Solving Equations of Random Convex Functions via Anchored Regression

Sohail Bahmani

sohail.bahmani@ece.gatech.edu

Justin Romberg

jrom@ece.gatech.edu

October 3, 2017

Abstract

We consider the question of estimating a solution to a system of equations that involve convex nonlinearities, a problem that is common in machine learning and signal processing. Because of these nonlinearities, conventional estimators based on empirical risk minimization generally involve solving a non-convex optimization program. We propose *anchored regression*, a new approach based on convex programming that amounts to maximizing a linear functional (perhaps augmented by a regularizer) over a convex set. The proposed convex program is formulated in the natural space of the problem, and avoids the introduction of auxiliary variables, making it computationally favorable. Working in the native space also provides great flexibility as structural priors (e.g., sparsity) can be seamlessly incorporated.

For our analysis, we model the equations as being drawn from a fixed set according to a probability law. Our main results provide guarantees on the accuracy of the estimator in terms of the number of equations we are solving, the amount of noise present, a measure of statistical complexity of the random equations, and the geometry of the regularizer at the true solution. We also provide recipes for constructing the anchor vector (that determines the linear functional to maximize) directly from the observed data.

1 Introduction

We consider the problem of (approximately) solving a system of nonlinear equations with convex nonlinearities. In particular, we observe

$$
\begin{aligned}
y_1 &= f_1\left(\boldsymbol{x}_\star\right) + \xi_1 \\
y_2 &= f_2\left(\boldsymbol{x}_\star\right) + \xi_2 \\
&\vdots \\
y_M &= f_M\left(\boldsymbol{x}_\star\right) + \xi_M,
\end{aligned}
\tag{1.1}
$$

where $\boldsymbol{x}_\star \in \mathbb{R}^N$ contains the ground-truth parameters to be estimated, the functions $f_1, f_2, \ldots, f_M$ are convex and known, and the $\xi_1, \xi_2, \ldots, \xi_M$ are additive noise terms. Given

1

the observations (1.1), we propose as an estimator for $\boldsymbol{x}_\star$ the solution of a convex program that balances consistency with the observations against structure induced by a (convex) regularization term. This convex formulation means that if the functions f_m and their first few derivatives can be computed efficiently, the proposed estimator is computationally tractable.

Our main results give error bounds on the quality of the produced estimate, specified by the Euclidean distance to $\boldsymbol{x}_\star$, under a model where the functions f_m are drawn at random from a set $\mathcal{F}$ according to some probability law. The estimation error, and the sufficient number of equations M needed to achieve it, depend on the *Rademacher complexity* of a set $\mathcal{A}$ with respect to the induced probability law on the gradients ∇f_m — $\mathcal{A}$ is essentially the set of all ascent directions of the functional being maximized at $\boldsymbol{x}_\star$, and depends on the geometry of the regularizer.

One of these results, Corollary 2.2, shows that if the f_m are "generic" convex functions (in that their gradients are sufficiently diverse in $\mathbb{R}^N$), then $\boldsymbol{x}_\star$ can be recovered from $M \sim N$ observations $y_m = f_m(\boldsymbol{x}_\star)$.

1.1 Motivating examples

To illustrate the broad applicability of the observation model (1.1), we provide a few motivating examples.

Example 1.1 (Convex function of linear predictor)**.** Perhaps the simplest example is the case that

$$f_m(\boldsymbol{x}) = \phi(\boldsymbol{a}_m^\top \boldsymbol{x}) \qquad\qquad m = 1, 2, \ldots, M\,, \qquad\qquad (1.2)$$

for a given convex function $\phi : \mathbb{R} \to \mathbb{R}$, and data samples $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_M \in \mathbb{R}^N$. This form of observation appears in many instances of *generalized linear models* (GLM) [21] with ϕ being the *mean function* (or the *inverse link function*) of the GLM. In some cases, the mean function may not be convex itself, but it has a property (e.g., it is concave, log-convex, or log-concave) that allows us to easily convert it to a convex function. In such cases, ϕ would be the appropriate transformation of the mean function.

Equations of the form (1.2) also appear in the *phase retrieval* problem from computational imaging. The goal in phase retrieval is to estimate an signal/image (up to a global phase) from intensity of some linear observations. More precisely, given (random) *measurement vectors* $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_M \in \mathbb{C}^N$ the goal is to estimate $\boldsymbol{x}_\star \in \mathbb{C}^N$ from noisy intensity measurements $y_m = |\boldsymbol{a}_m^* \boldsymbol{x}_\star|^2 + \xi_m$ for $m = 1, 2, \ldots, M$. Aside from minor technicalities for treating complex-valued vectors, clearly the observation model in phase retrieval is a special case of (1.2) with $\phi(z) = |z|^2$ being the nonlinearity.

Example 1.2 (A simple neural net with partially known weights)**.** The second example is related to learning with a *single hidden-layer neural network* [12, 30]. Given prescribed weights $w_1, w_2, \ldots, w_K \geq 0$ for the hidden layer of a single hidden-layer neural network, and an activation function $\phi : \mathbb{R} \to \mathbb{R}$, the output of the network can be expressed as

$$f_m(\boldsymbol{x}) = \sum_{k=1}^{K} w_k \phi(\boldsymbol{a}_m^\top \boldsymbol{x}_k)\,,$$

2

where $\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_1^\mathsf{T} & \boldsymbol{x}_2^\mathsf{T} & \cdots & \boldsymbol{x}_K^\mathsf{T} \end{bmatrix}^\mathsf{T}$ denotes the input layer's weight parameters and $\boldsymbol{a}_m$ denotes the mth data sample. For convex activation functions such as the popular $\phi(u) = \max(u, 0)$ and $\phi(u) = \log_2(1 + e^u)$, the functions $f_m(\boldsymbol{x})$ are also convex and the model assumed in (1.1) applies.

Example 1.3 (A simple neural net with unknown weights)**.** In the previous example, if the activation function is also positive homogeneous, then the knowledge of the output weights can be dropped from the assumptions. In fact, each non-negative weight w_k of the output layer can be absorbed into the corresponding input layer weight $\boldsymbol{x}_k$ because they interact only through $w_k \phi(\boldsymbol{a}_m^\mathsf{T} \boldsymbol{x}_k) = \phi(\boldsymbol{a}_m^\mathsf{T}(w_k \boldsymbol{x}_k))$. Therefore, this case reduces to the previous example with $w_1 = w_2 = \cdots = w_K = 1$. We may need to adapt the regularizer on the w_k and $\boldsymbol{x}_k$ to account for this combination of variables.

Example 1.4 (Dimensionality reduction)**.** Another example where our framework can be used is the problem of learning a dimensionality reduction. The goal of this problem is to find a linear embedding of data samples $\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_M \in \mathbb{R}^N$ to $\mathbb{R}^K$ with $K < N$ that still allows accurate prediction of the corresponding labels $y_1, y_2, \ldots, y_M$. More precisely, given functions $\mu_1, \mu_2, \ldots, \mu_T : \mathbb{R}^{N \times K} \to \mathbb{R}$, we want to estimate $\boldsymbol{X}$ such that the functions

$$f_m(\boldsymbol{X}) = \sum_{t=1}^{T} w_t \mu_t(\boldsymbol{X}^\mathsf{T} \boldsymbol{a}_m)$$

for some weights $w_1, w_2, \ldots, w_T$, accurately predict the corresponding labels y_m for $m = 1, 2, \ldots, M$. By naturally flattening $\boldsymbol{X}$ into a vector in $\mathbb{R}^{NK}$, and assuming that the functions $\mu_t(\cdot)$ are convex with known weights $w_t \geq 0$, our framework applies to this model. Similar to the previous example, we may also relax the knowledge of w_t's assuming that μ_t's are positive homogeneous so that with $\boldsymbol{w} = \begin{bmatrix} w_1 & w_2 & \cdots & w_T \end{bmatrix}^\mathsf{T}$ we can write

$$f_m(\boldsymbol{w}, \boldsymbol{X}) = \sum_{t=1}^{T} \mu_t(w_t \boldsymbol{X}^\mathsf{T} \boldsymbol{a}_m).$$

Again we can flatten $\boldsymbol{X}$ to $\boldsymbol{x} \in \mathbb{R}^{NK}$ and observe that the functions $f_m(\boldsymbol{w}, \boldsymbol{X})$ only depend on the rank-one matrix $\boldsymbol{Z} = \boldsymbol{x}\boldsymbol{w}^\mathsf{T}$. Namely, we can write

$$f_m(\boldsymbol{w}, \boldsymbol{X}) = \sum_{t=1}^{T} \mu_t\left((\boldsymbol{I}_{K \times K} \otimes \boldsymbol{a}_m^\mathsf{T}) \boldsymbol{z}_t\right),$$

where $\otimes$ denotes the Kronecker product and $\boldsymbol{z}_t$ is the tth column of $\boldsymbol{Z}$. As a function of $\boldsymbol{Z}$, the observation functions $f_m(\boldsymbol{Z})$ are convex and compatible with the observation model (1.1). The only remaining concern is inducing the rank-one structure of $\boldsymbol{Z}$ which can be done through *nuclear norm* regularization in the estimator.

The last two examples can be extended further to address composition of non-negative mixtures of positive homogeneous functions by lifting the unknown parameters to a higher order low-rank tensor. However, estimating low-rank tensors usually involves computationally prohibitive operations which renders these mentioned extensions less interesting from practical point of view.

3

1.2 Anchored regression

In this paper, we propose *anchored regression* to estimate $\boldsymbol{x}_\star$ in the parametric model described by (1.1). An *anchor vector* $\boldsymbol{a}_0 \in \mathbb{R}^N$ is a unit vector (i.e., $\|\boldsymbol{a}_0\|_2 = 1$) that obeys

$$\langle \boldsymbol{a}_0, \boldsymbol{x}_\star \rangle \geq \delta \|\boldsymbol{x}_\star\|_2 \,, \tag{1.3}$$

for an absolute constant $\delta \in (0, 1]$. In words, the anchor vector has a non-vanishing correlation with a ground truth. Given an anchor vector $\boldsymbol{a}_0$, our proposed estimator for (1.1) is the convex program

$$\begin{aligned} \underset{\boldsymbol{x}}{\operatorname{argmax}} \ \ &\langle \boldsymbol{a}_0, \boldsymbol{x} \rangle - \Omega(\boldsymbol{x}) \\ \text{subject to } &R_M^+(\boldsymbol{x}) \leq R_M^+(\boldsymbol{x}_\star) + \varepsilon \,, \end{aligned} \tag{1.4}$$

where $\Omega(\boldsymbol{x})$ is a convex regularizer, $\varepsilon > 0$ is a small constant, and $R_M^+(\cdot)$ is the empirical one-sided error

$$R_M^+(\boldsymbol{x}) \overset{\text{def}}{=} \frac{1}{M} \sum_{m=1}^{M} (f_m(\boldsymbol{x}) - y_m)_+ \,, \quad (\cdot)_+ = \max(\cdot, 0),$$

which is also convex. Note that when there is no noise in the observations and we take $\varepsilon = 0$, the constraint $R_M^+(\boldsymbol{x}) \leq 0$ is equivalent to $f_m(\boldsymbol{x}) \leq y_m$ for all m. Of course, when there is noise, the value of

$$R_M^+(\boldsymbol{x}_\star) = \frac{1}{M} \sum_{m=1}^{M} (-\xi_m)_+ \,,$$

is unknown in general, but depending on the noise model we may assume $R_M^+(\boldsymbol{x}_\star)$ can be estimated and the absolute error of such an estimate is captured by ε in (1.4).

The anchor vector $\boldsymbol{a}_0$ can be interpreted as a "rough guess" for the solution $\boldsymbol{x}_\star$. This guess might come from some kind of a priori information about the solution, or it might be formed directly from the observations. In Section 3 below, we describe some schemes by which an $\boldsymbol{a}_0$ obeying (1.3) can be constructed in a data-driven manner.

1.3 Advantages of anchored regression

Solving the system of nonlinear equations (1.1) is in general a hard problem. In the particular case where the f_m are multivariate polynomials, a standard approach is to *lift* the variables to a higher-dimensional space that linearizes the equations. The key idea is that any homogeneous polynomial function of the vector $\boldsymbol{x} \in \mathbb{R}^N$ of degree d can be expressed as a linear function of the tensor $\boldsymbol{X} = \boldsymbol{x}^{\otimes d}$. In the lifted domain, then, solving (1.1) amounts to finding a *rank-one tensor* solution to a system of linear equations. For quadratic polynomials ($d = 2$), there is a natural convex relaxation to this problem using semi-definite programming; the effectiveness of this relaxation for solving generic quadratic equations was studied in [4], and analysis for several types of structured equations that arise in engineering problems can be found in [1, 5, 20]. For $d \geq 3$, it is not clear how the lifted problem can be relaxed, as the required tensor computational primitives are computationally hard [13]. Even in the $d = 2$ case where

the relaxation can be cast as an SDP, the computational cost in squaring the number of variables can be prohibitive for medium- to large-scale problems.

