, $\mathbf{h}^{(d)}$.

A sequence $\{\mathbf{G}^{(d)}, \mathbf{g}^{(d)}, \mathbf{h}^{(d)}, \mathcal{S}_{\mathbf{w}}^{(d)}, \mathcal{S}_{\mathbf{u}}^{(d)}, \psi^{(d)}\}_{d\in\mathbb{N}}$ defines a sequence of min-max problems

$$\Phi^{(d)}(\mathbf{G}) := \min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}^{(d)}} \max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}^{(d)}} \mathbf{u}^T\mathbf{G}\mathbf{w} + \psi^{(d)}(\mathbf{w}, \mathbf{u}), \tag{10a}$$

$$\phi^{(d)}(\mathbf{g}, \mathbf{h}) := \min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}^{(d)}} \max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}^{(d)}} \|\mathbf{w}\|_2\mathbf{g}^T\mathbf{u} + \|\mathbf{u}\|_2\mathbf{h}^T\mathbf{w} + \psi^{(d)}(\mathbf{w}, \mathbf{u}). \tag{10b}$$

We refer to those as the Primary optimization (PO), and, the Auxiliary Optimization (AO) problems, respectively. Also, denote their optimal minimizers as $\mathbf{w}_{\Phi}^{(d)}(\mathbf{G})$ and $\mathbf{w}_{\phi}^{(d)}(\mathbf{g}, \mathbf{h})$, respectively. Then, define $\upsilon^{(d)} : \mathcal{S}_{\mathbf{w}}^{(d)} \to \mathbb{R}$ as follows,

$$\upsilon^{(d)}(\mathbf{w}; \mathbf{g}, \mathbf{h}) := \max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}^{(d)}} \|\mathbf{w}\|_2\mathbf{g}^T\mathbf{u} + \|\mathbf{u}\|_2\mathbf{h}^T\mathbf{w} + \psi^{(d)}(\mathbf{w}, \mathbf{u}). \tag{11}$$

Clearly, $\phi^{(d)}(\mathbf{g}, \mathbf{h}) = \min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}^{(d)}} \upsilon^{(d)}(\mathbf{w}; \mathbf{g}, \mathbf{h})$.

For a sequence of random variables $\{\mathcal{X}^{(d)}\}_{d\in\mathbb{N}}$ and constant $c \in \mathbb{R}$ (independent of d), we write $\mathcal{X}^{(d)} \xrightarrow{P} c$, to denote convergence in probability, i.e. $\forall \epsilon > 0$, $\lim_{d\to\infty} \mathbb{P}\left(|\mathcal{X}^{(d)} - c| > \epsilon\right) = 0$. Similarly, for a deterministic sequence $\{x^{(d)}\}_{d\in\mathbb{N}}$ we write $x^{(d)} \to c$ if $\lim_{d\to\infty} x^{(d)} = c,\ c \in \mathbb{R}$.

Theorem 3 (Convex GMT (CGMT)) *Let $\{\mathbf{G}^{(d)}, \mathbf{g}^{(d)}, \mathbf{h}^{(d)}, \mathcal{S}_{\mathbf{w}}^{(d)}, \mathcal{S}_{\mathbf{u}}^{(d)}, \psi^{(d)}\}_{d\in\mathbb{N}}$ be a GMT admissible sequence as in Definition 2, for which additionally the entries of $\mathbf{G}, \mathbf{h}$ and $\mathbf{g}$ are i.i.d. $\mathcal{N}(0,1)$. Let $\Phi^{(d)}(\mathbf{G})$, $\phi^{(d)}(\mathbf{g},\mathbf{h})$ be the optimal costs, and, $\mathbf{w}_{\Phi}^{(d)}(\mathbf{G})$, $\mathbf{w}_{\phi}^{(d)}(\mathbf{g},\mathbf{h})$ the corresponding optimal minimizers of the (PO) and (AO) problems in (10a) and (10b). The following three statements hold.*

(i) For any $d \in \mathbb{N}$ and $c \in \mathbb{R}$,

$$\mathbb{P}\left(\Phi^{(d)}(\mathbf{G}) < c\right) \le 2\mathbb{P}\left(\phi^{(d)}(\mathbf{g},\mathbf{h}) \le c\right). \tag{12}$$

(ii) Fix any $d \in \mathbb{N}$. If $\mathcal{S}_{\mathbf{w}}^{(d)}$, $\mathcal{S}_{\mathbf{u}}^{(d)}$ are convex, and, $\psi^{(d)}(\cdot,\cdot)$ is convex-concave[7] on $\mathcal{S}_{\mathbf{w}}^{(d)} \times \mathcal{S}_{\mathbf{u}}^{(d)}$, then, for any $\mu \in \mathbb{R}$ and $t > 0$,

$$\mathbb{P}\left(|\Phi^{(d)}(\mathbf{G}) - \mu| > t\right) \le 2\mathbb{P}\left(|\phi^{(d)}(\mathbf{g},\mathbf{h}) - \mu| > t\right). \tag{13}$$

(iii) Assume the conditions of (ii) hold for all $d \in \mathbb{N}$. Let $\|\cdot\|$ denote some norm in $\mathbb{R}^d$ and recall (11). If, there exist constants (independent of d) κ_*, α_* and $\tau > 0$ such that

(a) $\phi^{(d)}(\mathbf{g},\mathbf{h}) \xrightarrow{P} \kappa_*$,

(b) $\|\mathbf{w}_{\phi}^{(d)}(\mathbf{g},\mathbf{h})\| \xrightarrow{P} \alpha_*$,

(c) with probability one in the limit $d \to \infty$,

$$\{\upsilon^{(d)}(\mathbf{w};\mathbf{g},\mathbf{h}) \ge \phi^{(d)}(\mathbf{g},\mathbf{h}) + \tau(\|\mathbf{w}\| - \|\mathbf{w}_{\phi}^{(d)}(\mathbf{g},\mathbf{h})\|)^2,\ \forall \mathbf{w} \in \mathcal{S}_{\mathbf{w}}^{(d)}\},$$

then,

$$\|\mathbf{w}_{\Phi}^{(d)}(\mathbf{G})\| \xrightarrow{P} \alpha_*. \tag{14}$$

The probabilities in Theorem 3 are with respect to the randomness of $\mathbf{G}$, $\mathbf{g}$ and $\mathbf{h}$. The proof of the theorem is included in Appendix C.

2.1. Remarks

Concentration of the optimal cost: A main contribution of the convex GMT (CGMT) is inequality (13). It shows that in the presence of appropriate convexity assumptions the GMT is tight. In particular, choosing $\mu = \mathbb{E}\phi(\mathbf{g},\mathbf{h})$ in (13), we can deduce Corollary 4 below from the fact that $\phi(\mathbf{g},\mathbf{h})$ is Lipschitz in $(\mathbf{g},\mathbf{h})$ (see Lemma B.0.3) and from the Gaussian concentration of Lipschitz functions (e.g., Theorem B.0.1). (We drop the superscript (d) to enlighten notation).

Corollary 4 *Consider the same setup as in Theorem 3 and let the assumptions of statement (ii) hold. Further, define $R_{\mathbf{w}} := \max_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}} \|\mathbf{w}\|_2$ and $R_{\mathbf{u}} := \max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}} \|\mathbf{u}\|_2$. Then, for all $t > 0$,*

$$\mathbb{P}\left(\ |\Phi(\mathbf{G}) - \mathbb{E}\phi(\mathbf{g},\mathbf{h})| > t\ \right) \le 4\exp\left(-t^2/(4R_{\mathbf{w}}^2 R_{\mathbf{u}}^2)\right).$$

7. i.e., convex on its first argument and concave on its second argument. The result remains true under quasi-convexity/concavity (see Appendix C).

$\max\min = \min\max$*:* It turns out from the proof that what is critical for the statement to hold is that the min-max in (10a) can be flipped into a max-min. Statement (ii) provides sufficient conditions for this to occur which also appear in practice (e.g. analysis of (2)). It is conceivable that there exist cases in which the min-max operation can be flipped under relaxed assumptions, in which (13) would still hold.

Statement (iii): The statement is crucial for error analysis of Regression Optimization, since, in contrast to statements (i) & (ii), it concludes on the properties of the actual minimizer of the (PO). Note that it requires that d be large enough (the previous two statements hold for all dimensions). A few comments on the required conditions: First, the same convexity assumptions as in statement (ii) are present. Next, it is required that as $d \to \infty$ both $\phi^{(d)}(\mathbf{g}, \mathbf{h})$ and $\|\mathbf{w}_\phi^{(d)}(\mathbf{g}, \mathbf{h})\|$ converge to constants, say, κ_* and α_* (this may require for example proper normalization with d, e.g. Section 3.2.2). It is important to remark that $\mathbf{w}_\phi^{(d)}(\mathbf{g}, \mathbf{h})$ denotes *any* optimal minimizer in (10b). We do *not* require that $\mathbf{w}_\phi^{(d)}(\mathbf{g}, \mathbf{h})$ is unique; there might be multiple such optima, but they all have norms that converge to α_*. The last condition guarantees that any other feasible $\boldsymbol{\beta}$ with norm that is far from the optimal α_* results in a strictly positive increase (uniform over d) of the objective value. A sufficient (but not necessary) condition that often occurs in applications (e.g. Appendix D) and satisfies this is that the function $\upsilon^{(d)}(\cdot; \mathbf{g}, \mathbf{h})$ be *strongly convex* with respect to the norm $\|\cdot\|$.

Analysis of the (AO): Satisfying the conditions of the third statement of the theorem requires thorough analysis of the (AO) problem in (10b). Of course, the premise of the theorem is that the (AO) optimization is simpler to analyze that the (PO). Intuitively, this is the case since the bilinear term that includes a random matrix in (10a) is "decoupled" in (10b) into two terms which only involve independent random vectors instead. Practically, we illustrate this in Section 3.3.2 through a detailed example.

3. Precise Performance Analysis of the Regression Optimization

3.1. Preliminaries

Conjugate pairs: The Fenchel conjugate of $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$ is the function $\mathcal{L}^* : \mathbb{R}^d \to (-\infty, +\infty]$[8] defined as $\mathcal{L}^*(\mathbf{u}) := \sup_{\mathbf{v}} \mathbf{v}^T\mathbf{u} - \mathcal{L}(\mathbf{u})$. It is always convex and lower semi-continuous. Furthermore, by the Fenchel–Moreau theorem, if $\mathcal{L}$ is convex and continuous, then $\mathcal{L}(\mathbf{v}) = \sup_{\mathbf{u}}\{\mathbf{u}^T\mathbf{v} - \mathcal{L}^*(\mathbf{u})\}$ for all $\mathbf{v} \in \operatorname{dom}\mathcal{L}$ (Rockafellar, 1997, Thm. 12.2). In the last maximization, $\mathbf{u}_*$ is optimal iff $\mathbf{u}_* \in \partial\mathcal{L}(\mathbf{v})$ (e.g., (Rockafellar, 1997, Thm. 23.5)). Here, $\partial\mathcal{L}(\mathbf{v})$ denotes the subdifferential of $\mathcal{L}$ at $\mathbf{v}$; if $\mathbf{v} \in \operatorname{int}\operatorname{dom}\mathcal{L}$, then $\partial\mathcal{L}(\mathbf{v})$ is a non-empty, closed and bounded set. Standard examples of conjugate pairs of continuous convex functions, also relevant to our analysis, are the following:

$$\mathcal{L}(\mathbf{v}) = (1/2)\|\mathbf{v}\|^2 \leftrightarrow \mathcal{L}^*(\mathbf{u}) = (1/2)\|\mathbf{u}\|^2 \quad \text{and} \quad \mathcal{L}(\mathbf{v}) = \|\mathbf{v}\| \leftrightarrow \mathcal{L}^*(\mathbf{u}) = \begin{cases} 0 & \|\mathbf{u}\|_* \le 1, \\ +\infty & \text{else.} \end{cases}$$

Here, $\|\mathbf{u}\|_* = \sup_{\|\mathbf{v}\|\le 1} \mathbf{v}^T\mathbf{u}$ denotes the dual-norm of $\|\cdot\|$. For instance, $\|\cdot\|_\infty$ is the dual-norm of $\|\cdot\|_1$, while $\|\cdot\|_2$ is self-dual.

Assumptions: Let both $\mathcal{L}$ and f in (2) be continuous proper *convex* functions. In addition, we assume that $\mathcal{L}^*$ is continuous on its effective domain $\operatorname{dom}\mathcal{L}^* := \{\mathbf{u} | \mathcal{L}^*(\mathbf{u}) < \infty\}$. (We have not made any particular effort to relax this latter technical assumption, partly because it appears to be mild for our interests.) . Finally, the entries of $\mathbf{X}$ are drawn i.i.d. $\mathcal{N}(0, 1)$.

8. Following the common practice (e.g. as in (Rockafellar, 1997, Ch. 12) and (Bertsekas et al., 2003, Ch. 7)) we define $\mathcal{L}^*$ as an extended real-valued function that takes the value $+\infty$ whenever $\mathbf{u} \notin \operatorname{dom}\mathcal{L}$.

