Gaussian Regression with Convex Constraints

Matey Neykov
Carnegie Mellon University
Department of Statistics & Data Science

Abstract

The focus of this paper is the linear model with Gaussian design under convex constraints. Specifically, we study the performance of the constrained least squares estimate. We derive two general results characterizing its performance – one requiring a tangent cone structure, and one which holds in a general setting. We use our general results to analyze three functional shape constrained problems where the signal is generated from an underlying Lipschitz, monotone or convex function. In each of the examples we show specific classes of functions which achieve fast adaptive estimation rates, and we also provide non-adaptive estimation rates which hold for any function. Our results demonstrate that the Lipschitz, monotone and convex constraints allow one to analyze regression problems even in high-dimensional settings where the dimension may scale as the square or fourth degree of the sample size respectively.

1 INTRODUCTION

Recently there has been a flurry of work on the shape constrained Gaussian sequence model [see, e.g., Chatterjee et al., 2014, Bellec et al., 2018, Chatterjee et al., 2015, Guntuboyina and Sen, 2015, Zhang et al., 2002, among others]. These works study the model

$$Y_i = \theta_i^* + \varepsilon_i,$$

for $i \in [n] = \{1, \dots, n\}$ where the vector $\boldsymbol{\theta}^* \in \mathbb{R}^n$ is known to belong to a convex set K. In the present paper we are interested in the related, but different problem of linear regression

$$Y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta}^* + \varepsilon_i,$$

for $i \in [n]$ where the vector $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is known to belong to a convex set K. We study the constrained least squares estimate given by

$$\widehat{\boldsymbol{\beta}} := \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{n} \sum_{i \in [n]} (Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta})^2, \text{ given that } \boldsymbol{\beta} \in K. \tag{1.1}$$

This model is also related to the works of Thrampoulidis et al., Plan and Vershynin [2016], Plan et al. [2017] where the authors consider a single index model (SIM) formulation of the problem, i.e.,

$$Y_i = F(\boldsymbol{X}_i^\top \boldsymbol{\beta}^*, \varepsilon_i)$$

and let K be a star-shaped set[1]. In the latter model one loses the scaling of the vector $\boldsymbol{\beta}^*$ for identifiability (i.e., $\|\boldsymbol{\beta}^*\|_2$ can be absorbed in F), and the nice properties of convexity for the larger generality of star-shaped sets.

1.1 Convex Sets of Interest

Although in Section 2 we derive general results under the assumptions of Gaussian design $\boldsymbol{X}_i$, we are specifically interested in three types of convex sets K. All of our examples can be motivated by assuming that the vector $\boldsymbol{\beta}^*$ satisfies the following condition $\beta_i^* = f(\frac{i}{p})$ for $i \in [p] = \{1, \dots, p\}$, where $f : [0, 1] \mapsto \mathbb{R}$ is a function, which is constrained to be Lipschitz, monotone or convex. In this framework, one may think about the vectors $\boldsymbol{X}_i$ as compressing the discretized functional values to a potentially lower dimensional space ($n < p$), and the ε_i as noise occurring during transmission of the compressed values. The goal then is to recover the original functional values as closely as possible having the information $(Y_i, \boldsymbol{X}_i)_{i \in [n]}$. We will see

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

[1] I.e., a set S such that $\lambda S \subset S$ for all $\lambda \in [0, 1]$. Every convex set containing 0 is star-shaped.

that for the examples of convex sets K that we consider, it is indeed possible to compress to a lower dimensional space and then accurately recover the functional values. Such examples have not been previously analyzed in a regression setting, and some of them are more commonly associated with the Gaussian sequence model.

The first example of such set is when f is an L-Lipschitz function. In this case it is simple to see that

$$K := \{\boldsymbol{\beta} \in \mathbb{R}^p : |\beta_i - \beta_{i-1}| \leq \tfrac{L}{p}, i \in \{2, \ldots, p\}\},$$

where the Lipschitz constant L is a tuning parameter. Here the set K is convex, but it is not a cone unlike the following two examples. It is noteworthy to mention that the constraints on $\boldsymbol{\beta}^*$ do take advantage of $\beta_i^* = f(\frac{i}{p})$, i.e., that the points over which f is evaluated at are equispaced on $[0, 1]$.

The second example takes f as a monotone function. The convex set K is

$$K := \{\boldsymbol{\beta} \in \mathbb{R}^p : \beta_{i-1} \leq \beta_i, i \in \{2, \ldots, p\}\},$$

i.e., $\boldsymbol{\beta}^*$ forms a monotone sequence in the given order. Importantly, unlike in the Lipschitz case, this example is independent of the underlying design where the function f is evaluated, although one can still think of $\beta_i^* = f(\frac{i}{p})$.

The final example takes f as a convex function. The set K is

$$K := \{\boldsymbol{\beta} \in \mathbb{R}^p : \beta_i - \beta_{i-1} \leq \beta_{i+1} - \beta_i, i \in \{2, \ldots, p-1\}\}.$$

Just as in the Lipschitz f example, the conditions on K above assume that the design on which f is evaluated is equispaced on $[0, 1]$.

