A Geometric View on Constrained M-Estimators

Yen-Huan Li, Ya-Ping Hsieh, Nissim Zerbib and Volkan Cevher

Laboratory for Information and Inference Systems
École Polytechnique Fédérale de Lausanne

Abstract

We study the estimation error of constrained M-estimators, and derive explicit upper bounds on the expected estimation error determined by the Gaussian width of the constraint set. Both of 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) are considered. For both cases, we derive novel universal estimation error bounds for regression in a generalized linear model with the canonical link function. Our error bound for the mismatched constraint case is minimax optimal in terms of its dependence on the sample size, for Gaussian linear regression by the Lasso.

1 Introduction

Consider a general statistical estimation problem. Let $(y_1, \ldots, y_n)$ be a sample following a probability distribution $\mathbb{P}_{\theta^\natural}$ in a given class $\mathcal{P} := \{\mathbb{P}_\theta : \theta \in \mathbb{R}^p\}$. We are interested in estimating the parameter $\theta^\natural$, given $(y_1, \ldots, y_n)$ and $\mathcal{P}$, under the high-dimensional setting where $n < p$.

If $\theta^\natural$ is known to satisfy $g(\theta^\natural) \leq c$ for some continuous convex function g and positive constant c, we can consider a constrained M-estimator of the form

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

We assume that f_n is a continuously differentiable convex function, and the constraint set $\mathcal{G}$ is non-empty. For example, the Lasso [32] corresponds to

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

for some $a_1, \ldots, a_n \in \mathbb{R}^p$ and positive constant c . A matrix $\Theta \in \mathbb{R}^{d \times d}$ can be vectorized as a corresponding vector $\theta \in \mathbb{R}^p$, $d^2 = p$. In the low-rank matrix recovery problem [7, 13], a popular estimator corresponds to

$$f_n(\Theta) := \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \operatorname{Tr}\left(A_i^T \Theta \right) \right)^2, \quad \mathcal{G} := \left\{ \|\Theta\|_* \leq c \right\}, \tag{3}$$

for some $A_1, \ldots, A_n \in \mathbb{R}^{d \times d}$ and positive constant c, where $\|\cdot\|_*$ denotes the nuclear norm. In general, f_n can be the normalized negative log-likelihood function, or any properly defined function, and g depends on the *a priori* information on the structure of the parameter $\theta^\natural$ [1, 8, 9].

One can also consider a penalized M-estimator, given by

$$\hat{\theta}_{\text{penalized}} \in \arg\min_{\theta \in \mathbb{R}^p} \left\{ f_n(\theta) + \rho_n g(\theta) \right\}, \tag{4}$$

for some positive constant ρ_n. The penalized M-estimator can be computed by fast proximal methods, provided that the proximal mapping of g is easy to compute [3, 22]. This condition, however, is not always satisfied. For example, if g is the nuclear norm, computing the corresponding proximal mapping requires a full singular value decomposition (SVD) in the first few iterations, and hence is not scalable with the parameter dimension. In contrast, if we consider a constrained M-estimator and compute it by the Frank-Wolfe algorithm, each iteration of the algorithm requires a linear minimization oracle (LMO), which can be approximated efficiently by Lanczos' algorithm [15]. The paper [39] also shows that when g is a structured sparsity regularizer, the LMO can be much easier to compute than the proximal mapping.

If we consider a constrained M-estimator, setting the value of the constant c in (1) becomes a practical issue. For the case $c < g(\theta^\natural)$, the estimation error is obviously bounded below by the distance between $\theta^\natural$ and the constraint set $\mathcal{G}$, and hence estimation consistency is impossible. Ideally we would like to set $c = g(\theta^\natural)$, while in practice $g(\theta^\natural)$ is seldom known. The last case is when we have some estimate on $g(\theta^\natural)$, and choose c such that $c > g(\theta^\natural)$. Some natural questions arise: Is estimation consistency possible? How fast will the estimation error decay with the sample size n? Does setting $c > g(\theta^\natural)$ result in larger estimation error than setting $c = g(\theta^\natural)$? We review related works in Section 2, which shows that answers existed only for specific cases even when $c = g(\theta^\natural)$.

In this paper, we provide a unified analysis for constrained M-estimators. Specifically,

- We propose an elementary framework for analyzing any M-estimator applied to any statistical model in Section 3.
- We obtain universal error bounds in terms of the Gaussian width, valid *for all* canonical GLMs. We consider the matched constraint case ($c = g(\theta^\natural)$) in Section 4, and the mismatched constraint case ($c > g(\theta^\natural)$) in Section 5.
- To illustrate the universal error bounds, we specialize the universal error bound to Gaussian linear regression with arbitrary convex constraint, and regression in canonical GLMs with the ℓ_1-constraint in Section 6, and obtain explicit results.
- Our error bound for the Lasso applied to the Gaussian linear model is optimal in the minimax sense (cf. Section 7).

Existing results for penalized M-estimators [2, 4, 5, 14, 16, 21, 35], which are for deterministic ρ_n's, cannot directly recover our results, and vice versa. Indeed, by Lagrange duality, there exists some $\rho_n > 0$ such that the constrained M-estimator in (1) is equivalent to the penalized M-estimator in (4). This correspondence, however, holds *only for given realization of the sample* $(y_1, \ldots, y_n)$, and hence ρ_n is a random variable depending on the sample. Conversely, for any penalized M-estimator $\hat{\theta}_{\text{penalized}}$ for some $\rho_n > 0$, there exists a constant $c = g(\hat{\theta}_{\text{penalized}})$ such that the corresponding constrained M-estimator (4) is equivalent to $\hat{\theta}_{\text{penalized}}$. Note that $c = g(\hat{\theta}_{\text{penalized}})$ is again a random variable and dependent on the sample. We are not aware of any existing work on characterizing the correspondence between the two formulations.