3.2. Applying the Framework

3.2.1. (RO)→(PO)→(AO)

Recall the (RO) optimization in (2) and the goal of characterizing the squared error $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2$. As in Section 1.4, we introduce the new variable $\mathbf{w} := \boldsymbol{\beta} - \boldsymbol{\beta}_0$ and apply the Fenchel–Moreau theorem to equivalently express the optimization as follows,

$$\min_{\mathbf{w}} \max_{\mathbf{u}} \mathbf{u}^T\mathbf{X}\mathbf{w} - \mathbf{u}^T\boldsymbol{\epsilon} - \mathcal{L}^*(\mathbf{u}) + \lambda f(\boldsymbol{\beta}_0 + \mathbf{w}). \tag{15}$$

This can be immediately recognized to be in the form of the (PO) problem in (10a), with $\psi(\mathbf{w}, \mathbf{u}) := -\mathbf{u}^T\boldsymbol{\epsilon} - \mathcal{L}^*(\mathbf{u}) + \lambda f(\boldsymbol{\beta}_0 + \mathbf{w})$. Also, ψ is appropriately convex in $\mathbf{w}$ and concave in $\mathbf{u}$. However, both the constraint sets in (15) appear to be unbounded. In order to apply the framework of the CGMT (Theorem 3) we further need to impose compact constraint sets in (15), which otherwise appear to be unbounded. We proceed along the following strategy.

In agreement with the notation introduced in Section 2 let $\mathbf{w}_\Phi := \mathbf{w}_\Phi(\mathbf{X})$ be any minimizer in (15). Recall our end goal is evaluating a limit (if it exists) of $\|\mathbf{w}_\Phi\|_2$[9]. We will assume that with probability approaching one in the limit of $d \to \infty$, there exists an absolute constant (in particular independent of d), say $K_{\mathbf{w}} > 0$, such that $\|\mathbf{w}_\Phi\|_2 \le K_{\mathbf{w}}$. The exact value of $K_{\mathbf{w}}$ will be determined later in the proof, in particular, after the analysis of the (AO); we elaborate on this shortly, but for now assume that such a constant can be found. Once this is the case (after conditioning on the event), we can impose the additional constraint $\|\mathbf{w}\|_2 \le K_{\mathbf{w}}$ in (15), without altering the optimization. Next, we consider imposing a constraint $\mathbf{u} \in \mathcal{S}_{\mathbf{u}}$ in (15), for appropriately chosen compact $\mathcal{S}_{\mathbf{u}}$. Recall that the optimal $\mathbf{u}_*$ satisfies $\mathbf{u}_* \in \partial\mathcal{L}(\mathbf{X}\mathbf{w} - \boldsymbol{\epsilon})$. In the simplest case where $\operatorname{dom}\mathcal{L}^*$ is a (closed) bounded set, it suffices to choose $\mathcal{S}_{\mathbf{u}} = \operatorname{dom}\mathcal{L}_*$. This covers for example all norms, say $\mathcal{L} = \|\cdot\|$, as $\operatorname{dom}\mathcal{L}^* = \{\mathbf{u} \mid \|\mathbf{u}\|_* \le 1\}$. For the general case, we need to condition on the high-probability event that $\|\mathbf{X}\|_2 \le c(\sqrt{n} + \sqrt{d})$ for constant $c > 1$. Under this event, for all $\|\mathbf{w}\| \le K_{\mathbf{w}}$ and bounded $\boldsymbol{\epsilon}$, the set of optima $\bigcup\{\partial\mathcal{L}(\mathbf{X}\mathbf{w} - \boldsymbol{\epsilon}) | \mathbf{w} \in \mathcal{S}_{\mathbf{w}}\}$ is bounded, thus, there exists (sufficiently large, but finite) $K_{\mathbf{u}} > 0$ such that constraining the maximization in (15) over $\mathcal{S}_{\mathbf{u}} := \{\|\mathbf{u}\|_2 \le K_{\mathbf{u}}\}$ does not affect the optimization[10].

These suggest analyzing the following (AO) problem:

$$\begin{aligned}\phi(\mathbf{g}, \mathbf{h}) &= \min_{\|\mathbf{w}\|_2 \le K_{\mathbf{w}}} \max_{\|\mathbf{u}\|_2 \le K_{\mathbf{u}}} \|\mathbf{w}\|_2\mathbf{g}^T\mathbf{u} + \|\mathbf{u}\|_2\mathbf{h}^T\mathbf{w} - \boldsymbol{\epsilon}^T\mathbf{u} - \mathcal{L}^*(\mathbf{u}) + \lambda f(\boldsymbol{\beta}_0 + \mathbf{w}) \\ &= \min_{\|\mathbf{w}\|_2 \le K_{\mathbf{w}}} \max_{\|\mathbf{u}\|_2 \le K_{\mathbf{u}}} (\|\mathbf{w}\|_2\mathbf{g} - \boldsymbol{\epsilon})^T\mathbf{u} - \mathcal{L}^*(\mathbf{u}) + \|\mathbf{u}\|_2\mathbf{h}^T\mathbf{w} + \lambda f(\boldsymbol{\beta}_0 + \mathbf{w}). \end{aligned} \tag{16}$$

In view of Theorem 3, the analysis of (16) involves studying the convergence of its optimal cost and of (the norm of) the minimizer, say $\mathbf{w}_\phi := \mathbf{w}_\phi(\mathbf{g}, \mathbf{h})$. Recall that the exact values of $K_{\mathbf{w}}, K_{\mathbf{u}}$ have

9. Further, recall the asymptotic setting of Section 2: (15) actually defines a sequence of optimization problems (indexed by d), and thus, a sequence of minimizers $\mathbf{w}_\Phi^{(d)}$. Strictly speaking, we consider a sequence $\{\mathbf{X}^{(d)}, \boldsymbol{\epsilon}^{(d)}, \boldsymbol{\beta}_0^{(d)}, f^{(d)}(\cdot)\}$, such that $\mathbf{X}^{(d)} \in \mathbb{R}^{m\times n}$ with entries i.i.d. $\mathcal{N}(0,1)$, $\boldsymbol{\epsilon}^{(d)} \in \mathbb{R}^n$, $\boldsymbol{\beta}_0^{(d)} \in \mathbb{R}^d$ and $f^{(d)} : \mathbb{R}^d \to \mathbb{R}$ a convex function. We avoid explicitly introducing this notation to keep the presentation simple, but the statements made are to be interpreted in such a setting.

10. In consideration of the randomness of the noise vector $\boldsymbol{\epsilon}$, in order for $K_{\mathbf{u}}$ to be independent of it, we may further need to assume a high-probability upper bound on the noise vector, say $\|\epsilon\|_2 \le B_\epsilon$ w.h.p. (e.g. $B_\epsilon = c\sqrt{n}$ for noise with (sub)-gaussian i.i.d entries). Then $K_{\mathbf{u}}$ can be chosen to only depend on B_ϵ. Also, with proper normalization of the loss function we can guarantee that $K_{\mathbf{u}}$ is constant independent of d. In particular, this requires scaling $\mathcal{L}$ such that for all constants $c > 0$, there exists constant $C > 0$ with $\|\partial\mathcal{L}(\mathbf{v})\|_2 \le C$, for all $\mathbf{v} : \|\mathbf{v}\|_2 \le c(\sqrt{d} + \sqrt{n}) + B_\epsilon$. See (17) for an example.

not been (yet) fixed, thus, they can be treated as arbitrarily large, but finite, during this analysis. The precise values of $K_{\mathbf{w}}, K_{\mathbf{u}}$ can be determined after applying this analysis. By that time, we will have found the value that $\|\mathbf{w}_\phi\|_2$ converges to. If this value, say $\alpha_* > 0$, can be made independent of $K_{\mathbf{w}}, K_{\mathbf{u}}$, then we can choose $K_{\mathbf{w}} = 2\alpha_*$ making our initial assumption was correct. However, if we cannot find a $K_{\mathbf{w}}$ such that the norm of the optimizer is independent of it, then the initial problem could not have had a bounded optimizer. We provide a brief example to better illustrate these ideas in the next section. A detailed example is included in Section 3.3.2.

3.2.2. NORMALIZATION: AN EXAMPLE

To fix the ideas of the framework, let us consider the popular ℓ_2^2-LASSO which solves

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1.$$

We analyze the limiting behavior of $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2$ in the high dimensional proportional regime, where $n/d \to \delta \in (0, \infty)$. Further, we assume white noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ and $\|\beta_0\|_2 = \mathcal{O}(1)$ for simplicity. We start by appropriately normalizing the loss function and the regularizer to make sure that the optimal cost is $\mathcal{O}(1)$ (we need this for condition (a) of CGMT) . Thus, we consider

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2n}\|\mathbf{X}\mathbf{w} - \boldsymbol{\epsilon}\|_2^2 + \frac{\lambda}{\sqrt{n}}\|\boldsymbol{\beta}_0 + \mathbf{w}\|_1,$$

which we equivalently express as

$$\hat{\mathbf{w}} = \arg\min_{\|\mathbf{w}\|_2 \le K_{\mathbf{w}}} \max_{\|\mathbf{u}\|_2 \le K_{\mathbf{u}}} \frac{1}{2\sqrt{n}}\mathbf{u}^T\mathbf{X}\mathbf{w} - \frac{1}{2\sqrt{n}}\mathbf{u}^T\boldsymbol{\epsilon} - \frac{1}{2}\|\mathbf{u}\|_2^2 + \frac{\lambda}{\sqrt{n}}\|\boldsymbol{\beta}_0 + \mathbf{w}\|_1. \tag{17}$$

Here, as discussed $K_{\mathbf{w}}, K_{\mathbf{u}}$ are to be fixed after the analysis of the corresponding (AO) problem, i.e. after finding α_*. Also, note that we have normalized the loss function so that both $K_{\mathbf{w}}, K_{\mathbf{u}}$ are constants independent of d, i.e. $\mathcal{O}(1)$. Please refer to Thrampoulidis et al. (2015b) for a further analysis of the corresponding (AO) problem. The proper normalization differs case by case.

3.2.3. ANALYSIS OF THE (AO)

The analysis of the (AO) problem (cf. (16)) is typically performed in the following two steps. First, comes a deterministic analysis with the goal of simplifying the (AO): in many cases it is possible to reduce the optimizations involved into ones involving only scalar quantities. Next, follows the probabilistic study of the convergence properties of the optimal cost and the norm of the optimal solution of the (AO) as required in the third statement of Theorem 3. For this, we typically require a probabilistic model[11] for $\boldsymbol{\epsilon}$ and $\boldsymbol{\beta}_0$, the choice of which depends on the specific instance of the (RO) in consideration. For example, for the LASSO we assume that $\boldsymbol{\epsilon}$ is Gaussian, while a sparse noise model is more reasonable for the LAD. Also, an ℓ_1-regularizer is typically associated with a sparse $\boldsymbol{\beta}_0$, while nuclear-norm regularization corresponds to a low-rank $\boldsymbol{\beta}_0$. Thus, the analysis of (16) is problem specific, Thrampoulidis et al. (2015a); Thrampoulidis and Hassibi (2014); Thrampoulidis et al. (2015b); Thrampoulidis and Hassibi (2015). To make these ideas concrete we include a detailed example in Section 3.3.2.

11. Note, however, that the probabilistic relation established by Theorem 3 between (15) and (16) holds for all $\boldsymbol{\epsilon}$ and all $\boldsymbol{\beta}_0$. Thus, provided that $\mathbf{X}$ is statistically independent from them, Theorem 3 continues to hold even after interpreting the probabilities to be over the joint distribution of $\mathbf{X}$, $\boldsymbol{\epsilon}$ and $\boldsymbol{\beta}_0$.

3.3. Examples

3.3.1. HIGH-SNR REGIME

Although the framework is not restrictive to this, it is often common to model the noise $\boldsymbol{\epsilon}$ as having entries i.i.d. of variance, say, σ^2. Then, the Normalized Squared Error (NSE) essentially corresponds to the quantity $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2^2/\sigma^2$. Predicting the NSE for arbitrary values of the noise variance is of course the ultimate goal, but a significant special case often becomes that of studying the high-SNR regime corresponding to $\sigma^2\to 0$. The significance is due to the fact that, in several instances, this captures the *worst-case noise sensitivity behavior*, i.e. $\lim_{\sigma^2\to 0}\mathrm{NSE}=\sup_{\sigma^2>0}\mathrm{NSE}$ (e.g. Donoho et al. (2011b); Oymak and Hassibi (2013); Oymak et al. (2013); Wu and Verdú (2012)). It turns out that when $\sigma^2\to 0$, the analysis of (16) is somewhat simplified, owning to the fact that f can then be approximated on the first-order ((Rockafellar, 1997, Thm. 23.4)) by $f(\boldsymbol{\beta}_0+\mathbf{w})\approx f(\boldsymbol{\beta}_0)+\max_{\mathbf{s}\in\partial f(\boldsymbol{\beta}_0)}\mathbf{s}^T\mathbf{w}$[12]. With this, the analysis only depends on f and $\boldsymbol{\beta}_0$ through a "first-order surrogate", namely the subdifferential $\partial f(\boldsymbol{\beta}_0)$. For example, in sparse recovery with ℓ_1-regularization, the high-SNR NSE depends only on the sparsity of the unknown signal $\boldsymbol{\beta}_0$. Similarly, in low-rank recovery with nuclear-norm regularization, the high-SNR NSE depends only on the rank of $\boldsymbol{\beta}_0$. On the other hand, the NSE in the finite-SNR regime depends on the specific statistics of $\boldsymbol{\beta}_0$. The framework of this paper is, of course, applicable in both regimes. We discuss a few specific examples next.