In contrast, our estimator works for observations that are general convex functions of the unknown variables.[1] Our method works in the natural domain of the unknown variables, which is computationally efficient and provides us with the flexibility to incorporate prior structural information about the ground truth through regularization.

The anchor $\boldsymbol{a}_0$ also allows us to avoid one of the pitfalls of convexification. It can be that there are multiple equally valid solutions to (1.1), especially since the f_m are convex. For instance, in the phase retrieval problem where $f_m(\boldsymbol{x}) = |\boldsymbol{a}_m^* \boldsymbol{x}|^2$ for a set of complex-valued vectors $\boldsymbol{a}_m$, a global change of phase to $\boldsymbol{x}_\star$ will not affect the measurements at all. Minimizing a convex loss function that assigns an equal score to each of these equivalent solutions is bound to fail, as all of the points in the convex hull of the equivalence set will score equally well or better. The lifting approach for quadratic equations gets around this by mapping every equivalent solution to the same point in the lifted space. Our method also handles this in a straightforward way by introducing a bias towards the solutions best aligned with the anchor $\boldsymbol{a}_0$.

1.4   Geometry of convex equations

The analysis of (1.4) for solving (1.1) has a very clean geometric interpretation. Suppose that the observations are noise-free, $y_m = f_m(\boldsymbol{x}_\star)$, and we solve (1.4) without the regularization term and with $\varepsilon = 0$. Since $R_M^+(\boldsymbol{x}_\star) = 0$, this is equivalent to solving

$$\begin{aligned} \underset{\boldsymbol{x} \in \mathbb{R}^N}{\text{maximize}} \ & \langle \boldsymbol{x}, \boldsymbol{a}_0 \rangle \\ \text{subject to } & f_m(\boldsymbol{x}) \leq y_m, \quad m = 1, \ldots, M \,. \end{aligned} \tag{1.5}$$

Since all of the constraints in the program above are active at $\boldsymbol{x}_\star$, the KKT conditions for this program tell us that $\boldsymbol{x}_\star$ is indeed the solution when

$$-\boldsymbol{a}_0 + \sum_{m=1}^M \lambda_m \nabla f_m(\boldsymbol{x}_\star) = \boldsymbol{0},$$

for some $\lambda_1, \ldots, \lambda_M \geq 0$. More succinctly, (1.5) is successful when

$$\boldsymbol{a}_0 \in \text{cone}\left(\{\nabla f_m(\boldsymbol{x}_\star)\}_{m=1}^M\right).$$

This is illustrated in Figure 1. If we knew the $\nabla f_m(\boldsymbol{x}_\star)$ in addition to the $f(\boldsymbol{x}_\star)$, then generating such a $\boldsymbol{a}_0$ would be straightforward. But in general, we do not have knowledge of these gradients.

The main results in this paper say that it is enough to find a $\boldsymbol{a}_0$ that is roughly aligned with $\boldsymbol{x}_\star$. That is, if the f_m (and hence the $\nabla f_m(\boldsymbol{x}^\star)$) are generated at random, then with high probability

$$\langle \boldsymbol{x}_\star, \boldsymbol{a}_0 \rangle \geq \delta \quad \Rightarrow \quad \boldsymbol{a}_0 \in \text{cone}\left(\{\nabla f_m(\boldsymbol{x}_\star)\}_{m=1}^M\right),$$

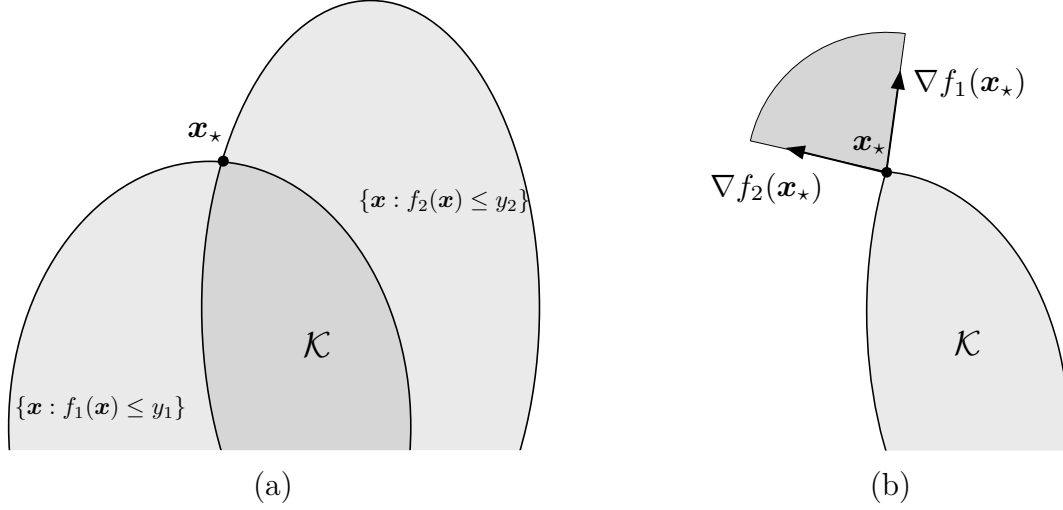[1]Of course, we need to be able to evaluate the f_m and some number of its derivatives to actually solve (1.4).

Figure 1: *A simple set of two nonlinear equations, $y_1 = f_1(\boldsymbol{x}_\star)$ and $y_2 = f_2(\boldsymbol{x}_\star)$, in $\mathbb{R}^2$. (a) The unknown $\boldsymbol{x}_\star$ is an extreme point of $\mathcal{K} = \{\boldsymbol{x} \; : \; f_m(\boldsymbol{x}) \leq y_m, \; m = 1, \ldots, M\}$. (b) The program (1.5) will recover $\boldsymbol{x}_\star$ when $\boldsymbol{a}_0$ is in the cone generated by the two gradients.*

provided that the number of equations M is large enough. In fact, we show that *every* roughly aligned vector will work. Geometrically, this means that the cone of all valid anchor vectors

$$\mathcal{C}_\delta = \{\boldsymbol{z} \; : \; \langle \boldsymbol{z}, \boldsymbol{x}_\star \rangle \geq \delta \|\boldsymbol{x}_\star\|_2 \|\boldsymbol{z}\|_2\},$$

is included in the cone generated by the gradients

$$\mathcal{C}_\delta \subseteq \operatorname{cone}\left(\{\nabla f_m(\boldsymbol{x}_\star)\}_{m=1}^M\right),$$

again when M is large enough. As δ becomes smaller (meaning that the anchor is more aligned with the true solution) the cone $\mathcal{C}_\delta$ shrinks, and the inclusion above may be satisfied with smaller M.

Adding a convex regularizer $\Omega(\boldsymbol{x})$ to the program makes the optimality conditions weaker. If we again observe $y_m = f_m(\boldsymbol{x}_\star)$ and then solve

$$\underset{\boldsymbol{x} \in \mathbb{R}^N}{\operatorname{maximize}} \; \langle \boldsymbol{a}_0, \boldsymbol{x} \rangle - \Omega(\boldsymbol{x})$$
$$\text{subject to } f_m(\boldsymbol{x}) \leq y_m, \qquad m = 1, 2, \ldots, M\,,$$

then $\boldsymbol{x}_\star$ is the solution when

$$\boldsymbol{a}_0 \in \operatorname{cone}\left(\{\nabla f_m(\boldsymbol{x}^\star)\}_{m=1}^M\right) + \partial\Omega(\boldsymbol{x}^\star),$$

where $\partial\Omega(\boldsymbol{x}_\star)$ is the subdifferential of Ω at $\boldsymbol{x}_\star$,

$$\partial\Omega(\boldsymbol{x}) = \left\{\boldsymbol{g} \; : \; \Omega(\boldsymbol{x} + \boldsymbol{z}) \geq \Omega(\boldsymbol{x}) + \boldsymbol{g}^\mathsf{T}\boldsymbol{z} \text{ for all } \boldsymbol{z} \in \mathbb{R}^N\right\}.$$

Finding such a $\boldsymbol{a}_0$ is now easier than in the unregularized case, as the convex cone that needs to enfold it is larger. The (subdifferential of the) regularization term is effectively working as an additional set of observations — if the subgradients in $\partial\Omega(\boldsymbol{x}^\star)$ are not aligned with the $\nabla f_m(\boldsymbol{x}^\star)$, we may be able to get away with many fewer equations.

1.5 Related work

As mentioned in Section 1.1 our approach can be applied to most GLMs. Another relevant model is the semiparametric *single index model* (see [14], for example) where, again, the model assumes linear predictors, but the nonlinear function μ is not known. Under some regularity assumption on μ, in [28, 29] simple estimators based on convex programming are shown to produce accurate estimates (up to some scaling factor). Because μ is unknown, our framework does not apply to the single index model. However, there are interesting instances of (1.1) (e.g., phase retrieval) where the results of [28, 29] do not apply as the assumed regularity conditions lead to trivial bounds.

The most relevant results to our work are the recent methods proposed for phase retrieval in [2] and independently in [8] that exploit anchor vectors. As mentioned above, the phase retrieval problem can be described by the model (1.1) with $f_m(\boldsymbol{x}) = |\boldsymbol{a}_m^* \boldsymbol{x}|^2$ for some random measurement vector $\boldsymbol{a}_m$. It is shown in [2, 8] that, using an anchor, the ground truth $\boldsymbol{x}_\star$ can be estimated accurately and with optimal sample complexity through a convex program analogous to (1.4). The analyses in [8] and [2] differ in that the former assumes the anchor is independent of the measurements whereas the latter does not make this assumption. Alternative proofs and variations of this phase retrieval method also appeared later in [9–11].

2 Main result

To show that (1.4) produces an accurate estimate of $\boldsymbol{x}_\star$, it suffices to show that the set of ascent directions,

$$\mathcal{A} = \{\boldsymbol{h} \, : \, \langle \boldsymbol{a}_0 - \boldsymbol{g}, \boldsymbol{h} \rangle \geq 0, \text{ for all } \boldsymbol{g} \in \partial\Omega(\boldsymbol{x}_\star)\} \, ,$$

does not contain any vector with large ℓ_2 norm that is consistent with the constraint. Namely, if $\boldsymbol{h} \in \mathcal{A}$ has a large ℓ_2 norm, then $R_M^+(\boldsymbol{x}_\star + \boldsymbol{h}) > R_M^+(\boldsymbol{x}_\star) + \varepsilon$.

As explained further in Section 3, we will be interested in anchor vectors $\boldsymbol{a}_0$ constructed from the observations y_m (and knowledge of the f_m). However, the dependence of the set $\mathcal{A}$ on the observations would complicate the probabilistic analysis. We avoid this dependence by an expansion of the set $\mathcal{A}$ that allows us to decouple it from $\boldsymbol{a}_0$.