We will reveal that in all three examples above, estimate (1.1) works in two vastly different regimes, deemed the adaptive and non-adaptive regimes, depending on the underlying vector $\boldsymbol{\beta}^*$. In all three examples there exist vectors $\boldsymbol{\beta}^*$ with special structure which admit fast (or adaptive), nearly parametric rates of convergence, which adapt to the special structure of $\boldsymbol{\beta}^*$. These adaptive rates are typically much faster than the general non-adaptive rates which hold for any arbitrary $\boldsymbol{\beta}^* \in K$. Furthermore, both the adaptive and non-adaptive rates allow one to work in a high dimensional setting where $p > n$. Specifically, for the Lipschitz and monotone examples p can be nearly of order n^2, while for the convex functions example p can be of order n^4. The latter shows that it is possible to effectively compress Lipchitz, monotone and convex functional values and then recover them consistently.

1.2 Innovation and Related Work

As indicated earlier, our work is related to two separate branches of statistics and signal processing – the Gaussian sequence model and compressive sensing. On the compressive sensing side related works include [Thrampoulidis et al., Oymak et al., 2013, Plan and Vershynin, 2016, Plan et al., 2017, Cai et al., 2016]. Out of those works the most closely related work is Plan and Vershynin [2016]. In contrast to Plan and Vershynin [2016] however, we do not restrict the norm of the vector $\boldsymbol{\beta}^*$, and consider a convex set K, which helps us to carry through a more refined analysis compared to a star-shaped set.

Works on the Gaussian sequence model which are relevant are [Chatterjee et al., 2014, Bellec et al., 2018, Chatterjee et al., 2015, Guntuboyina and Sen, 2015, Bellec and Tsybakov, 2015]. Specifically, they consider a sequence model under convex constraints, and examples which are similar to the type of examples that we consider. In particular, they have focused on the monotone and convex function cones.

The novelty of this paper is two-fold. First we derive two general results regarding the performance of the convex constrained regression in a Gaussian design setting. Our results have advantages over existing works in that they allow for selecting a target vector different from the underlying true vector, and do not restrict the norm of the parameter. Next we propose and analyze the example of Lipschitz regression which to the best of our knowledge has not appeared either in the Gaussian sequence model or in the compressive sensing literature. Furthermore, to the best of our knowledge, the adaptive and non-adaptive rates derived for the three examples of Lipschitz, monotone and convex functions have not been analyzed in a regression setting previously and are novel.

1.3 Notation

We now briefly outline some commonly used notation. Other notation will be defined as needed throughout the paper. For a vector $\mathbf{v} = (v_1, \ldots, v_p)^\top$ let $\|\mathbf{v}\|_q$ denote the ℓ_q norm. For a set K and a given vector $\mathbf{v}$, the difference $K - \mathbf{v} := \{\mathbf{u} - \mathbf{v} : \mathbf{u} \in K\}$. We often use $\mathbf{I}_p$ to denote a $p \times p$ identity matrix. For any integer $k \in \mathbb{N}$ we use the shorthand notation $[k] = \{1, \ldots, k\}$. We also use standard asymptotic notations. Given two sequences $\{a_n\}, \{b_n\}$ we write $a_n \lesssim b_n$ if there exists an absolute constant $C < \infty$ such that $a_n \leq C b_n$. The inequality $\gtrsim$ is similarly defined.

1.4 Organization

For convenience of the reader here we outline the structure of the paper. In Section 2 we give our general results for the case when $\boldsymbol{X}_i$ have a Gaussian distribution. In Section 3 we consider in detail the three examples which we outlined above. Section 4 is dedicated to numerical experiments and the discussion is left to the final Section 5.

2 MODEL BACKGROUND AND GENERAL RESULTS

Suppose we have n independent and identically distributed (i.i.d.) observations from the model

$$Y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta}^* + \varepsilon_i$$

where $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is known to belong to a convex set K, the noise ε_i is independent of $\boldsymbol{X}_i$ with $\mathbb{E}\varepsilon_i = 0$, variance $\operatorname{Var}\varepsilon_i = \sigma^2 > 0$ and we assume $\mathbb{E}\varepsilon_i^4 < \infty$. Let $\mathbf{X}$ be the matrix stacking $\boldsymbol{X}_i^\top$ into its rows, and $\boldsymbol{Y}$ be the vector with entries equal to Y_i. In view of this notation the constrained least squares estimator (1.1) can be rewritten compactly as:

$$\widehat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2, \text{ given that } \boldsymbol{\beta} \in K.^{2} \quad (2.1)$$

There are multiple examples of convex sets K where the above formulation is very meaningful. In the simplest of examples, K can be taken a subspace of $\mathbb{R}^p$. Another example is to take $K = \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}\|_1 \le L\}$, which produces the vanilla LASSO procedure as defined by Tibshirani [1996]. Motivated by a scenario where $\boldsymbol{\beta}^*$ takes discretized functional values, in Section 3 we consider three examples of sets K restricting the shape of the function generating $\boldsymbol{\beta}^*$ as we detailed in Section 1.1.

Before we proceed with our general results, we require several definitions. We will first introduce the concept of a tangent cone.