2 Related Works

In [23, 24], the authors derived sharp estimation error bounds for regression in the linear model by constrained least squares (LS) estimators. The analysis in [38] provides a minimax estimation error bound for the same setting . There are some related works on learning a function in a function class [17, 18]. When the function class is linearly parametrized by vectors in $\mathbb{R}^p$, and the function corresponding to $\theta^\natural$ is in the function class, the L_2-estimation error in the function class may be translated into the ℓ_2-estimation error with respect to $\theta^\natural$. A common limitation of [17, 18, 24, 23, 38] is that the results are not extendable to general non-linear statistical models.

Another research direction considers constrained estimation in possibly non-linear statistical models [25, 26, 27]. A constrained M-estimator for logistic regression was proposed and analyzed in [25]. In [27], the authors proposed and analyzed a universal projection-based estimator for regression in generalized linear models (GLMs). In [26], the authors analyzed the performance of the constrained LS estimator in GLMs. A common limitation of [25, 26, 27] is that the results are valid only for the specific proposed estimators, and they do not even apply to the constrained maximum-likelihood (ML) estimator, which is the most popular approach in practice. Moreover, the proposed estimators in [25, 26, 27] can only recover the true parameter up to a scale ambiguity.

We say that the constraint is *matched* if $\theta^\natural$ lies on the boundary of $\mathcal{G}$ in (1) (or $c = g(\theta^\natural)$), and *mismatched* if $\theta^\natural$ lies strictly in $\mathcal{G}$ (or $c < g(\theta^\natural)$). The analyses in [23, 24] require the constraint to be matched, while in practice the exact value of $g(\theta^\natural)$ is seldom known. The constraint in [17] is always matched due to the special structure of quantum density operators. The error bounds in [25, 38] can be overly pessimistic, because they hold for all $\theta^\natural \in \mathcal{G}$. The results in [18, 26, 27] do not require a matched constraint and depend on $\theta^\natural$; our result is of this kind. Recall that, however, [18] is limited to specific statistical models, and [26, 27] are limited to specific M-estimators.

3 A Geometric Framework

3.1 Basic Idea

To illustrate the basic idea of our framework, let us start with a simple setting, where f_n is strongly convex with parameter $\mu > 0$, i.e.,

$$\langle \nabla f_n(y) - \nabla f_n(x), y - x \rangle \geq \mu \|y - x\|_2^2,$$

for any $x, y \in \operatorname{dom} f$. Note that then $\hat{\theta}$ is uniquely defined.

Define $\iota_g : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ as the indicator function of the constraint set $\mathcal{G}$; that is, $\iota_{\mathcal{G}}(\theta) = 0$ if $\theta \in \mathcal{G}$, and $\iota_{\mathcal{G}}(\theta) = +\infty$ otherwise. 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, or the monotonicity of the subdifferential mapping, 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 $\hat{z} \in \partial \iota_g(\hat{\theta})$. By the optimality condition of $\hat{\theta}$, there exists some $\hat{z} \in \partial \iota_{\mathcal{G}}(\hat{\theta})$ such that

$$0 = \nabla f_n(\hat{\theta}) + \hat{z}, \tag{7}$$

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 can choose $z^\natural = 0$ and obtain

$$\left\langle -\nabla f_n(\theta^\natural), \hat{\theta} - \theta^\natural \right\rangle \geq \mu \left\| \hat{\theta} - \theta^\natural \right\|_2^2. \tag{8}$$

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

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

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

$$\mathbb{E} \left\| \hat{\theta} - \theta^\natural \right\|_2 \leq \frac{1}{\mu} \mathbb{E} \left\| \nabla f_n(\theta^\natural) \right\|_2. \tag{10}$$

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

The simple error bound (10) is not desirable for two reasons:

1. In the high-dimensional setting where $n < p$, f_n cannot be strongly convex even for the basic LS estimator.
2. It 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. Similar (but not exactly the same) ideas had appeared in [8, 21], and can be traced back to [4, 34].

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

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

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

$$\left\langle \nabla f_n(\theta^\natural + e) - \nabla f_n(\theta^\natural), e \right\rangle \geq \mu \left\| e \right\|_2^2, \tag{11}$$

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

If f_n is twice continuously differentiable, we have a sufficient condition.

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

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 sub-section.

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 or $\{0\}$). This implies, by (8),

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

The left-hand side, however, is not tractable due to its dependence on $\hat{\theta}$. As $\hat{\theta} - \theta^\natural \in \overline{\mathcal{F}_g(\theta^\natural)}$ by definition, we consider a looser bound:

$$\left\| \Pi_{\overline{\mathcal{F}_g(\theta^\natural)}} (-\nabla f_n(\theta^\natural)) \right\|_2 \geq \mu \left\| \hat{\theta} - \theta^\natural \right\|_2, \tag{12}$$

where $\Pi_{\overline{\mathcal{F}_g(\theta^\natural)}}(\cdot)$ denotes projection 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.

4 Estimation Error Bound in Terms of the Gaussian Width

We apply Lemma 3.2 to constrained ML estimators in a GLM with the canonical link function. Examples of a canonical GLM include the Gaussian linear, logistic, gamma, and Poisson regression models.

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

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

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

$$f_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} L(y_i, \theta), \tag{13}$$

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 $c > g(\theta^\natural)$ in Section 5.

We specialize Lemma 3.2 to the canonical GLM and obtain the following theorem.

Definition 4.1 (Gaussian width [8, 19, 33])**.** Let $\mathcal{C} \subseteq \mathbb{R}^p$. The *Gaussian width* of $\mathcal{C}$ is given by

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

where $h := (h_1, \ldots, h_p)$ is a vector of i.i.d. standard Gaussian random variables, and $\mathcal{S}^{p-1}$ denotes the unit ℓ_2-sphere in $\mathbb{R}^p$.