Let $\boldsymbol{h}_\perp$ and $\boldsymbol{a}_{0\perp}$ denote the projection of $\boldsymbol{h}$ and $\boldsymbol{a}_0$ onto the hyperplane orthogonal to $\boldsymbol{x}_\star$, respectively. Using the property of the anchor vector (1.3) we can write

$$\begin{aligned}
\langle \boldsymbol{a}_0, \boldsymbol{h} \rangle &= \langle \boldsymbol{a}_{0\perp}, \boldsymbol{h} \rangle + \langle \frac{\boldsymbol{x}_\star \boldsymbol{x}_\star^\top \boldsymbol{a}_0}{\|\boldsymbol{x}_\star\|_2^2}, \boldsymbol{h} \rangle \\
&\leq \|\boldsymbol{a}_{0\perp}\|_2 \|\boldsymbol{h}_\perp\|_2 + \left(\frac{\boldsymbol{x}_\star^\top \boldsymbol{a}_0}{\|\boldsymbol{x}_\star\|_2} - \delta \right) \langle \frac{\boldsymbol{x}_\star}{\|\boldsymbol{x}_\star\|_2}, \boldsymbol{h} \rangle + \delta \langle \frac{\boldsymbol{x}_\star}{\|\boldsymbol{x}_\star\|_2}, \boldsymbol{h} \rangle \\
&\leq \sqrt{\|\boldsymbol{a}_{0\perp}\|_2^2 + \left(\frac{\boldsymbol{x}_\star^\top \boldsymbol{a}_0}{\|\boldsymbol{x}_\star\|_2} - \delta \right)^2} \sqrt{\|\boldsymbol{h}_\perp\|_2^2 + \langle \frac{\boldsymbol{x}_\star}{\|\boldsymbol{x}_\star\|_2}, \boldsymbol{h} \rangle^2} + \delta \langle \frac{\boldsymbol{x}_\star}{\|\boldsymbol{x}_\star\|_2}, \boldsymbol{h} \rangle \\
&\leq \sqrt{1 - \delta^2} \|\boldsymbol{h}\|_2 + \delta \langle \frac{\boldsymbol{x}_\star}{\|\boldsymbol{x}_\star\|_2}, \boldsymbol{h} \rangle \, .
\end{aligned}$$

This implies the inclusion

$$\mathcal{A} \subset \mathcal{A}_\delta \overset{\text{def}}{=} \left\{ \boldsymbol{h} \, : \, \sqrt{1 - \delta^2} \, \|\boldsymbol{h}\|_2 + \langle \frac{\delta \boldsymbol{x}_\star}{\|\boldsymbol{x}_\star\|_2} - \boldsymbol{g}, \boldsymbol{h} \rangle \geq 0, \text{ for all } \boldsymbol{g} \in \partial\Omega\left(\boldsymbol{x}_\star\right) \right\}. \tag{2.1}$$

It suffices to show that $\mathcal{A}_\delta$ does not contain any point with (relatively) large ℓ_2 norm consistent with the constraints as mentioned above.

As it becomes clear in the sequel, it is critical that $\mathcal{A}_\delta$ excludes a sufficiently large subset of $\mathbb{R}^N$. Therefore, we may require δ to be bounded away from zero. For example, in unstructured phase retrieval, where $f_m(\boldsymbol{x}) = (\boldsymbol{a}_m^\intercal \boldsymbol{x})^2$ and $\Omega(\boldsymbol{x}) = 0$, the set $\mathcal{A}_\delta$ should not contain $-\boldsymbol{x}_\star$. Therefore, in this case we must have $\sqrt{1 - \delta^2} < \delta$ or equivalently $\delta > 1/\sqrt{2}$. Throughout, we implicitly assume that such required lower bounds on δ hold. Furthermore, instead of (2.1) we could have used the tighter approximation

$$\mathcal{A} \subseteq \left\{ \boldsymbol{h} \, : \, \langle \frac{\boldsymbol{x}_\star}{\|\boldsymbol{x}_\star\|_2}, \boldsymbol{h} \rangle < \delta \, \|\boldsymbol{h}\|_2 \text{ and } \sqrt{1 - \delta^2} \, \|\boldsymbol{h}_\perp\|_2 + \langle \frac{\delta \boldsymbol{x}_\star}{\|\boldsymbol{x}_\star\|_2} - \boldsymbol{g}, \boldsymbol{h} \rangle \geq 0 \text{ for all } \boldsymbol{g} \in \partial\Omega\left(\boldsymbol{x}_\star\right) \right\}$$
$$\bigcup \left\{ \boldsymbol{h} \, : \, \langle \frac{\boldsymbol{x}_\star}{\|\boldsymbol{x}_\star\|_2}, \boldsymbol{h} \rangle \geq \delta \, \|\boldsymbol{h}\|_2 \text{ and } \|\boldsymbol{h}\|_2 - \langle \boldsymbol{g}, \boldsymbol{h} \rangle \geq 0 \text{ for all } \boldsymbol{g} \in \partial\Omega\left(\boldsymbol{x}_\star\right) \right\},$$

where $\boldsymbol{h}_\perp$ denotes part of $\boldsymbol{h}$ that is orthogonal to $\boldsymbol{x}_\star$. While using this approximation improves the dependence of our result on δ, we prefer (2.1) merely for simpler notation and derivations.

Our main theorem below provides a sample complexity for establishing the desired sufficient condition and thus accuracy of (1.4) in terms of a Rademacher complexity and a probability bound for $\nabla f(\boldsymbol{x}_\star)$ being in certain half-spaces. Let us pause here to describe these two quantities first.

For a set $\mathcal{H} \subset \mathbb{R}^N$ define the Rademacher complexity[2] with respect to $\nabla f_m(\boldsymbol{x}_\star)$'s as

$$\mathfrak{C}_M\left(\mathcal{H}\right) \overset{\text{def}}{=} \mathbb{E} \sup_{\boldsymbol{h} \in \mathcal{H}} \frac{1}{\sqrt{M}} \sum_{m=1}^{M} \epsilon_m \langle \nabla f_m\left(\boldsymbol{x}_\star\right), \frac{\boldsymbol{h}}{\|\boldsymbol{h}\|_2} \rangle, \tag{2.2}$$

where $\epsilon_1, \epsilon_2, \ldots, \epsilon_M$ are i.i.d. Rademacher random variables independent of everything else. If $\mathcal{H}$ is non-negative homogeneous (e.g., it is a convex cone), then $\mathfrak{C}_M(\mathcal{H})$ is a measure of "wideness" of $\mathcal{H}$ near the origin; the notion of wideness here depends on the law of $\nabla f(\boldsymbol{x}_\star)$. For example, in the case that $\nabla f(\boldsymbol{x}_\star)$ is a standard Gaussian random vector, the vector $M^{-1/2} \sum_{m=1}^{M} \epsilon_m \nabla f_m(\boldsymbol{x}_\star)\rangle$ is also distributed like a standard Gaussian vector and $\mathfrak{C}_M(\mathcal{H})$ reduces to the *Gaussian width* of $\mathcal{H}$ which is defined as

$$\gamma(\mathcal{H}) \overset{\text{def}}{=} \mathbb{E} \sup_{\boldsymbol{h} \in \mathcal{H}} \langle \boldsymbol{g}, \frac{\boldsymbol{h}}{\|\boldsymbol{h}\|_2} \rangle, \qquad \boldsymbol{g} \sim \text{Normal}(\boldsymbol{0}, \boldsymbol{I}).$$

The reduction of the Rademacher complexity $\mathfrak{C}_M(\mathcal{H})$ to the Gaussian width $\gamma(\mathcal{H})$ is also possible for $\nabla f(\boldsymbol{x}_\star)$ that is not Gaussian but has a sufficiently regular law: for any fixed $\boldsymbol{h}$, assuming bounded moments of sufficiently high order the random quantity $M^{-1/2} \sum_{m=1}^{M} \epsilon_m \langle \nabla f_m(\boldsymbol{x}_\star), \boldsymbol{h} \rangle$ can be approximated by a Gaussian using Berry-Esséen theorem; we only need to make

---

[2]Unlike conventional definition of Rademacher complexities, we use a normalization by square root of the number of samples.

this approximation uniform over the entire $\mathcal{H}$. While the Gaussian width of $\mathcal{H}$ can provide sharp approximations for $\mathfrak{C}_M(\mathcal{H})$, it is generally a difficult quantity to compute. In Section 2.2 we use different techniques to bound the Rademacher complexity in the special cases of unstructured and sparse regression problems.

The second quantity that affects the sample complexity of our method is a probability lower bound defined for $\mathcal{H}$ and a positive parameter τ as

$$p_\tau(\mathcal{H}) \stackrel{\text{def}}{=} \inf_{\boldsymbol{h} \in \mathcal{H}} \mathbb{P}\left(\langle \nabla f(\boldsymbol{x}_\star), \boldsymbol{h} \rangle \geq \tau \|\boldsymbol{h}\|_2\right). \tag{2.3}$$

Intuitively, $p_\tau(\mathcal{H})$ measures how "well-spread" the random vector $\nabla f(\boldsymbol{x}_\star)$ is in the space. A smaller value of $p_\tau(\mathcal{H})$ indicates that realizations of $\nabla f(\boldsymbol{x}_\star)$ are often confined to some half-space whose normal vector belongs to $\mathcal{H}$.

We now state our main theorem in terms of the quantities $\mathfrak{C}_M(\mathcal{A}_\delta)$ and $p_\tau(\mathcal{A}_\delta)$. In Sections 2.1 and 2.2 below, we show how these quantities can be bounded in a way that shows their dependence on the probability law for $\nabla f(\boldsymbol{x}_\star)$ more clearly.

Theorem 2.1. *Let $\mathcal{A}_\delta$ be defined as in (2.1) for which $\mathfrak{C}_M(\mathcal{A}_\delta)$ and $p_\tau(\mathcal{A}_\delta)$ can be determined using (2.2) and (2.3), respectively. For any $t > 0$, if*

$$M \geq 4\left(\frac{2\mathfrak{C}_M(\mathcal{A}_\delta) + t\tau}{\tau p_\tau(\mathcal{A}_\delta)}\right)^2, \tag{2.4}$$

then with probability $\geq 1 - \exp(-2t^2)$ any solution $\widehat{\boldsymbol{x}}$ of (1.4) obeys

$$\|\widehat{\boldsymbol{x}} - \boldsymbol{x}_\star\|_2 \leq \frac{2}{\tau p_\tau(\mathcal{A}_\delta)}\left(\frac{1}{M}\sum_{m=1}^{M}|\xi_m| + \varepsilon\right).$$

The factor 4 in (2.4) can be made arbitrarily close to 1 at the cost of increasing the constant factor in the error bound. Proof of Theorem 2.1, provided below in Section 4, is based on the idea of *small-ball method* introduced in [16, 23] and further developed in [17, 18, 24].

2.1 Simplifying $p_\tau(\mathcal{A}_\delta)$ and the dependence on τ

The sample complexity in the statement of Theorem 2.1 depends critically depends on the parameters τ and $p_\tau(\mathcal{A}_\delta)$. Approximating $p_\tau(\mathcal{A}_\delta)$ is challenging in general and may require detailed calculations even for specific cases (see, for example, [2, Lemmas 3 and 5] where this is done for the phase retrieval problem $f_m(\boldsymbol{x}) = |\langle \boldsymbol{x}, \boldsymbol{a}_m \rangle|^2$ with Gaussian vectors $\boldsymbol{a}_m$). In this section, we show how $p_\tau(\mathcal{A}_\delta)$ and τ can be approximated by expressions that capture the interplay between the probability distribution on the f_m and the geometry of the set $\mathcal{A}_\delta$. These may still be very difficult to calculate, but they will allow us to state more interpretable corollaries to our main theorem,

We start by observing that

$$\mathbb{P}\left(\langle \nabla f(\boldsymbol{x}_\star), \boldsymbol{h} \rangle \geq \tau \|\boldsymbol{h}\|_2\right) = \mathbb{P}\left((\langle \nabla f(\boldsymbol{x}_\star), \boldsymbol{h} \rangle)_+ \geq \tau \|\boldsymbol{h}\|_2\right),$$

9

and then define

$$\tau(\boldsymbol{h}) \stackrel{\text{def}}{=} \frac{\mathbb{E}\left(\langle \nabla f(\boldsymbol{x}_\star), \boldsymbol{h}\rangle\right)_+}{\|\boldsymbol{h}\|_2}, \quad \tau(\mathcal{A}_\delta) \stackrel{\text{def}}{=} \inf_{\boldsymbol{h} \in \mathcal{A}_\delta \backslash \{\boldsymbol{0}\}} \tau(\boldsymbol{h}). \tag{2.5}$$

Clearly, both of these quantities are non-negative. However, to ensure that $p_\tau(\mathcal{A}_\delta) > 0$ for some $\tau > 0$, it is necessary that $\tau(\mathcal{A}_\delta)$ is strictly positive. The only way this does not occur is when there is some $\boldsymbol{h} \in \mathcal{A}_\delta \backslash \{\boldsymbol{0}\}$ for which $\langle \nabla f(\boldsymbol{x}_\star), \boldsymbol{h}\rangle \leq 0$ almost surely. This means that with probability 1, $\nabla f(\boldsymbol{x}_\star)$ lies in a half-space determined whose normal vector is in $\mathcal{A}_\delta$. Qualitatively, if the distribution for $\nabla f(\boldsymbol{x}_\star)$ is well-spread, then this situation does not occur, and $\tau(\mathcal{A}_\delta)$ is positive.