3.3.2. GENERALIZED-LASSO

We study the popular Generalized LASSO. For simplicity, we focus on the high-SNR regime. Also, we restrict attention to the ℓ_2-LASSO, although the results can be extended to the ℓ_2^2-LASSO (see Thrampoulidis et al. (2015b)). Specializing (16), the corresponding (AO) optimization for the high-SNR regime becomes:

$$\phi_{\text{GLASSO}}(\mathbf{g},\mathbf{h}):=\min_{\|\mathbf{w}\|\le K_{\mathbf{w}}}\max_{\substack{\|\mathbf{u}\|_2\le 1\\ \mathbf{s}\in\partial f(\boldsymbol{\beta}_0)}}\frac{1}{\sqrt{d}}\{(\|\mathbf{w}\|_2\mathbf{g}-\epsilon)^T\mathbf{u}-(\|\mathbf{u}\|_2\mathbf{h}-\lambda\mathbf{s})^T\mathbf{w}\},\tag{18}$$

where we have approximated f in the first order and have properly normalized the objective. Assuming white noise $\boldsymbol{\epsilon}\sim\mathcal{N}(0,\sigma^2\mathbf{I})$, $\|\mathbf{w}\|_2\mathbf{g}-\boldsymbol{\epsilon}$ above is statistically identical to a random vector with entries i.i.d $\mathcal{N}(0,\|\mathbf{w}\|_2^2+\sigma^2)$. Thus, with some abuse of notation it becomes equivalent to substitute the first-term in the objective with $\sqrt{\|\mathbf{w}\|_2^2+\sigma^2}\,\mathbf{g}^T\mathbf{u}$, for $\mathbf{g}\sim\mathcal{N}(0,\mathbf{I})$. Then, we can easily maximize over the *direction* of $\mathbf{u}$ to equivalently express the optimization as

$$\min_{\|\mathbf{w}\|_2\le K_{\mathbf{w}}}\max_{\substack{0\le\beta\le 1\\ \mathbf{s}\in\partial f(\boldsymbol{\beta}_0)}}\frac{1}{\sqrt{d}}\{\sqrt{\|\mathbf{w}\|_2^2+\sigma^2}\,\|\mathbf{g}\|_2\beta-(\beta\mathbf{h}-\lambda\mathbf{s})^T\mathbf{w}\}\tag{19}$$

The objective is now convex in $\mathbf{w}$ and (jointly) concave in $\beta,\mathbf{s}$, and, the constraint sets are bounded. Thus, as in (Rockafellar, 1997, Corollary 37.3.2) we can flip the order of min-max. Then, it is easy to minimize over the direction of $\mathbf{w}$ to find

$$\max_{\substack{0\le\beta\le 1\\ \mathbf{s}\in\partial f(\boldsymbol{\beta}_0)}}\min_{0\le\alpha\le K_{\mathbf{w}}}\frac{1}{\sqrt{d}}\{\sqrt{\alpha^2+\sigma^2}\,\|\mathbf{g}\|_2\beta-\alpha\|\beta\mathbf{h}-\lambda\mathbf{s}\|_2\}.$$

12. The idea here being that the error $\|\hat{\mathbf{w}}\|_2$ also tends to zero as $\sigma^2\to 0$. Please refer to (Oymak et al., 2013, Sec. 9.1).

As a last step, it takes flipping the order of min-max once more. Maximization over $\mathbf{s}$ results in the distance term below, (defined as $\text{dist}(\mathbf{v}, \lambda\partial f(\boldsymbol{\beta}_0)) := \min_{\mathbf{s}\in\partial f(\boldsymbol{\beta}_0)} \|\mathbf{v}-\lambda\mathbf{s}\|_2$):

$$\max_{0\le\beta\le 1} \min_{0\le\alpha\le K_{\mathbf{w}}} \frac{1}{\sqrt{d}}\{\sqrt{\alpha^2+\sigma^2}\,\|\mathbf{g}\|_2\beta - \alpha\cdot\text{dist}(\beta\mathbf{h}, \lambda\partial f(\boldsymbol{\beta}_0))\}. \tag{20}$$

In just a few lines we were able to reduce the (AO) problem to an equivalent optimization in (20) that now only involves two scalar variables, out of which, α, plays the role of $\|\mathbf{w}\|_2$. Also, the objective is strongly convex with respect to α (this can be used to show condition (c) of Theorem 3). Furthermore, it is now easier to get a handle on the random components: both $\|\mathbf{g}\|$ and $\text{dist}(\mathbf{h}, \lambda\partial f(\boldsymbol{\beta}_0))$ are Lipschitz , thus, they normally concentrate around their means. In particular, it can be shown (e.g. (Oymak et al., 2013, Lem. B.2)) that $\|\mathbf{g}\|$ concentrates around $\sqrt{n}$ and $\text{dist}(\mathbf{h}, \lambda\partial f(\boldsymbol{\beta}_0))$ around $\sqrt{D(\lambda)}$, where $D(\lambda)$ is the *Gaussian squared distance* to the scaled subdifferential:

$$D(\tau) = D_{f,\boldsymbol{\beta}_0}(\tau) = \mathbb{E}_{\mathbf{h}\sim\mathcal{N}(0,\mathbf{I})}\left[\text{dist}^2(\mathbf{h}, \tau\partial f(\boldsymbol{\beta}_0))\right]. \tag{21}$$

With these and assuming a high-dimensional proportional regime where $\frac{n}{d}\to\delta\in(0,\infty)$ and $\frac{D(\tau)}{n}\to\bar{D}(\tau)\in(0,1)$, we show in Appendix D that the optimal cost and the optimal minimizer of (20), they both converge to the corresponding quantities of the following deterministic optimization:

$$\max_{0\le\beta\le 1} \min_{0\le\alpha\le K_{\mathbf{w}}} \beta\sqrt{\alpha^2+\sigma^2}\,\sqrt{\delta} - \alpha\beta\sqrt{\bar{D}(\lambda/\beta)}. \tag{22}$$

It only remains to analyze the optimality conditions of this to find α_*. We defer this step to the Appendix D. With all these we conclude with Theorem 5 below, where we define

$$\lambda_{\text{best}} := \arg\min_{\tau\ge 0} D(\tau), \quad \text{and} \quad \bar{C}(\tau) := \bar{C}_{f,\boldsymbol{\beta}_0}(\tau) = -(\tau/2)\partial\bar{D}(\tau)/\partial\tau. \tag{23}$$

Theorem 5 (Generalized LASSO: high-SNR regime) [13] *Let* $\frac{n}{d}\to\delta\in(0,\infty)$ *and* $\frac{D(\tau)}{d}\to\bar{D}(\tau)\in(0,1)$. *If* $\delta<1$, *define* λ_{crit} *as the unique solution of the equation* $\delta-\bar{D}(\tau)-\bar{C}(\tau)=0$ *for* $\tau\in[0,\lambda_{best}]$. *Otherwise, set* $\lambda_{crit}:=0$. *For any* $\lambda>0$, *let* $\hat{\lambda}=\min\{\lambda_{crit},\lambda\}$. *If* $\delta>\bar{D}(\hat{\lambda})$, *then,*

$$\lim_{\sigma^2\to 0}\frac{\|\hat{\boldsymbol{\beta}}^{\lambda,\sigma}-\boldsymbol{\beta}_0\|_2^2}{\sigma^2} \xrightarrow{P} \frac{\bar{D}(\hat{\lambda})}{\delta-\bar{D}(\hat{\lambda})}. \tag{24}$$

A few remarks are in place regarding (24) (we refer the reader to the relevant discussion in (Thrampoulidis et al., 2015b, Sec. II.C). First, the theorem holds for general convex regularizers and corresponding structures; the structure induced by f and the particular $\boldsymbol{\beta}_0$ are summarized in the geometric parameter D, which admits explicit closed form expressions for several popular regularizers and corresponding structures (e.g., (50) in D.1.1). Importantly, evaluating D only requires knowledge of the particular structure of the unknown signal $\boldsymbol{\beta}_0$ (e.g. sparsity), and not the explicit unknown signal itself. Besides, (24) characterizes the NSE for all values of λ. Minimizing the formula with respect to λ can lead to useful guidelines for tuning the regularizer parameter. It can be shown that the minimum is achieved for $\lambda=\lambda_{\text{best}}$ as defined in (23). Calculating λ_{best} does not require explicit knowledge of $\boldsymbol{\beta}_0$ itself, but only knowledge of the particular structure, e.g. of the sparsity level, or the rank (in practice approximate knowledge on these quantities might suffice).

13. The high-SNR NSE of the ℓ_2-LASSO was first studied (in a non-asymptotic setting) by the authors in Oymak et al. (2013). Theorem 3.2 therein recovers (24) for $\lambda\ge\lambda_{\text{crit}}$. Theorem 5 completes the proof for all values of λ. Most importantly, thanks to the transparent framework offered by Theorem 3, the analysis here is significantly simplified, shortened and insightful.

3.3.3. FINITE-SNR ANALYSIS

In the previous section, we studied the high-SNR NSE of the Generalized-LASSO. Our framework allows extending the analysis to all values of the noise variance. One needs to consider the following (AO) problem (we restrict attention to ℓ_1-regularization for concreteness):

$$\phi_{\ell_2\text{-LASSO}}(\mathbf{g},\mathbf{h}) = \min_{\|\mathbf{w}\|_2\leq K_{\mathbf{w}}} \max_{\|\mathbf{u}\|_2\leq 1} \sqrt{\|\mathbf{w}\|_2^2+\sigma^2}\, \mathbf{g}^T\mathbf{u} + \|\mathbf{u}\|_2\mathbf{h}^T\mathbf{w} + \lambda\|\boldsymbol{\beta}_0+\mathbf{w}\|_1. \tag{25}$$

Although more involved when compared to (18), the analysis of (25) is completely do-able. Please refer to Thrampoulidis et al. (2015a) for the details.

3.3.4. LAD

Thus far, our examples involved loss function of the forms $\|\cdot\|_2$ or $\|\cdot\|_2^2$. Here, we consider the error analysis of the LAD in the presence of sparse noise. To study this, we assume that $\boldsymbol{\epsilon}$ is s-sparse with its non-zero entries i.i.d. $\mathcal{N}(0,\sigma^2)$. With these, it is not hard to see that the (AO) problem corresponding to the LAD estimator becomes (say, in the high-SNR regime):

$$\phi_{\text{LAD}}(\mathbf{g},\mathbf{h}) = \min_{\|\mathbf{w}\|_2\leq K_{\mathbf{w}}} \max_{\substack{\|\mathbf{u}\|_\infty\leq 1\\ \mathbf{s}\in\partial f(\boldsymbol{\beta}_0)}} \sqrt{\|\mathbf{w}\|_2^2+\sigma^2}\sum_{i=1}^{s}\mathbf{g}_i\mathbf{u}_i + \|\mathbf{w}\|_2\sum_{i=s+1}^{m}\mathbf{g}_i\mathbf{u}_i + (\|\mathbf{u}\|_2\mathbf{h}+\lambda\mathbf{s})^T\mathbf{w}$$

This has been analyzed in Thrampoulidis and Hassibi (2014). An interesting consequence of the analysis is an exact performance comparison between the LASSO and the LAD.

4. Conclusion

Starting with the work of Rudelson and Vershynin (2008), Gaussian comparison theorems have played instrumental role in developing a clear understanding of linear inverse problems when the measurement matrix follows the standard Gaussian distribution, Stojnic (2009); Oymak and Hassibi (2010); Chandrasekaran et al. (2012), etc. All works prior to Stojnic (2013a) use Gordon's original Theorem 1 to give "lower-bounds". Stojnic is attributed with the idea of using strong duality to obtain upper-bounds. However, all statements and proofs of our main Theorem 3 (CGMT) appear to be novel. First, we use a symmetrization trick to identify $\Phi(\mathbf{G})$ as the (PO) (which is slightly different than $\Phi(\mathbf{G},\mathbf{g})$ of Theorem 1). This is critical, and leads to identifying precise convexity conditions for the concentration result of (13) (and Corollary 4) to hold. As mentioned, the most important contribution is the conditions and the result of Theorem 3-(iii). Also, expressing the (RO) as in (5) seems novel. These, when combined allow the use of GMT for the error analysis of (RO) problems with loss functions other than the classically used $\mathcal{L}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2$. Using the framework of the CGMT, we analyzed the high-SNR squared error of the Generalized LASSO in Section 3.3.2. Our accompanying series of work applies the framework to more general settings, e.g. finite-SNR regime, LAD, etc.. Each of these cases involves a different (AO) problem that needs to be analyzed. The level of difficulty of the analysis varies from problem to problem, but, in all instances we have considered it turns out to be doable, and, significantly simpler than a direct analysis of the original (PO).