Theorem 4.1. *Consider the canonical GLM and the corresponding ML estimator described above for $c = g(\theta^\natural)$. Assume that the entries of $a_1, \ldots, a_n$ are either all i.i.d. standard Gaussian or all i.i.d. Rademacher random variables (random variables taking values in $\{+1, -1\}$ with equal probability), and f_n satisfies the RSC condition for $\mu > 0$ with probability at least $1/2$. Then*

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

where $\sigma_{\max} := \max_i \sqrt{\operatorname{var} y_i}$.

Remark. Note that the expectation is with respect to A and ε, conditioned on the event that the RSC condition holds.

The feasible cone $\overline{\mathcal{F}_g(\theta^\natural)}$ coincides with the tangent cone of g at $\theta^\natural$ defined in [8]. Therefore, to evaluate the estimation error bound, we only need to evaluate the Gaussian width of the corresponding tangent cone. We note that there are already many results for a variety of commonly used regularization functions, such as the ℓ_1-norm, nuclear norm, total variation semi-norm, and general atomic norms [6, 8, 11, 25, 30, 38]. Therefore, for most of the applications, we only need to *plug in* an existing bound on the Gaussian width.

Finally, we would like to emphasize that the Gaussian width in Theorem 4.1 comes from bounding the random process induced by the random gradient $\nabla f_n(\theta^\natural)$ (cf. the proof of Theorem 4.1), instead of being a consequence of applying Gordon's Lemma. That is, our result is essentially different from those in [8, 23, 24].

5 Effect of a Mismatched Constraint

In this section, we discuss the effect of a mismatched constraint for ML regression in a canonical GLM. Recall that the constraint set $\mathcal{G}$ is called *mismatched* if $c > g(\theta^\natural)$ in (1).

The notion of the RSC in Definition 3.2 is no longer meaningful when the constraint set is mismatched. Take ML regression in the Gaussian linear model for example, for which the corresponding f_n is given in (2). Let $A \in \mathbb{R}^{n\times p}$ be defined as in Theorem 4.1. The RSC condition requires

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

for some $\mu > 0$ and all $e \in \overline{\mathcal{F}_g(\theta^\natural)}$, where we say $e \in \overline{\mathcal{F}_g(\theta^\natural)}$ instead of $e \in \mathcal{F}_g(\theta^\natural)$ because A is a linear operator. Since when the constraint is mismatched, $\overline{\mathcal{F}_g(\theta^\natural)}$ is the whole space $\mathbb{R}^p$, the RSC condition requires A to be a non-singular matrix. This cannot be true in the high-dimensional setting, where $A \in \mathbb{R}^{n\times p}$ and $n < p$.

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

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

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 it is possible to 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, i.e., t. Finally, we can bound the estimation error from 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 holds for arbitrary constrained M-estimators of the form (1) and statistical models.

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

$$\left\langle \nabla f_n(\theta^\natural + e) - \nabla f_n(\theta^\natural), e \right\rangle \geq \mu \left\| e \right\|_2^2, \tag{14}$$

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

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

We can also prove an analogue of Theorem 4.1 for constrained ML regression in a canonical GLM.

Corollary 5.2. *Consider the canonical GLM and the corresponding ML estimator described in Section 4, for $c > g(\theta^\natural)$. Let A be defined as in Theorem 4.1*

and let $t > 0$. Suppose that (14) holds true with for some $\mu > 0$ with probability at least $1/2$. Then we have

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

where $\sigma_{\max}$ is defined as in Theorem 4.1.

The proofs of Lemma 5.1 and Corollary 5.2 are similar to the proofs of Lemma 3.2 and Theorem 4.1, respectively.

6 Applications

Once the conditions (11) and (14) are verified, our results Theorem 4.1 and Corollary 5.2 immediately follow. We explicitly verify the conditions for two applications and obtain the corresponding estimation error bounds.

The first application is regression by the constrained LS estimator in a Gaussian linear model. Let $\theta^\natural \in \mathbb{R}^p$ and $a_1, \ldots, a_n$ be vectors in $\mathbb{R}^p$. The sample is given by

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

for some $\sigma > 0$, where $w_1, \ldots, w_n$ are i.i.d. standard Gaussian random variables. We consider the constrained LS estimator, for which f_n is given by (2), and $\mathcal{G} := \{\theta : g(\theta) \leq c\}$ for some $c \geq g(\theta^\natural)$, where g can be any convex continuous function.

Corollary 6.1. *Consider the Gaussian linear model and the constrained LS estimator described above. Assume that the entries of $a_1, \ldots, a_n$ are either all i.i.d. standard Gaussian or all i.i.d. Rademacher random variables. Let $\epsilon \in (0, 1)$. For any $t \geq 0$, there exist positive constants c_1 and c_2 such that if*

$$\sqrt{n} \geq \frac{c_1\alpha^2\omega_1(\overline{\mathcal{F}_g(\theta^\natural) \setminus t\mathcal{B}})}{\epsilon}, \tag{15}$$

then we have

$$\mathbb{E}\left\|\hat{\theta} - \theta^\natural\right\|_2 \leq t + 2\sqrt{2\pi}\sigma\,\frac{\omega_1(\overline{\mathcal{F}_g(\theta^\natural) \setminus t\mathcal{B}})}{(1-\epsilon)\sqrt{n}}, \tag{16}$$

with probability at least $1 - \exp(-c_2\epsilon^2 n) > 1/2$ when n is large enough.

Remark. When the constraint is matched, we can simply set $t = 0$. Recall that t cannot be zero for the mismatched constraint case when $n < p$ (cf. Section 5). This remark also applies to Corollary 6.2 below.

Remark. For the mismatched constraint case, Corollary (6.1) is minimax optimal for the Lasso in the Gaussian linear model. We address this in Section 7.

Corollary 6.1 is consistent with [24]. The result in [24] is sharper, while Corollary 6.1 is more general as it also covers the mismatched constraint case.

The second application is ℓ_1-constrained ML regression in a canonical GLM.