Definition 2.1 (Tangent Cone)**.** For a convex set K and a vector $\mathbf{v} \in K$, define the tangent cone of K at $\mathbf{v}$: $\mathcal{T}_{K,\mathbf{v}}$ as the closure of the set

$$\mathcal{T}_{K,\mathbf{v}} := \operatorname{cl}\{t\mathbf{u} : t \ge 0, \mathbf{u} \in K - \mathbf{v}\}.$$

The tangent cone collects all possible directions at arbitrary scales, along which centered at the vector $\mathbf{v}$ one remains in the set K. Next we present the concept of local Gaussian mean width, which is a (localized) measure of size for arbitrary sets in $\mathbb{R}^p$.

[2] If there exist multiple solutions, pick any of them.

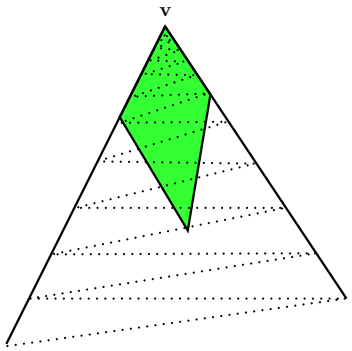


Figure 1: A depiction of a tangent cone. The set K is depicted in green. The tangent cone $\mathcal{T}_{K,\mathbf{v}}$ consists of all vectors between the two rays (it is also marked by a dotted line).

Definition 2.2 (Local Gaussian Mean Width)**.** For a set $S \subset \mathbb{R}^p$ and a real number $x > 0$ define its local Gaussian mean width as

$$w_x(S) := \mathbb{E} \sup_{\mathbf{v} \in S, \|\mathbf{v}\|_2 \le x} \mathbf{v}^\top \boldsymbol{\xi},$$

where $\boldsymbol{\xi} \sim \mathcal{N}(0, \mathbf{I})$ is a standard Gaussian vector.

We have the following result

Theorem 2.3 (Tangent Cone Bound)**.** Let $\boldsymbol{X}_i \sim \mathcal{N}(0, \mathbf{I})$. Suppose that $\boldsymbol{\beta}^* \in K$ and fix any $\mathbf{v} \in K$. Let $n \gtrsim w_1^2(\mathcal{T}_{K,\mathbf{v}})$. Then the estimate of program (2.1) satisfies

$$\begin{aligned}\|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}\|_2 &\le \left(\frac{4(\sqrt{n} + w_1(\mathcal{T}_{K,\mathbf{v}}) + t)}{(\sqrt{n-1} - w_1(\mathcal{T}_{K,\mathbf{v}}) - t)} + 1\right)\|\boldsymbol{\beta}^* - \mathbf{v}\|_2 \\ &+ \frac{\sqrt{2}(w_1(\mathcal{T}_{K,\mathbf{v}}) + \sqrt{2t})\sigma}{\sqrt{n}\left(\sqrt{\frac{n-1}{n}} - \frac{w_1(\mathcal{T}_{K,\mathbf{v}})+t}{\sqrt{n}}\right)^2},\end{aligned}$$

with probability at least $1 - e^{-t} - 3e^{-t^2/2} - \frac{\operatorname{Var}\varepsilon^2}{n\sigma^4}$.

In the above, we note that the parameter t is a "free" parameter and can be set to arbitrary values, where higher values of t correspond to higher probabilities. The assumption $\boldsymbol{X}_i \sim \mathcal{N}(0, \mathbf{I})$ is required for technical reasons which help us relate bounds on the coefficients to geometric characteristics of the underlying convex set K. We show how this condition can be relaxed to the more general condition $\boldsymbol{X}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ below. We remark that Theorem 2.3 is related to Theorem 1 of Oymak et al. [2013]. However, unlike the latter result, our bound allows one to pick a different vector $\mathbf{v}$ in place of $\boldsymbol{\beta}^*$ such that $\mathbf{v}$ is close to $\boldsymbol{\beta}^*$ but also has a simple tangent cone structure. This is especially useful in cases where the original signal $\boldsymbol{\beta}^*$ might not have

a small tangent cone, and the Gaussian width of its cone might approach its maximum value of order $\sqrt{p}$. We would also like to stress that this result utilizes the fact that K is convex, and we do not expect it to hold for star-shaped sets as in Plan and Vershynin [2016], Plan et al. [2017]. Below we give a direct Corollary to Theorem 2.3 which ignores the precise constant terms for the sake of readability.

Corollary 2.4 (Tangent Cone Bound)**.** Let $\boldsymbol{X}_i \sim \mathcal{N}(0, \mathbf{I})$ and fix any $\mathbf{v} \in K$. Suppose that $n \gtrsim w_1^2(\mathcal{T}_{K,\mathbf{v}})$, and assume that $\boldsymbol{\beta}^* \in K$. Then conditionally on the error term with probability at least .99 we have

$$\|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}\|_2 \lesssim \|\boldsymbol{\beta}^* - \mathbf{v}\|_2 + \frac{w_1(\mathcal{T}_{K,\mathbf{v}})}{\sqrt{n}}\sigma + \frac{\sigma}{\sqrt{n}}.$$

We will apply Corollary 2.4 to derive both the adaptive and non-adaptive rates for Lipschitz regression, and the adaptive rates for monotone and convex regressions. We can also give a bound without requiring a tangent cone structure and without even requiring that $\boldsymbol{\beta}^* \in K$.