Summarizing, the CGMT offers a powerful machinery for the precise performance analysis of non-smooth convex optimization algorithms when they are used to recover structured signals from noisy linear observations. At the same time, we expect that it finds applications under more general settings and even to different problem setups. We briefly discuss a potential example in Section E.

References

Radoslaw Adamczak, Alexander E Litvak, Alain Pajor, and Nicole Tomczak-Jaegermann. Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling. *Constructive Approximation*, 34(1):61–88, 2011.

Per Kragh Andersen and Richard D Gill. Cox's regression model for counting processes: a large sample study. *The annals of statistics*, pages 1100–1120, 1982.

Arindam Banerjee, Sheng Chen, Farideh Fazayeli, and Vidyashankar Sivakumar. Estimation with norm regularization. In *Advances in Neural Information Processing Systems*, pages 1556–1564, 2014.

Mohsen Bayati and Andrea Montanari. The lasso risk for gaussian matrices. *Information Theory, IEEE Transactions on*, 58(4):1997–2017, 2012.

Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

Dimitri P Bertsekas, Angelia Nedić, and Asuman E Ozdaglar. *Convex analysis and optimization*. Athena Scientific Belmont, 2003.

Peter J Bickel, Yaacov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2009.

Emmanuel Candès and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351, 2007.

Emmanuel J Candès. Mathematics of sparsity (and few other things). 2014.

Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *Information Theory, IEEE Transactions on*, 52(2):489–509, 2006.

Emmanuel J Candès, Yonina C Eldar, Deanna Needell, and Paige Randall. Compressed sensing with coherent and redundant dictionaries. *Applied and Computational Harmonic Analysis*, 31 (1):59–73, 2011.

Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

David L Donoho. Compressed sensing. *Information Theory, IEEE Transactions on*, 52(4):1289–1306, 2006.

David L Donoho, Arian Maleki, and Andrea Montanari. The noise-sensitivity phase transition in compressed sensing. *Information Theory, IEEE Transactions on*, 57(10):6920–6941, 2011a.

David L Donoho, Arian Maleki, and Andrea Montanari. The noise-sensitivity phase transition in compressed sensing. *Information Theory, IEEE Transactions on*, 57(10):6920–6941, 2011b.

Theodoros Evgeniou, Massimiliano Pontil, and Tomaso Poggio. Regularization networks and support vector machines. *Advances in computational mathematics*, 13(1):1–50, 2000.

Rina Foygel and Lester Mackey. Corrupted sensing: Novel guarantees for separating structured signals. *Information Theory, IEEE Transactions on*, 60(2):1223–1247, 2014.

Yehoram Gordon. Some inequalities for gaussian processes and applications. *Israel Journal of Mathematics*, 50(4):265–289, 1985.

Yehoram Gordon. Elliptically contoured distributions. *Probability theory and related fields*, 76(4): 429–438, 1987.

Yehoram Gordon. *On Milman's inequality and random subspaces which escape through a mesh in* $\mathbb{R}^n$. Springer, 1988.

Guillaume Lecué and Shahar Mendelson. Sparse recovery under weak moment assumptions. *arXiv preprint arXiv:1401.2188*, 2014.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer, 1991.

Shahar Mendelson. Learning without concentration. In *Proceedings of The 27th Conference on Learning Theory*, pages 25–39, 2014.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245, 1994.

Samet Oymak and Babak Hassibi. New null space results and recovery thresholds for matrix rank minimization. *arXiv preprint arXiv:1011.6326*, 2010.

Samet Oymak and Babak Hassibi. Sharp mse bounds for proximal denoising. *arXiv preprint arXiv:1305.2714*, 2013.

Samet Oymak, Christos Thrampoulidis, and Babak Hassibi. The squared-error of generalized lasso: A precise analysis. *arXiv preprint arXiv:1311.0830*, 2013.

Mert Pilanci and Martin J Wainwright. Randomized sketches of convex programs with sharp guarantees. In *Information Theory (ISIT), 2014 IEEE International Symposium on*, pages 921–925. IEEE, 2014.

Calyampudi Radhakrishna Rao and Helge Toutenburg. *Linear models*. Springer, 1995.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 99:2241–2259, 2010.

R Tyrell Rockafellar. *Convex analysis*, volume 28. Princeton university press, 1997.

Mark Rudelson and Roman Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics*, 61(8):1025–1045, 2008.

Maurice Sion et al. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

Mihailo Stojnic. Various thresholds for ℓ_1-optimization in compressed sensing. *arXiv preprint arXiv:0907.3666*, 2009.

Mihailo Stojnic. A framework to characterize performance of lasso algorithms. *arXiv preprint arXiv:1303.7291*, 2013a.

Mihailo Stojnic. Meshes that trap random subspaces. *arXiv preprint arXiv:1304.0003*, 2013b.

Mihailo Stojnic. Spherical perceptron as a storage memory with limited errors. *arXiv preprint arXiv:1306.3809*, 2013c.

Mihailo Stojnic. Upper-bounding ℓ_1-optimization weak thresholds. *arXiv preprint arXiv:1303.7289*, 2013d.

Christos Thrampoulidis and Babak Hassibi. Estimating structured signals in sparse noise: A precise noise sensitivity analysis. In *Communication, Control, and Computing (Allerton), 2014 52nd Annual Allerton Conference on*, pages 866–873. IEEE, 2014.

Christos Thrampoulidis and Babak Hassibi. Isotropically random orthogonal matrices: Performance of lasso and minimum conic singular values. *arXiv preprint arXiv:1503.07236, accepted to ISIT 2015*, 2015.

Christos Thrampoulidis, Ashkan Panahi, Daniel Guo, and Babak Hassibi. Precise error analysis of the lasso. In *in 40th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2015, available on arXiv:1502.04977*, 2015a.

Christos Thrampoulidis, Ashkan Panahi, and Babak Hassibi. Asymptotically exact error analysis for the generalized ℓ_2^2-lasso. *arXiv preprint arXiv:1502.04977, accepted to ISIT 2015*, 2015b.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Roman Vershynin. Estimation in high dimensions: a geometric perspective. *arXiv preprint arXiv:1405.5103*, 2014.

Lie Wang. The l1 penalized lad estimator for high dimensional linear regression. *Journal of Multivariate Analysis*, 120:135–151, 2013.

John Wright and Yi Ma. Dense error correction via-minimization. *Information Theory, IEEE Transactions on*, 56(7):3540–3560, 2010.

Yihong Wu and Sergio Verdú. Optimal phase transitions in compressed sensing. *Information Theory, IEEE Transactions on*, 58(10):6241–6263, 2012.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Tong Zhang et al. Some sharp performance bounds for least squares regression with l1 regularization. *The Annals of Statistics*, 37(5A):2109–2144, 2009.

Appendix A. Gordon's Gaussian Min-max Theorem

Gaussian comparison theorems are powerful tools in probability theory Ledoux and Talagrand (1991). A particularly useful such comparison inequality is described by Gordon's comparison theorem. In fact. Gordon's theorem is a generalization of the classical Slepian's lemma and Fernique's theorem Gordon (1985). It was first proved by Y. Gordon in Gordon (1985), where it was also shown how it can be used as an alternative to (re)-derive other well-known results in the field. See also Gordon (1987) for slight generalized versions of the theorem and the classical reference (Ledoux and Talagrand, 1991, Chapter 3.3) for an introduction to gaussian comparison theorems and some applications.

Theorem A.0.1 (Gordon's Gaussian comparison theorem, Gordon (1985)) *Let* $\{X_{ij}\}$ *and* $\{Y_{ij}\}$, $1 \leq i \leq I$, $1 \leq j \leq J$, *be centered Gaussian processes such that*

$$\begin{cases} \mathbb{E}X_{ij}^2 = \mathbb{E}Y_{ij}^2, & \text{for all } i,j, \\ \mathbb{E}X_{ij}X_{ik} \geq \mathbb{E}Y_{ij}Y_{ik}, & \text{for all } i,j,k, \\ \mathbb{E}X_{ij}X_{\ell k} \leq \mathbb{E}Y_{ij}Y_{\ell k}, & \text{for all } i \neq \ell \text{ and } j,k. \end{cases}$$

Then, for all $\lambda_{ij} \in \mathbf{R}$,

$$\mathbb{P}\left(\bigcap_{i=1}^{I}\bigcup_{j=1}^{J}[Y_{ij} \geq \lambda_{ij}]\right) \geq \mathbb{P}\left(\bigcap_{i=1}^{I}\bigcup_{j=1}^{J}[X_{ij} \geq \lambda_{ij}]\right).$$

Gordon's Theorem A.0.1 establishes a probabilistic comparison between two abstract Gaussian processes $\{X_{ij}\}$ and $\{Y_{ij}\}$ based on conditions on their corresponding covariance structures. Theorem 1 is a corollary of Theorem A.0.1 when applied to specific Gaussian processes.

We begin with using Theorem A.0.1 to prove an analogue of Theorem 1 for discrete sets. The proof is almost identical to the proof of Gordon's original Lemma 3.1 in Gordon (1988). Nevertheless, we include it here for completeness. Theorem 1 then follows from Lemm A.0.1 by a compactness argument.

Onwards, we suppress notation and write $\|\cdot\|$ instead of $\|\cdot\|_2$.

Lemma A.0.1 (Gordon's Gaussian Min-max Theorem: Discrete Sets) *Let* $\mathbf{X} \in \mathbb{R}^{n\times d}$, $g \in \mathbb{R}$, $\mathbf{g} \in \mathbb{R}^n$ *and* $\mathbf{h} \in \mathbb{R}^d$ *have entries i.i.d.* $\mathcal{N}(0,1)$ *and be independent of each other. Also, let* $\mathcal{I}_1 \subset \mathbb{R}^d$, $\mathcal{I}_2 \subset \mathbb{R}^n$ *be finite sets of vectors and* $\psi(\cdot,\cdot)$ *be a finite function defined on* $\mathcal{I}_1 \times \mathcal{I}_2$. *For all* $c > 0$,

$$\mathbb{P}\left(\min_{\mathbf{w}\in\mathcal{I}_1}\max_{\mathbf{u}\in\mathcal{I}_2}\ \left\{\mathbf{u}^T\mathbf{X}\mathbf{w} + g\|\mathbf{w}\|\|\mathbf{u}\| + \psi(\mathbf{w},\mathbf{u})\right\} \geq c\right) \geq \mathbb{P}\left(\min_{\mathbf{w}\in\mathcal{I}_1}\max_{\mathbf{u}\in\mathcal{I}_2}\left\{\|\mathbf{w}\|\mathbf{g}^T\mathbf{u} + \|\mathbf{u}\|\mathbf{h}^T\mathbf{w} + \psi(\mathbf{w},\mathbf{u})\right\} \geq c\right)$$

Proof Define two Gaussian processes indexed on the set $\mathcal{I}_1 \times \mathcal{I}_2$:

$$Y_{\mathbf{w},\mathbf{u}} = \mathbf{w}^T\mathbf{G}\mathbf{u} + g\|\mathbf{u}\|\|\mathbf{w}\| \quad \text{and} \quad X_{\mathbf{w},\mathbf{u}} = \|\mathbf{w}\|\mathbf{g}^T\mathbf{u} - \|\mathbf{u}\|\mathbf{h}^T\mathbf{w}.$$

First, we show that the processes defined satisfy the conditions of Gordon's Theorem A.0.1. Clearly, they are both centered. Furthermore, for all $\mathbf{w},\mathbf{w}' \in \mathcal{I}_1$ and $\mathbf{u},\mathbf{u}' \in \mathcal{I}_2$:

$$\mathbb{E}[X^2_{\mathbf{w},\mathbf{u}}] = \|\mathbf{w}\|^2\|\mathbf{u}\|^2 + \|\mathbf{u}\|^2\|\mathbf{w}\|^2 = \mathbb{E}[Y^2_{\mathbf{w},\mathbf{u}}],$$

and

$$\begin{aligned}\mathbb{E}[X_{\mathbf{w},\mathbf{u}}X_{\mathbf{w}',\mathbf{u}'}] - \mathbb{E}[Y_{\mathbf{w},\mathbf{u}}Y_{\mathbf{w}',\mathbf{u}'}] &= \|\mathbf{w}\|\|\mathbf{w}'\|(\mathbf{u}^T\mathbf{u}') + \|\mathbf{u}\|^2(\mathbf{w}^T\mathbf{w}') - (\mathbf{w}^T\mathbf{w}')(\mathbf{u}^T\mathbf{u}') - \|\mathbf{u}\|\|\mathbf{u}'\|\|\mathbf{w}\|\|\mathbf{w}'\| \\ &= \left(\underbrace{\|\mathbf{w}\|\|\mathbf{w}'\| - (\mathbf{w}^T\mathbf{w}')}_{\geq 0}\right)\left(\underbrace{(\mathbf{u}^T\mathbf{u}') - \|\mathbf{u}\|\|\mathbf{u}'\|}_{\leq 0}\right),\end{aligned}$$

which is non positive and equal to zero when $\mathbf{w} = \mathbf{w}'$.