Corollary 6.2. *Consider the canonical GLM and the constrained ML estimator described in Section 4, for $g(\theta) := \|\theta\|_1$ and $c \geq \left\|\theta^\natural\right\|$. Assume that f_n in (13) is twice continuously differentiable, and the entries of $a_1, \ldots, a_n$ are i.i.d. Rademacher random variables. Let $\epsilon \in (0,1)$. For any $t \geq 0$, there exist positive constants c_1, and c_2 such that if (15) is satisfied, then we have*

$$\mathbb{E}\left\|\hat{\theta} - \theta^\natural\right\|_2 \leq t + 2\sqrt{2\pi}\,\sigma_{\max}\frac{\omega_1(\overline{\mathcal{F}_g(\theta^\natural) \setminus t\mathcal{B}})}{(1-\epsilon)\sqrt{n}}, \tag{17}$$

with probability at least $1 - \exp(c_2\epsilon^2 n) > 1/2$ when n is large enough, where $\sigma_{\max} := \max_i \sqrt{\operatorname{var} y_i}$ is bounded above by a constant independent of n.

To the best of our knowledge, there are not existing results for ℓ_1-constrained ML regression in GLMs. Here we compare Corollary 6.2 with [20], which provides an error bound for ℓ_1-penalized ML estimators in GLMs . Recall that, however, the correspondence between the constrained and penalized estimators is currently unclear. When the constraint is matched and $\theta^\natural$ is s-sparse, Corollary 6.2 states that when $n = \Omega(s\log(p/s))$,

$$\mathbb{E}\left\|\hat{\theta} - \theta^\natural\right\|_2 = O\left(\sqrt{\frac{s}{n}\log\left(\frac{p}{s}\right)}\right)$$

by Proposition 3.10 in [8], which essentially coincides with Corollary 5 in [20][1]. We note that [20] only provides an error bound for the ℓ_1-penalization case.

7 Sharpness of Our Error Bound

It has been shown that in a Gaussian linear model with $\mathcal{G}$ being an ℓ_1-ball, *any* estimator $\hat{\theta}_{\text{arbitrary}}$ must satisfy, with probability larger than $1/2$,

$$\max_{\theta^\natural \in \mathcal{G}}\left\|\hat{\theta}_{\text{arbitrary}} - \theta^\natural\right\|_2 = \Omega(n^{-1/4}),$$

under some technical conditions [31]. Now we show our error bound for the Lasso in Corollary 6.1 actually achieves the error decaying rate $O(n^{-1/4})$ in the mismatched constraint case, and hence cannot be essentially improved.

By the definition of the Gaussian width, we have, for any $t > 0$,

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

[1] We cite [20] instead of the published version [21], because the estimation error bound only appears in [20].

and hence the estimation error bound in Corollary 16 can be written as

$$\mathbb{E}\left\|\hat{\theta}-\theta^{\natural}\right\|_2 \leq t+\frac{C}{t}\frac{\omega_t(\mathcal{F}_g(\theta^{\natural}))}{\sqrt{n}}, \tag{18}$$

for some $C>0$, when n is large enough such that (15) is satisfied.

Define the *global Gaussian width*:

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

where $h\in\mathbb{R}^p$ is a vector of i.i.d. standard Gaussian random variables. By definition, $\omega_t(\mathcal{F}_g(\theta^{\natural}))$ is bounded above by $\omega(\mathcal{F}_g(\theta^{\natural}))$, independent of n. Replacing $\omega_t(\mathcal{F}_g(\theta^{\natural}))$ by $\omega(\mathcal{F}_g(\theta^{\natural}))$ in (18), we have a looser error upper bound:

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

Minimizing this bound over all $t>0$, we obtain the $O(n^{-1/4})$ error decaying rate. Similar discussion can be found in [27].

8 Discussion

Note that by the elementary argument in Section 3, we arrive at an estimation error bound (12) that holds *surely*. It is possible to derive a concentration-type error guarantee based on this sure error bound, which we are working on.

Our framework is not restricted to constraint sets of the form (1); it applies to any non-empty closed convex set $\mathcal{G}$, as we only require $\iota_{\mathcal{G}}(\cdot)$ to be proper closed convex in the proof. This observation is crucial to applying our framework to analyze constrained estimators for quantum tomography [10, 12] and photon-limited imaging systems [29], which we are studying.

In this paper, we consider a random matrix A, and discuss the expected estimation error with respect to both A and the sample $(y_1,\ldots,y_n)$. The extension to the the case where A is deterministic is technically non-trivial, and we have not obtained a satisfactory result. We address this in the remark following the proof of Theorem 4.1 in the appendix.

Acknowledgements

This work was supported in part by the European Commission under Grant MIRG-268398, ERC Future Proof, SNF 200021-132548, SNF 200021-146750 and SNF CRSII2-147633.

A Proof of Proposition 3.1

We have

$$\left\langle \nabla f_n(\theta^\natural + e) - \nabla f_n(\theta^\natural), e \right\rangle = \int_0^1 \left\langle e, \nabla^2 f_n(\theta^\natural + \lambda e) e \right\rangle \, d\lambda.$$

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

B Proof of Theorem 4.1

The main goal of the proof is to evaluate $\mathbb{E} \left\| \Pi_{\overline{\mathcal{F}_g(\theta^\natural)}} \left(-\nabla f_n(\theta^\natural) \right) \right\|_2$. Here the expectation is with respect to both A and the sample $(y_i)_{i=1,\ldots,n}$.

We start with an equivalent formulation:

$$\mathbb{E} \left\| \Pi_{\overline{\mathcal{F}_g(\theta^\natural)}} \left(-\nabla f_n(\theta^\natural) \right) \right\|_2 = \mathbb{E} \sup_{v \in \overline{\mathcal{F}_g(\theta^\natural)} \cap \mathcal{S}^{p-1}} \left\{ \left\langle -\nabla f_n(\theta^\natural), v \right\rangle \right\}, \tag{19}$$

where $\mathcal{S}^{p-1}$ denotes the unit ℓ_2-sphere in $\mathbb{R}^p$. It is well known that in a canonical GLM, we have

$$\nabla f_n(\theta^\natural) = -\frac{1}{n} A^T \varepsilon, \tag{20}$$

where $\varepsilon := (y_i - \mathbb{E}\, y_i)_{i=1,\ldots,n}$, and hence

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

To proceed, we need the following symmetrization inequality. The symmetrization inequality is different from the well-known symmetrization inequality by a Rademacher process, so we show it here for completeness.

