A Geometric View on Constrained M-Estimators

Yen-Huan Li YEN-HUAN.LI@EPFL.CH
Ya-Ping Hsieh YA-PING.HSIEH@EPFL.CH
Volkan Cevher VOLKAN.CEVHER@EPFL.CH
Laboratory for Information and Inference Systems, École Polytechnique Fédérale de Lausanne, Switzerland

Abstract

We study the estimation error of constrained M-estimators from a geometric perspective. We present a novel general geometric framework for analyzing constrained M-estimators, and show how this geometric framework leads to explicit upper bounds on the expected estimation error determined by the (squared) Gaussian complexity of the constraint set. We study the cases where the true parameter is on the boundary of the constraint set (matched constraint), and where the true parameter is strictly in the constraint set (mismatched constraint). For both cases, we derive novel universal bounds for regression in a generalized linear model with the canonical link function.

1. Introduction and Problem Formulation

We consider constrained M-estimators of the form

$$\hat{\theta} \in \arg\min_{\theta\in\mathbb{R}^p} \left\{ f_n(\theta) : \theta \in \mathcal{G} \right\}, \tag{1}$$

where f_n is some continuously differentiable convex function, and $\mathcal{G} \subseteq \mathbb{R}^p$ is a compact convex set given by

$$\mathcal{G} := \{\theta \in \mathbb{R}^p : g(\theta) \leq c\} \tag{2}$$

with some continuous convex function g and some constant $c > 0$. For example, if we choose

$$f_n(\theta) := \frac{1}{2n}\sum_{i=1}^{n} \left(y_i - \langle a_i, \theta\rangle\right)^2, \tag{3}$$
$$g(\theta) := \|\theta\|_1,$$

for some given $y_1, \ldots, y_n \in \mathbb{R}$ and $a_1, \ldots, a_n \in \mathbb{R}^p$, the estimator $\hat{\theta}$ corresponds to the famous Lasso proposed in (Tibshirani, 1996). We may also consider each $\theta \in \mathbb{R}^p$ as the vectorization of a corresponding matrix $\Theta \in \mathbb{R}^{d\times d}$ with $d^2 = p$. If we choose

$$f_n(\Theta) := \frac{1}{2n}\sum_{i=1}^{n} \left(y_1 - \mathrm{Tr}\left(A_i^T\Theta\right)\right)^2,$$
$$g(\Theta) := \|\Theta\|_*,$$

for some given $y_1, \ldots, y_n \in \mathbb{R}$ and $A_1, \ldots, A_n \in \mathbb{R}^{d\times d}$, where $\|\cdot\|_*$ denotes the nuclear norm, the estimator $\hat{\theta}$ corresponds to the estimator proposed in (Davenport et al., 2014; Flammia et al., 2012; Gunasekar et al., 2014) for noisy matrix completion and compressive quantum tomography. In general, f_n can be the normalized negative log-likelihood function, or any properly defined function; g is usually designed to match the *a priori* information on the structure of the true parameter $\theta^\natural$ (Bach, 2013; Chandrasekaran et al., 2012; El Halabi & Cevher, 2015).

For convenience, we define $\iota_g : \mathbb{R}^p \to \mathbb{R}\cup\{+\infty\}$ as the indicator function of the set $\left\{\theta : g(\theta) \leq g(\theta^\natural)\right\}$; that is

$$\iota_g(\theta) := \begin{cases} 0 & , g(\theta) \leq c \\ +\infty & , \text{otherwise} \end{cases}.$$

Then $\hat{\theta}$ has an equivalent definition:

$$\hat{\theta} \in \arg\min_{\theta\in\mathbb{R}^p} \left\{ f_n(\theta) + \iota_g(\theta) \right\};$$

that is, $\hat{\theta}$ can be viewed as a regularized M-estimator with the regularization function ι_g. Note that ι_g is by definition a proper closed convex function. If we choose $c = g(\theta^\natural)$ in (2), then

$$\partial\iota_g(\theta^\natural) = \mathrm{cone}\left(\partial g(\theta^\natural)\right). \tag{4}$$

In general $\partial\iota_g(\theta^\natural)$ is always a closed convex cone.

In this paper, we study the estimation error of such constrained M-estimators from a geometric perspective, under the high-dimensional setting where $n < p$.

2. Related Works and Our Contribution

The estimation error of constrained estimators had been studied in various ways (Chatterjee, 2014; Oymak et al.,

2013a;b; Plan & Vershynin, 2013; Plan et al., 2014; Vershynin, 2014). The work (Chatterjee, 2014) studies the constrained least squares estimator for estimating the mean of a Gaussian process under constraint, and provides a concentration inequality for the estimation error. In (Oymak et al., 2013a;b), the authors derived very sharp error bounds for the estimation error for regression in the Gaussian linear model. However, the approaches in (Chatterjee, 2014; Oymak et al., 2013a;b) are not obviously extendable to non-linear statistical models, e.g., the logistic regression model. The constrained M-estimator for sparse logistic regression was analyzed in (Plan & Vershynin, 2013), but the estimator is not the popular maximum-likelihood estimator and requires knowledge of the sparsity level of the true parameter. The analysis in (Vershynin, 2014) is minimax in essence, and hence the error bound can be very loose for most of the cases; moreover, the framework is only for the Gaussian linear model, as it is based on ideas similar to Dvoretzky's theorem. In (Plan et al., 2014), the authors proposed a universal estimator for possibly non-linear statistical models, and derived a universal error bound for the proposed estimator. However, the analysis in (Plan et al., 2014) is specific to the proposed estimator. In general, even the error performance of the very common maximum-likelihood approach is unclear, to the best of our knowledge.

We summarize our contributions as follows.

- We propose a general geometric framework for analyzing an arbitrarily given constrained M-estimator in Section 3.
- Based on the geometric framework, we derive universal error bounds for constrained maximum-likelihood estimators in generalized linear models in Section 4 and Section 5.
- We identify an interesting coincidence between our result and the result in (Plan et al., 2014) in Section 5 and Section 6.3, though the estimator considered in (Plan et al., 2014) is very different from the maximum-likelihood estimator in general.

By the theory of Lagrange duality, there exists some $\rho_n > 0$ such that the constrained estimator given in (1) is equivalent to the regularized estimator

$$\hat{\theta}_{\text{regularized}} \in \arg\min_{\theta\in\mathbb{R}^p} \left\{ f_n(\theta) + \rho_n g(\theta) \right\},$$

which is also closely related to basis pursuit-like estimators in (Chandrasekaran et al., 2012). We show in Section 6 that our derived error bounds are actually compatible to those derived in (Bickel et al., 2009; Bühlmann & van de Geer, 2011; Chandrasekaran et al., 2012; Negahban et al., 2012; Honorio & Jaakkola, 2014; Kakade et al., 2010) for regularized and basis pursuit-like estimators in Section 6.

3. A Geometric Framework

3.1. Basic Idea

To illustrate the basic idea of the geometric framework, let us start with a simple setting, where f_n is strongly convex with parameter $\mu > 0$. Note that then $\hat{\theta}$ is uniquely defined.