Next, for each $(\mathbf{w},\mathbf{u}) \in \mathcal{I}_1 \times \mathcal{I}_2$, let $\lambda_{\mathbf{w},\mathbf{u}} = -\psi(\mathbf{w},\mathbf{u}) + c$ and apply Theorem A.0.1. This completes the proof by observing that

$$\left[\min_{\mathbf{w}\in\mathcal{I}_1}\max_{\mathbf{u}\in\mathcal{I}_2}\{Y_{\mathbf{w},\mathbf{u}} + \psi(\mathbf{w},\mathbf{u})\} \geq c\right] = \bigcap_{\mathbf{w}\in\mathcal{I}_1}\bigcup_{\mathbf{u}\in\mathcal{I}_2}\left[Y_{\mathbf{w},\mathbf{u}} \geq \lambda_{\mathbf{w},\mathbf{u}}\right],$$

and similar for the process $X_{\mathbf{w},\mathbf{u}}$. ■

Proof (of Theorem 1) Denote $R_1 := \max_{\mathbf{w}\in\mathcal{S}_\mathbf{w}}\|\mathbf{w}\|$ and $R_2 := \max_{\mathbf{u}\in\mathcal{S}_\mathbf{u}}\|\mathbf{u}\|$. Fix any $\epsilon > 0$. Since $\psi(\cdot,\cdot)$ is continuous and the sets $\mathcal{S}_\mathbf{w}$,$\mathcal{S}_\mathbf{u}$ are compact, $\psi(\cdot,\cdot)$ is uniformly continuous on $\mathcal{S}_\mathbf{w} \times \mathcal{S}_\mathbf{u}$. Thus, there exists $\delta := \delta(\epsilon) > 0$ such that for every $(\mathbf{w},\mathbf{u}),(\tilde{\mathbf{w}},\tilde{\mathbf{u}}) \in \mathcal{S}_\mathbf{w} \times \mathcal{S}_\mathbf{u}$ with $\left\|\begin{bmatrix}\mathbf{w} & \mathbf{u}\end{bmatrix} - \begin{bmatrix}\tilde{\mathbf{w}} & \tilde{\mathbf{u}}\end{bmatrix}\right\| \leq \delta$, we have that $|\psi(\mathbf{w},\mathbf{u}) - \psi(\tilde{\mathbf{w}},\tilde{\mathbf{u}})| \leq \epsilon$. Let $\mathcal{S}^\delta_\mathbf{w}$,$\mathcal{S}^\delta_\mathbf{u}$ be δ-nets of the sets $\mathcal{S}_\mathbf{w}$ and $\mathcal{S}_\mathbf{u}$, respectively. Then, for any $\mathbf{w} \in \mathcal{S}_\mathbf{w}$, there exists $\mathbf{w}' \in \mathcal{S}^\delta_\mathbf{w}$ such that $\|\mathbf{w} - \mathbf{w}'\| \leq \delta$ and an analogous statement holds for $\mathcal{S}_\mathbf{u}$. In what follows, for any vector $\mathbf{v}$ in a set $\mathcal{S}$, we denote $\mathbf{v}'$ the element in the δ-net of $\mathcal{S}$ that is the closest to $\mathbf{v}$ in the usual ℓ_2-metric. To simplify notation, denote

$$\alpha(\mathbf{w},\mathbf{u}) := \mathbf{u}^T\mathbf{X}\mathbf{w} + g\|\mathbf{w}\|\|\mathbf{u}\| + \psi(\mathbf{w},\mathbf{u}) \quad \text{and} \quad \beta(\mathbf{w},\mathbf{u}) := \|\mathbf{w}\|\mathbf{g}^T\mathbf{u} + \|\mathbf{u}\|\mathbf{h}^T\mathbf{w} + \psi(\mathbf{w},\mathbf{u}).$$

From Lemma A.0.1, we know that for all $c \in \mathbb{R}$:

$$\mathbb{P}\left(\min_{\mathbf{w}\in\mathcal{S}^\delta_\mathbf{w}}\max_{\mathbf{u}\in\mathcal{S}^\delta_\mathbf{u}}\alpha(\mathbf{w},\mathbf{u}) \geq c\right) \geq \mathbb{P}\left(\min_{\mathbf{w}\in\mathcal{S}^\delta_\mathbf{w}}\max_{\mathbf{u}\in\mathcal{S}^\delta_\mathbf{u}}\beta(\mathbf{w},\mathbf{u}) \geq c\right). \tag{26}$$

In what follows we show that constraining the minimax optimizations over only the δ-nets $\mathcal{S}_{\mathbf{w}}^{\delta}, \mathcal{S}_{\mathbf{u}}^{\delta}$ instead of the entire sets $\mathcal{S}_{\mathbf{w}}$,$\mathcal{S}_{\mathbf{u}}$, changes the achieved optimal values by only a small amount.

First, we calculate an upper bound on

$$\begin{aligned}
\min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}^{\delta}}\max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}^{\delta}}\alpha(\mathbf{w},\mathbf{u}) - \min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}}\max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}}\alpha(\mathbf{w},\mathbf{u}) &\le \min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}^{\delta}}\max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}^{\delta}}\alpha(\mathbf{w},\mathbf{u}) - \min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}}\max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}^{\delta}}\alpha(\mathbf{w},\mathbf{u}) =: \alpha(\mathbf{w}_1,\mathbf{u}_1) - \alpha(\mathbf{w}_2,\mathbf{u}_2)\\
&\le \max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}^{\delta}}\alpha(\mathbf{w}_2',\mathbf{u}) - \alpha(\mathbf{w}_2,\mathbf{u}_2) =: \alpha(\mathbf{w}_2',\mathbf{u}_*) - \alpha(\mathbf{w}_2,\mathbf{u}_2)\\
&\le \alpha(\mathbf{w}_2',\mathbf{u}_*) - \alpha(\mathbf{w}_2,\mathbf{u}_*)\\
&= \mathbf{u}_*^T\mathbf{X}(\mathbf{w}_2'-\mathbf{w}_2) + g\|\mathbf{u}_*\|(\|\mathbf{w}_2'\|-\|\mathbf{w}_2\|) + (\psi(\mathbf{w}_2',\mathbf{u}_*)-\psi(\mathbf{w}_2,\mathbf{u}_*))\\
&\le (\|\mathbf{X}\|_2+|g|)\underbrace{\|\mathbf{u}_*\|}_{\le R_2}\underbrace{\|\mathbf{w}_2'-\mathbf{w}_2\|}_{\le\delta} + \underbrace{|\psi(\mathbf{w}_2',\mathbf{u}_*)-\psi(\mathbf{w}_2,\mathbf{u}_*)|}_{\le\epsilon}\\
&\le (\|\mathbf{X}\|_2+|g|)R_2\delta+\epsilon.
\end{aligned}$$

From this, we have that

$$\mathbb{P}\left(\min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}}\max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}}\alpha(\mathbf{w},\mathbf{u})\ge c\right) \ge \mathbb{P}\left(\min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}^{\delta}}\max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}^{\delta}}\alpha(\mathbf{w},\mathbf{u})\ge c+(\|\mathbf{X}\|_2+|g|)R_2\delta+\epsilon\right). \quad (27)$$

Using standard concentration results on Gaussians, it is shown in Lemma B.0.2 that for all $t>0$,

$$\mathbb{P}(\|\mathbf{X}\|_2+|g|\le\sqrt{n}+\sqrt{d}+1+t)\ge 1-2\exp(-t^2/4).$$

This, when combined with (27) yileds:

$$\mathbb{P}\left(\min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}}\max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}}\alpha(\mathbf{w},\mathbf{u})\ge c\right) \ge \mathbb{P}\left(\min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}^{\delta}}\max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}^{\delta}}\alpha(\mathbf{w},\mathbf{u})\ge c+(\sqrt{d}+\sqrt{n}+1+t)R_2\delta+\epsilon\right) - 2e^{-t^2/4}. \quad (28)$$

Similarly,

$$\begin{aligned}
\min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}^{\delta}}\max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}^{\delta}}\beta(\mathbf{w},\mathbf{u}) - \min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}}\max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}}\beta(\mathbf{w},\mathbf{u}) &\ge \min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}^{\delta}}\max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}^{\delta}}\beta(\mathbf{w},\mathbf{u}) - \min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}^{\delta}}\max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}}\beta(\mathbf{w},\mathbf{u}) =: \beta(\mathbf{w}_1,\mathbf{u}_1) - \beta(\mathbf{w}_2,\mathbf{u}_2)\\
&\ge \beta(\mathbf{w}_1,\mathbf{u}_1) - \max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}}\beta(\mathbf{w}_1,\mathbf{u}) =: \beta(\mathbf{w}_1,\mathbf{u}_1) - \beta(\mathbf{w}_1,\mathbf{u}_*)\\
&\ge \beta(\mathbf{w}_1,\mathbf{u}_*') - \beta(\mathbf{w}_1,\mathbf{u}_*)\\
&= \|\mathbf{w}_1\|\mathbf{g}^T(\mathbf{u}_*'-\mathbf{u}_*) + (\|\mathbf{u}_*'\|-\|\mathbf{u}_*\|)\mathbf{h}^T\mathbf{w}_1 + (\psi(\mathbf{w}_1,\mathbf{u}_*')-\psi(\mathbf{w}_1,\mathbf{u}_*))\\
&\ge -(\|\mathbf{g}\|+\|\mathbf{h}\|)\underbrace{\|\mathbf{w}_1\|}_{\le R_1}\underbrace{\|\mathbf{u}_*'-\mathbf{u}_*\|}_{\le\delta} - \underbrace{|\psi(\mathbf{w}_1,\mathbf{u}_*')-\psi(\mathbf{w}_1,\mathbf{u}_*)|}_{\le\epsilon}\\
&\ge -(\|\mathbf{g}\|+\|\mathbf{h}\|)R_1\delta-\epsilon.
\end{aligned}$$

Thus,

$$\mathbb{P}\left(\min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}}\max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}}\beta(\mathbf{w},\mathbf{u})\ge c+(\|\mathbf{g}\|+\|\mathbf{h}\|)R_1\delta+\epsilon\right) \le \mathbb{P}\left(\min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}^{\delta}}\max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}^{\delta}}\beta(\mathbf{w},\mathbf{u})\ge c\right),$$

and a further application of Lemma B.0.2 shows that for all $t > 0$:

$$\mathbb{P}\left(\min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}}\max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}}\beta(\mathbf{w},\mathbf{u}) \geq c + (\sqrt{d}+\sqrt{n}+t)R_2\delta + \epsilon\right) - 2e^{-t^2/4} \leq \mathbb{P}\left(\min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}^{\delta}}\max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}^{\delta}}\beta(\mathbf{w},\mathbf{u}) \geq c\right), \tag{29}$$

Now, we can apply (26) in order to combine (28) and (29) to yield the following:

$$\mathbb{P}\left(\min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}}\max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}}\alpha(\mathbf{w},\mathbf{u}) \geq c\right) \geq \\ \mathbb{P}\left(\min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}}\max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}}\beta(\mathbf{w},\mathbf{u}) \geq c + (\sqrt{d}+\sqrt{n}+1+t)(R_1+R_2)\delta + 2\epsilon\right) - 4e^{-t^2/4}.$$

This holds for all $\epsilon > 0$ and all $t > 0$. In particular, set $t = \delta^{-\frac{1}{2}}$ and take the limit of the right-hand side as $\epsilon \to 0$. Then, $t \to \infty$ and we can of course choose $\delta \to 0$, which proves that

$$\mathbb{P}\left(\min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}}\max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}}\alpha(\mathbf{w},\mathbf{u}) \geq c\right) \geq \mathbb{P}\left(\min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}}\max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}}\beta(\mathbf{w},\mathbf{u}) > c\right).$$

■

Appendix B. Auxiliary Results

Definition B.0.1 (Lipschitz) *We say that a function $f : \mathbb{R}^d \to \mathbb{R}$ is Lipschitz with constant L or is L-Lipschitz if $|f(\mathbf{w}) - f(\mathbf{u})| \leq L\|\mathbf{w}-\mathbf{u}\|$ for all $\mathbf{w}, \mathbf{u} \in \mathbb{R}^d$.*

Proposition B.0.1 (Gaussian Lipschitz concentration) *(e.g.,(Boucheron et al., 2013, Theorem 5.6)) Let $\mathbf{w} \in \mathbb{R}^d$ have entries i.i.d. $\mathcal{N}(0,1)$ and $f : \mathbb{R}^d \to \mathbb{R}$ be L-Lipschitz. Then, each one of the events $\{f(\mathbf{w}) > \mathbb{E}f(\mathbf{w}) + t\}$ and $\{f(\mathbf{w}) < \mathbb{E}f(\mathbf{w}) - t\}$ occurs with probability no greater than $\exp\left(-t^2/(2L^2)\right)$.*

Lemma B.0.2 *Let $\mathbf{X} \in \mathbb{R}^{n\times d}$, $g \in \mathbb{R}$, $\mathbf{g} \in \mathbb{R}^n$ and $\mathbf{h} \in \mathbb{R}^d$ have entries i.i.d. $\mathcal{N}(0,1)$ and be independent of each other. Then, for all $t > 0$, each one of the events*

$$\{\|\mathbf{X}\|_2 + |g| \leq \sqrt{d}+\sqrt{n}+1+t\} \quad \textit{and} \quad \{\|\mathbf{h}\|_2 + \|\mathbf{g}\|_2 \leq \sqrt{d}+\sqrt{n}+t\}, \tag{30}$$

holds with probability at least $1 - 2\exp(-t^2/4)$.