Lemma B.1 ([36]). *Let $\xi_1, \ldots, \xi_n$ be independent real-valued random variables, and let $\mathcal{F}$ be a class of real functions. We have*

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left\{ \sum_{i=1}^n \left[f(\xi_i) - \mathbb{E}\, f(\xi_i) \right] \right\} \le \sqrt{2\pi}\, \mathbb{E} \sup_{f \in \mathcal{F}} \left\{ \sum_{i=1}^n h_i f(\xi_i) \right\},$$

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

Remark. In [36], the lemma is stated for the case when $\xi_1, \ldots, \xi_n$ are i.i.d. The case when $\xi_1, \ldots, \xi_n$ are not necessarily identical can be proved in a similar way, as noted in [28].

By Lemma B.1, we have

$$\begin{aligned} \mathbb{E} \sup_{v \in \overline{\mathcal{F}_g(\theta^\natural)} \cap \mathcal{S}^{p-1}} \left\{ \left\langle A^T \varepsilon, v \right\rangle \right\} &= \mathbb{E} \sup_{v \in \overline{\mathcal{F}_g(\theta^\natural)} \cap \mathcal{S}^{p-1}} \left\{ \langle \varepsilon, Av \rangle \right\} \\ &\le \sqrt{2\pi}\, \mathbb{E} \sup_{v \in \overline{\mathcal{F}_g(\theta^\natural)} \cap \mathcal{S}^{p-1}} \left\{ \langle h \cdot \varepsilon, Av \rangle \right\}, \end{aligned}$$

where $h \cdot \varepsilon := (h_i \varepsilon_i)_{i=1,\dots,n}$, and $h_1, \dots, h_n$ are i.i.d. standard Gaussian random variables. Note that $h \cdot \varepsilon$ is a random Gaussian vector with zero mean and covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$ which is dependent on A in general; moreover, since the entries in ε are independent, Σ is a diagonal matrix with diagonal entries given by $\Sigma_{i,i} := \operatorname{var} y_i$. Define $\tilde{h} := (\tilde{h}_i)_{i=1,\dots,n}$, where $\tilde{h}_i := \Sigma_{i,i}^{-1/2} h_i \varepsilon_i$. Then $\tilde{h}$ is a vector of i.i.d. standard Gaussian random variables; furthermore, it is still a vector of i.i.d. standard Gaussian random variables condition on A, and hence it is statistically independent of A.

Since $h \cdot \varepsilon$ and $\sqrt{\Sigma}\tilde{h}$ have the same probability distribution, we can write

$$\mathbb{E} \sup_{v \in \overline{\mathcal{F}_g(\theta^\natural)} \cap \mathcal{S}^{p-1}} \{\langle h \cdot \varepsilon, Av \rangle\} = \mathbb{E} \sup_{v \in \overline{\mathcal{F}_g(\theta^\natural)} \cap \mathcal{S}^{p-1}} \left\{ \left\langle \sqrt{\Sigma}\tilde{h}, Av \right\rangle \right\}.$$

Let $\mathcal{T} := \overline{\mathcal{F}_g(\theta^\natural)} \cap \mathcal{S}^{p-1}$. Condition on any given A (and hence Σ), we consider two mean-zero Gaussian processes $\{X_t\}_{t \in \mathcal{T}}$ and $\{Y_t\}_{t \in \mathcal{T}}$ defined as

$$X_t := \left\langle \sqrt{\Sigma}\tilde{h}, At \right\rangle, \quad Y_t := \sigma_{\max} \left\langle \tilde{h}, At \right\rangle,$$

where $\sigma_{max} := \max_i \Sigma_{i,i} = \max_i \sqrt{\operatorname{var} \varepsilon_i}$. We have, for any $t_1, t_2 \in \mathcal{T}$,

$$\mathbb{E} \left| X_{t_1} - X_{t_2} \right|^2 = \left\| \Sigma A (t_1 - t_2) \right\|_2^2 \le \sigma_{\max}^2 \left\| A(t_1 - t_2) \right\|_2^2 = \mathbb{E} \left| Y_{t_1} - Y_{t_2} \right|^2.$$

By Slepian's lemma, this implies

$$\mathbb{E} \sup_{t \in \mathcal{T}} X_t \le \mathbb{E} \sup_{t \in \mathcal{T}} Y_t.$$

Since the inequality holds given any realization of A, we have

$$\begin{aligned} \mathbb{E} \sup_{v \in \overline{\mathcal{F}_g(\theta^\natural)} \cap \mathcal{S}^{p-1}} \left\{ \left\langle A^T \varepsilon, v \right\rangle \right\} &\le \sqrt{2\pi}\, \sigma_{\max}\, \mathbb{E} \sup_{v \in \overline{\mathcal{F}_g(\theta^\natural)} \cap \mathcal{S}^{p-1}} \left\{ \left\langle \tilde{h}, Av \right\rangle \right\} \\ &= \sqrt{2\pi}\, \sigma_{\max}\, \mathbb{E} \sup_{v \in \overline{\mathcal{F}_g(\theta^\natural)} \cap \mathcal{S}^{p-1}} \left\{ \left\langle A^T \tilde{h}, v \right\rangle \right\}. \end{aligned}$$