By the optimality condition of $\hat{\theta}$, we have

$$0 \in \nabla f_n(\hat{\theta}) + \partial \iota_g(\hat{\theta}).$$

By the strong convexity of f_n, we have

$$\left\langle \nabla f_n(\hat{\theta}) - \nabla f_n(\theta^\natural), \hat{\theta} - \theta^\natural \right\rangle \geq \mu \left\| \hat{\theta} - \theta^\natural \right\|_2^2. \tag{5}$$

By the convexity of ι_g, we have

$$\left\langle \hat{z} - z^\natural, \hat{\theta} - \theta^\natural \right\rangle \geq 0, \tag{6}$$

for any $\hat{z} \in \partial \iota_g(\hat{\theta})$, and any $z^\natural \in \partial \iota_g(\theta^\natural)$. Summing up (5) and (6), we obtain

$$\left\langle \nabla f_n(\hat{\theta}) + \hat{z} - \nabla f_n(\theta^\natural) - z^\natural, \hat{\theta} - \theta^\natural \right\rangle \geq \mu \left\| \hat{\theta} - \theta^\natural \right\|_2^2,$$

for any $\hat{z} \in \partial \iota_g(\hat{\theta})$, and any $z^\natural \in \partial \iota_g(\theta^\natural)$. By the optimality condition, there exists some $\hat{z} \in \partial \iota_g(\hat{\theta})$ such that $\nabla f_n(\hat{\theta}) + \hat{z} = 0$, and hence we have

$$\left\langle -\nabla f_n(\theta^\natural) - z^\natural, \hat{\theta} - \theta^\natural \right\rangle \geq \mu \left\| \hat{\theta} - \theta^\natural \right\|_2^2,$$

for any $z^\natural \in \partial \iota_g(\theta^\natural)$. Since $\partial \iota_g(\theta^\natural)$ is always a closed convex cone, we may choose $z^\natural = 0$ and obtain

$$\left\langle -\nabla f_n(\theta^\natural), \hat{\theta} - \theta^\natural \right\rangle \geq \mu \left\| \hat{\theta} - \theta^\natural \right\|_2^2. \tag{7}$$

Applying the Cauchy-Schwarz inequality to the left-hand side, we obtain

$$\left\| \nabla f_n(\theta^\natural) \right\|_2 \left\| \hat{\theta} - \theta^\natural \right\|_2 \geq \mu \left\| \hat{\theta} - \theta^\natural \right\|_2^2,$$

or

$$\left\| \hat{\theta} - \theta^\natural \right\|_2 \leq \frac{1}{\mu} \left\| \nabla f_n(\theta^\natural) \right\|_2. \tag{8}$$

Taking expectations on both sides, we immediately obtain the following estimation error bound:

$$\mathbb{E} \left\| \hat{\theta} - \theta^\natural \right\|_2 \leq \frac{1}{\mu} \mathbb{E} \left\| \nabla f_n(\theta^\natural) \right\|_2. \tag{9}$$

The gradient at the true parameter $\nabla f_n(\theta^\natural)$ usually concentrates around 0 with high probability.

The simple error bound (9) is not desirable for two reasons:

1. In the high-dimensional setting where $n < p$, usually we do not have strong convexity of f_n.
2. The simple error bound does not depend on the choice of g.

We address the first issue in Section 3.2, and the second issue in Section 3.3.

3.2. Restricted Strong Convexity

Note that in order to facilitate the arguments in the previous sub-section, we only require (5) to hold for $\hat{\theta}$ and $\theta^\natural$, instead of any two vectors in $\mathbb{R}^p$. Therefore, we only need f_n to satisfy some *restricted* notion of strong convexity.

Definition 3.1 (Feasible Set and Feasible Cone)**.** The *feasible set* of g at $\theta^\natural$, denoted by $\mathcal{F}_g(\theta^\natural)$, is given by

$$\mathcal{F}_g(\theta^\natural) := \mathcal{G} - \theta^\natural.$$

The *feasible cone* of g at $\theta^\natural$, denoted by $\overline{\mathcal{F}_g(\theta^\natural)}$, is defined as the conic hull of $\mathcal{F}_g(\theta^\natural)$.

By the definition of $\hat{\theta}$, the estimation error must satisfy $\hat{\theta} - \theta^\natural \in \mathcal{F}_g(\theta^\natural)$.

Definition 3.2 (Restricted Strong Convexity)**.** The function f_n satisfies the restricted strong convexity (RSC) condition with parameter $\mu > 0$ if

$$\left\langle \nabla f_n(\theta^\natural + e) - \nabla f_n(\theta^\natural), e \right\rangle \geq \mu \left\| e \right\|_2^2, \qquad (10)$$

for any $e \in \mathcal{F}_g(\theta^\natural)$.

If f_n is twice continuously differentiable, we have another characterization.

Proposition 3.1. *The function f_n satisfies the RSC condition with parameter $\mu > 0$ if*

$$\left\langle e, \nabla^2 f_n(\theta^\natural + \lambda e) e \right\rangle \geq \mu \left\| e \right\|_2^2,$$

for all $\lambda \in [0, 1]$ and all $e \in \mathcal{F}_g(\theta^\natural)$.

Proof. We have

$$\begin{aligned} &\left\langle \nabla f_n(\theta^\natural + e) - \nabla f_n(\theta^\natural), e \right\rangle \\ &\quad = \int_0^1 \left\langle e, \nabla^2 f_n(\theta^\natural + \lambda e) e \right\rangle d\lambda. \end{aligned}$$

The right-hand side is always larger than $\mu \left\| e \right\|_2^2$ by assumption. □

The uniqueness of $\hat{\theta}$ and the derivation of the error bound in Section 3.1 are still valid even when $n < p$, as long as f_n satisfies the RSC condition with some parameter $\mu > 0$.

3.3. Refined Error Bound

We address the dependence of the estimation error on the choice of g, and derive a refined error bound in this subsection.

We note that

$$\left\langle -\nabla f_n(\theta^\natural), \hat{\theta} - \theta^\natural \right\rangle = \left\| \Pi_{\overline{\hat{\theta} - \theta^\natural}} \left(-\nabla f_n(\theta^\natural) \right) \right\|_2 \left\| \hat{\theta} - \theta^\natural \right\|_2,$$

where $\Pi_{\overline{\hat{\theta} - \theta^\natural}}(\cdot)$ denotes the projection onto the conic hull of $\left\{ \hat{\theta} - \theta^\natural \right\}$ (which is a half-line). This implies, by (7),

$$\left\| \Pi_{\overline{\hat{\theta} - \theta^\natural}} \left(-\nabla f_n(\theta^\natural) \right) \right\|_2 \geq \mu \left\| \hat{\theta} - \theta^\natural \right\|_2.$$

The left-hand side, however, is not tractable due to its dependence on $\hat{\theta}$, so we use a looser bound:

$$\sup_{e \in \mathcal{F}_g(\theta^\natural)} \left\{ \left\| \Pi_{\overline{e}} \left(-\nabla f_n(\theta^\natural) \right) \right\|_2 \right\} \geq \mu \left\| \hat{\theta} - \theta^\natural \right\|_2,$$

where $\Pi_{\overline{e}}(\cdot)$ denotes the projection onto the conic hull of $\{e\}$. The left-hand side is simply the projection of $-\nabla f_n(\theta^\natural)$ onto the feasible cone $\overline{\mathcal{F}_g(\theta^\natural)}$. Taking expectations on both sides, we obtain the following lemma.