Proof A well-known non-asymptotic bound on the largest singular value of an $n \times d$ Gaussian matrix shows (e.g. (Vershynin, 2010, Corollary 5.35)) that for all $t > 0$:

$$\mathbb{P}\left(\|\mathbf{X}\|_2 > \sqrt{n}+\sqrt{d}+t\right) \leq \exp(-t^2/2).$$

Also, $\|\cdot\|_2$ is an 1-Lipschitz function and for a standard gaussian vector $\mathbf{v} \in \mathbb{R}^d$: $\mathbb{E}\|\mathbf{v}\|_2 \leq \sqrt{d}$. Applying Proposition B.0.1 we have that for all $t > 0$ the events $\{|g| > 1 + t\}$, $\{\|\mathbf{g}\|_2 > \sqrt{n} + t\}$

and $\{\|\mathbf{h}\|_2 > \sqrt{d}+t\}$, each one occurs with probability no larger than $\exp(-t^2/2)$. Combining those,

$$\begin{aligned}\mathbb{P}\left(\|\mathbf{X}\|_2+|g|\le\sqrt{d}+\sqrt{n}+1+t\right) &\ge \mathbb{P}\left(\|\mathbf{X}\|_2\le\sqrt{d}+\sqrt{n}+t/2\,,\ |g|\le 1+t/2\right)\\ &\ge 1-\mathbb{P}\left(\|\mathbf{X}\|_2>\sqrt{d}+\sqrt{n}+t/2\right)-\mathbb{P}\left(\,|g|>1+t/2\right)\\ &\ge 1-2\exp(-t^2/4).\end{aligned}$$

The proof of the second statement is identical and is omitted for brevity. ■

Lemma B.0.3 (Lipschitzness of the AO problem) *Let $\mathcal{S}_{\mathbf{w}}\subset\mathbb{R}^d$, $\mathcal{S}_{\mathbf{u}}\subset\mathbb{R}^n$ be compact sets and function $\phi:\mathbb{R}^n\times\mathbb{R}^d\to\mathbb{R}$:*

$$\phi(\mathbf{g},\mathbf{h}):=\min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}}\max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}}\|\mathbf{w}\|_2\mathbf{g}^T\mathbf{u}+\|\mathbf{u}\|_2\mathbf{h}^T\mathbf{w}+\psi(\mathbf{w},\mathbf{u}).$$

Further let $R_1=\max_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}}\|\mathbf{w}\|_2$ and $R_2=\max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}}\|\mathbf{u}\|_2$. Then, $\phi(\mathbf{g},\mathbf{h})$ is Lipschitz with constant $\sqrt{2}R_1R_2$.

Proof Fix any two pairs $(\mathbf{g}_1,\mathbf{h}_1)$ and $(\mathbf{g}_2,\mathbf{h}_2)$ and let

$$(\mathbf{w}_2,\mathbf{u}_2)=\arg\min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}}\max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}}\|\mathbf{w}\|\mathbf{g}_2^T\mathbf{u}+\|\mathbf{u}\|\mathbf{h}_2^T\mathbf{w}+\psi(\mathbf{w},\mathbf{u}),$$

and

$$\mathbf{u}_*=\arg\max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}}\|\mathbf{w}_2\|\mathbf{g}_1^T\mathbf{u}+\|\mathbf{u}\|\mathbf{h}_1^T\mathbf{w}_2+\psi(\mathbf{w}_2,\mathbf{u}).$$

Clearly,

$$\phi(\mathbf{g}_1,\mathbf{h}_1)\le\|\mathbf{w}_2\|\mathbf{g}_1^T\mathbf{u}_*+\|\mathbf{u}_*\|\mathbf{h}_1^T\mathbf{w}_2+\psi(\mathbf{w}_2,\mathbf{u}_*),$$

and

$$\phi(\mathbf{g}_2,\mathbf{h}_2)\ge\|\mathbf{w}_2\|\mathbf{g}_2^T\mathbf{u}_*+\|\mathbf{u}_*\|\mathbf{h}_2^T\mathbf{w}_2+\psi(\mathbf{w}_2,\mathbf{u}_*),$$

Without loss of generality, assume $\phi(\mathbf{g}_1,\mathbf{h}_1)\ge\phi(\mathbf{g}_2,\mathbf{h}_2)$. Then,

$$\begin{aligned}\phi(\mathbf{g}_1,\mathbf{h}_1)-\phi(\mathbf{g}_2,\mathbf{h}_2) &\le \|\mathbf{w}_2\|\mathbf{g}_1^T\mathbf{u}_*+\|\mathbf{u}_*\|\mathbf{h}_1^T\mathbf{w}_2+\psi(\mathbf{w}_2,\mathbf{u}_*)-(\|\mathbf{w}_2\|\mathbf{g}_2^T\mathbf{u}_*+\|\mathbf{u}_*\|\mathbf{h}_2^T\mathbf{w}_2+\psi(\mathbf{w}_2,\mathbf{u}_*))\\ &\le \|\mathbf{w}_2\|\mathbf{u}_*^T(\mathbf{g}_1-\mathbf{g}_2)+\|\mathbf{u}_*\|\mathbf{w}_2^T(\mathbf{h}_1-\mathbf{h}_2)\\ &\le \sqrt{\|\mathbf{w}_2\|^2\|\mathbf{u}_*\|^2+\|\mathbf{u}_*\|^2\|\mathbf{w}_2\|^2}\sqrt{\|\mathbf{g}_1-\mathbf{g}_2\|^2+\|\mathbf{h}_1-\mathbf{h}_2\|^2}\\ &\le R_1R_2\sqrt{2}\sqrt{\|\mathbf{g}_1-\mathbf{g}_2\|^2+\|\mathbf{h}_1-\mathbf{h}_2\|^2},\end{aligned}$$

where the penultimate inequality follows from Cauchy-Schwarz. ■

Appendix C. Proof of Theorem 3

For the proof of (12) and (13), we fix arbitrary $d \in \mathbb{N}$ and drop the superscript (d) to simplify notation.

Proof of (12): As discussed inequality (12) is an almost direct consequence of Theorem 1. Yet we need to get rid of the term "$g\|\mathbf{w}\|_2\|\mathbf{u}\|_2$" in (7) in Gordon's Theorem 1. The argument is rather simple but critical for the rest of the statements of Theorem 3. We will show that

$$\mathbb{P}\left(\Phi(\mathbf{G}) \leq c\right) \leq 2\mathbb{P}\left(\Phi(\mathbf{G}, g) \geq c\right). \tag{31}$$

Once this is established, (12) follows directly after applying Theorem 1. To prove (31), fix $\mathbf{G}$ and $g < 0$ and denote

$$f_1(\mathbf{w}, \mathbf{u}) = \mathbf{u}^T\mathbf{G}\mathbf{w} + \psi(\mathbf{w}, \mathbf{u}) \quad \text{and} \quad f_2(\mathbf{w}, \mathbf{u}) = \mathbf{u}^T\mathbf{G}\mathbf{w} + g\|\mathbf{w}\|_2\|\mathbf{u}\|_2 + \psi(\mathbf{w}, \mathbf{u}).$$

Clearly, $f_1(\mathbf{w}, \mathbf{u}) \geq f_2(\mathbf{w}, \mathbf{u})$ for all $(\mathbf{w}, \mathbf{u}) \in \mathcal{S}_\mathbf{w} \times \mathcal{S}_\mathbf{u}$. We may then write,

$$\begin{aligned}\min_{\mathbf{w}\in\mathcal{S}_\mathbf{w}} \max_{\mathbf{u}\in\mathcal{S}_\mathbf{u}} f_1(\mathbf{w}, \mathbf{u}) = f_1(\mathbf{w}_1, \mathbf{u}_1) &\geq f_1(\mathbf{w}_1, \mathbf{u}) \text{ for all } \mathbf{u} \in \mathcal{S}_\mathbf{u} \\ &\geq \max_{\mathbf{u}\in\mathcal{S}_\mathbf{u}} f_2(\mathbf{w}_1, \mathbf{u}) \geq \min_{\mathbf{w}\in\mathcal{S}_\mathbf{w}} \max_{\mathbf{u}\in\mathcal{S}_\mathbf{u}} f_2(\mathbf{w}, \mathbf{u}).\end{aligned}$$

This proves $\Phi(\mathbf{G}) \geq \Phi(\mathbf{G}, g)$, when $g < 0$. From this and from the independence of g and $\mathbf{G}$, for all $c \in \mathbb{R}$:

$$\mathbb{P}\left(\Phi(\mathbf{G}, g) \leq c \mid g < 0\right) \geq \mathbb{P}\left(\Phi(\mathbf{G}) \leq c \mid g < 0\right) = \mathbb{P}(\Phi(\mathbf{G}) \leq c).$$

When combined with $g \sim \mathcal{N}(0, 1)$, the above yields the desired inequality (31):

$$\mathbb{P}\left(\Phi(\mathbf{G}, g) \leq c\right) = \frac{1}{2}\mathbb{P}\left(\Phi(\mathbf{G}, g) \leq c \mid g > 0\right) + \frac{1}{2}\mathbb{P}\left(\Phi(\mathbf{G}, g) \leq c \mid g < 0\right) \geq \frac{1}{2}\mathbb{P}(\Phi(\mathbf{G}) \leq c).$$

Proof of (13): The additional convexity assumptions imposed in statement (ii) of the theorem are critical for the proof of (13). By assumption, the sets $\mathcal{S}_\mathbf{w}, \mathcal{S}_\mathbf{u}$ are non-empty, compact and convex. Furthermore, the function $\mathbf{u}^T\mathbf{G}\mathbf{w}+\psi(\mathbf{w}, \mathbf{u})$ is continuous, finite[14] and convex-concave on $\mathcal{S}_\mathbf{w} \times \mathcal{S}_\mathbf{u}$. Thus, we can apply the minimax result in (Rockafellar, 1997, Corollary 37.3.2) to exchange "min-max" with a "max-min" in (10a)[15]:

$$\Phi(\mathbf{G}) = \max_{\mathbf{u}\in\mathcal{S}_\mathbf{u}} \min_{\mathbf{w}\in\mathcal{S}_\mathbf{w}} \mathbf{u}^T\mathbf{G}\mathbf{w} + \psi(\mathbf{w}, \mathbf{u}).$$

It is convenient to rewrite the above as

$$-\Phi(\mathbf{G}) = \min_{\mathbf{u}\in\mathcal{S}_\mathbf{u}} \max_{\mathbf{w}\in\mathcal{S}_\mathbf{w}} -\mathbf{u}^T\mathbf{G}\mathbf{w} - \psi(\mathbf{w}, \mathbf{u}).$$

Then, using the symmetry of $\mathbf{G}$, we have that for any $c \in \mathbb{R}$:

$$\mathbb{P}\left(-\Phi(\mathbf{G}) \leq c\right) = \mathbb{P}\left(\min_{\mathbf{u}\in\mathcal{S}_\mathbf{u}} \max_{\mathbf{w}\in\mathcal{S}_\mathbf{w}} \left\{\mathbf{u}^T\mathbf{G}\mathbf{w} - \psi(\mathbf{w}, \mathbf{u})\right\} \leq c\right).$$

14. A continuous function on a compact set is bounded from Weierstrass extreme value theorem.
15. Flipping the order of min-max remains valid even under the weaker assumption of a *quasi*-convex-concave function ψ, (Sion et al., 1958, Thm. 3.4). Hence, (13) holds in this case too by the same argument.