Theorem 2.5 (Misspecified Estimation Bound without Tangent Cone Structure)**.** Let $\boldsymbol{X}_i \sim \mathcal{N}(0, \mathbf{I})$. Fix $\mathbf{v} \in K$ such that for some $x > 0$ we have $n \gtrsim w_x^2(K - \mathbf{v})/x^2$. Then, the estimate of program (2.1) satisfies

$$\begin{aligned}
&\|\mathbf{v} - \widehat{\boldsymbol{\beta}}\|_2 \\
&\quad\lesssim \frac{\|\mathbf{X}(\boldsymbol{\beta}^* - \mathbf{v})\|_2}{\sqrt{n}} + \frac{w_x(K - \mathbf{v}) + x\sqrt{t}}{\sqrt{n}x}\sigma + x \\
&\quad\lesssim \left(1 + \sqrt{\frac{p}{n}} + \sqrt{\frac{t}{n}}\right)\|\boldsymbol{\beta}^* - \mathbf{v}\|_2 \\
&\quad\quad + \frac{w_x(K - \mathbf{v}) + x\sqrt{t}}{\sqrt{n}x}\sigma + x,
\end{aligned}$$

with probability at least $1 - \exp(-n/8) - 2\exp(-t) - \frac{\operatorname{Var}\varepsilon^2}{n\sigma^4}$.

We apply Theorem 2.5 (with $\mathbf{v} = \boldsymbol{\beta}^*$) to derive the non-adaptive rates of monotone and convex regressions utilizing calculations of the local mean width which have appeared in the literature.

2.1 What if there is a Covariance?

We have so far supposed that $\boldsymbol{X}_i \sim \mathcal{N}(0, \mathbf{I})$. We now derive two corollaries which are useful in the more general case when $\boldsymbol{X}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma})$. We have

Corollary 2.6 (Tangent Cone Bound General Covariance)**.** Let $\boldsymbol{X}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ and fix any $\mathbf{v} \in K$. Let $n \gtrsim w_1^2(\boldsymbol{\Sigma}^{1/2}\mathcal{T}_{K,\mathbf{v}})$ and suppose $\boldsymbol{\beta}^* \in K$. Then conditionally on the error term with probability at least .99 we have

$$\begin{aligned}
\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})\|_2 &\lesssim \|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}^* - \mathbf{v})\|_2 \\
&\quad + \frac{w_1(\boldsymbol{\Sigma}^{1/2}\mathcal{T}_{K,\mathbf{v}})}{\sqrt{n}}\sigma + \frac{\sigma}{\sqrt{n}}.
\end{aligned}$$

It is not hard to check (see Remark 1.7 [Plan and Vershynin, 2016]) that

$$w_1(\boldsymbol{\Sigma}^{1/2}\mathcal{T}_{K,\mathbf{v}}) \le \|\boldsymbol{\Sigma}^{-1/2}\|_2\|\boldsymbol{\Sigma}^{1/2}\|_2 w_1(\mathcal{T}_{K,\mathbf{v}}),$$

and therefore if $\boldsymbol{\Sigma}^{1/2}$ is well conditioned the mean width would be of the same order as the one without using $\boldsymbol{\Sigma}$. On the other hand there could exist matrices that are not well conditioned but do not change the tangent cone too much in some examples; such matrices will result in similar mean widths as in the one of independent Gaussians.

Plugging in $\mathbf{v} = \boldsymbol{\beta}^*$ in Theorem 2.5 we state an analogous corollary below.

Corollary 2.7 (Estimation Bound without Tangent Cone Structure General Covariance)**.** Let $\boldsymbol{X}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma})$. Suppose that $\boldsymbol{\beta}^* \in K$ and that for some $x > 0$ we have $n \gtrsim w_x^2(\boldsymbol{\Sigma}^{1/2}(K - \boldsymbol{\beta}^*))/x^2$. Then the estimate of program (2.1), satisfies

$$\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}})\|_2 \lesssim \frac{w_x(\boldsymbol{\Sigma}^{1/2}(K - \boldsymbol{\beta}^*)) + x\sqrt{t}}{\sqrt{n}x}\sigma + x,$$

with probability at least $1 - \exp(-n/8) - \exp(-t) - \frac{\operatorname{Var}\varepsilon^2}{n\sigma^4}$.

3 EXAMPLES

One of the most well studied examples of the set K is the LASSO, i.e., $K = \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}\|_1 \le L\}$. We will not focus on this classical example since it has received plenty of attention and results can be found in Oymak et al. [2013] among others. Instead below we consider three examples where the parameter $\boldsymbol{\beta}$ can be thought of evaluating a function $f : [0, 1] \mapsto \mathbb{R}$ at equispaced partition of $[0, 1]$, i.e., $\beta_i^* = f(\frac{i}{p})$ for $i \in [p]$. We will focus on the independent Gaussian design setting for the sake of a clear presentation.