By the Paley-Zygmund inequality [27][6, Corollary 3.3.2],

$$\mathbb{P}\left((\langle \nabla f(\boldsymbol{x}_\star), \boldsymbol{h}\rangle)_+ \geq \frac{1}{2}\tau(\boldsymbol{h})\|\boldsymbol{h}\|_2\right) \geq \frac{\tau^2(\boldsymbol{h})\|\boldsymbol{h}\|_2^2}{\mathbb{E}[(\langle \nabla f(\boldsymbol{x}_\star), \boldsymbol{h}\rangle)_+^2]}.$$

Since $(\langle \nabla f(\boldsymbol{x}_\star), \boldsymbol{h}\rangle)_+^2 \leq |\nabla f(\boldsymbol{x}_\star)^\mathsf{T}\boldsymbol{h}|^2 = \boldsymbol{h}^\mathsf{T}\nabla f(\boldsymbol{x}_\star)\nabla f(\boldsymbol{x}_\star)^\mathsf{T}\boldsymbol{h}$, we have for $\tau = \tau(\mathcal{A}_\delta)/2$,

$$p_\tau(\mathcal{A}_\delta) \geq \frac{\tau^2(\mathcal{A}_\delta)}{\varsigma^2(\mathcal{A}_\delta)}, \tag{2.6}$$

where

$$\varsigma^2(\mathcal{A}_\delta) \stackrel{\text{def}}{=} \sup_{\boldsymbol{h} \in \mathcal{A}_\delta \backslash \{\boldsymbol{0}\}} \frac{\boldsymbol{h}^\mathsf{T}\boldsymbol{\Sigma}_\star\boldsymbol{h}}{\|\boldsymbol{h}\|_2^2}, \quad \text{with } \boldsymbol{\Sigma}_\star = \mathbb{E}\left[\nabla f(\boldsymbol{x}_\star)\nabla f(\boldsymbol{x}_\star)^\mathsf{T}\right]. \tag{2.7}$$

This gives us the following corollary to Theorem 2.1:

Corollary 2.1. *Let $\mathcal{A}_\delta$ be defined as in (2.1), and $\mathfrak{C}_M(\mathcal{A}_\delta), \tau(\mathcal{A}_\delta), \varsigma^2(\mathcal{A}_\delta)$ as in (2.2), (2.5), and (2.7). For any $t > 0$, if*

$$M \geq \frac{16\,\varsigma^4(\mathcal{A}_\delta)}{\tau^4(\mathcal{A}_\delta)}\left(\frac{2\mathfrak{C}_M(\mathcal{A}_\delta)}{\tau(\mathcal{A}_\delta)} + \frac{t}{2}\right)^2, \tag{2.8}$$

then with probability $\geq 1 - 2\exp(-2t^2)$, any solution $\widehat{\boldsymbol{x}}$ of (1.4) obeys

$$\|\widehat{\boldsymbol{x}} - \boldsymbol{x}_\star\|_2 \leq \left(\frac{4\,\varsigma^2(\mathcal{A}_\delta)}{\tau^3(\mathcal{A}_\delta)}\right)\left(\frac{1}{M}\sum_{m=1}^{M}|\xi_m| + \varepsilon\right). \tag{2.9}$$

A simple dimensional analysis can help us understand how the bound above scale qualitatively. The entries in the gradient vector $\nabla f(\boldsymbol{x})$ represent quantities in "units of f" divided by "units of $\boldsymbol{x}$". The constants $\varsigma(\mathcal{A}_\delta), \tau(\mathcal{A}_\delta)$, and $\mathfrak{C}_M(\mathcal{A}_\delta)$ all have the same units as $\nabla f(\boldsymbol{x})$. Thus the lower bound on the number of equations M in (2.8) is dimensionless, as we expect. As the noise variables ξ_m are also in "units of f", both sides of the inequality in (2.9) are in "units of $\boldsymbol{x}$".

2.2 Bounding the Rademacher complexity $\mathfrak{C}_M(\mathcal{A}_\delta)$

To make the result of Theorem 2.1 more explicit we consider the special cases of unstructured regression (i.e., $\boldsymbol{x}_\star$ is arbitrary and $\Omega(\boldsymbol{x}) = 0$) and sparse regression (i.e., $\boldsymbol{x}_\star$ is sparse and $\Omega(\boldsymbol{x}) = \lambda\|\boldsymbol{x}\|_1$ for some $\lambda > 0$). The specific choice of $\Omega(\boldsymbol{x})$ helps to simplify the Rademacher complexity $\mathfrak{C}_M(\mathcal{A}_\delta)$. Of course, $\mathfrak{C}_M(\mathcal{A}_\delta)$ would depend on the law of $f_1, f_2, \ldots, f_M$ as well.

2.2.1 Unstructured regression

In the first example, we approximate $\mathfrak{C}_M(\mathcal{A}_\delta)$ when no regularization is applied, i.e., $\Omega(\boldsymbol{x}) = 0$. In this case, (2.1) reduces to

$$\mathcal{A}_\delta = \left\{ \boldsymbol{h} \; : \; \sqrt{1-\delta^2}\,\|\boldsymbol{h}\|_2 + \langle \frac{\delta \boldsymbol{x}_\star}{\|\boldsymbol{x}_\star\|_2}, \boldsymbol{h} \rangle \geq 0 \right\},$$

which we approximate by the entire space (i.e., $\mathcal{A}_\delta \subseteq \mathbb{R}^N$). Because $\mathfrak{C}_M(\cdot)$ is monotonic, we obtain

$$\begin{aligned}
\mathfrak{C}_M(\mathcal{A}_\delta) &\leq \mathfrak{C}_M\left(\mathbb{R}^N\right) \\
&= \mathbb{E}\sup_{\boldsymbol{h}} \frac{1}{\sqrt{M}} \sum_{m=1}^{M} \epsilon_m \langle \nabla f_m(\boldsymbol{x}_\star), \frac{\boldsymbol{h}}{\|\boldsymbol{h}\|_2} \rangle \\
&= \mathbb{E}\left\| \frac{1}{\sqrt{M}} \sum_{m=1}^{M} \epsilon_m \nabla f_m(\boldsymbol{x}_\star) \right\|_2 .
\end{aligned}$$

Applying the Cauchy-Schwarz for the square-root function which is concave, we deduce that

$$\begin{aligned}
\mathfrak{C}_M(\mathcal{A}_\delta) &\leq \sqrt{\mathbb{E}\left\| \frac{1}{\sqrt{M}} \sum_{m=1}^{M} \epsilon_m \nabla f_m(\boldsymbol{x}_\star) \right\|_2^2} \\
&= \sqrt{\frac{1}{M} \mathbb{E} \sum_{m=1}^{M} \|\nabla f_m(\boldsymbol{x}_\star)\|_2^2} \\
&= \sqrt{\mathbb{E}\|\nabla f(\boldsymbol{x}_\star)\|_2^2},
\end{aligned}$$

where the second and third lines respectively hold because ϵ_ms are independent zero-mean random variables and f_ms are i.i.d. copies of f. The following corollary of Theorem 2.1 is the immediate result of the bound above.

Corollary 2.2. *Let $\mathcal{A}_\delta$ and $p_\tau(\mathcal{A}_\delta)$ be defined by (2.1) and (2.3), respectively. For any $t > 0$, if*

$$M \geq 4 \left(\frac{2\sqrt{\mathbb{E}\|\nabla f(\boldsymbol{x}_\star)\|_2^2} + t\tau}{\tau p_\tau(\mathcal{A}_\delta)} \right)^2, \tag{2.10}$$

then with probability $\geq 1 - \exp(-2t^2)$ any solution $\hat{\boldsymbol{x}}$ of (1.4) obeys

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}_\star\|_2 \leq \frac{2}{\tau p_\tau(\mathcal{A}_\delta)} \left(\frac{1}{M} \sum_{m=1}^{M} |\xi_m| + \varepsilon \right).$$

Applying the bounds (2.6) and (2.7) from the last section, (2.10) becomes

$$M \;\geq\; 4\frac{\|\boldsymbol{\Sigma}_\star\|^2}{\tau^4(\mathcal{A}_\delta)} \left(\frac{4\sqrt{\text{tr}(\boldsymbol{\Sigma}_\star)}}{\tau(\mathcal{A}_\delta)} + t \right)^2,$$

11

where again $\boldsymbol{\Sigma}_\star$ is the correlation matrix of the gradient of f_m, and we have applied the bound $\varsigma^2(\mathcal{A}_\delta) \leq \|\boldsymbol{\Sigma}_\star\|$. Since $\operatorname{tr}(\boldsymbol{\Sigma}_\star) \leq N\|\boldsymbol{\Sigma}_\star\|$, it suffices to take

$$M \geq 4\frac{\|\boldsymbol{\Sigma}_\star\|^2}{\tau^4(\mathcal{A}_\delta)}\left(4\sqrt{N}\,\frac{\sqrt{\|\boldsymbol{\Sigma}_\star\|}}{\tau(\mathcal{A}_\delta)} + t\right)^2.$$

When $\sqrt{\|\boldsymbol{\Sigma}_\star\|}/\tau(\mathcal{A}_\delta)$ is on the order of a constant, this means we can take $M \gtrsim N$. Provided we have an anchor vector that obeys (1.3), we can robustly recover a vector $\boldsymbol{x}_\star$ of length N through observations of slightly more than N convex functions of $\boldsymbol{x}_\star$

2.2.2 Sparse regression

Next, we approximate $\mathfrak{C}_M(\mathcal{A}_\delta)$ for the case of ℓ_1-regularized anchored regression, i.e., $\Omega(\boldsymbol{x}) = \lambda\|\boldsymbol{x}\|_1$ for some $\lambda > 0$. Let $\mathcal{S}_\star$ denote the support set of $\boldsymbol{x}_\star$ (i.e., $\mathcal{S}_\star = \{i \in [1, N] : x_{\star i} \neq 0\}$) and $s = |\mathcal{S}_\star|$. The subdifferential of the $\Omega(\cdot)$ at $\boldsymbol{x}_\star$ can be expressed as

$$\partial\Omega(\boldsymbol{x}_\star) = \left\{\boldsymbol{g} : \|\boldsymbol{g}\|_\infty \leq \lambda \text{ and } \boldsymbol{g}|_{\mathcal{S}_\star} = \lambda\operatorname{sgn}\left(\boldsymbol{x}_\star|_{\mathcal{S}_\star}\right)\right\}.$$

Therefore, we have

$$\mathcal{A}_\delta = \left\{\boldsymbol{h} : \sqrt{1-\delta^2}\,\|\boldsymbol{h}\|_2 - \lambda\left\|\boldsymbol{h}|_{\mathcal{S}_\star^c}\right\|_1 + \langle\frac{\delta\boldsymbol{x}_\star|_{\mathcal{S}_\star}}{\|\boldsymbol{x}_\star\|_2} - \lambda\operatorname{sgn}\left(\boldsymbol{x}_\star|_{\mathcal{S}_\star}\right), \boldsymbol{h}|_{\mathcal{S}_\star}\rangle \geq 0\right\}.$$

It follows from the Cauchy-Schwarz inequality that

$$\langle\frac{\delta\boldsymbol{x}_\star|_{\mathcal{S}_\star}}{\|\boldsymbol{x}_\star\|_2} - \lambda\operatorname{sgn}\left(\boldsymbol{x}_\star|_{\mathcal{S}_\star}\right), \boldsymbol{h}|_{\mathcal{S}_\star}\rangle \leq \left\|\frac{\delta\boldsymbol{x}_\star|_{\mathcal{S}_\star}}{\|\boldsymbol{x}_\star\|_2} - \lambda\operatorname{sgn}\left(\boldsymbol{x}_\star|_{\mathcal{S}_\star}\right)\right\|_2 \left\|\boldsymbol{h}|_{\mathcal{S}_\star}\right\|_2$$

$$= \left\|\frac{\delta\boldsymbol{x}_\star}{\|\boldsymbol{x}_\star\|_2} - \lambda\operatorname{sgn}\left(\boldsymbol{x}_\star\right)\right\|_2 \left\|\boldsymbol{h}|_{\mathcal{S}_\star}\right\|_2,$$

thereby

$$\mathcal{A}_\delta \subseteq \mathcal{A}_\delta' \stackrel{\text{def}}{=} \left\{\boldsymbol{h} : \sqrt{1-\delta^2}\,\|\boldsymbol{h}\|_2 - \lambda\left\|\boldsymbol{h}|_{\mathcal{S}_\star^c}\right\|_1 + \left\|\frac{\delta\boldsymbol{x}_\star}{\|\boldsymbol{x}_\star\|_2} - \lambda\operatorname{sgn}\left(\boldsymbol{x}_\star\right)\right\|_2 \left\|\boldsymbol{h}|_{\mathcal{S}_\star}\right\|_2 \geq 0\right\}.$$