Thus, we may apply[16] statement (i) of Theorem 3 (with the roles of $\mathbf{w}$ and $\mathbf{u}$ flipped):

$$\begin{aligned}\mathbb{P}\left(-\Phi(\mathbf{G})<c\right) &\leq 2\mathbb{P}\left(\min_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}}\max_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}}\left\{\|\mathbf{u}\|_2\mathbf{h}^T\mathbf{w}+\|\mathbf{w}\|_2\mathbf{g}^T\mathbf{u}-\psi(\mathbf{w},\mathbf{u})\right\}\leq c\right)\\ &= 2\mathbb{P}\left(\min_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}}\max_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}}\left\{-\|\mathbf{u}\|_2\mathbf{h}^T\mathbf{w}-\|\mathbf{w}\|_2\mathbf{g}^T\mathbf{u}-\psi(\mathbf{w},\mathbf{u})\right\}\leq c\right), \qquad (32)\end{aligned}$$

where the last equation follows because of the symmetry of $\mathbf{g}$ and $\mathbf{h}$. To continue, note that

$$\min_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}}\max_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}}\left\{-\|\mathbf{u}\|_2\mathbf{h}^T\mathbf{w}-\|\mathbf{w}\|_2\mathbf{g}^T\mathbf{u}-\psi(\mathbf{w},\mathbf{u})\right\} = -\max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}}\min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}}\left\{\|\mathbf{u}\|_2\mathbf{h}^T\mathbf{w}+\|\mathbf{w}\|_2\mathbf{g}^T\mathbf{u}+\psi(\mathbf{w},\mathbf{u})\right\},$$

and further apply the minimax inequality (Rockafellar, 1997, Lemma 36.1) which requires that for all $\mathbf{g},\mathbf{h}$

$$\max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}}\min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}}\left\{\|\mathbf{w}\|_2\mathbf{g}^T\mathbf{u}+\|\mathbf{u}\|_2\mathbf{h}^T\mathbf{w}+\psi(\mathbf{w},\mathbf{u})\right\}\leq\min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}}\max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}}\left\{\|\mathbf{w}\|_2\mathbf{g}^T\mathbf{u}+\|\mathbf{u}\|_2\mathbf{h}^T\mathbf{w}+\psi(\mathbf{w},\mathbf{u})\right\}:=\phi(\mathbf{g},\mathbf{h}).$$

These, when combined with (32), give $\mathbb{P}\left(-\Phi(\mathbf{G})<c\right)\leq 2\mathbb{P}\left(-\phi(\mathbf{g},\mathbf{h})\leq c\right)$. Apply this for $c=-(\mu+t)$ and combine with (12) for $c=\mu-t$, to conclude with (13) as desired.

Proof of (14): We start with some notation that simplifies the exposition. In what follows, $\mathbf{w}$ is always constrained to belong to the set $\mathcal{S}_{\mathbf{w}}^{(d)}$; we simply write $\min_{\mathbf{w}}$ instead of $\min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}^{(d)}}$. We will say that a sequence of events $\mathcal{E}^{(d)}$ holds/occurs with probability approaching (w.p.a.) 0 (or 1), if $\lim_{d\to\infty}\mathbb{P}(\mathcal{E}^{(d)})=0$, (or 1). Denote

$$\ell(\eta):=\{\alpha\mid|\alpha-\alpha_*|>\eta\}.$$

We will prove that for all $\eta>0$, the event $\|\mathbf{w}_\Phi^{(d)}(\mathbf{G})\|\in\ell(\eta)$ holds w.p.a. 1.

Consider the function $\Upsilon^{(d)}:\mathcal{S}_{\mathbf{w}}^{(d)}\to\mathbb{R}$:

$$\Upsilon^{(d)}(\mathbf{w};\mathbf{G})=\max_{\mathbf{u}\in\mathcal{S}_{\mathbf{u}}^{(d)}}\mathbf{u}^T\mathbf{G}\mathbf{w}+\psi(\mathbf{w},\mathbf{u}).$$

Observe that $\Phi^{(d)}(\mathbf{G})=\min_{\mathbf{w}}\Upsilon^{(d)}(\mathbf{w};\mathbf{G})=\Upsilon^{(d)}(\mathbf{w}_\Phi^{(d)}(\mathbf{G});\mathbf{G})$. It is not hard to see that it suffices to prove that for all $\eta>0$ there exists $\delta:=\delta(\eta)>0$ such that

$$\min_{\|\mathbf{w}\|\in\ell(\eta)}\Upsilon^{(d)}(\mathbf{w};\mathbf{G})<\min_{\mathbf{w}}\Upsilon^{(d)}(\mathbf{w};\mathbf{G})+\delta \qquad (33)$$

occurs w.p.a. 0.

In what follows, fix any $\eta>0$. Proving (33) takes the following two steps: (i) upper bound $\min_{\mathbf{w}\in\mathcal{S}_{\mathbf{w}}^{(d)}}\Upsilon^{(d)}(\mathbf{w};\mathbf{G})$, and (ii) lower bound $\min_{\|\mathbf{w}\|\in\ell(\eta)}\Upsilon^{(d)}(\mathbf{w};\mathbf{G})$.

Step 1: Fix some $\epsilon_1>0$ and consider the following event

$$\mathcal{E}^{(d)}(\epsilon_1)=\{\min_{\mathbf{w}}\Upsilon^{(d)}(\mathbf{w};\mathbf{G})>\kappa_*+\epsilon_1\}. \qquad (34)$$

16. Observe that the signs of $\mathbf{u}^T\mathbf{G}\mathbf{w}$, $\mathbf{g}^T\mathbf{u}$ and $\mathbf{h}^T\mathbf{w}$ do not matter because of the assumed symmetry in the distributions of $\mathbf{G},\mathbf{g}$ and $\mathbf{h}$.

Then, we may use statement (ii) of the theorem (cf. (13)) to show that

$$\mathbb{P}(\mathcal{E}^{(d)}(\epsilon_1)) = \mathbb{P}(\Phi^{(d)}(\mathbf{G}) > \kappa_* + \epsilon_1) \leq 2\mathbb{P}(\phi^{(d)}(\mathbf{g}, \mathbf{h}) \geq \kappa_* + \epsilon_1)$$

But, $\phi^{(d)}(\mathbf{g}, \mathbf{h}) \xrightarrow{P} \kappa_*$ by hypothesis of the theorem. Therefore, $\mathcal{E}^{(d)}$ occurs w.p.a. 0.

Step 2: Fix some $\epsilon_2 > 0$ and consider the following event:

$$\mathcal{H}(\epsilon_2) := \{\min_{\|\mathbf{w}\| \in \ell(\eta)} \Upsilon^{(d)}(\mathbf{w}; \mathbf{G}) < \kappa_* + \epsilon_2\}. \tag{35}$$

Using statement (i) of the theorem (cf. (13)) we have

$$\mathbb{P}(\mathcal{H}(\epsilon_2)) \leq 2\mathbb{P}(\min_{\|\mathbf{w}\| \in \ell(\eta)} \Upsilon^{(d)}(\mathbf{w}; \mathbf{G}) \leq \kappa_* + \epsilon_2). \tag{36}$$

We will upper bound the probability on the right hand side by conditioning on the event

$$\{\|\mathbf{w}_\phi^{(d)}(\mathbf{g}, \mathbf{h})\| \notin \ell(\eta/2)\},$$

which occurs w.p.a. 1, by assumption. In this event, it is not hard to see that

$$\|\mathbf{w}\| \in \ell(\eta) \Rightarrow |\|\mathbf{w}\| - \|\mathbf{w}_\phi^{(d)}(\mathbf{g}, \mathbf{h})\|| \geq \eta/2.$$

That is, conditioned on $\mathcal{E}^{(d)}$, the probability in (36) is further upper bounded by

$$\mathbb{P}(\min_{|\|\mathbf{w}\| - \|\mathbf{w}_\phi^{(d)}(\mathbf{g}, \mathbf{h})\|| \geq \eta/2} \Upsilon^{(d)}(\mathbf{w}; \mathbf{G}) \leq \kappa_* + \epsilon_2\ ). \tag{37}$$

We will condition once more, only this time it will be on the event

$$\{\phi^{(d)}(\mathbf{g}, \mathbf{h}) \geq \kappa_* - \epsilon_2/2\},$$

which occurs w.p.a. 1, by assumption. In this event, the probability in (37) is further upper bounded by

$$\mathbb{P}(\min_{|\|\mathbf{w}\| - \|\mathbf{w}_\phi^{(d)}(\mathbf{g}, \mathbf{h})\|| \geq \eta/2} \Upsilon^{(d)}(\mathbf{w}; \mathbf{G}) \leq \phi^{(d)}(\mathbf{g}, \mathbf{h}) + \epsilon_2/2\ ). \tag{38}$$

Finally, we condition on the event

$$\{\Upsilon^{(d)}(\mathbf{w}; \mathbf{G}) \geq \phi^{(d)}(\mathbf{g}, \mathbf{h}) + \tau(\|\mathbf{w}\| - \|\mathbf{w}_\phi^{(d)}(\mathbf{g}, \mathbf{h})\|)^2,\ \forall \mathbf{w} \in \mathcal{S}_\mathbf{w}\},$$

which also occurs w.p.a. 1, by assumption. In this event,

$$\min_{|\|\mathbf{w}\| - \|\mathbf{w}_\phi^{(d)}(\mathbf{g}, \mathbf{h})\|| \geq \eta/2} \Upsilon^{(d)}(\mathbf{w}; \mathbf{G}) \geq \phi^{(d)}(\mathbf{g}, \mathbf{h}) + \tau(\eta/2)^2.$$

Thus, the probability in (37) is further upper bounded by

$$\mathbb{P}(\ \tau(\eta/2)^2 \leq \epsilon_2/2\ ), \tag{39}$$

which is of course a deterministic event. To sum up, following the chain of inequalities implied by (36)-(39), we find that

$$\mathbb{P}(\mathcal{H}(\epsilon_2)) \leq 2\mathbb{P}(\ \tau(\eta/2)^2 \leq \epsilon_2/2\ ) + p^{(d)}(\epsilon_2),$$

where $p^{(d)}(\epsilon_2)$ converges to 0 as $d \to \infty$. In particular, $\mathcal{H}(\epsilon_2)$ occurs w.p.a. 0, for all ϵ_2 such that $\epsilon_2 < 2\tau(\eta/2)^2$.

We are now ready to conclude the proof. For any $\eta > 0$, choose $\epsilon_2 := \epsilon(\eta) := \tau(\eta/2)^2 > 0$, $\epsilon_1 := \epsilon_2/2$ and $\delta := \epsilon_2/4 > 0$. Consider the events $\mathcal{E}(\epsilon_1)$ and $\mathcal{H}(\epsilon_2)$ as defined in (34) and (35), respectively. For the particular choice of ϵ_1, ϵ_2 both events occur w.p.a. 0. Condition on both the complements of these events. Then, the probability of the event in (33) is upper bounded by

$$\mathbb{P}(\kappa_* + \epsilon_2 < \kappa_* + \epsilon_1 + \delta) + p^{(d)}(\epsilon_1, \epsilon_2) = \mathbb{P}(2 < 1) + p^{(d)}(\epsilon_1, \epsilon_2) = p^{(d)}(\epsilon_1, \epsilon_2),$$

where $p^{(d)}(\epsilon_1, \epsilon_2)$ converges to 0 as $d \to \infty$. This concludes the proof.

Appendix D. Proof of Theorem 5

In this section, we complete the analysis of Section 3.3.2 and the proof of Theorem 5. Recall that the (AO) problem of interest is given by (18), which we repeat here for convenience:

$$\phi(\mathbf{g}, \mathbf{h}) := \min_{\|\mathbf{w}\| \leq K_{\mathbf{w}}} \max_{\substack{\|\mathbf{u}\|_2 \leq 1 \\ \mathbf{s} \in \partial f(\boldsymbol{\beta}_0)}} \frac{1}{\sqrt{d}} \{(\|\mathbf{w}\|_2 \mathbf{g} - \epsilon)^T \mathbf{u} - (\|\mathbf{u}\|_2 \mathbf{h} - \lambda \mathbf{s})^T \mathbf{w}\}. \tag{40}$$

In agreement with the notation of Theorem 3, let $\mathbf{w}_\phi := \mathbf{w}_\phi(\mathbf{g}, \mathbf{h})$ denote any minimizer of (40). Also, as in Section 3.2.1, $K_{\mathbf{w}}$ is an (arbitrarily large) finite constant the value of which will be fixed later in the proof. It was shown that (40) simplifies to the following optimization, which only involves scalar variables:

$$\phi(\mathbf{g}, \mathbf{h}) = \min_{0 \leq \alpha \leq K_{\mathbf{w}}} \max_{0 < \beta \leq 1} \phi(\alpha, \beta; \mathbf{g}, \mathbf{h}) := \frac{1}{\sqrt{d}} \{\sqrt{\alpha^2 + \sigma^2}\ \|\mathbf{g}\|_2 \beta - \alpha\beta \cdot \text{dist}(\mathbf{h}, \frac{\lambda}{\beta} \partial f(\boldsymbol{\beta}_0))\}, \tag{41}$$

where, compared to (20) we have flipped the order of min-max: the objective is (strongly) convex in α and concave in β. To conclude the latter, identify the second term as the perspective function of the distance function which can be shown to be convex. Also, α in (41) plays the role of $\|\mathbf{w}\|$. Thus, if we let $\alpha_*(\mathbf{g}, \mathbf{h})$ denote the optimal above, then $\alpha_*(\mathbf{g}, \mathbf{h}) = \|\mathbf{w}_\phi(\mathbf{g}, \mathbf{h})\|_2$. Next, we consider the convergence properties of (41). From standard concentration inequalities of Lipschitz functions (e.g. (Oymak et al., 2013, Lem. B.2)) it can be shown

$$\frac{\|\mathbf{g}\|_2}{\sqrt{d}} \xrightarrow{P} \sqrt{\delta} \quad \text{and} \quad \frac{\text{dist}(\mathbf{h}, \tau)}{\sqrt{d}} \xrightarrow{P} \sqrt{D(\tau)}, \tag{42}$$

where, recall the definition of the Gaussian squared distance $D(\tau)$ in (21). In particular, convergence here is at an exponential rate. The objective function in (41) then converges point wise in α, β to

$$\kappa(\alpha, \beta) := \beta\sqrt{\alpha^2 + \sigma^2}\ \sqrt{\delta} - \alpha\beta\sqrt{\bar{D}(\lambda/\beta)}. \tag{43}$$