3.1 Lipschitz Regression

Suppose that $\boldsymbol{\beta}^*$ satisfies $|\beta_i^* - \beta_{i-1}^*| \le \frac{L}{p}$ for some fixed constant L. This is equivalent to the setting where $\beta_i^* = f(\frac{i}{p})$ for $f : [0, 1] \mapsto \mathbb{R}$ being an L-Lipschitz function. The convex set K therefore consists of the following collection of vectors:

$$K := \{\boldsymbol{\beta} \in \mathbb{R}^p : |\beta_i - \beta_{i-1}| \le \tfrac{L}{p}, i \in \{2, \dots, p\}\},$$

where $L > 0$ is a tuning parameter. The constraints in K and the least squares objective (2.1) produce a quadratic program which can be implemented efficiently via interior point methods.

Our goal for this section is to apply Corollary 2.4 to derive both adaptive and non-adaptive bounds. In order to derive the non-adaptive bound, we would like to find a vector $\mathbf{v}$ which is close to any given vector $\boldsymbol{\beta} \in K$ but has a small tangent cone structure. Hence we start by first showing a lemma regarding approximations of arbitrary $\boldsymbol{\beta} \in K$.

Lemma 3.1. Any vector $\boldsymbol{\beta} \in K$ can be approximated by a vector $\mathbf{v} \in K$ with at most $\ell+1$ affine pieces and slopes equal to $\pm\frac{L}{p}$ and satisfying $|v_i - v_{i-1}| = \frac{L}{p}$ for all i, so that

$$\|\boldsymbol{\beta} - \mathbf{v}\|_2 \leq \frac{2L\sqrt{p}}{\ell}.$$

Next we will upper bound the mean width of a tangent cone of a vector $\mathbf{v}$ with $\ell+1$ affine pieces and satisfying $|v_i - v_{i+1}| = \frac{L}{p}$ for all i. Before stating the result we define the monotone sequence cone $\mathcal{S}_k^{\uparrow} := \{\mathbf{v} \in \mathbb{R}^k : v_1 \leq v_2 \leq \ldots \leq v_k\}$. We have the following result

Lemma 3.2. Let $\mathbf{v}$ be a vector as described above, and let $T_1, \ldots, T_{\ell+1}$ be a disjoint partition of $[p]$ such that $\mathbf{v}$ is affine on each T_i with sign of the slope on T_i equal to s_i. Then $\mathcal{T}_{K,\mathbf{v}} \subset (-s_1\mathcal{S}^{\uparrow}_{|T_1|}) \times \ldots \times (-s_{\ell+1}\mathcal{S}^{\uparrow}_{|T_{\ell+1}|})$.

We will now use Lemma 3.2 to derive an upper bound of the mean width $w_1(\mathcal{T}_{K,\mathbf{v}})$. Before we do so let us first introduce a closely related concept to the mean width of a cone – the statistical dimension.

Definition 3.3 (Statistical Dimension)**.** For a given cone $\mathcal{C}$ define the statistical dimension as:

$$\delta(\mathcal{C}) := \mathbb{E}\big[(\sup_{\mathbf{v}\in\mathcal{C}, \|\mathbf{v}\|_2\leq 1} \mathbf{v}^\top \boldsymbol{\xi})^2\big],$$

where $\boldsymbol{\xi} \sim \mathcal{N}(0, \mathbf{I})$ is a standard Gaussian vector.

Simple properties of the statistical dimension include $\delta(\mathcal{C}_1) \leq \delta(\mathcal{C}_2)$ for $\mathcal{C}_1 \subseteq \mathcal{C}_2$, and $\delta(\mathcal{C}_1 \times \mathcal{C}_2) = \delta(\mathcal{C}_1) + \delta(\mathcal{C}_2)$. For proofs of these statements we refer the reader to Amelunxen et al. [2014]. Take a vector $\mathbf{v}$ which consists of $\ell+1$ affine functions with slopes $\pm\frac{L}{p}$ as in Lemma 3.2. It follows that the statistical dimension

$$\begin{aligned}\delta(\mathcal{T}_{K,\mathbf{v}}) &\leq \sum_{i=1}^{\ell+1} \delta(-s_i\mathcal{S}^{\uparrow}_{|T_i|}) \\ &= \sum_{i=1}^{\ell+1} \delta(\mathcal{S}^{\uparrow}_{|T_i|}) \leq \sum_{i=1}^{\ell+1} \log(e|T_i|) \\ &\leq (\ell+1)\log(ep/(\ell+1)).\end{aligned}$$

The second bound comes from a known result on the statistical dimension of the monotone cone [see Amelunxen et al., 2014, e.g.], while the last bound is due to Jensen's inequality. Since by Cauchy-Schwartz $\delta(\mathcal{T}_{K,\mathbf{v}}) \geq w_1^2(\mathcal{T}_{K,\mathbf{v}})$, we conclude an inequality on the local mean width:

$$w_1(\mathcal{T}_{K,\mathbf{v}}) \leq \sqrt{(\ell+1)\log(ep/(\ell+1))}. \tag{3.1}$$

Using Lemmas 3.1 and the bound in the preceding display, we arrive at the following Corollary of Corollary 2.4.