Lemma 3.2. *Assume that f_n satisfies the RSC condition with parameter $\mu > 0$. Then $\hat{\theta}$ is uniquely defined, and satisfies*

$$\mathbb{E} \left\| \hat{\theta} - \theta^\natural \right\|_2 \leq \frac{1}{\mu} \mathbb{E} \left\| \Pi_{\overline{\mathcal{F}_g(\theta^\natural)}} (-\nabla f_n(\theta^\natural)) \right\|_2.$$

Since $-\nabla f_n(\theta^\natural)$ is a descent direction of f_n, if its direction is coherent with the feasible cone $\overline{\mathcal{F}_g(\theta^\natural)}$, we may find some point $\hat{\theta}'$ far away from $\theta^\natural$ in the feasible set $\mathcal{F}_g(\theta^\natural)$ such that $f_n(\hat{\theta}')$ is much smaller than $f_n(\theta^\natural)$, and hence the estimation error can be large. This provides an intuitive interpretation of the lemma.

Since projection onto a closed convex set is a non-expansive mapping, we have

$$\left\| \Pi_{\overline{\mathcal{F}_g(\theta^\natural)}} (-\nabla f_n(\theta^\natural)) \right\|_2 \leq \left\| \nabla f_n(\theta^\natural) \right\|_2,$$

so the error bound is always no larger than the one in Section 3.1.

Lemma 3.2 is the theoretical foundation of the rest of this paper. We shall show, based on Lemma 3.2, that the error bound is closely related to the *Gaussian squared-complexity* of the feasible cone in the next section.

4. Estimation Error Bound in Terms of the Gaussian Squared Complexity

When applying Lemma 3.2 to constrained maximum-likelihood (ML) estimators in a generalized linear model

(GLM) with the canonical link function, we obtain a universal error bound in terms of the Gaussian squared complexity, valid *for all* such GLMs. Examples of a GLM with the canonical link function include Gaussian linear regression, logistic regression, and Poisson regression.

Let $\theta^\natural \in \mathbb{R}^p$ be the parameter to be estimated, or the unknown vector of regression coefficients. In a GLM with the canonical link function, the negative log-likelihood of a sample y, given $\theta^\natural$, is of the form (up to scaling and shifting by some constants)

$$\ell(y;\theta^\natural) = y\left\langle a_i, \theta^\natural\right\rangle - b(\left\langle a_i, \theta^\natural\right\rangle),$$

where $a_1, \ldots, a_n \in \mathbb{R}^p$ are given, and we assume that b is some given concave function. Let $y_1, \ldots, y_n \in \mathbb{R}^p$ be the samples. The constrained ML estimator is given by (1) with

$$f_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \theta), \tag{11}$$

and g being some continuous convex function. For simplicity, we consider the case where $c = g(\theta^\natural)$ in this section; we address the case where $c > g(\theta^\natural)$, i.e., where we only have a upper bound on the value of $g(\theta^\natural)$, in Section 5.

The following theorem is by a direct application of Lemma 3.2.

Theorem 4.1. *Consider the constrained M-estimator defined by (1) with $c = g(\hat{\theta})$. Assume that f_n satisfies the RSC condition with parameter $\mu > 0$, and*

$$\sup_{\theta \in \overline{\mathcal{F}_g(\theta^\natural)}} \left\| \frac{A}{\sqrt{n}}\theta \right\|_2 \leq \sqrt{L}\left\|\theta\right\|_2, \tag{12}$$

for some $L > 0$, where the i-th row of $A \in \mathbb{R}^{n\times p}$ is given by a_i. Then

$$\mathbb{E}\left\|\hat{\theta} - \theta^\natural\right\|_2 \leq \frac{\sqrt{L}\,\Omega\left(A\overline{\mathcal{F}_g(\theta^\natural)}\right)}{\mu\sqrt{n}},$$

where

$$\Omega\left(A\overline{\mathcal{F}_g(\theta^\natural)}\right) := \mathbb{E}\sup_{v \in \left(A\overline{\mathcal{F}_g(\theta^\natural)}\right)\cap \mathcal{S}^{n-1}} \{\langle \varepsilon, v\rangle\}$$

with $\varepsilon := (y_1 - \mathbb{E}\,y_1, \ldots, y_n - \mathbb{E}\,y_n)$.

If ε is a sequence of independent identically distributed (i.i.d.) Rademacher random variables, the quantity $\Omega\left(A\overline{\mathcal{F}_g(\theta^\natural)}\right)$ is the Rademacher complexity for the function class

$$\mathscr{F} := \left\{\langle e, \cdot\rangle : e \in A\overline{\mathcal{F}_g(\theta^\natural)} \cap \mathcal{S}^{n-1}\right\},$$

given A; if ε is a sequence of i.i.d. standard Gaussian random variables, the quantity $\Omega\left(A\overline{\mathcal{F}_g(\theta^\natural)}\right)$ corresponds to the Gaussian complexity of $\mathscr{F}$ given A (Bartlett & Mendelson, 2002).

However, two issues appear:

1. The entries in ε are in general independent but not necessarily identically distributed, so the quantity $\Omega\left(A\overline{\mathcal{F}_g(\theta^\natural)}\right)$ is not obviously related to the Rademacher and Gaussian complexities.
2. There do not exist known results for the value of $\Omega\left(A\overline{\mathcal{F}_g(\theta^\natural)}\right)$ for commonly used regularization functions, such as the ℓ_1-norm and the nuclear norm discussed in Section 1.

The following corollary provides a simple error bound in terms of the well-studied *Gaussian squared-complexity* of the feasible cone $\overline{\mathcal{F}_g(\theta^\natural)}$.

Definition 4.1 (Gaussian Squared-Complexity (Chandrasekaran & Jordan, 2013))**.** The *Gaussian squared-complexity* of a set $\mathcal{C} \subseteq \mathbb{R}^p$ is given by

$$\omega_t(\mathcal{C}) := \mathbb{E}\sup_{v \in \mathcal{C}\cap t\mathcal{S}^{p-1}} \left\{\langle h, v\rangle^2\right\},$$

for any $t \geq 0$, where the entries of $h \in \mathbb{R}^p$ are i.i.d. standard Gaussian random variables.

Corollary 4.2. *Under the same assumptions as in Theorem 4.1, we have*

$$\mathbb{E}\left\|\hat{\theta} - \theta^\natural\right\|_2 \leq C_{\max}^{1/2}\left(\frac{L}{\mu}\right)\sqrt{\frac{2\pi\,\omega_1\left(\overline{\mathcal{F}_g(\theta^\natural)}\right)}{n}},$$

where $C_{\max} := \max_i\left\{\mathbb{E}\,y_i^2\right\}$.

The feasible cone $\overline{\mathcal{F}_g(\theta^\natural)}$ coincides with the tangent cone of g at $\theta^\natural$ defined in (Chandrasekaran et al., 2012). Therefore, to evaluate the estimation error bound, we only need to evaluate the Gaussian squared-complextiy of the corresponding tangent cone. We note that there are already many results for a variety of commonly used regularization functions (Cai & Xu, 2013; Foygel & Mackey, 2014; Chandrasekaran et al., 2012; Plan & Vershynin, 2013; Rao et al., 2012; Vershynin, 2014)[1].