It remains to prove

$$\mathbb{E} \sup_{v \in \overline{\mathcal{F}_g(\theta^\natural)} \cap \mathcal{S}^{p-1}} \left\{ \left\langle A^T \tilde{h}, v \right\rangle \right\} \le \sqrt{n}\, \omega_1(\overline{\mathcal{F}_g(\theta^\natural)}) := \sqrt{n}\, \mathbb{E} \sup_{v \in \overline{\mathcal{F}_g(\theta^\natural)} \cap \mathcal{S}^{p-1}} \left\{ \left\langle \tilde{h}, v \right\rangle \right\}. \tag{21}$$

We consider two cases:

Case 1: If A has i.i.d. standard Gaussian entries, then condition on $\tilde{h}$, $A^T\tilde{h}$ is a vector of mean-zero Gaussian random variables with covariance matrix $\left\| \tilde{h} \right\|_2 I$, and hence has the same probablity distribution as $\left\| \tilde{h} \right\| \bar{h}$, where $\bar{h}$ is a vector of

i.i.d. standard Gaussian random variables independent of $\tilde{h}$. Therefore,

$$\begin{aligned}\mathbb{E} \sup_{v \in \overline{\mathcal{F}_g(\theta^\natural)} \cap \mathcal{S}^{p-1}} \left\{ \left\langle A^T \tilde{h}, v \right\rangle \right\} &= \mathbb{E} \sup_{v \in \overline{\mathcal{F}_g(\theta^\natural)} \cap \mathcal{S}^{p-1}} \left\{ \left\langle \left\| \tilde{h} \right\| \bar{h}, v \right\rangle \right\} \\ &= \left(\mathbb{E}_{\tilde{h}} \left\| \tilde{h} \right\|_2 \right) \mathbb{E}_{\bar{h}} \sup_{v \in \overline{\mathcal{F}_g(\theta^\natural)} \cap \mathcal{S}^{p-1}} \left\{ \left\langle \bar{h}, v \right\rangle \right\} \\ &\leq \sqrt{n}\, \omega_1(\overline{\mathcal{F}_g(\theta^\natural)}).\end{aligned}$$

Case 2: If A has i.i.d. Rademacher entries, then condition on A, $A^T \tilde{h}$ is a vector of mean-zero Gaussian random variables with covariance matrix nI, and hence has the same probability distribution as $\sqrt{n}\bar{h}$, where $\bar{h}$ is a vector of i.i.d. standard Gaussian random variables. Therefore,

$$\begin{aligned}\mathbb{E} \sup_{v \in \overline{\mathcal{F}_g(\theta^\natural)} \cap \mathcal{S}^{p-1}} \left\{ \left\langle A^T \tilde{h}, v \right\rangle \right\} &= \mathbb{E} \sup_{v \in \overline{\mathcal{F}_g(\theta^\natural)} \cap \mathcal{S}^{p-1}} \left\{ \left\langle \sqrt{n} \bar{h}, v \right\rangle \right\} \\ &= \sqrt{n}\, \omega_1(\overline{\mathcal{F}_g(\theta^\natural)}).\end{aligned}$$

In summary, we obtain

$$\mathbb{E} \left\| \Pi_{\overline{\mathcal{F}_g(\theta^\natural)}}(-\nabla f_n(\theta^\natural)) \right\|_2 \leq \sqrt{2\pi}\, \sigma_{\max} \frac{\omega_1(\overline{\mathcal{F}_g(\theta^\natural)})}{\sqrt{n}},$$

if the entries of A are i.i.d. standard Gaussian or Rademacher random variables, for a canonical GLM, where the expectation is with respect to both A and the sample $(y_i)_{i=1,\ldots,n}$.

Let $\mathcal{E}$ denote that event that the RSC condition holds. Then we have

$$\begin{aligned}\mathbb{E} \left\| \Pi_{\overline{\mathcal{F}_g(\theta^\natural)}}(-\nabla f_n(\theta^\natural)) \right\|_2 = &\mathbb{P}(\mathcal{E})\, \mathbb{E}_{A,(y_i)|\mathcal{E}} \left\| \Pi_{\overline{\mathcal{F}_g(\theta^\natural)}}(-\nabla f_n(\theta^\natural)) \right\|_2 \\ &+ \mathbb{P}(\mathcal{E}^C)\, \mathbb{E}_{A,(y_i)|\mathcal{E}^C} \left\| \Pi_{\overline{\mathcal{F}_g(\theta^\natural)}}(-\nabla f_n(\theta^\natural)) \right\|_2,\end{aligned}$$

and hence

$$\begin{aligned}\mathbb{E}_{A,(y_i)|\mathcal{E}} \left\| \Pi_{\overline{\mathcal{F}_g(\theta^\natural)}}(-\nabla f_n(\theta^\natural)) \right\|_2 &\leq \frac{\mathbb{E} \left\| \Pi_{\overline{\mathcal{F}_g(\theta^\natural)}}(-\nabla f_n(\theta^\natural)) \right\|_2}{\mathbb{P}(\mathcal{E})} \\ &\leq 2\mathbb{E} \left\| \Pi_{\overline{\mathcal{F}_g(\theta^\natural)}}(-\nabla f_n(\theta^\natural)) \right\|_2,\end{aligned}$$

where we applied the assumption that $\mathbb{P}(\mathcal{E}) \geq 1/2$. By Lemma 3.2, this implies

$$\begin{aligned}\mathbb{E}_{A,\varepsilon|\mathcal{E}} \left\| \hat{\theta} - \theta^\natural \right\|_2 &\leq \frac{1}{\mu} \mathbb{E}_{A,(y_i)|\mathcal{E}} \left\| \Pi_{\overline{\mathcal{F}_g(\theta^\natural)}}(-\nabla f_n(\theta^\natural)) \right\|_2 \\ &\leq 2\sqrt{2\pi}\, \sigma_{\max} \frac{\omega_1(\overline{\mathcal{F}_g(\theta^\natural)})}{\mu\sqrt{n}}.\end{aligned}$$

This completes the proof.