Corollary 3.4 (Lipschitz Regression Adaptive and Non-Adaptive Rates)**.** Suppose that $\boldsymbol{\beta}^*$ satisfies $|\beta_i^* - \beta_{i-1}^*| \leq \frac{L}{p}$ for a fixed constant L. Let $n \gtrsim Lp^{1/2}\log p$. Then conditionally on the error term with probability at least .99 we have

$$\|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}\|_2 \lesssim \frac{[\log(ep)L]^{1/3}p^{1/6}}{n^{1/3}}(1+\sigma) + \frac{\sigma}{\sqrt{n}}. \tag{3.2}$$

Furthermore if $\boldsymbol{\beta}^*$ consists of ℓ affine functions with slopes $\pm\frac{L}{p}$ as in Lemma 3.2, we have the following adaptive rate

$$\|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}\|_2 \lesssim \sqrt{\frac{\ell\log ep/\ell}{n}}\sigma + \frac{\sigma}{\sqrt{n}},$$

with probability at least .99 given that $n \gtrsim \ell\log ep/\ell$.

Corollary 3.4 makes it apparent that Lipschitz regression works in two major regimes in the high dimensional setting when $p \geq n$. In the first regime, if one has a function which consists of ℓ affine pieces of slopes precisely equal to $\pm\frac{L}{p}$ the rate is nearly parametric; while in the second regime Lipschitz regression works as long as $n \gg Lp^{1/2}\log p$. Importantly, the bound (3.2) cannot be derived using results from previous works such as [Oymak et al., 2013], since it requires evaluation of a tangent cone which is not centered at the true value $\boldsymbol{\beta}^*$. Finally we remark that it may be possible to remove the $\log p$ factor from (3.2) via an application of Theorem 2.5. Since this requires a delicate calculation of the mean width we defer this improvement for future work.

3.2 Monotone Regression

Monotone regression uses the set of vectors $K := \mathcal{S}_p^{\uparrow}$, where the set of monotone sequences $\mathcal{S}_p^{\uparrow}$ was defined in the previous section. The monotone constraint combined with the least squares loss function produce a quadratic program which can be implemented via interior point methods. Additionally, monotone regression can be implemented efficiently by using projected gradient descent. This is so since at each step the

projection boils down to running isotonic regression, which can be done by the fast pool adjacent violators algorithm (PAVA) [Mair et al., 2009].

Below we give a Corollary of Corollary 2.4.

Corollary 3.5 (Constant Pieces Adaptive Rate)**.** Suppose that the vector $\boldsymbol{\beta}^* \in K$ consists of ℓ monotone constant pieces. Then we have the following bound

$$\|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}\|_2 \lesssim \sqrt{\frac{\ell \log ep/\ell}{n}}\sigma + \frac{\sigma}{\sqrt{n}}.$$

This is the same rate as the nearly parametric rate in Lipschitz sequences from the previous section, however it is achieved at different type of vectors $\boldsymbol{\beta}^*$. Similar adaptive to the number of constant pieces of $\boldsymbol{\beta}^*$ phenomenon has been observed in the Gaussian sequence setting where the behavior of isotonic regression has been well studied [Bellec et al., 2018, Chatterjee et al., 2015]. Next, to obtain a sharp bound in the setting where $\boldsymbol{\beta}^*$ need not consist of constant pieces, we will derive a Corollary of Theorem 2.5.

Corollary 3.6 (General Non-Adaptive Rate)**.** Suppose that $n \gtrsim (\beta_n - \beta_1)p^{1/2}$ with $\boldsymbol{\beta}^* \in K$. Then with probability at least .99 we have

$$\|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}\|_2 \lesssim \frac{(\beta_n - \beta_1)^{1/3}p^{1/6}}{n^{1/3}}\sigma^{2/3} + \frac{\sigma}{\sqrt{n}}.$$

The bound in the preceding display resembles closely the one we obtained in (3.2). Even when $p \geq n$ the monotone constraint helps and one can obtain precise estimate provided that $n \gtrsim (\beta_n - \beta_1)p^{1/2}$.

3.3 Convex Regression

Suppose the parameter $\boldsymbol{\beta}^*$ satisfies $\beta_i^* = f(\frac{i}{p})$ for $i \in [p]$, where f is a convex function. Equivalently one can express this constraint as $\beta_i^* - \beta_{i-1}^* \leq \beta_{i+1}^* - \beta_i^*$ for all $i \in \{2, \ldots, p-1\}$. Let the set K be

$$K := \{\boldsymbol{\beta} \in \mathbb{R}^p : \beta_i - \beta_{i-1} \leq \beta_{i+1} - \beta_i, i \in \{2, \ldots, p-1\}\}.$$

It is simple to check that K is indeed a convex set, and hence (2.1) has efficient implementation. We first give a Corollary of Corollary 2.4.

Corollary 3.7 (Affine Pieces Adaptive Rate)**.** Suppose that the vector $\boldsymbol{\beta}^* \in K$ consists of ℓ affine pieces. Then we have the following bound

$$\|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}\|_2 \lesssim \sqrt{\frac{\ell \log ep/\ell}{n}}\sigma + \frac{\sigma}{\sqrt{n}}.$$

Finally we provide a Corollary to Theorem 2.5 which discusses general convex functions.