[1] In some of the cited works, the authors focus on the *Gaussian width*, denoted by ℓ_t (cf. Definition 5.1), instead of the conic Gaussian squared complexity of the feasible cone. We note that the cited works always upper-bounded the conic Gaussian width by

$$\ell_t(\overline{\mathcal{F}_g(\theta^\natural)}) \leq \sqrt{\omega_t(\overline{\mathcal{F}_g(\theta^\natural)})},$$

based on Jensen's inequality. Therefore, we can find upper bounds on the Gaussian squared complexities in the cited works.

5. Effect of a Mismatched Constraint

We call the constraint set $\mathcal{G}$ *mismatched* if $c > g(\theta^\natural)$ in (2). The case where $c = g(\theta^\natural)$ has been discussed in Section 4. If $c < g(\theta^\natural)$, the estimation error is always lower bounded by the distance between $\theta^\natural$ and $\mathcal{G}$, a constant; therefore, the estimation error will not converge to zero with increasing number of samples n. In this section, we discuss the effect of a mismatched constraint for ML regression in a GLM with the canonical link function.

The notion of RSC in Definition 3.2 is no longer useful when the constraint set is mismatched. Take ML regression in the Gaussian linear model for example, for which the corresponding f_n is given by (3). Let $A \in \mathbb{R}^{n\times p}$ be defined as in Theorem 4.1. By direct calculations, we get

$$\nabla f_n(\theta) = -\frac{1}{n}A^T(y - Ax).$$

To have the RSC condition hold true, we need

$$\left\langle \nabla f_n(\theta^\natural + e) - \nabla f_n(\theta^\natural), e\right\rangle = \frac{1}{n}\left\|Ae\right\|_2^2 \geq \mu \left\|e\right\|_2^2,$$

for some $\mu > 0$ and all $e \in \mathcal{F}_g(\theta^\natural)$. Since when the constraint is mismatched, $\overline{\mathcal{F}_g(\theta^\natural)}$ can be the whole space $\mathbb{R}^p$, the RSC condition holds if and only if A is a non-singular matrix. This cannot be true in the high-dimensional setting, where $n < p$.

Our Approach: Let $t > 0$ and denote by $\mathcal{B}$ the unit ℓ_2-ball in $\mathbb{R}^p$. We partition the feasible set $\mathcal{F}_g(\theta^\natural)$ as

$$\mathcal{F}_g(\theta^\natural) = t\mathcal{B} \cup (\mathcal{F}_g(\theta^\natural) \setminus t\mathcal{B}).$$

When t is large enough, the conic hull of $(\mathcal{F}_g(\theta^\natural) \setminus t\mathcal{B})$ will not be the whole space $\mathbb{R}^p$, so we may have restricted strong convexity on $(\mathcal{F}_g(\theta^\natural) \setminus t\mathcal{B})$ when $n < p$. If the error vector $\hat{\theta} - \theta^\natural$ lies in $(\mathcal{F}_g(\theta^\natural) \setminus t\mathcal{B})$, we can obtain an error bound, say, $\tilde{t}$, as in Section 4; otherwise, if the error vector lies in $\mathcal{B}_t$, a naïve error bound is the radius of the ball t. Finally, we can bound the estimation error above by the maximum of $\tilde{t}$ and t. Note that $\tilde{t}$ is implicitly dependent on t.

The arguments in the previous paragraph can be made precise as in Lemma 5.1, which is an analogue of Lemma 3.2 in the mismatched case.

Lemma 5.1. *Suppose that for some $t > 0$, we have*

$$\left\langle \nabla f_n(\theta^\natural + e) - \nabla f_n(\theta^\natural), e\right\rangle \geq \mu \left\|e\right\|_2^2, \qquad (13)$$

for some $\mu > 0$ and all $e \in \mathcal{F}_g(\theta^\natural) \setminus t\mathcal{B}$. Then

$$\mathbb{E}\left\|\hat{\theta} - \theta^\natural\right\|_2 \leq t + \mathbb{E}\left\|\Pi_{\overline{\mathcal{F}_g(\theta^\natural)\setminus t\mathcal{B}}}\left(-\nabla f_n(\theta^\natural)\right)\right\|_2.$$

Similarly, we can prove an analogue of Corollary 4.2 for ML regression in a GLM with the canonical link function.

Corollary 5.2. *Consider the constrained M-estimator defined by (1) with $c > g(\theta^\natural)$. Let A be defined as in Theorem 4.1 and let $t > 0$. Suppose that (13) holds true with some $\mu > 0$, and A satisfies*

$$\sup_{e\in\overline{\mathcal{F}_g(\theta^\natural)\setminus t\mathcal{B}}}\left\|\frac{A}{\sqrt{n}}e\right\|_2 \leq \sqrt{L}\left\|e\right\|_2, \qquad (14)$$

with some $L > 0$. Then we have

$$\mathbb{E}\left\|\hat{\theta} - \theta^\natural\right\|_2 \leq t + \frac{C_{\max}^{1/2}}{t}\left(\frac{L}{\mu}\right)\sqrt{\frac{2\pi\,\omega_t\left(\mathcal{F}_g(\theta^\natural)\right)}{n}},$$

where $C_{\max}$ is defined as in Corollary 4.2.

Due to space limit, we put the proofs of Lemma 5.1 and Corollary 5.2 in the supplementary material. Conceptually the proofs are similar to the proofs of Lemma 3.2 and Corollary 4.2, respectively.

We compare Corollary 5.2 with the result in (Plan et al., 2014). Ignoring the constants, our result is of the form

$$\mathbb{E}\left\|\hat{\theta} - \theta^\natural\right\|_2 \leq t + C\frac{\sqrt{\omega_t(\mathcal{F}_g(\theta^\natural))}}{t\sqrt{n}},$$

for some constant $C > 0$ and all t such that (13) and (14) hold. In (Plan et al., 2014), a universal estimator for a very general statistical model is proposed, where only $y_i \sim \mathbb{P}_{\langle a_i,\theta^\natural\rangle}$ is assumed. The error guarantee in (Plan et al., 2014) is of the form

$$\mathbb{E}\left\|\tilde{\hat{\theta}} - \gamma\theta^\natural\right\|_2 \leq t + \tilde{C}\frac{\ell_t(\mathcal{F}_g(\theta^\natural))}{t\sqrt{n}},$$

for some other constant $\tilde{C} > 0$, some scaling parameter γ and all $t > 0$, where $\tilde{\hat{\theta}} := \Pi_{\mathcal{F}_g(\theta^\natural)}\left(A^T y\right)$, and $\ell_t(\mathcal{F}_g(\theta^\natural))$ is the Gaussian width of $\mathcal{F}_g(\theta^\natural)$. We note that the scaling parameter γ is unknown as it is dependent on the true parameter $\theta^\natural$; therefore, the estimator $\tilde{\hat{\theta}}$ can only recover $\theta^\natural$ up to a scale ambiguity.