Remark. If we want to adapt this proof to the deterministic A case, a technical issue arises when bounding the right-hand side of (21). As the random process $\left\{\tilde{X}_v := \left\langle \tilde{h}, v \right\rangle\right\}_{v\in\mathcal{V}}$, where $\mathcal{V} := \overline{\mathcal{F}_g(\theta^\natural)}\cap\mathcal{S}^{p-1}$, is a mean-zero Gaussian process, a standard approach is to bound $\sup_{v\in\mathcal{V}} \tilde{X}_v$ by Slepian's lemma. Note that, for any $v_1, v_2 \in \mathcal{V}$,

$$\mathbb{E}\left|\tilde{X}_{v_1} - \tilde{X}_{v_2}\right|^2 = \|A(v_1 - v_2)\|_2^2,$$

and hence an upper-bound on $\mathbb{E}\left|\tilde{X}_{v_1} - \tilde{X}_{v_2}\right|^2$ would depend on the largest eigenvalue of A. The largest eigenvalue of A, however, cannot be bounded above by a constant independent of n under the high-dimensional setting. Although we can weaken the requirement on A to a restricted smoothness condition as

$$\|Av\|_2 \le \sqrt{1+\epsilon}\,\|v\|_2, \quad \text{for all } v \in \overline{\mathcal{F}_g(\theta^\natural)} \cap \mathcal{S}^{p-1},$$

which, by Theorem E.1, holds with high probability. This condition does not imply

$$\|A(v_1 - v_2)\|_2^2 \le C\,\|v_1 - v_2\|_2^2,$$

for some dimension-independent constant $C > 0$, for all $v_1, v_2 \in \mathcal{V}$.

C Proof of Lemma 5.1

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

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

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: $\|e\|_2 \le t$. Therefore,

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

The lemma follows by taking expectations on both sides.

D Proof of Corollary 5.2

Let $e := \hat{\theta} - \theta^\natural$. If $e \in \mathcal{F}_g(\theta^\natural) \setminus t\mathcal{B}$, following the proof of Theorem 4.1, we can obtain

$$\mathbb{E}\,\|e\|_2 \le 2\sqrt{2\pi}\,\sigma_{\max}\frac{\omega_1(\overline{\mathcal{F}_g(\theta^\natural) \setminus t\mathcal{B}})}{\mu\sqrt{n}};$$

otherwise, we can bound the expected estimation error from above by t. Therefore,

$$\begin{aligned}\mathbb{E}\,\|e\|_2 &\le \max\left\{t, 2\sqrt{2\pi}\,\sigma_{\max}\frac{\omega_1(\overline{\mathcal{F}_g(\theta^\natural)\setminus t\mathcal{B}})}{\mu\sqrt{n}}\right\}\\ &\le t + 2\sqrt{2\pi}\,\sigma_{\max}\frac{\omega_1(\overline{\mathcal{F}_g(\theta^\natural)\setminus t\mathcal{B}})}{\mu\sqrt{n}}.\end{aligned}$$

E Proof of Corollary 6.1 and Corollary 6.2

The proofs in this section rely on the following theorem [19].

Theorem E.1 ([19]). *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 $\alpha\ge 1$, and let $\epsilon\in(0,1)$. Then there exist constants c_1 and c_2 such that for all $x\in\mathcal{T}$ satisfying*

$$\|x\|_2 \ge \gamma_n^*\left(\frac{\epsilon}{c_1\alpha^2},\mathcal{T}\right) := \inf\left\{t>0 : t\ge\frac{c_1\alpha^2\omega_t(\mathcal{T})}{\epsilon\sqrt{n}}\right\}, \tag{22}$$

we have

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

with probability at least $1-\exp\left(-c_2\epsilon^2 n/\alpha^4\right)$.

We note that the sub-Gaussian norm of a vector of i.i.d. standard Gaussian entries or i.i.d. Rademacher entries is bounded above by a constant [37].

E.1 Proof of Corollary 6.1

We prove by Corollary 5.2.

Let A be defined as in Theorem 4.1. We verify the condition (14) by Theorem E.1. Since $\omega_t(\overline{\mathcal{F}_g(\theta^\natural)\setminus t\mathcal{B}}) = t\omega_1(\overline{\mathcal{F}_g(\theta^\natural)\setminus t\mathcal{B}})$, the condition (22) is equivalent to requiring

$$\sqrt{n}\ge\frac{c_1\alpha^2\omega_1(\overline{\mathcal{F}_g(\theta^\natural)\setminus t\mathcal{B}})}{\epsilon}.$$

Once this inequality is satisfied, we can set $\mu=1-\epsilon$, and the condition (14) hold with probability at least $1-\exp\left(-c_2\epsilon^2 n/\alpha^4\right)$. Note that $\sigma_{\max}=\sqrt{\mathbb{E}\,w_i^2}=\sigma$. This completes the proof.

E.2 Proof of Corollary 6.2

We prove the corollary by Corollary 5.2.

It is known that

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

for the ML estimator in a canonical GLM, where A is defined as in Theorem 4.1, and $D(\theta)$ is a diagonal matrix; furthermore, there exists a continuous strictly positive function ϕ such that the (i,i)-th entry of $D(\theta)$ is given by $\phi(\langle a_i, \theta \rangle)$. Since the entries of A are i.i.d. Rademacher random variables, for any $\theta \in \mathcal{G}$,

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

By the extreme value theorem, the diagonal entries of $D(\theta)$ are bounded below by a constant $\nu > 0$ for all $\theta \in \mathcal{G}$, which is independent of n. Similarly, $\sigma_{\max}$ is bounded above by a constant independent of n.

The rest of the proof is similar to the last paragraph in the previous subsection. By Theorem E.1, if we choose n such that

$$\sqrt{n} \geq \frac{c\alpha^2 \omega_1(\overline{\mathcal{F}_g(\theta^\natural) \setminus t\mathcal{B}})}{\epsilon},$$

then the condition (14) holds with probability at least $1 - \exp\left(-c_2 \epsilon^2 n / \alpha^4\right)$ with $\mu = \nu(1 - \epsilon)$.