Corollary 3.8 (General Non-Adaptive Rate)**.** Let $n \gtrsim ((\max \beta_i - \min \beta_i) + 1)^{1/2}p^{1/4}$. Then with probability at least .99 we have

$$\|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}\|_2 \lesssim \frac{\sigma^{4/5}((\max \beta_i - \min \beta_i) + 1)^{1/5}p^{1/10}}{n^{2/5}} + \sigma\sqrt{\frac{\log p}{n}}.$$

It follows that convex regression with equispaced design can be consistent even when p is of the order of n^4. This contrasts the two previous examples, where p could be of the order of n^2.

4 NUMERICAL EXPERIMENTS

In this section we present numerical evidence in support of our theoretical findings. We consider six different scenarios of $\boldsymbol{\beta}^*$ two per each example – one adaptive and one general $\boldsymbol{\beta}^*$. Specifically, in the Lipschitz example we take $\boldsymbol{\beta}^*$ as $\beta_i^* = f_{\text{adapt}}(\frac{i}{p})$ or $\beta_i^* = f_{\text{nonadapt}}(\frac{i}{p})$ where

$$f_{\text{adapt}}(x) = x\mathbb{1}(0 \leq x < \tfrac{1}{3}) + (\tfrac{2}{3} - x)\Big[\mathbb{1}(\tfrac{1}{3} \leq x < \tfrac{2}{3}) - \mathbb{1}(\tfrac{2}{3} \leq x \leq 1)\Big],$$
$$f_{\text{nonadapt}}(x) = \frac{\sin(10x)}{10}.$$

Similarly for monotone regression we take,

$$f_{\text{adapt}}(x) = \tfrac{1}{2}\mathbb{1}(\tfrac{1}{3} \leq x < \tfrac{2}{3}) + \mathbb{1}(\tfrac{2}{3} \leq x \leq 1),$$
$$f_{\text{nonadapt}}(x) = x,$$

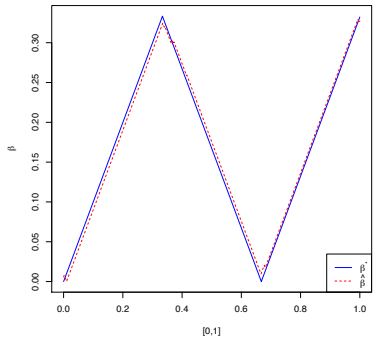
and for convex regression we take

$$f_{\text{adapt}}(x) = x, \quad f_{\text{nonadapt}}(x) = x^2.$$

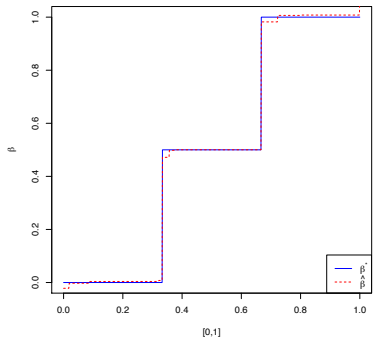
The noise is $\varepsilon_i \sim \mathcal{N}(0, 1)$. We plot the examples of the parameters $\boldsymbol{\beta}^*$ along with typical estimates in Figs 2 and 3. Additional numerical simulations which verify our results can be found in the supplementary material.

5 DISCUSSION

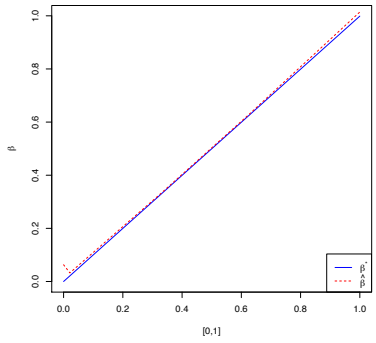
In this manuscript we considered Gaussian linear regression under convex constraints. We gave two types of general results – one under the presence of tangent cone structure and one without the need of such structure. We analyzed three examples where the vector $\boldsymbol{\beta}^*$ is generated by an underlying function whose shape is constrained by the set K. The examples we considered, showed that Lipschitz, monotone and convex functions may be compressed to a low dimensional