For now, fix α. The (sequence of) objective functions in (43) are concave with respect to β. We combine this with the fact that point-wise convergence in probability of convex functions implies uniform convergence on compact subspaces ((Andersen and Gill, 1982, Cor. II.1)) to conclude that

$$\max_{0<\beta\leq 1} \phi(\alpha,\beta;\mathbf{g},\mathbf{h}) \xrightarrow{P} \max_{0<\beta\leq 1} \kappa(\alpha,\beta). \tag{44}$$

Now, we view the functions in the panel above as functions of α. The first (strictly, this is a sequence of such functions over d) is convex in α, as the minima of convex functions. The latter is strongly convex, thus, it has a unique minimizer, say α_*. Thus, as in (Newey and McFadden, 1994, Thm. 2.7),

$$\phi(\mathbf{g},\mathbf{h}) = \min_{0\leq\alpha\leq K_{\mathbf{w}}} \max_{0<\beta\leq 1} \phi(\alpha,\beta;\mathbf{g},\mathbf{h}) \xrightarrow{P} \kappa_* := \min_{0\leq\alpha\leq K_{\mathbf{w}}} \max_{0<\beta\leq 1} \kappa(\alpha,\beta) \tag{45}$$

and the minimizer of the former converges to the unique minimizer α_* of the latter.

We calculate α_* via analyzing the deterministic min-max optimization in the RHS of (45). Start by flipping the order of min-max (everything is appropriately convex; in particular $\sqrt{\bar{D}(\tau)}$ is convex as the point wise limit of convex functions) and perform the optimization over α first. Also, we ignore the constraint $\alpha \leq K_{\mathbf{w}}$ for the moment; we return to that once we have performed the unconstrained minimization. Via direct differentiation, we find

$$\alpha_*(\beta) = \sigma\sqrt{\frac{\bar{D}(\lambda/\beta)}{\delta-\bar{D}(\lambda/\beta)}} \tag{46}$$

(In particular, this requires β such that $\delta > \bar{D}(\lambda/\beta)$.) Thus,

$$\kappa(\alpha_*(\beta),\beta) = \sigma\beta\sqrt{\delta-\bar{D}(\lambda/\beta)}.$$

Differentiating this with respect to β and using (23), gives

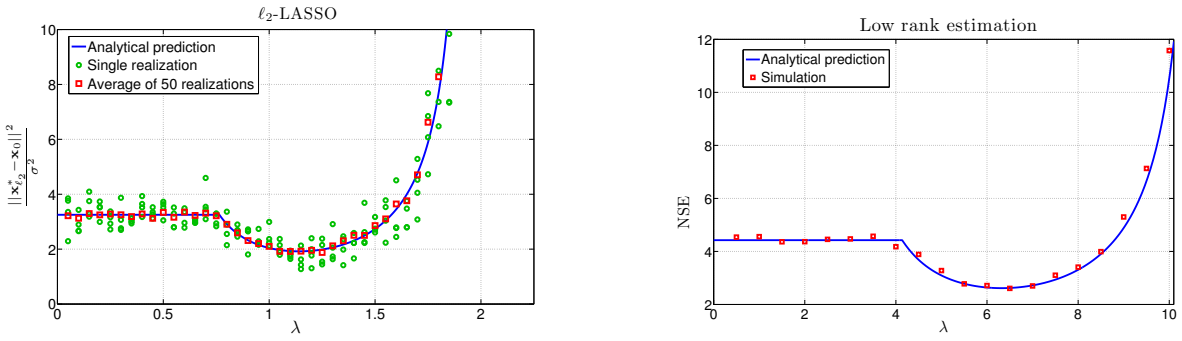
$$\frac{\partial\kappa}{\partial\beta} = \frac{\sigma}{\sqrt{\delta-\bar{D}(\lambda/\beta)}}\left(\delta-\bar{D}(\lambda/\beta)-\bar{C}(\lambda/\beta)\right). \tag{47}$$

Recall the assumption of the theorem that $\delta > \bar{D}(\hat{\lambda})$. Here, we prove the result for the underdetermined case where $\delta < 1$. The overdetermined case follows easily along the same arguments. First, if $\lambda \leq \lambda_{\text{crit}}$ such that $\hat{\lambda} = \lambda_{\text{crit}}$, then $\beta_* = \lambda/\lambda_{\text{crit}} \leq 1$ makes (47) zero (by definition of λ_{crit}) and from (46), $\alpha_* = \sigma\sqrt{\frac{\bar{D}(\lambda_{\text{crit}})}{\delta-\bar{D}(\lambda_{\text{crit}})}}$ (this is well defined by assumption $\delta > \bar{D}(\lambda_{\text{crit}})$). Next, if $\lambda > \lambda_{\text{crit}}$ (but such that $\delta > \bar{D}(\lambda)$), the derivative in (47) is non-negative at $\beta = 1$ (see (Oymak et al., 2013, Lem. 8.3)). From concavity, this implies optimality of $\beta_* = 1$. Substituting in (46), gives $\alpha_* = \sigma\sqrt{\frac{\bar{D}(\lambda)}{\delta-\bar{D}(\lambda)}}$. To conclude, if $\bar{D}(\hat{\lambda}) < \delta < 1$, the minimizer α_* of the unconstrained (disregarding $\alpha \leq K_{\mathbf{w}}$) optimization in (44) is

$$\alpha_* = \sigma\sqrt{\frac{\bar{D}(\hat{\lambda})}{\delta-\bar{D}(\hat{\lambda})}}. \tag{48}$$

Finally, we may now choose a value for $K_{\mathbf{w}}$ as promised. Setting, $K_{\mathbf{w}} = 2\alpha_*$, does not change the optimal. Combining these with Theorem 3–(iii), we have shown (under the assumptions of Theorem 5) that in the limit of $d \to \infty$, any minimizer $\hat{\mathbf{w}}$ of

$$\min_{\|\mathbf{w}\|\leq 2\alpha_*} \|\mathbf{X}\mathbf{w}-\boldsymbol{\epsilon}\|_2 + \lambda \max_{\mathbf{s}\in\partial f(\boldsymbol{\beta}_0)} \mathbf{s}^T\mathbf{w}, \tag{49}$$



(*a*) (*b*)

Figure 1: Illustration of the prediction of Theorem 5 for sparse and low-rank recovery. Plots of the NSE in high-SNR as a function of the regularizer parameter λ. Each simulation point represents an average over 50 realizations of $\mathbf{X}, \boldsymbol{\epsilon}, \mathbf{B}_0$. In both cases the noise variance is set to $\sigma^2 = 10^{-5}$. (a) $d = 1500$, $n = 750$, $k = \rho d = 150$, (b) $\sqrt{d} = 45$, $n = 0.6d$, $r = 6$.

is such that $\|\hat{\mathbf{w}}\|_2 \xrightarrow{P} \alpha_* > 0$. Using standard convexity argument, the conclusion remains unchanged for the original LASSO problem, i.e. the one without the (artificial) constraint on $\mathbf{w}$. This completes the proof.

D.1. Empirical Simulations

For completeness, we include two plots that illustrate the accuracy of Theorem 5 via numerical simulations. For more figures please refer to Oymak et al. (2013). Also, see Thrampoulidis and Hassibi (2014); Thrampoulidis et al. (2015a,b); Thrampoulidis and Hassibi (2015) for corresponding evidence regarding error predictions for other Regression Optimization problems.

D.1.1. SPARSE RECOVERY

Assume sparse signal $\boldsymbol{\beta}_0 \in \mathbb{R}^d$ with normalized sparsity level $\rho \in (0, 1)$, i.e. only $\rho \cdot d$ of its entries are non-zero. Consider solving the LASSO with ℓ_1-regularization:

$$\hat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + \lambda\|\boldsymbol{\beta}\|_1.$$

It can be easily shown (e.g. (Oymak et al., 2013, App. H)) that:

$$\begin{aligned}\bar{D}(\tau) &= \rho(1+\tau)^2 + (1-\rho)(2(1+\tau^2)Q(\tau) - \sqrt{2/\pi}\tau e^{-\frac{\tau^2}{2}}) \\ \bar{C}(\tau) &= -\rho\tau^2 + (1-\rho)(2\tau^2 Q(\tau) - \sqrt{2/\pi}\tau e^{-\frac{\tau^2}{2}}),\end{aligned} \tag{50}$$

where Q denotes the standard Q-function. With these expressions, we can numerically evaluate the formula of Theorem 3. An instance is shown in Figure 1(*a*), where the NSE is plotted as function of the regularizer parameter λ. To obtain the empirical points on the plot, we solve LASSO using CVX. The noise variance was chosen small enough to approximate $\sigma^2 \to 0$. (In particular, $\sigma^2 = 10^{-5}$ and $\|\boldsymbol{\beta}_0\|_2 = 1$). The prediction accuracy of Theorem 5 is clear.

D.1.2. LOW-RANK RECOVERY

Consider a low rank matrix $\mathbf{B}_0 \in \mathbb{R}^{\sqrt{d}\times\sqrt{d}}$. Then, $\boldsymbol{\beta}_0 = \text{vec}(\mathbf{B}_0)$ is the vector representation of $\mathbf{B}_0$ and $\mathbf{X}$ is a Gaussian linear map $\mathbb{R}^{\sqrt{d}\times\sqrt{d}} \to \mathbb{R}^n$. We solve,

$$\min_{\mathbf{B}\in\mathbb{R}^{d\times d}} \|\mathbf{y} - \mathbf{X}\cdot\text{vec}(\mathbf{B})\| + \lambda\sqrt{d}\|\mathbf{B}\|_\star.$$

where $\mathbf{y} = \mathbf{X}\cdot\text{vec}(\mathbf{B}_0) + \boldsymbol{\epsilon}$. Observe that we have appropriately normalized the regularizer (i.e. $f(\mathbf{B}) = \sqrt{d}\|\mathbf{B}\|_*$). This is necessary to satisfy the condition of Theorem 5 that $D(\tau)/d$ be constant independent of d. Please see (Oymak et al., 2013, Sec. H2) for explicit expressions of $D(\tau), C(\tau)$. Simulation results are shown in Figure 1(*b*). In the simulations we generate $\mathbf{B}_0$ as follows: we pick i.i.d. standard normal matrices $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{d\times r}$ and set $\mathbf{B}_0 = \frac{\mathbf{U}\mathbf{V}^T}{\|\mathbf{U}\mathbf{V}^T\|_F}$ which ensures $\mathbf{B}_0$ is unit norm and rank r.

Appendix E. Sketching Linear Regression or Sparsity in a Dictionary

We consider two problems that differ from the classical Regression Optimization setup considered in the main body of the paper, and, briefly discuss how Theorem 3 could prove useful for their analysis. In a first scenario, suppose $\boldsymbol{\beta}_0$ is a structured sparse signal, $\mathbf{D}$ is a large deterministic matrix, and, we observe $\mathbf{D}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}$, Pilanci and Wainwright (2014). Alternatively, $\boldsymbol{\beta}_0$ may be a sparse representation of the signal $\mathbf{D}\boldsymbol{\beta}_0$ under a dictionary $\mathbf{D}$, Candès et al. (2011). In the first case, instead of solving the LASSO with observations $\mathbf{y} = \mathbf{D}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}$, one can reduce the problem dimensionality by multiplying both sides with a Gaussian matrix $\mathbb{R}^{n\times d}$. In the latter case, we can consider estimation of the sparse features from a few linear observations of $\mathbf{D}\boldsymbol{\beta}_0$. It is desirable to give guarantees for this new problem which takes the following variation form of the LASSO:

$$\hat{\mathbf{w}}_{SLR} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{G}(\mathbf{D}\mathbf{w} + \boldsymbol{\epsilon})\|^2 + \lambda f(\boldsymbol{\beta}_0 + \mathbf{w}).$$

To predict the behavior of the residual $\hat{\mathbf{w}}_{SLR}$ one would need to analyze the corresponding (AO) problem, which takes the form

$$\phi_{\text{SLR}}(\mathbf{g}, \mathbf{h}) = \min_{\mathbf{w}} \frac{1}{2}(\|\mathbf{g}\|\|\mathbf{D}\mathbf{w} + \boldsymbol{\epsilon}\| + \mathbf{h}^T(\mathbf{D}\mathbf{w} + \boldsymbol{\epsilon}))^2 + f(\boldsymbol{\beta}_0 + \mathbf{w}).$$