For any non-zero $\boldsymbol{h}$ and any $\boldsymbol{z} \in \mathbb{R}^N$, it follows from Hölder's inequality that

$$\langle\boldsymbol{z}, \frac{\boldsymbol{h}}{\|\boldsymbol{h}\|_2}\rangle \leq \left\|\boldsymbol{z}|_{\mathcal{S}_\star}\right\|_2 \frac{\left\|\boldsymbol{h}|_{\mathcal{S}_\star}\right\|_2}{\|\boldsymbol{h}\|_2} + \left\|\boldsymbol{z}|_{\mathcal{S}_\star^c}\right\|_\infty \frac{\left\|\boldsymbol{h}|_{\mathcal{S}_\star^c}\right\|_1}{\|\boldsymbol{h}\|_2}. \tag{2.11}$$

Therefore, recalling the definition (2.2), we can write

$$\mathfrak{C}_M(\mathcal{A}_\delta) \leq \mathfrak{C}_M(\mathcal{A}_\delta')$$

$$= \mathbb{E}\sup_{\boldsymbol{h}\in\mathcal{A}_\delta'}\frac{1}{\sqrt{M}}\sum_{m=1}^M \epsilon_m\langle\nabla f_m(\boldsymbol{x}_\star), \frac{\boldsymbol{h}}{\|\boldsymbol{h}\|_2}\rangle$$

$$\leq \mathbb{E}\sup_{\boldsymbol{h}\in\mathcal{A}_\delta'}\left\|\frac{1}{\sqrt{M}}\sum_{m=1}^M \epsilon_m\nabla f_m(\boldsymbol{x}_\star)|_{\mathcal{S}_\star}\right\|_2 \frac{\left\|\boldsymbol{h}|_{\mathcal{S}_\star}\right\|_2}{\|\boldsymbol{h}\|_2}$$

$$+ \mathbb{E}\sup_{\boldsymbol{h}\in\mathcal{A}_\delta'}\left\|\frac{1}{\sqrt{M}}\sum_{m=1}^M \epsilon_m\nabla f_m(\boldsymbol{x}_\star)|_{\mathcal{S}_\star^c}\right\|_\infty \frac{\left\|\boldsymbol{h}|_{\mathcal{S}_\star^c}\right\|_1}{\|\boldsymbol{h}\|_2}$$

12

where the first inequality follows from the fact that $\mathcal{A}_\delta \subseteq \mathcal{A}'_\delta$ and the second inequality is an application of (2.11) for $\boldsymbol{z} = \frac{1}{\sqrt{M}} \sum_{m=1}^{M} \epsilon_m \nabla f_m(\boldsymbol{x}_\star)$ followed by a triangle inequality. We always have $\frac{\left\| \boldsymbol{h}|_{\mathcal{S}_\star} \right\|_2}{\|\boldsymbol{h}\|_2} \leq 1$ and the definition of $\mathcal{A}'_\delta$ implies that

$$
\begin{aligned}
\frac{\left\| \boldsymbol{h}|_{\mathcal{S}_\star^{\mathrm{c}}} \right\|_1}{\|\boldsymbol{h}\|_2} &\leq \frac{\sqrt{1 - \delta^2}}{\lambda} + \frac{1}{\lambda} \left\| \frac{\delta \boldsymbol{x}_\star}{\|\boldsymbol{x}_\star\|_2} - \lambda \operatorname{sgn}(\boldsymbol{x}_\star) \right\|_2 \frac{\left\| \boldsymbol{h}|_{\mathcal{S}_\star} \right\|_2}{\|\boldsymbol{h}\|_2} \\
&\leq \frac{\sqrt{1 - \delta^2}}{\lambda} + \frac{1}{\lambda} \left\| \frac{\delta \boldsymbol{x}_\star}{\|\boldsymbol{x}_\star\|_2} - \lambda \operatorname{sgn}(\boldsymbol{x}_\star) \right\|_2.
\end{aligned}
$$

Hence, we deduce that

$$
\begin{aligned}
\mathfrak{C}_M(\mathcal{A}_\delta) \leq{}& \mathbb{E} \left\| \frac{1}{\sqrt{M}} \sum_{m=1}^{M} \epsilon_m \nabla f_m(\boldsymbol{x}_\star)|_{\mathcal{S}_\star} \right\|_2 \\
&+ \left(\frac{\sqrt{1 - \delta^2}}{\lambda} + \frac{1}{\lambda} \left\| \frac{\delta \boldsymbol{x}_\star}{\|\boldsymbol{x}_\star\|_2} - \lambda \operatorname{sgn}(\boldsymbol{x}_\star) \right\|_2 \right) \mathbb{E} \left\| \frac{1}{\sqrt{M}} \sum_{m=1}^{M} \epsilon_m \nabla f_m(\boldsymbol{x}_\star)|_{\mathcal{S}_\star^{\mathrm{c}}} \right\|_\infty.
\end{aligned}
$$

Next, we bound the terms with the expectation. For the first term, the Cauchy-Schwarz inequality and the fact that ϵ_ms are independent and zero-mean guarantees that

$$
\begin{aligned}
\mathbb{E} \left\| \frac{1}{\sqrt{M}} \sum_{m=1}^{M} \epsilon_m \nabla f_m(\boldsymbol{x}_\star)|_{\mathcal{S}_\star} \right\|_2 &\leq \sqrt{\mathbb{E} \left\| \frac{1}{\sqrt{M}} \sum_{m=1}^{M} \epsilon_m \nabla f_m(\boldsymbol{x}_\star)|_{\mathcal{S}_\star} \right\|_2^2} \\
&= \sqrt{\frac{1}{M} \sum_{m=1}^{M} \mathbb{E} \left\| \nabla f_m(\boldsymbol{x}_\star)|_{\mathcal{S}_\star} \right\|_2^2} \\
&= \sqrt{\mathbb{E} \left\| \nabla f(\boldsymbol{x}_\star)|_{\mathcal{S}_\star} \right\|_2^2}.
\end{aligned}
$$

For the second term, a similar application of the Cauchy-Schwarz inequality, followed by the Nemirovski's inequality (see [25, Lemma 5.2.2] and [7, Theorem 2.2]) with the minor assumption that $N - s \geq 3$, yield

$$
\begin{aligned}
\mathbb{E} \left\| \frac{1}{\sqrt{M}} \sum_{m=1}^{M} \epsilon_m \nabla f_m(\boldsymbol{x}_\star)|_{\mathcal{S}_\star^{\mathrm{c}}} \right\|_\infty &\leq \sqrt{\mathbb{E} \left\| \frac{1}{\sqrt{M}} \sum_{m=1}^{M} \epsilon_m \nabla f_m(\boldsymbol{x}_\star)|_{\mathcal{S}_\star^{\mathrm{c}}} \right\|_\infty^2} \\
&\leq \sqrt{2e \log(N - s) \cdot \frac{1}{M} \sum_{m=1}^{M} \mathbb{E} \left\| \nabla f_m(\boldsymbol{x}_\star)|_{\mathcal{S}_\star^{\mathrm{c}}} \right\|_\infty^2} \\
&\leq \sqrt{2e \log N \, \mathbb{E} \left\| \nabla f(\boldsymbol{x}_\star)|_{\mathcal{S}_\star^{\mathrm{c}}} \right\|_\infty^2}.
\end{aligned}
$$

Therefore, we obtain the following corollary of Theorem 2.1.

Corollary 2.3. *Let $\mathcal{A}_\delta$ and $p_\tau(\mathcal{A}_\delta)$ be defined by (2.1) and (2.3), respectively. Furthermore, let $C^\star_{\delta,\lambda} = 2\sqrt{2e}\left(\frac{\sqrt{1-\delta^2}}{\lambda} + \frac{1}{\lambda}\left\|\frac{\delta \boldsymbol{x}_\star}{\|\boldsymbol{x}_\star\|_2} - \lambda \operatorname{sgn}(\boldsymbol{x}_\star)\right\|_2\right)$. For any $t > 0$, if*

$$M \geq 4\left(\frac{2\sqrt{\mathbb{E}\left\|\nabla f(\boldsymbol{x}_\star)|_{\mathcal{S}_\star}\right\|_2^2} + C^\star_{\delta,\lambda}\sqrt{\log(N)\,\mathbb{E}\left\|\nabla f(\boldsymbol{x}_\star)|_{\mathcal{S}_\star^c}\right\|_\infty^2} + t\tau}{\tau p_\tau(\mathcal{A}_\delta)}\right)^2,$$

then with probability $\geq 1 - \exp(-2t^2)$ any solution $\widehat{\boldsymbol{x}}$ of (1.4) obeys

$$\|\widehat{\boldsymbol{x}} - \boldsymbol{x}_\star\|_2 \leq \frac{2}{\tau p_\tau(\mathcal{A}_\delta)}\left(\frac{1}{M}\sum_{m=1}^{M}|\xi_m| + \varepsilon\right).$$

Let us consider the particular case of regression with linear predictors to make the result of Corollary 2.3 more explicit.

Example 2.1 (Sparse regression with linear predictor). Suppose that $f(\boldsymbol{x}) = \phi(\boldsymbol{a}^\mathsf{T}\boldsymbol{x})$ for $\boldsymbol{a} \sim \operatorname{Normal}(\boldsymbol{0}, \boldsymbol{I})$ and a convex $\phi(\cdot)$ that has derivatives of sufficiently high order. Moreover, in the calculations below we implicitly assume that the expectations that involve $\phi(\cdot)$ or its derivatives exist. Clearly, we can write $\nabla f(\boldsymbol{x}_\star) = \phi'(\boldsymbol{a}^\mathsf{T}\boldsymbol{x}_\star)\boldsymbol{a}$ and obtain

$$\begin{aligned}
\mathbb{E}\left\|\nabla f(\boldsymbol{x}_\star)|_{\mathcal{S}_\star}\right\|_2^2 &= \mathbb{E}\left(\phi'^2(\boldsymbol{a}^\mathsf{T}\boldsymbol{x}_\star)\left\|\boldsymbol{a}|_{\mathcal{S}_\star}\right\|_2^2\right)\\
&= \mathbb{E}\left(\phi'^2(\|\boldsymbol{x}_\star\|_2\,g)(g^2 + |\mathcal{S}_\star| - 1)\right), \qquad g \sim \operatorname{Normal}(0,1)\\
&= s\,\mathbb{E}\left(\phi'^2(\|\boldsymbol{x}_\star\|_2\,g)\right)\\
&\quad + 2\mathbb{E}\left(\phi''^2(\|\boldsymbol{x}_\star\|_2\,g) + \phi'(\|\boldsymbol{x}_\star\|_2\,g)\phi'''(\|\boldsymbol{x}_\star\|_2\,g)\right)\|\boldsymbol{x}_\star\|_2^2.
\end{aligned}$$

where the second equation follows from independence of projections of $\boldsymbol{a}|_{\mathcal{S}_\star}$ onto subspaces parallel and orthogonal to $\boldsymbol{x}_\star$. The third equation follows by applying the Stein's lemma [31] twice. Furthermore, we have

$$\begin{aligned}
\mathbb{E}\left\|\nabla f(\boldsymbol{x}_\star)|_{\mathcal{S}_\star^c}\right\|_\infty^2 &= \mathbb{E}\left(\phi'^2(\boldsymbol{a}^\mathsf{T}\boldsymbol{x}_\star)\max_{i\in\mathcal{S}_\star^c}a_i^2\right)\\
&= \mathbb{E}\left(\phi'^2(\boldsymbol{a}^\mathsf{T}\boldsymbol{x}_\star)\right)\mathbb{E}\left(\max_{i\in\mathcal{S}_\star^c}a_i^2\right)\\
&\leq \mathbb{E}\left(\phi'^2(\|\boldsymbol{x}_\star\|_2\,g)\right)\mathbb{E}\left(\sum_{i\in\mathcal{S}_\star^c}a_i^{2q}\right)^{\frac{1}{q}}, \qquad g \sim \operatorname{Normal}(0,1)\\
&\leq \mathbb{E}\left(\phi'^2(\|\boldsymbol{x}_\star\|_2\,g)\right)\left((N-s)\mathbb{E}\left(g^{2q}\right)\right)^{\frac{1}{q}},
\end{aligned}$$

where the second line holds by independence of $\boldsymbol{a}^\mathsf{T}\boldsymbol{x}_\star$ and a_i for $i \in \mathcal{S}_\star^c$, the third line holds for any positive integer q, and the fourth line follows from concavity of $t \mapsto t^{1/q}$ and the Jensen's inequality. Hence, for $q = O(\log N)$ we obtain

$$\mathbb{E}\left\|\nabla f(\boldsymbol{x}_\star)|_{\mathcal{S}_\star^c}\right\|_\infty^2 \lesssim \mathbb{E}\left(\phi'^2(\|\boldsymbol{x}_\star\|_2\,g)\right)\log N.$$

Treating the terms that only depend on $\|\boldsymbol{x}_\star\|_2$ as constants we can summarize the derived bounds as

$$\mathbb{E}\left\|\nabla f\left(\boldsymbol{x}_\star\right)|_{\mathcal{S}_\star}\right\|_2^2 \lesssim s\,, \qquad \text{and} \qquad \mathbb{E}\left\|\nabla f\left(\boldsymbol{x}_\star\right)|_{\mathcal{S}_\star^c}\right\|_\infty^2 \lesssim \log N\,.$$

If we choose the regularization parameter as $\lambda = O(1/\sqrt{s})$ for a suitable constant factor that depends on δ, then we obtain

$$C_{\delta,\lambda}^\star \le c_\delta \sqrt{s}\,,$$

where c_δ is a constant decreasing in δ. Therefore, we can conclude that having

$$M \gtrsim \left(\frac{\sqrt{s} + c_\delta \sqrt{s} \log\left(N\right) + t\tau}{\tau p_\tau\left(\mathcal{A}_\delta\right)}\right)^2$$

is sufficient to meet the requirements in Corollary 2.3. This result suggests a sample complexity of $O\left(s \log^2 N\right)$ which is optimal up to the dependence on $\log N$. A more refined bound can be obtained using a tighter bound for $\mathbb{E}\left\|\frac{1}{\sqrt{M}} \sum_{m=1}^M \epsilon_m \nabla f_m\left(\boldsymbol{x}_\star\right)|_{\mathcal{S}_\star^c}\right\|_\infty$ in the proof of Corollary 2.3.