References

[1] F. Bach, "Learning with submodular functions: A convex optimization perspective," *Found. Trends Mach. Learn.*, vol. 6, no. 2–3, pp. 145–373, 2013.

[2] A. Banerjee, S. Chen, F. Fazayeli, and V. Sivakumar, "Estimation with norm regularization," 2015, arXiv:1505.02294v1 [stat.ML].

[3] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[4] P. Bickel, Y. Ritov, and A. B. Tsybakov, "Simultaneous analysis of Lasso and Dantzig selector," *Ann. Stat.*, vol. 37, no. 4, pp. 1705–1732, 2009.

[5] P. Bühlmann and S. van de Geer, *Statistics for High-Dimensional Data*. Berlin: Springer, 2011.

[6] J.-F. Cai and W. Xu, "Guarantees of total variation minimization for signal recovery," 2013, arXiv:1301.6791v6.

[7] E. J. Candès and Y. Plan, "Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 2342–2359, 2011.

[8] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Found. Comput. Math.*, vol. 12, pp. 805–849, 2012.

[9] M. El Halabi and V. Cevher, "A totally unimodular view of structured sparsity," in *18th Int. Conf. Artificial Intelligence and Statistics*, 2015.

[10] S. T. Flammia, D. Gross, Y.-K. Liu, and J. Eisert, "Quantum tomography via compressed sensing: Error bounds, sample complexity and efficient estimators," *New J. Phys.*, vol. 14, 2012.

[11] R. Foygel and L. Mackey, "Corrupted sensing: Novel guarantees for separating structured signals," 2014, arXiv:1305.2524v2 [cs.IT].

[12] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert, "Quantum state tomography via compressed sensing," *Phys. Rev. Lett.*, vol. 105, 2010.

[13] S. Gunasekar, P. Ravikumar, and J. Ghosh, "Exponential family matrix completion under structural constraints," in *Proc. 31st Int. Conf. Machine Learning*, 2014.

[14] J. Honorio and T. Jaakkola, "A unified framework for consistency of regularized loss minimizers," in *Proc. 31st Int. Conf. Machine Learning*, 2014.

[15] M. Jaggi, "Revisiting Frank-Wolfe: Projection-free sparse convex optimization," in *Proc. 30th Int. Conf. Machine Learning*, 2013.

[16] S. M. Kakade, O. Shamir, K. Sridharan, and A. Tewari, "Learning exponential families in high-dimensions: Strong convexity and sparsity," in *Proc. 13th Int. Conf. Artificial Intelligence and Statistics*, 2010.

[17] V. Koltchinskii, "A remark on low rank matrix recovery and noncommutative Bernstein type inequalities," in *From Probability to Statistics and Back: High-Dimensional Models and Processes*, M. Banerjee, F. Bunea, J. Huang, V. Koltchinskii, and M. H. Maathuis, Eds. Inst. Math. Stat., 2013, pp. 213–226.

[18] S. Mendelson, "Learning without concentration," 2014, arXiv:1401.0304v2 [cs.LG].

[19] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann, "Reconstruction and subgaussian operators in asymptotic geometric analysis," *Geom. Funct. Anal.*, vol. 17, pp. 1248–1282, 2007.

[20] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," 2010, arXiv:1010.2731v1 [math.ST].

[21] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," *Stat. Sci.*, vol. 27, no. 4, pp. 538–557, 2012.

[22] Y. Nesterov, "Gradient methods for minimizing composite functions," *Math. Program., Ser. B*, vol. 140, pp. 125–161, 2013.

[23] S. Oymak, C. Thrampoulidis, and B. Hassibi, "Simple bounds for noisy linear inverse problems with exact side information," 2013, arXiv:1312.0641v2 [cs.IT].

[24] ——, "The squared-error of generalized LASSO: A precise analysis," 2013, arXiv:1311.0830v2 [cs.IT].

[25] Y. Plan and R. Vershynin, "Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach," *IEEE Trans. Inf. Theory*, vol. 59, no. 1, pp. 482–494, Jan. 2013.

[26] ——, "The generalized Lasso with non-linear observations," 2015, arXiv:1502.0407v1 [cs.IT].

[27] Y. Plan, R. Vershynin, and E. Yudovina, "High-dimensional estimation with geometric constraints," 2014, arXiv:1404.3749v1 [math.PR].

[28] D. Pollard, *Convergence of Stochastic Processes*. New York, NY: Springer-Verl., 1984.

[29] M. Raginsky, R. M. Willett, Z. T. Harmany, and R. F. Marcia, "Compressed sensing performance bounds under Poisson noise," *IEEE Trans. Inf. Theory*, vol. 58, no. 8, pp. 3990–2010, Aug. 2010.

[30] N. Rao, B. Recht, and R. Nowak, "Universal measurement bounds for structured sparse signal recovery," in *Adv. Neural Information Processing Systems*, 2012.

[31] G. Raskutti, M. J. Wainwright, and B. Yu, "Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6976–6994, Oct. 2011.

[32] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. R. Stat. Soc., Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.

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

[34] S. van de Geer, "The deterministic Lasso," Seminar für Statistik, Eidgenössische Technische Hochschule, Research Report No. 140, 2007.

[35] ——, "Weakly decomposable regularization penalties and structured sparsity," *Scand. J. Stat.*, vol. 41, pp. 72–86, 2013.

[36] R. van Handel, *Probability in High Dimension*, Jun. 2014. [Online]. Available: http://www.princeton.edu/~rvan/

[37] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," in *Compressed Sensing: Theory and Applications*, Y. Eldar and G. Kutyniok, Eds. Cambridge Univ. Press, 2012, ch. 5, pp. 210–268.

[38] ——, "Estimation in high dimensions: A geometric perspective," 2014, arXiv:1405.5103v1 [math.ST].

[39] X. Zhang, Y. Yu, and D. Schuurmans, "Polar operators for structured sparse estimation," in *Adv. Neural Information Processing Systems 26*, 2013.