(a) Lipschitz Regression

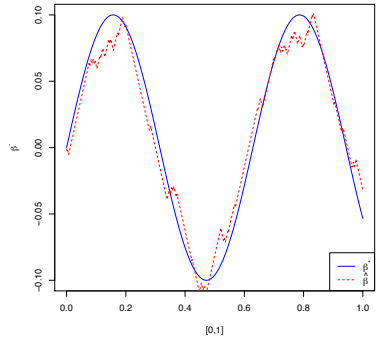


(b) Monotone Regression

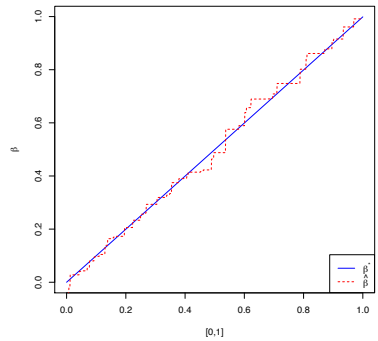


(c) Convex Regression

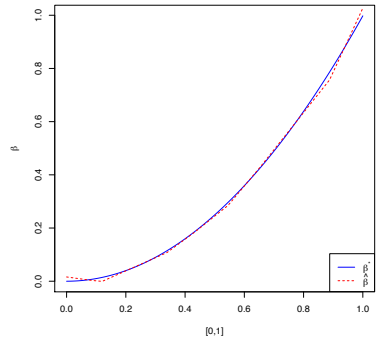
Figure 2: Three typical examples of signals $\boldsymbol{\beta}^*$ (in blue solid line) which have "small" tangent cone structure with their corresponding estimates $\widehat{\boldsymbol{\beta}}$ (in red dashed line). The dimension $p = 900$, the samples $n = 100$, and we have linearly interpolated the $\boldsymbol{\beta}^*$ and $\widehat{\boldsymbol{\beta}}$ values. For Lipschitz regression, we have chosen a piecewise affine function with a fixed and known slope. For monotone and convex regression we used a piecewise constant and a linear function respectively.



(a) Lipschitz Regression



(b) Monotone Regression



(c) Convex Regression

Figure 3: Three examples of signals $\boldsymbol{\beta}^*$ which do not have a "small" tangent cone with their corresponding estimates $\widehat{\boldsymbol{\beta}}$. The dimension $p = 900$, the samples $n = 100$, and we have linearly interpolated the $\boldsymbol{\beta}^*$ and $\widehat{\boldsymbol{\beta}}$ values. We see that in comparison to Fig 2 the estimates $\widehat{\boldsymbol{\beta}}$ are not as close to $\boldsymbol{\beta}^*$ and appear more jagged. Convex regression seems to give a closer fit as compared to the Lipschitz and monotone regressions. This corroborates our finding that convex regression with equispaced design has a faster rate of convergence.

space with a Gaussian matrix and then consistently recovered.

While our results are derived in a noisy setting, it is not hard to see that in the noiseless setting exact recovery is possible for vectors $\boldsymbol{\beta}^*$ with tangent cone structure as long as $n \gtrsim w_1^2(\mathcal{T}_{K,\boldsymbol{\beta}^*})$. We leave for future work the question of whether exact recovery is possible for vectors lacking tangent cone structure. The conjecture is that exact recovery is impossible, but the same rates as we derived in the noisy setting continue to hold.

ACKNOWLEDGEMENTS

The author is grateful to four anonymous reviewers and the area chair for their comments and suggestions which helped to improve this manuscript.

References

D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294, 2014.

P. C. Bellec and A. B. Tsybakov. Sharp oracle bounds for monotone and convex regression through aggregation. *Journal of Machine Learning Research*, 16: 1879–1892, 2015.

P. C. Bellec et al. Sharp oracle inequalities for least squares estimators in shape restricted regression. *The Annals of Statistics*, 46(2):745–780, 2018.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. OUP Oxford, 2013.

T. T. Cai, T. Liang, A. Rakhlin, et al. Geometric inference for general high-dimensional linear inverse problems. *The Annals of Statistics*, 44(4):1536–1563, 2016.

S. Chatterjee, A. Guntuboyina, B. Sen, et al. On risk bounds in isotonic and other shape restricted regression problems. *The Annals of Statistics*, 43(4): 1774–1800, 2015.

S. Chatterjee et al. A new perspective on least squares under convex constraint. *The Annals of Statistics*, 42 (6):2340–2381, 2014.

S. Chatterjee et al. An improved global risk bound in concave regression. *Electronic Journal of Statistics*, 10(1):1608–1629, 2016.

Y. Gordon. On milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$. In *Geometric Aspects of Functional Analysis*, pages 84–106. Springer, 1988.

A. Guntuboyina and B. Sen. Global risk bounds and adaptation in univariate convex regression. *Probability Theory and Related Fields*, 163(1-2):379–411, 2015.

P. Mair, K. Hornik, and J. de Leeuw. Isotone optimization in r: pool-adjacent-violators algorithm (pava) and active set methods. *Journal of statistical software*, 32(5):1–24, 2009.

S. Oymak, C. Thrampoulidis, and B. Hassibi. Simple bounds for noisy linear inverse problems with exact side information. *arXiv preprint arXiv:1312.0641*, 2013.

Y. Plan and R. Vershynin. The generalized lasso with non-linear observations. *IEEE Transactions on information theory*, 62(3):1528–1537, 2016.

Y. Plan, R. Vershynin, and E. Yudovina. High-dimensional estimation with geometric constraints. *Information and Inference: A Journal of the IMA*, 6(1):1–40, 2017.

C. Thrampoulidis, E. Abbasi, and B. Hassibi. The lasso with non-linear measurements is equivalent to one with linear measurements. 2015.

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

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. C. Eldar and G. Kutyniok, editors, *Compressed Sensing: Theory and Applications*. Cambridge University Press, 2012.

C.-H. Zhang et al. Risk bounds in isotonic regression. *The Annals of Statistics*, 30(2):528–555, 2002.