3 Recipes for creating anchors

While it may be assumed that the anchor vector is provided by an oracle, it is more realistic to have a data-driven method to construct the anchor vector. This requires us to impose new assumptions, albeit implicitly, on the class of functions $\mathcal{F}$ where the random samples $f_1, f_2, \ldots, f_M$ are drawn from. A natural assumption is that for a certain sample loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ the corresponding *risk* $R\left(\boldsymbol{x}\right) \stackrel{\text{def}}{=} \mathbb{E}\ell\left(f\left(\boldsymbol{x}\right), f\left(\boldsymbol{x}_\star\right)\right)$ "encodes" the information about $\boldsymbol{x}_\star$ in its derivatives at some reference point $\boldsymbol{x}_0$, i.e., $\nabla^k R\left(\boldsymbol{x}_0\right)$ for $k = 1, 2, 3, \ldots$. For simplicity, henceforth we take the origin as the reference point, i.e., $\boldsymbol{x}_0 = \boldsymbol{0}$. This assumption is reasonable, for instance, if $R\left(\boldsymbol{x}\right)$ has a Taylor series approximation around the origin. Note that we are implicitly assuming that the random functions drawn from the class $\mathcal{F}$ as well as the sample loss $\ell\left(\cdot, \cdot\right)$ are k-times differentiable. Furthermore, to allow the derivative operator and expectation to commute, the required regularity conditions for the law of f are assumed.

Definition 3.1 (Spiked Derivatives). We say that the risk $R\left(\boldsymbol{x}\right) = \mathbb{E}\ell\left(f\left(\boldsymbol{x}\right), f\left(\boldsymbol{x}_\star\right)\right)$ has a *spiked kth-order derivative* if $\boldsymbol{x}_\star$ is a *simple principal eigenvector* of $-\nabla^k R\left(\boldsymbol{0}\right) \ne \boldsymbol{0}$. Namely,

$$\frac{\boldsymbol{x}_\star}{\|\boldsymbol{x}_\star\|_2} \in \underset{\boldsymbol{u} \in \mathbb{S}^{N-1}}{\operatorname{argmax}} \left\langle -\nabla^k R\left(\boldsymbol{0}\right), \boldsymbol{u}^{\otimes k}\right\rangle, \tag{3.1}$$

and every other maximizer is parallel to $\boldsymbol{x}_\star$.

For computational considerations, we focus only on risks with *spiked gradient* (i.e., 1st-order derivative) or *spiked Hessian* (i.e., 2nd-order derivative). In the case of spiked gradients the condition (3.1) reduces to $-\nabla R\left(\boldsymbol{0}\right)$ being perfectly aligned with $\boldsymbol{x}_\star$, that is,

$$\frac{\boldsymbol{x}_\star}{\|\boldsymbol{x}_\star\|_2} = -\frac{\nabla R\left(\boldsymbol{0}\right)}{\|\nabla R\left(\boldsymbol{0}\right)\|_2}. \tag{3.2}$$

15

Similarly, for spiked Hessians, the condition (3.1) reduces to $\boldsymbol{x}_\star$ being a *simple principal eigenvector* of $-\nabla^2 R(\mathbf{0})$. Specifically, we have

$$
\begin{aligned}
\frac{\boldsymbol{x}_\star}{\|\boldsymbol{x}_\star\|_2} &\in \underset{\boldsymbol{u}\in\mathbb{S}^{N-1}}{\operatorname{argmax}} \ -\boldsymbol{u}^\mathsf{T}\nabla^2 R(\mathbf{0})\,\boldsymbol{u} \\
\gamma_\star &\overset{\text{def}}{=} \lambda_1\left(-\nabla^2 R(\mathbf{0})\right) - \lambda_2\left(-\nabla^2 R(\mathbf{0})\right) > 0
\end{aligned}
\tag{3.3}
$$

where $\lambda_i(\cdot)$ denotes the ith largest (multiplicity inclusive) eigenvalue of its argument. For example, with the squared error $\ell(s,t) = \frac{1}{2}(s-t)^2$ as the sample loss function, if the risk $R(\boldsymbol{x}) = \mathbb{E}\,\frac{1}{2}(f(\boldsymbol{x}) - f(\boldsymbol{x}_\star))^2$ has a spiked gradient then

$$
\frac{\boldsymbol{x}_\star}{\|\boldsymbol{x}_\star\|_2} = \frac{1}{\|\mathbb{E}\left((f(\boldsymbol{x}_\star) - f(\mathbf{0}))\nabla f(\mathbf{0})\right)\|_2}\mathbb{E}\left((f(\boldsymbol{x}_\star) - f(\mathbf{0}))\nabla f(\mathbf{0})\right).
$$

and if it has spiked Hessian then

$$
\frac{\boldsymbol{x}_\star}{\|\boldsymbol{x}_\star\|_2} \in \underset{\boldsymbol{u}\in\mathbb{S}^{N-1}}{\operatorname{argmax}} \ \boldsymbol{u}^\mathsf{T}\mathbb{E}\left(f(\boldsymbol{x}_\star)\nabla^2 f(\mathbf{0}) - \nabla f(\mathbf{0})\nabla^\mathsf{T} f(\mathbf{0})\right)\boldsymbol{u},
$$

and the Hessian has a positive spectral gap.

Let us again consider the problem of regression with linear predictors as a concrete example.

Example 3.1 (Spiked derivatives in regression with linear predictor). Let $f(\boldsymbol{x}) = \phi(\boldsymbol{a}^\mathsf{T}\boldsymbol{x})$ for $\boldsymbol{a} \sim \mathrm{Normal}(\mathbf{0}, \boldsymbol{I})$ and a twice-differentiable convex function $\phi(\cdot)$. Without loss of generality we may assume that $\phi(0) = 0$. Suppose that we observe M i.i.d. samples of $f(\boldsymbol{x}_\star)$, i.e., $y_m = f_m(\boldsymbol{x}_\star) = \phi(\boldsymbol{a}_m^\mathsf{T}\boldsymbol{x}_\star)$ for $m = 1, 2, \ldots, M$ with $\boldsymbol{a}_m$s being i.i.d. copies of $\boldsymbol{a}$. The risk with respect to the squared loss $\ell(u,v) = \frac{1}{2}(u-v)^2$ is

$$
R(\boldsymbol{x}) = \mathbb{E}\left(\phi(\boldsymbol{a}^\mathsf{T}\boldsymbol{x}) - \phi(\boldsymbol{a}^\mathsf{T}\boldsymbol{x}_\star)\right)^2.
$$

Therefore, we can calculate $\nabla R(\mathbf{0})$ as

$$
\begin{aligned}
\nabla R(\mathbf{0}) &= \mathbb{E}\left(\phi'(0)\left(\phi(0) - \phi(\boldsymbol{a}^\mathsf{T}\boldsymbol{x}_\star)\right)\boldsymbol{a}\right) \\
&= -\phi'(0)\mathbb{E}\left(\phi(\boldsymbol{a}^\mathsf{T}\boldsymbol{x}_\star)\boldsymbol{a}\right) \\
&= -\phi'(0)\mathbb{E}\phi'(\|\boldsymbol{x}\|_2\,g)\boldsymbol{x}_\star && , g \sim \mathrm{Normal}(0, 1),
\end{aligned}
$$

where the third line follows from Stein's lemma. Similarly, we can write

$$
\begin{aligned}
\nabla^2 R(\mathbf{0}) &= \mathbb{E}\left(\left(\phi'^2(0) - \phi''(0)\phi(\boldsymbol{a}^\mathsf{T}\boldsymbol{x}_\star)\right)\boldsymbol{a}\boldsymbol{a}^\mathsf{T}\right) \\
&= \left(\phi'^2(0) - \phi''(0)\mathbb{E}\phi(\boldsymbol{a}^\mathsf{T}\boldsymbol{x}_\star)\right)\boldsymbol{I} - \phi''(0)\mathbb{E}\phi''(\boldsymbol{a}^\mathsf{T}\boldsymbol{x}_\star)\boldsymbol{x}_\star\boldsymbol{x}_\star^\mathsf{T} \\
&= \left(\phi'^2(0) - \phi''(0)\mathbb{E}\phi(\|\boldsymbol{x}_\star\|_2\,g)\right)\boldsymbol{I} - \phi''(0)\mathbb{E}\phi''(\|\boldsymbol{x}_\star\|_2\,g)\boldsymbol{x}_\star\boldsymbol{x}_\star^\mathsf{T}.
\end{aligned}
$$

If $\phi'(0) \neq 0$ then $R(\cdot)$ clearly has a spiked gradient. If $\phi'(0) = 0$, however, $\nabla R(\mathbf{0}) = \mathbf{0}$ and we need to inspect the second derivative. In that case, if $\phi''(0) \neq 0$ then we have

$$
\nabla^2 R(\mathbf{0}) = -\phi''(0)\mathbb{E}\phi''(\|\boldsymbol{x}_\star\|_2\,g)\boldsymbol{x}_\star\boldsymbol{x}_\star^\mathsf{T} - \phi''(0)\mathbb{E}\phi(\|\boldsymbol{x}_\star\|_2\,g)\boldsymbol{I}.
$$

Because $\phi(\cdot)$ is convex, $\phi''(\cdot)$ is non-negative and the matrix above is clearly a spiked Hessian with spectral gap $\gamma_\star = \phi''(0) \|\boldsymbol{x}_\star\|_2^2 \mathbb{E}\phi''(\|\boldsymbol{x}_\star\|_2 g)$. By a simple change of variable the derivations above can be extended to the case where the Gaussian random vector $\boldsymbol{a}$ is still zero-mean, but has an arbitrary covariance matrix.

Of course, in practice we do not have access to $R(\boldsymbol{x})$. With finite number of observations, however, the empirical risk

$$R_M(\boldsymbol{x}) = \frac{1}{M} \sum_{m=1}^{M} \ell\left(f_m(\boldsymbol{x}), y_m\right),$$

may provide a sufficiently good approximation for $R(\boldsymbol{x})$. Therefore, with sufficient number of observations, $\nabla R_M(\boldsymbol{0})$, as an approximation to $\nabla R(\boldsymbol{0})$, or $\nabla^2 R_M(\boldsymbol{0})$, as an approximation to $\nabla^2 R(\boldsymbol{0})$, can be used to find an anchor vector under the spiked gradient or spiked Hessian conditions, respectively.