Definition 5.1 (Gaussian width (Chandrasekaran et al., 2012; Mendelson et al., 2007; Tropp, 2014))**.** Let $\mathcal{C} \subseteq \mathbb{R}^p$. The *Gaussian width* of $\mathcal{C}$ is given by

$$\ell_t(\mathcal{C}) := \mathbb{E}\sup_{v\in\mathcal{C}\cap t\mathcal{S}^{p-1}}\left\{\langle h, v\rangle\right\}.$$

By Jensen's inequality, we have $\ell_t(\mathcal{F}_g(\theta^\natural)) \leq \sqrt{\omega_t(\mathcal{F}_g(\theta^\natural))}$ (Chandrasekaran & Jordan, 2013), and hence the result in (Plan et al., 2014) implies

$$\mathbb{E}\left\|\tilde{\hat{\theta}} - \gamma\theta^\natural\right\|_2 \leq t + \tilde{C}\frac{\sqrt{\omega_t(\mathcal{F}_g(\theta^\natural))}}{t\sqrt{n}}.$$

The only major difference in the two error bounds is that our result is not for all $t > 0$. We will discuss the similarity of our result and the result in (Plan et al., 2014) further in the next section.

6. Applications and Discussions

Once the conditions (10), (12), (13), and (14) are verified, our results Theorem 4.1, Corollary 4.2, and Corollary 5.2 immediately follow. We show how we can verify the conditions for two classes of applications, and compare our results with existing ones. Finally, we discuss the relation between our result and the result in (Plan et al., 2014) in a more precise way.

The discussions in this section rely on the following theorem (Mendelson et al., 2007).

Theorem 6.1 ((Mendelson et al., 2007)). *Let $\mathcal{T} \subseteq \mathbb{R}^p$ be star-shaped. Let $A \in \mathbb{R}^{n\times p}$, $n < p$, whose rows are i.i.d. isotropic subgaussian random vectors with subgaussian norm $\alpha \geq 1$, and let $\varepsilon \in (0,1)$. Then there exist constants c and $\bar{c}$ such that for all $x \in \mathcal{T}$ satisfying*

$$\|x\|_2 \geq \gamma_n^*\left(\frac{\varepsilon}{c\alpha^2}, \mathcal{T}\right) := \inf\left\{t > 0 : t \geq \frac{c\alpha^2 \ell_t(\mathcal{T})}{\varepsilon\sqrt{n}}\right\}, \tag{15}$$

we have

$$(1-\varepsilon)\|x\|_2^2 \leq \frac{\|Ax\|_2^2}{n} \leq (1+\varepsilon)\|x\|_2^2$$

with probability at least $1 - \exp\left(-\bar{c}\varepsilon^2 n/\alpha^4\right)$.

Recall that $\ell_t(\cdot)$ is the Gaussian width defined in Definition 5.1. Examples of an isotropic subgaussian random vector include a vector of i.i.d. standard Gaussian random variables, and a vector of i.i.d. Rademacher random variables (Vershynin, 2012).

6.1. Gaussian Linear Regression with Arbitrary Constraint Function

Let $\theta^\natural \in \mathbb{R}^p$ be the unknown vector of regression coefficients, and let $a_1, \ldots, a_n \in \mathbb{R}^p$. We are interested in estimating $\theta^\natural$ given $a_1, \ldots, a_n$ and

$$y_i = \langle a_i, \theta^\natural \rangle + w_i, \quad i = 1, \ldots, n,$$

where $w_1, \ldots, w_n$ are i.i.d. standard Gaussian random variables.

We consider constrained least squares (LS) estimators of the form (1) with f_n given by (3), and g can be any continuous convex function. For simplicity, we consider the matched constraint case, where $c = g(\theta^\natural)$ in (1).

Suppose $a_1, \ldots, a_n$ are i.i.d. isotropic subgaussian random vectors of subgaussian norm α. Set $\mathcal{T} = \overline{\mathcal{F}_g(\theta^\natural)}$ in Theorem 6.1. Since then $\ell_t(\overline{\mathcal{F}_g(\theta^\natural)}) = t\ell_1(\overline{\mathcal{F}_g(\theta^\natural)})$, the condition (15) is equivalent to requiring

$$\sqrt{n} \geq \frac{c\alpha^2 \ell_1(\overline{\mathcal{F}_g(\theta^\natural)})}{\varepsilon}.$$

Once this inequality is satisfied, we can set $\mu = 1-\varepsilon$ and $L = 1+\varepsilon$ and the conditions (10) and (12) hold with probability at least $1 - \exp\left(-\bar{c}\varepsilon^2 n/\alpha^4\right)$. We note that $\Omega(A\overline{\mathcal{F}_g(\theta^\natural)}) = \ell_1(A\overline{\mathcal{F}_g(\theta^\natural)})$ in Theorem 4.1. Furthermore, by Jensen's inequality and Lemma B.2,

$$\ell_1(A\overline{\mathcal{F}_g(\theta^\natural)}) \leq \sqrt{\omega_1(A\overline{\mathcal{F}_g(\theta^\natural)})} \leq \sqrt{L\,\omega_1(\overline{\mathcal{F}_g(\theta^\natural)})}.$$

In conclusion, we obtain the following corollary of Theorem 4.1.

Corollary 6.2. *Consider the Gaussian linear regression model and the corresponding constrained LS estimator described in this sub-section. Then there exist constants c and $\bar{c}$ such that, if*

$$\sqrt{n} \geq \frac{c\alpha^2 \ell_1(\overline{\mathcal{F}_g(\theta^\natural)})}{\varepsilon}, \tag{16}$$

we have

$$\mathbb{E}\left\|\hat{\theta} - \theta^\natural\right\|_2 \leq \left(\frac{1+\varepsilon}{1-\varepsilon}\right)\sqrt{\frac{\omega_1(\overline{\mathcal{F}_g(\theta^\natural)})}{n}},$$

with probability at least $1 - \exp\left(-\bar{c}\varepsilon^2 n/\alpha^4\right)$.

This corollary coincides with the results in (Chandrasekaran et al., 2012; Oymak et al., 2013a;b) up to some constant scaling, though the problem formulations and the estimators considered are not exactly the same. We note that though our framework is more general, the results in (Oymak et al., 2013a;b) are sharper in terms of the constants, by exploiting the specific structure of the linear regression model and the loss function f_n given by (3).

6.2. ML Regression in a GLM with ℓ_1-Norm Constraint

In this sub-section, we consider a GLM with the canonical link function as described in Section 4. Let f_n be defined as in (11) and assume that f_n is twice continuously differentiable. Then it is known that

$$\nabla^2 f_n(\theta) = \frac{1}{n} A^T D(\theta) A,$$

where A is defined as in Theorem 4.1, and $D(\theta)$ is a diagonal matrix, whose (i,i)-th entry is a continuous function of the inner product $\langle a_i, \theta\rangle$. We focus on the case where the constraint set is given by the ℓ_1-norm, i.e., $\mathcal{G} := \{\theta \in \mathbb{R}^p : \|\theta\|_1 \leq c\}$. We set $c = \left\|\theta^\natural\right\|_1$ for simplicity in (1).