In case of the spiked gradient, it suffices to have sufficiently large number of samples (i.e., M) such that

$$\boldsymbol{a}_0 = -\frac{\nabla R_M(\boldsymbol{0})}{\|\nabla R_M(\boldsymbol{0})\|_2},$$

obeys anchor vectors' required property (1.3). Suppose that the law of f is such that $\nabla R_M(\boldsymbol{0})$ concentrates around $\nabla R(\boldsymbol{0})$, namely

$$\|\nabla R_M(\boldsymbol{0}) - \nabla R(\boldsymbol{0})\|_2 \leq \epsilon \|\nabla R(\boldsymbol{0})\|_2,$$

with high probability for a small $\epsilon \in (0, 1)$. If $R(\cdot)$ has a spiked gradient, then obtain

$$\begin{aligned}
\langle \boldsymbol{a}_0, \frac{\boldsymbol{x}_\star}{\|\boldsymbol{x}_\star\|_2}\rangle &= \langle \frac{\nabla R_M(\boldsymbol{0})}{\|\nabla R_M(\boldsymbol{0})\|_2}, \frac{\nabla R(\boldsymbol{0})}{\|\nabla R(\boldsymbol{0})\|_2}\rangle \\
&\geq \frac{(1 - \epsilon^2)\|\nabla R(\boldsymbol{0})\|_2^2 + \|\nabla R_M(\boldsymbol{0})\|_2^2}{2\|\nabla R_M(\boldsymbol{0})\|_2 \|\nabla R(\boldsymbol{0})\|_2} \\
&\geq \sqrt{1 - \epsilon^2},
\end{aligned}$$

where the third line follows from the AM-GM inequality. Therefore, $\boldsymbol{a}_0$ would satisfy (1.3) with $\delta = \sqrt{1 - \epsilon^2}$.

In the case that $R(\cdot)$ has a spiked Hessian, using a variant of the Davis–Kahan's theorem [34, Corollary 3], we can show that for

$$\boldsymbol{a}_0 \in \underset{\boldsymbol{u} \in \mathbb{S}^{N-1}}{\operatorname{argmax}} \; -\boldsymbol{u}^\mathsf{T} \nabla^2 R_M(\boldsymbol{0})\, \boldsymbol{u}, \tag{3.4}$$

we have

$$\left\| \boldsymbol{a}_0 \boldsymbol{a}_0^\mathsf{T} - \frac{1}{\|\boldsymbol{x}_\star\|_2^2} \boldsymbol{x}_\star \boldsymbol{x}_\star^\mathsf{T} \right\| \leq \frac{2\left\|\nabla^2 R_M(\boldsymbol{0}) - \nabla^2 R(\boldsymbol{0})\right\|}{\gamma_\star},$$

where $\gamma_\star$ is the spectral gap defined in (3.3). Therefore, if $\|\nabla^2 R_M(\boldsymbol{0}) - \nabla^2 R(\boldsymbol{0})\|$ is sufficiently small relative to the spectral gap $\gamma_\star$, the inequality above implies $\boldsymbol{a}_0$ (or $-\boldsymbol{a}_0$) can obey the required property (1.3) for some $\delta > 0$. Depending on the law of f, we can bound $\|\nabla^2 R_M(\boldsymbol{0}) - \nabla^2 R(\boldsymbol{0})\|$ using matrix concentration inequalities as $\nabla^2 R_M(\boldsymbol{0})$ can be written as a sum of independent random matrices.

3.1 Anchors for structured regression

How does construction of the anchor change if we need to estimate a structured ground truth (e.g., a sparse vector)? In this scenario, the situation is more complicated compared to what described above mainly because computationally efficient methods may not achieve optimal sample complexity.

Ideally, we would have a method for constructing the anchor vector from a number of observations that does not dominate the sample complexity of the estimator. This may be possible if we impose an explicit structural constraint $\boldsymbol{u} \in \mathcal{S}$ with $\mathcal{S}$ denoting a prescribed set of structured vectors; namely we would have

$$\boldsymbol{a}_0 \in \underset{\boldsymbol{u} \in \mathbb{S}^{N-1}, \boldsymbol{u} \in \mathcal{S}}{\operatorname{argmax}} \ \langle -\nabla^k R_M\left(\boldsymbol{0}\right), \boldsymbol{u}^{\otimes k} \rangle. \tag{3.5}$$

Regardless of whether this construction can achieve (1.3) at an appropriate sample complexity, there is no guarantee that solving (3.5) is computationally tractable. For some of the problems discussed above, there are relaxations of (3.5) that produce an anchor, but at the cost of increasing M well beyond the required sample complexity (2.4) of the main estimation procedure.

As a concrete example, consider phase retrieval for a (real-valued) sparse target $\boldsymbol{x}_\star$ using noiseless measurements obtained through i.i.d. copies of $\boldsymbol{a} \sim \mathrm{Normal}\left(\boldsymbol{0}, \boldsymbol{I}\right)$ as the measurement vectors. As in the ordinary phase retrieval, the risk $R\left(\boldsymbol{x}\right) = \mathbb{E}\frac{1}{2}\left(\left|\boldsymbol{a}^\mathsf{T}\boldsymbol{x}\right|^2 - \left|\boldsymbol{a}^\mathsf{T}\boldsymbol{x}_\star\right|^2\right)^2$ has a spiked Hessian because $-\nabla^2 R\left(\boldsymbol{0}\right) = \mathbb{E}\left|\boldsymbol{a}^\mathsf{T}\boldsymbol{x}_\star\right|^2 \boldsymbol{a}\boldsymbol{a}^\mathsf{T} = \|\boldsymbol{x}_\star\|_2^2\,\boldsymbol{I} + 2\boldsymbol{x}_\star\boldsymbol{x}_\star^\mathsf{T}$ whose spectral gap is $\gamma_\star = 2\|\boldsymbol{x}_\star\|_2^2$. Ignoring the computational restrictions, if $\mathcal{S}$ is the set of (sufficiently) sparse vectors, a brute force estimator generally solves (3.5) for $k = 2$ by searching over all possible sparse support sets. This estimator can produce the desired $\boldsymbol{a}_0$ with (near) optimal sample complexity as we only need small submatrices of $-\nabla^2 R_M\left(\boldsymbol{0}\right)$ to concentrate around their expected values. In the case of sparse phase retrieval, the desired concentration occurs when M (i.e., the number of samples) grows (nearly) linearly with the sparsity of $\boldsymbol{x}_\star$ (i.e., $\|\boldsymbol{x}_\star\|_0$). However, computationally tractable relaxations of (3.5), which rely on diagonal thresholding or mixed nuclear-norm ℓ_1-norm regularization, require M to grow quadratically with $\|\boldsymbol{x}_\star\|_0$ to guarantee accuracy. This situation is similar to the case of *sparse principal component analysis* (SPCA) [15, 35], where the goal is to estimate the sparse principal component of a (covariance) matrix from an empirical (covariance) matrix. It is widely believed that computationally efficient estimators cannot achieve the optimal sample complexity in SPCA [3]. Suboptimality of mixed nuclear-norm ℓ_1-norm regularization is also shown in estimation of simultaneously sparse and low-rank matrices [26].

4 Proof of the main result

In this section we provide a proof of Theorem 2.1. The argument is based on the small-ball method introduced in [16, 23] with minor modifications. Our derivations mostly follow the exposition of this method in [32].

Proof of Theorem 2.1. For $t \geq 0$, let $\psi_t\left(s\right) \stackrel{\text{def}}{=} \left(s\right)_+ - \left(s - t\right)_+$ which is a *contraction* (i.e.,

$|\psi_\tau(s_2) - \psi_\tau(s_1)| \leq |s_2 - s_1|$ for all $s_1, s_2 \in \mathbb{R}$). We can bound $R_M^+(\boldsymbol{x}_\star + \boldsymbol{h})$ from below as

$$
\begin{aligned}
R_M^+(\boldsymbol{x}_\star + \boldsymbol{h}) &= \frac{1}{M} \sum_{m=1}^M (f_m(\boldsymbol{x}_\star + \boldsymbol{h}) - y_m)_+ \\
&= \frac{1}{M} \sum_{m=1}^M (f_m(\boldsymbol{x}_\star + \boldsymbol{h}) - f_m(\boldsymbol{x}_\star) - \xi_m)_+ \\
&\geq \frac{1}{M} \sum_{m=1}^M (\langle \nabla f_m(\boldsymbol{x}_\star), \boldsymbol{h} \rangle - \xi_m)_+ \\
&\geq \frac{1}{M} \sum_{m=1}^M (\langle \nabla f_m(\boldsymbol{x}_\star), \boldsymbol{h} \rangle)_+ - \frac{1}{M} \sum_{m=1}^M (\xi_m)_+ \\
&\geq \frac{1}{M} \sum_{m=1}^M \psi_{\tau\|\boldsymbol{h}\|_2} (\langle \nabla f_m(\boldsymbol{x}_\star), \boldsymbol{h} \rangle) - \frac{1}{M} \sum_{m=1}^M (\xi_m)_+ \;, \quad (4.1)
\end{aligned}
$$

where the inequalities hold, respectively, because $f_m(\cdot)$ is convex, $(\cdot)_+$ is subadditive, and $(s)_+ \geq \psi_t(s)$. Furthermore, using the fact that $t\mathbb{1}(s \geq t) \leq \psi_t(s)$, we have

$$
\begin{aligned}
\tau \|\boldsymbol{h}\|_2 \, \mathbb{P}(\langle \nabla f(\boldsymbol{x}_\star), \boldsymbol{h} \rangle \geq \tau \|\boldsymbol{h}\|_2) &= \tau \|\boldsymbol{h}\|_2 \, \mathbb{E}(\mathbb{1}(\langle \nabla f(\boldsymbol{x}_\star), \boldsymbol{h} \rangle \geq \tau \|\boldsymbol{h}\|_2)) \\
&\leq \mathbb{E}\left(\psi_{\tau\|\boldsymbol{h}\|_2}(\langle \nabla f(\boldsymbol{x}_\star), \boldsymbol{h} \rangle)\right)
\end{aligned}
$$

Therefore, by adding and subtracting the sides of the inequality above in the right-hand side of (4.1) and some rearrangement we obtain

$$
\begin{aligned}
R_M^+(\boldsymbol{x}_\star + \boldsymbol{h}) \geq {}& \tau \|\boldsymbol{h}\|_2 \, \mathbb{P}(\langle \nabla f(\boldsymbol{x}_\star), \boldsymbol{h} \rangle \geq \tau \|\boldsymbol{h}\|_2) - \frac{1}{M} \sum_{m=1}^M (\xi_m)_+ \\
&+ \frac{1}{M} \sum_{m=1}^M \psi_{\tau\|\boldsymbol{h}\|_2}(\langle \nabla f_m(\boldsymbol{x}_\star), \boldsymbol{h} \rangle) - \mathbb{E}\left(\psi_{\tau\|\boldsymbol{h}\|_2}(\langle \nabla f(\boldsymbol{x}_\star), \boldsymbol{h} \rangle)\right) .
\end{aligned}
\qquad (4.2)
$$

We only need to establish a uniform lower bound over $\boldsymbol{h} \in \mathcal{A}_\delta$ for the expression in the second line of (4.2). It is easy to verify that for every $\alpha, t \geq 0$ and $s \in \mathbb{R}$ the identity $\psi_{\alpha t}(s) = t\psi_\alpha\left(\frac{s}{t}\right)$ holds.[3] Thus, we write

$$
\begin{aligned}
&\frac{1}{M} \sum_{m=1}^M \psi_{\tau\|\boldsymbol{h}\|_2}(\langle \nabla f_m(\boldsymbol{x}_\star), \boldsymbol{h} \rangle) - \mathbb{E}\left(\psi_{\tau\|\boldsymbol{h}\|_2}(\langle \nabla f(\boldsymbol{x}_\star), \boldsymbol{h} \rangle)\right) \\
&= -\frac{1}{M} \|\boldsymbol{h}\|_2 \sum_{m=1}^M \mathbb{E}\left(\psi_\tau\left(\langle \nabla f(\boldsymbol{x}_\star), \frac{\boldsymbol{h}}{\|\boldsymbol{h}\|_2} \rangle\right)\right) - \psi_\tau\left(\langle \nabla f_m(\boldsymbol{x}_\star), \frac{\boldsymbol{h}}{\|\boldsymbol{h}\|_2} \rangle\right) ,
\end{aligned}
$$

and we only need to find an upper bound for

$$
F_{\mathcal{A}_\delta}(f_1, f_2, \ldots, f_M) \stackrel{\text{def}}{=} \sup_{\boldsymbol{h} \in \mathcal{A}_\delta} \frac{1}{M} \sum_{m=1}^M \mathbb{E}\left(\psi_\tau\left(\langle \nabla f(\boldsymbol{x}_\star), \frac{\boldsymbol{h}}{\|\boldsymbol{h}\|_2} \rangle\right)\right) - \psi_\tau\left(\langle \nabla f_m(\boldsymbol{x}_\star), \frac{\boldsymbol{h}}{\|\boldsymbol{h}\|_2} \rangle\right) .
$$

---

[3]Because $\psi_\alpha(\cdot)$ is bounded, we can treat $t = 0$ as $t \to 0$ to avoid the issue of division by zero.