Suppose that the entries of A are i.i.d. Rademacher random variables. Then for any $\theta \in \mathcal{G}$,

$$|\langle a_i, \theta\rangle| \leq \|a_i\|_\infty \|\theta\|_1 = c,$$

and hence the diagonal entries of $D(\theta)$ are bounded below by a constant $\nu > 0$. Similarly, $C_{\max} := \max_i \{\mathbb{E}\, y_i^2\}$ is also a constant independent of n.

Following the arguments in the previous sub-section and Proposition 3.1, if we choose n such that (16) is satisfied, we may choose $L = 1 + \varepsilon$ and $\mu = \nu(1 - \varepsilon)$, and the conditions (10) and (12) hold with probability at least $1 - \exp\left(-\bar{c}\varepsilon^2 n/\alpha^4\right)$. Then we have the following result by Corollary 4.2.

Corollary 6.3. *Consider the Gaussian linear regression model and the corresponding ℓ_1-constrained ML estimator described in this sub-section. There exist constants c and $\bar{c}$ such that, if (16) is satisfied, we have*

$$\mathbb{E}\left\|\hat{\theta} - \theta^\natural\right\|_2 \le C_{\max}^{1/2}\left(\frac{1+\varepsilon}{1-\varepsilon}\right)\sqrt{\frac{2\pi\,\omega_1\left(\overline{\mathcal{F}_g(\theta^\natural)}\right)}{\nu^2 n}}, \tag{17}$$

with probability at least $1 - \exp\left(-\bar{c}\varepsilon^2 n/\alpha^4\right)$.

An upper bound of $\omega_1(\overline{\mathcal{F}_g(\theta^\natural)})$ when g is the ℓ_1-norm can be found in (Chandrasekaran et al., 2012), which says

$$\omega_1(\overline{\mathcal{F}_g(\theta^\natural)}) \le 2s\log\left(\frac{p}{s}\right) + \frac{5}{4}s,$$

where s denotes the number of non-zero entries in $\theta^\natural$. Ignoring the constants, our bound is essentially

$$\mathbb{E}\left\|\hat{\theta} - \theta^\natural\right\|_2 \le C\sqrt{\frac{s\log p}{n}},$$

for some $C > 0$. This is consistent with the results in (Bickel et al., 2009; Negahban et al., 2010). Though the estimator considered in (Bickel et al., 2009; Negahban et al., 2010) is the ℓ_1-regularized ML estimator, by Lagrange duality there always exists some regularization coefficient such that the regularized estimator satisfies the error bound (17).

6.3. Further Discussions on the Mismatched Case

Consider the setting in Section 6.2 again, and recall the comparison with (Plan et al., 2014) in the end of Section 5. Let us start by defining the *global Gaussian squared complexity*:

$$\omega(\mathcal{F}_g(\theta^\natural)) := \mathbb{E}\sup_{v\in\mathcal{F}_g(\theta^\natural)}\left\{\langle h, x\rangle^2\right\},$$

where $h \in \mathbb{R}^p$ is a vector of i.i.d. standard Gaussian random variables. By definition, $\omega(\mathcal{F}_g(\theta^\natural)) \ge \omega_t(\mathcal{F}_g(\theta^\natural))$ for any $t > 0$, and hence we obtain a looser upper bound for the estimator in (Plan et al., 2014):

$$\mathbb{E}\left\|\tilde{\hat{\theta}} - \gamma\theta^\natural\right\| \le t + \tilde{C}\frac{\sqrt{\omega(\mathcal{F}_g(\theta^\natural))}}{t\sqrt{n}},$$

for all $t > 0$. Note that $\omega(\mathcal{F}_g(\theta^\natural))$ is a constant and does not change with n. Optimizing over all possible values of t, we have

$$\mathbb{E}\left\|\tilde{\hat{\theta}} - \gamma\theta^\natural\right\|_2 = O(n^{-1/4}).$$

By Theorem 6.1 and replacing $\omega_t(\mathcal{F}_g(\theta^\natural))$ by the upper bound $\omega(\mathcal{F}_g(\theta^\natural))$, our result is essentially

$$\mathbb{E}\left\|\hat{\theta} - \theta^\natural\right\|_2 \le t + C\frac{\sqrt{\omega(\mathcal{F}_g(\theta^\natural))}}{t\sqrt{n}},$$

for all $t \ge r_n^*(\varepsilon/c\alpha^2, \mathcal{F}_g(\theta^\natural))$. Since $\mathcal{F}_g(\theta^\natural) \subseteq 2\mathcal{G}$ by definition, Lemma 3.4 in (Mendelson et al., 2007) says that

$$r_n^*\left(\frac{\varepsilon}{c\alpha^2}, \mathcal{F}_g(\theta^\natural)\right) \le r_n^*\left(\frac{\varepsilon}{c\alpha^2}, 2\mathcal{G}\right) = O(n^{-1/2});$$

hence $t = O(n^{-1/4})$ is allowed and we obtain the same error decaying rate:

$$\mathbb{E}\left\|\hat{\theta} - \theta^\natural\right\|_2 = O(n^{-1/4}).$$

It has been shown that for regression in the Gaussian linear model with an ℓ_1-norm constraint, the decaying rate $O(n^{-1/4})$ is optimal in the minimax sense (Raskutti et al., 2011). See also discussions in Section 4 of (Plan et al., 2014).

A. Proof of Theorem 4.1

The aim in this proof is to estimate the ℓ_2-norm of $\Pi_{\overline{\mathcal{F}_g(\theta^\natural)}}\left(-\nabla f_n(\theta^\natural)\right)$.

We start with an equivalent definition:

$$\Pi_{\overline{\mathcal{F}_g(\theta^\natural)}}\left(-\nabla f_n(\theta^\natural)\right) = \sup_{v\in\overline{\mathcal{F}_g(\theta^\natural)}\cap\mathcal{S}^{p-1}}\left\{\left\langle -\nabla f_n(\theta^\natural), v\right\rangle\right\},$$

where $\mathcal{S}^{p-1}$ denotes the unit sphere in $\mathbb{R}^p$. It is well known that in a GLM with the canonical link function, we have

$$\nabla f_n(\theta^\natural) = -\frac{1}{n}A^T(y - \mathbb{E}\, y) = -\frac{1}{n}A^T\varepsilon,$$

where $y := (y_1, \ldots, y_n)$. Then, by some direct calculations,

$$\begin{aligned}
&\left\|\Pi_{\overline{\mathcal{F}_g(\theta^\natural)}}\left(-\nabla f_n(\theta^\natural)\right)\right\|_2 \\
&= \sup_{v\in\overline{\mathcal{F}_g(\theta^\natural)}\cap\mathcal{S}^{p-1}}\left\{\left\langle\frac{A^T}{n}\varepsilon, v\right\rangle\right\} \\
&= \frac{1}{\sqrt{n}}\sup_{v\in\overline{\mathcal{F}_g(\theta^\natural)}\cap\mathcal{S}^{p-1}}\left\{\left\langle\varepsilon, \frac{A}{\sqrt{n}}v\right\rangle\right\} \\
&= \frac{1}{\sqrt{n}}\sup_{\tilde{v}\in\left(\frac{A}{\sqrt{n}}\overline{\mathcal{F}_g(\theta^\natural)}\right)\cap\left(\frac{A}{\sqrt{n}}\mathcal{S}^{p-1}\right)}\left\{\langle\varepsilon, \tilde{v}\rangle\right\}.
\end{aligned}$$

Since $\overline{\mathcal{F}_g(\theta^\natural)}$ is a cone, $\frac{A}{\sqrt{n}}\overline{\mathcal{F}_g(\theta^\natural)} = A\overline{\mathcal{F}_g(\theta^\natural)}$. By assumption (12), $\frac{A}{\sqrt{n}}(\overline{\mathcal{F}_g(\theta^\natural)} \cap \mathcal{S}^{p-1}) \subset \sqrt{L}(\overline{\mathcal{F}_g(\theta^\natural)} \cap \mathcal{B}^n)$, where $\mathcal{B}^n$ denotes the unit ball in $\mathbb{R}^n$. Hence,

$$\left\|\Pi_{\overline{\mathcal{F}_g(\theta^\natural)}}\left(-\nabla f_n(\theta^\natural)\right)\right\|_2 \leq \sqrt{\frac{L}{n}} \sup_{\tilde{v}\in A\overline{\mathcal{F}_g(\theta^\natural)}\cap\mathcal{B}^n} \{\langle \varepsilon, \tilde{v}\rangle\}.$$

As the function $\langle \varepsilon, \cdot\rangle$ is linear, whose extreme value must occur on the boundary of the constraint set $A\overline{\mathcal{F}_g(\theta^\natural)} \cap \mathcal{B}^n$, we have

$$\left\|\Pi_{\overline{\mathcal{F}_g(\theta^\natural)}}\left(-\nabla f_n(\theta^\natural)\right)\right\|_2 \leq \sqrt{\frac{L}{n}} \sup_{\tilde{v}\in A\overline{\mathcal{F}_g(\theta^\natural)}\cap\mathcal{S}^{n-1}} \{\langle \varepsilon, \tilde{v}\rangle\}.$$

The theorem follows by taking expectations on both sides.

B. Proof of Corollary 4.2

Our proof relies on the symmetrization trick. Instead of using the well-known symmetrization with a Rademacher process (see, e.g., (Koltchinskii, 2011)), we would like to do symmetrization with a Gaussian process. In this way, we are able to express the error bound in terms of the Gaussian squared-complexity.

Theorem B.1 (Gaussian Symmetrization (Pollard, 1989))**.** *Let $y_1, \ldots, y_n$ be independent random variables, and let $\mathscr{F}$ be a class of measurable functions. Then*

$$\mathbb{E}\sup_{f\in\mathscr{F}}\left\{\left|\sum_{i=1}^n (f(y_i) - \mathbb{E}\, f(y_i))\right|^2\right\} \leq 2\pi\,\mathbb{E}\sup_{f\in\mathscr{F}}\left\{\left|\sum_{i=1}^n h_i f(y_i)\right|^2\right\},$$

where $h_1, \ldots, h_n$ are i.i.d. standard Gaussian random variables.

Remark. In (Pollard, 1989), the symmetrization inequality is only proved for the case when $y_1, \ldots, y_n$ are i.i.d. random variables. Though in Theorem 4.1 the random variables $y_1, \ldots, y_n$ may not be identically distributed, the same arguments in Section 4 of (Pollard, 1989) still apply.

By Jensen's inequality, we have

$$\Omega\left(A\overline{\mathcal{F}_g(\theta^\natural)}\right) \leq \sqrt{\mathbb{E}\sup_{v\in(A\mathcal{F}_g(\theta^\natural))\cap\mathcal{S}^{n-1}}\left\{\langle \varepsilon, v\rangle^2\right\}}.$$

Applying Theorem B.1, we obtain

$$\Omega\left(A\overline{\mathcal{F}_g(\theta^\natural)}\right) \leq \sqrt{\mathbb{E}\sup_{v\in(A\mathcal{F}_g(\theta^\natural))\cap\mathcal{S}^{n-1}}\left\{\langle h\cdot y, v\rangle^2\right\}},$$

where $h\cdot y := (h_1y_1, \ldots, h_ny_n)$. Note that $h\cdot y$ is a vector of independent symmetric Gaussian random variables, whose covariance matrix $\Sigma \in \mathbb{R}^{n\times n}$ is given by $\Sigma_{i,j} = \delta_{i,j}\mathbb{E}\, y_i^2$, where δ denotes the Kronecker delta. Therefore, the distribution of $h\cdot y$ is the same as the Gaussian vector $\Sigma^{1/2}h$, where $h := (h_1, \ldots, h_n)$, and

$$\begin{aligned}\Omega\left(A\overline{\mathcal{F}_g(\theta^\natural)}\right) &\leq \sqrt{\mathbb{E}\sup_{v\in(A\mathcal{F}_g(\theta^\natural))\cap\mathcal{S}^{n-1}}\left\{\langle \Sigma^{1/2}h, v\rangle^2\right\}} \\ &= \sqrt{\mathbb{E}\sup_{v\in(A\mathcal{F}_g(\theta^\natural))\cap\mathcal{S}^{n-1}}\left\{\langle h, \Sigma^{1/2}v\rangle^2\right\}} \\ &\leq C_{\max}^{1/2}\,\omega_1\left(A\overline{\mathcal{F}_g(\theta^\natural)}\right),\end{aligned}$$

where $C_{\max} := \max_i\{\mathbb{E}\, y_i^2\}$.

The remaining work is to bound $\omega_1\left(A\overline{\mathcal{F}_g(\theta^\natural)}\right)$ from above by $L\,\omega_1\left(\overline{\mathcal{F}_g(\theta^\natural)}\right)$.

Lemma B.2. *Under the assumption (12), we have $\omega_t(A\overline{\mathcal{F}_g(\theta^\natural)}) \leq L\,\omega_t(\overline{\mathcal{F}_g(\theta^\natural)})$ for all $t > 0$.*

Proof. Define

$$\tilde{A} := \begin{bmatrix} A \\ 0^{(p-n)\times p}\end{bmatrix} \in \mathbb{R}^{p\times p},$$

where $0^{(p-n)\times p}$ is the all-zero matrix in $\mathbb{R}^{(p-n)\times p}$. Obviously, $\omega(A\overline{\mathcal{F}_g(\theta^\natural)}) = \omega(\tilde{A}\overline{\mathcal{F}_g(\theta^\natural)})$.

By definition, the Gaussian squared-complexity is rotation invariant, i.e., $\omega_t(R\overline{\mathcal{F}_g(\theta^\natural)}) = \omega_t(\overline{\mathcal{F}_g(\theta^\natural)})$ for any unitary matrix R. Let $\tilde{M} = USV'$ be the singular value decomposition of $\tilde{M}$, where U and V are unitary matrices and S is a diagonal matrix. We have

$$\begin{aligned}\omega_t(\tilde{A}\overline{\mathcal{F}_g(\theta^\natural)}) &= \omega_t(SV'\overline{\mathcal{F}_g(\theta^\natural)}) \\ &\leq L\,\omega_t(V'\overline{\mathcal{F}_g(\theta^\natural)}) \\ &= L\,\omega_t(\overline{\mathcal{F}_g(\theta^\natural)}),\end{aligned}$$

where the equalities are because of the rotation invariance of the Gaussian squared-complexity, and the inequality is because $S \leq \sqrt{L}I$, by the assumption (12). This proves the lemma. □

The corollary follows by applying Lemma B.2.

C. Proof of Lemma 5.1

If $e \in \mathcal{F}_g(\theta^\natural) \setminus t\mathcal{B}$, following the proof of Theorem 4.1, we obtain

$$\|e\|_2 \leq \frac{1}{\mu}\left\|\Pi_{\overline{\mathcal{F}_g(\theta^\natural)\setminus t\mathcal{B}}}\left(-\nabla f_n(\theta^\natural)\right)\right\|_2,$$

where $\overline{\mathcal{F}_g(\theta^\natural) \setminus t\mathcal{B}}$ denotes the conic hull of $\mathcal{F}_g(\theta^\natural) \setminus t\mathcal{B}$. If $e \in t\mathcal{B}$, we have the naïve bound: $\|e\|_2 \leq t$. Therefore,

$$\begin{aligned}\|e\|_2 &\leq \max\left\{t, \frac{1}{\mu}\left\|\Pi_{\overline{\mathcal{F}_g(\theta^\natural)\setminus t\mathcal{B}}}\left(-\nabla f_n(\theta^\natural)\right)\right\|_2\right\} \\ &\leq t + \frac{1}{\mu}\left\|\Pi_{\overline{\mathcal{F}_g(\theta^\natural)\setminus t\mathcal{B}}}\left(-\nabla f_n(\theta^\natural)\right)\right\|_2.\end{aligned}$$

The lemma follows by taking expectations on both sides.

D. Proof of Corollary 5.2

If $e \in \mathcal{F}_g(\theta^\natural) \setminus t\mathcal{B}$, following the proof of Corollary 4.2, we obtain

$$\mathbb{E}\left\|\hat{\theta} - \theta^\natural\right\|_2 \leq t + C_{\max}^{1/2}\sqrt{\frac{2\pi L\,\omega_1\left(\overline{\mathcal{F}_g(\theta^\natural) \setminus t\mathcal{B}}\right)}{\mu^2 n}}.$$

The corollary follows because by the definition of Gaussian squared complexity,

$$\omega_1\left(\overline{\mathcal{F}_g(\theta^\natural) \setminus t\mathcal{B}}\right) = \frac{\omega_t\left(\mathcal{F}_g(\theta^\natural)\right)}{t^2}.$$

References

Bach, Francis. Learning with submodular functions: A convex optimization perspective. *Found. Trends Mach. Learn.*, 6(2–3):145–373, 2013.

Bartlett, Peter L. and Mendelson, Shahar. Rademacher and Gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3, 2002.

Bickel, Peter, Ritov, Ya'acov, and Tsybakov, Alexandre B. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Stat.*, 37(4):1705–1732, 2009.

Bühlmann, Peter and van de Geer, Sara. *Statistics for High-Dimensional Data*. Springer, Berlin, 2011.

Cai, Jian-Feng and Xu, Weiyu. Guarantees of total variation minimization for signal recovery. 2013. arXiv:1301.6791v6.

Chandrasekaran, Venkat and Jordan, Michael I. Computational and statistical tradeoffs via convex relaxation. *Proc. Natl. Acad. Sci.*, 110(13):E1181–E1190, 2013.

Chandrasekaran, Venkat, Recht, Benjamin, Parrilo, Pablo A., and Willsky, Alan S. The convex geometry of linear inverse problems. *Found. Comput. Math.*, 12: 805–849, 2012.

Chatterjee, Sourav. A new perspective on least squares under convex constraint. *Ann. Stat.*, 42(6):2340–2381, 2014.

Davenport, Mark A., Plan, Yaniv, van den Berg, Ewout, and Wootters, Mary. 1-bit matrix completion. *Inf. Inference*, 3:189–223, 2014.

El Halabi, Marwa and Cevher, Volkan. A totally unimodular view of structured sparsity. In *18th Int. Conf. Artificial Intelligence and Statistics*, 2015.

Flammia, Steven T., Gross, David, Liu, Yi-Kai, and Eisert, Jens. Quantum tomography via compressed sensing: Error bounds, sample complexity and efficient estimators. *New J. Phys.*, 14, 2012.

Foygel, Rina and Mackey, Lester. Corrupted sensing: Novel guarantees for separating structured signals. 2014. arXiv:1305.2524v2 [cs.IT].

Gunasekar, Suriya, Ravikumar, Pradeep, and Ghosh, Joydeep. Exponential family matrix completion under structural constraints. In *Proc. 31st Int. Conf. Machine Learning*, 2014.

Honorio, Jean and Jaakkola, Tommi. A unified framework for consistency of regularized loss minimizers. In *Proc. 31st Int. Conf. Machine Learning*, 2014.

Kakade, Sham M., Shamir, Ohad, Sridharan, Karthik, and Tewari, Ambuj. Learning exponential families in high-dimensions: Strong convexity and sparsity. In *Proc. 13th Int. Conf. Artificial Intelligence and Statistics*, 2010.

Koltchinskii, Vladimir. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer-Verl., Berlin, 2011.

Mendelson, Shahar, Pajor, Alain, and Tomczak-Jaegermann, Nicole. Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geom. Funct. Anal.*, 17:1248–1282, 2007.

Negahban, Sahand, Ravikumar, Pradeep, Wainwright, Martin J., and Yu, Bin. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. 2010. arXiv:1010.2731v1 [math.ST].

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

Oymak, Samet, Thrampoulidis, Christos, and Hassibi, Babak. The squared-error of generalized LASSO: A precise analysis. 2013a. arXiv:1311.0830v2 [cs.IT].

Oymak, Samet, Thrampoulidis, Christos, and Hassibi, Babak. Simple bounds for noisy linear inverse problems with exact side information. 2013b. arXiv:1312.0641v2 [cs.IT].

Plan, Yaniv and Vershynin, Roman. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Trans. Inf. Theory*, 59(1): 482–494, January 2013.

Plan, Yaniv, Vershynin, Roman, and Yudovina, Elena. High-dimensional estimation with geometric constraints. 2014. arXiv:1404.3749v1 [math.PR].

Pollard, David. Asymptotics via empirical processes. *Stat. Sci.*, 4(4):341–366, 1989.

Rao, Nikhil, Recht, Benjamin, and Nowak, Robert. Universal measurement bounds for structured sparse signal recovery. In *Adv. Neural Information Processing Systems*, 2012.

Raskutti, Garvesh, Wainwright, Martin J., and Yu, Bin. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inf. Theory*, 57 (10):6976–6994, October 2011.

Tibshirani, Robert. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc., Ser. B*, 58(1):267–288, 1996.

Tropp, Joel A. Convex recovery of a structured signal from independent random linear measurements. 2014. arXiv:1405.1102v1 [cs.IT].

Vershynin, Roman. Introduction to the non-asymptotic analysis of random matrices. In Eldar, Y. and Kutyniok, G. (eds.), *Compressed Sensing: Theory and Applications*, chapter 5, pp. 210–268. Cambridge Univ. Press, 2012.

Vershynin, Roman. Estimation in high dimensions: A geometric perspective. 2014. arXiv:1405.5103v1 [math.ST].