Since $\psi_\tau(\cdot)$ is bounded by 0 and τ, a standard application of the *bounded difference inequality* [22] to $F_{\mathcal{A}_\delta}(\cdot)$ shows that for any $t > 0$ we have

$$F_{\mathcal{A}_\delta}(f_1, f_2, \ldots, f_M) \leq \mathbb{E}F_{\mathcal{A}_\delta}(f_1, f_2, \ldots, f_M) + \frac{t\tau}{\sqrt{M}}, \tag{4.3}$$

with probability $\geq 1 - \exp(-2t^2)$.

It remains only to upper bound $\mathbb{E}F_{\mathcal{A}_\delta}(f_1, f_2, \ldots, f_M)$. Writing the inner expectation in $\mathbb{E}F_{\mathcal{A}_\delta}(f_1, f_2, \ldots, f_M)$ with respect to random functions $\widetilde{f}_m$ that are i.i.d. copies of f, independent of everything else, we obtain

$$\mathbb{E}F_{\mathcal{A}_\delta}(f_1, f_2, \ldots, f_M) = \mathbb{E} \sup_{\boldsymbol{h} \in \mathcal{A}_\delta} \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}\left(\psi_\tau\left(\left\langle \nabla f(\boldsymbol{x}_\star), \frac{\boldsymbol{h}}{\|\boldsymbol{h}\|_2}\right\rangle\right)\right) - \psi_\tau\left(\left\langle \nabla f_m(\boldsymbol{x}_\star), \frac{\boldsymbol{h}}{\|\boldsymbol{h}\|_2}\right\rangle\right)$$

$$\leq \mathbb{E} \sup_{\boldsymbol{h} \in \mathcal{A}_\delta} \frac{1}{M} \sum_{m=1}^{M} \psi_\tau\left(\left\langle \nabla \widetilde{f}_m(\boldsymbol{x}_\star), \frac{\boldsymbol{h}}{\|\boldsymbol{h}\|_2}\right\rangle\right) - \psi_\tau\left(\left\langle \nabla f_m(\boldsymbol{x}_\star), \frac{\boldsymbol{h}}{\|\boldsymbol{h}\|_2}\right\rangle\right),$$

where in the second line the expectation with respect to the $\widetilde{f}_m$ is pulled outside of the supremum. The next step is the standard *symmetrization* argument [see e.g., 33, Lemma 2.3.1]. Since f_m and $\widetilde{f}_m$ are i.i.d., multiplying each summand in the right-hand side of the inequality above by a corresponding $\epsilon_m = \pm 1$ does not change the distribution of the random process and thereby the desired expected value. Take $\epsilon_1, \epsilon_2, \ldots, \epsilon_M$ to be i.i.d. Rademacher random variables independent of everything else. Therefore, we obtain

$$\mathbb{E}F_{\mathcal{A}_\delta}(f_1, f_2, \ldots, f_M) \leq \mathbb{E} \sup_{\boldsymbol{h} \in \mathcal{A}_\delta} \left[\frac{1}{M} \sum_{m=1}^{M} \epsilon_m \psi_\tau\left(\left\langle \nabla \widetilde{f}_m(\boldsymbol{x}_\star), \frac{\boldsymbol{h}}{\|\boldsymbol{h}\|_2}\right\rangle\right) \right.$$

$$\left. - \epsilon_m \psi_\tau\left(\left\langle \nabla f_m(\boldsymbol{x}_\star), \frac{\boldsymbol{h}}{\|\boldsymbol{h}\|_2}\right\rangle\right)\right]$$

$$\leq \mathbb{E} \sup_{\boldsymbol{h} \in \mathcal{A}_\delta} \frac{1}{M} \sum_{m=1}^{M} \epsilon_m \psi_\tau\left(\left\langle \nabla \widetilde{f}_m(\boldsymbol{x}_\star), \frac{\boldsymbol{h}}{\|\boldsymbol{h}\|_2}\right\rangle\right)$$

$$+ \mathbb{E} \sup_{\boldsymbol{h} \in \mathcal{A}_\delta} \frac{1}{M} \sum_{m=1}^{M} -\epsilon_m \psi_\tau\left(\left\langle \nabla f_m(\boldsymbol{x}_\star), \frac{\boldsymbol{h}}{\|\boldsymbol{h}\|_2}\right\rangle\right)$$

$$= 2 \mathbb{E} \sup_{\boldsymbol{h} \in \mathcal{A}_\delta} \frac{1}{M} \sum_{m=1}^{M} \epsilon_m \psi_\tau\left(\left\langle \nabla f_m(\boldsymbol{x}_\star), \frac{\boldsymbol{h}}{\|\boldsymbol{h}\|_2}\right\rangle\right). \tag{4.4}$$

Because $\psi_\tau(0) = 0$ and $\psi_\tau(\cdot)$ is a contraction, we can invoke the *Rademacher contraction principle* [19, Theorem 4.12] to show that

$$\mathbb{E} \sup_{\boldsymbol{h} \in \mathcal{A}_\delta} \frac{1}{M} \sum_{m=1}^{M} \epsilon_m \psi_\tau\left(\left\langle \nabla f_m(\boldsymbol{x}_\star), \frac{\boldsymbol{h}}{\|\boldsymbol{h}\|_2}\right\rangle\right) \leq \mathbb{E} \sup_{\boldsymbol{h} \in \mathcal{A}_\delta} \frac{1}{M} \sum_{m=1}^{M} \epsilon_m \left\langle \nabla f_m(\boldsymbol{x}_\star), \frac{\boldsymbol{h}}{\|\boldsymbol{h}\|_2}\right\rangle$$

$$= \frac{1}{\sqrt{M}} \mathfrak{C}_M(\mathcal{A}_\delta),$$

where $\mathfrak{C}_M(\mathcal{A}_\delta)$ is defined in (2.2) and can be interpreted as a measure of *complexity* of $\mathcal{A}_\delta$ with respect to the law of f. It follows from (4.4) that

$$\mathbb{E}F_{\mathcal{A}_\delta}(f_1, f_2, \ldots, f_M) \leq \frac{2}{\sqrt{M}} \mathfrak{C}_M(\mathcal{A}_\delta).$$

20

This bound together with (4.3) guarantees that

$$F_{\mathcal{A}_\delta}(f_1, f_2, \ldots, f_M) \le \frac{2}{\sqrt{M}} \, \mathfrak{C}_M(\mathcal{A}_\delta) + \frac{t\tau}{\sqrt{M}} \, ,$$

with probability $\ge 1 - \exp(-2t^2)$. Finally, on the same event, it follows from (4.2) that for all $\boldsymbol{h} \in \mathcal{A}_\delta$ we have

$$R_M^+(\boldsymbol{x}_\star + \boldsymbol{h}) \ge \tau \, \|\boldsymbol{h}\|_2 \, \mathbb{P}\left(\langle \nabla f(\boldsymbol{x}_\star), \boldsymbol{h} \rangle \ge \tau \, \|\boldsymbol{h}\|_2\right) - \frac{1}{M} \sum_{m=1}^{M} (\xi_m)_+$$

$$- \frac{2 \, \|\boldsymbol{h}\|_2}{\sqrt{M}} \, \mathfrak{C}_M(\mathcal{A}_\delta) - \frac{t\tau \, \|\boldsymbol{h}\|_2}{\sqrt{M}}$$

$$\ge \tau \, \|\boldsymbol{h}\|_2 \, p_\tau(\mathcal{A}_\delta) - \frac{1}{M} \sum_{m=1}^{M} (\xi_m)_+ - \frac{2 \, \|\boldsymbol{h}\|_2}{\sqrt{M}} \, \mathfrak{C}_M(\mathcal{A}_\delta) - \frac{t\tau \, \|\boldsymbol{h}\|_2}{\sqrt{M}} \, .$$

If $\boldsymbol{x}_\star + \boldsymbol{h}$ is feasible in (1.4), then

$$R_M^+(\boldsymbol{x}_\star + \boldsymbol{h}) \le R_M^+(\boldsymbol{x}_\star) + \varepsilon$$

$$= \frac{1}{M} \sum_{m=1}^{M} (-\xi_m)_+ + \varepsilon \, .$$

Therefore, if $\widehat{\boldsymbol{x}}$ denotes a solution to (1.4) and $\boldsymbol{h} = \widehat{\boldsymbol{x}} - \boldsymbol{x}_\star$, then we have

$$\tau \, \|\boldsymbol{h}\|_2 \, p_\tau(\mathcal{A}_\delta) - \frac{1}{M} \sum_{m=1}^{M} (\xi_m)_+$$

$$- \frac{2 \, \|\boldsymbol{h}\|_2}{\sqrt{M}} \, \mathfrak{C}_M(\mathcal{A}_\delta) - \frac{t\tau \, \|\boldsymbol{h}\|_2}{\sqrt{M}} \le \frac{1}{M} \sum_{m=1}^{M} (-\xi_m)_+ + \varepsilon \, ,$$

or equivalently

$$\left(\tau p_\tau(\mathcal{A}_\delta) - \frac{2\mathfrak{C}_M(\mathcal{A}_\delta) + t\tau}{\sqrt{M}}\right) \|\boldsymbol{h}\|_2 \le \frac{1}{M} \sum_{m=1}^{M} |\xi_m| + \varepsilon \, .$$

Applying the assumption (2.4) on the left-hand side completes the proof as

$$\|\widehat{\boldsymbol{x}} - \boldsymbol{x}_\star\|_2 = \|\boldsymbol{h}\|_2 \le \frac{2}{\tau p_\tau(\mathcal{A}_\delta)} \left(\frac{1}{M} \sum_{m=1}^{M} |\xi_m| + \varepsilon\right) \, .$$

$\square$

References

[1] A. Ahmed, B. Recht, and J. Romberg. Blind deconvolution using convex programming. *IEEE Trans. Inform. Theory*, 60(3):1711–1732, 2014.

[2] S. Bahmani and J. Romberg. Phase retrieval meets statistical learning theory: A flexible convex relaxation. In A. Singh and J. Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 252–260, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR.

[3] Q. Berthet and P. Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *Journal of Machine Learning Research W&CP*, volume 30 of *Proceedings of the 26th Conference on Learning Theory (COLT'13)*, pages 1046–1066, 2013.

[4] E. Candès and X. Li. Solving quadratic equations via PhaseLift when there are about as many equations as unknowns. *Found. of Comput. Math.*, 14:1017–1026, 2014.

[5] E. Candès, T. Strohmer, and V. Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Comm. Pure Appl. Math.*, 66 (8):1241–1274, 2013.

[6] V. H. de la Peña and E. Giné. *Decoupling: From dependence to independence.* Probability and its Applications. Springer-Verlag, New York, 1999.

[7] L. Dümbgen, S. A. van de Geer, M. C. Veraar, and J. A. Wellner. Nemirovski's inequalities revisited. *American mathematical monthly*, 117(2):138–160, 2010.

[8] T. Goldstein and C. Studer. PhaseMax: convex phase retrieval via basis pursuit. preprint `arXiv: 1610.07531 [cs.IT]`, Oct. 2016.

[9] P. Hand and V. Voroninski. Compressed sensing from phaseless gaussian measurements via linear programming in the natural parameter space. preprint `arXiv:1611.05985`, 2016.

[10] P. Hand and V. Voroninski. Corruption robust phase retrieval via linear programming. preprint `arXiv:1612.03547`, 2016.

[11] P. Hand and V. Voroninski. An elementary proof of convex phase retrieval in the natural parameter space via the linear program PhaseMax. preprint `arXiv:1611.03935`, 2016.

[12] S. Haykin. *Neural Networks and Learning Machines.* Pearson, Upper Saddle River, NJ, USA, 3rd edition, 2009.

[13] C. J. Hillar and L. Lim. Most tensor problems are NP-hard. *Journal of the ACM*, 60(6): 45:1–45:39, Nov. 2013.

[14] H. Ichimura. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, 58(1):71 – 120, 1993.

[15] I. M. Johnstone and A. Y. Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486): 682–693, 2009.

[16] V. Koltchinskii and S. Mendelson. Bounding the smallest singular value of a random matrix without concentration. *International Mathematics Research Notices*, 2015(23): 12991–13008, 2015.

[17] G. Lecué and S. Mendelson. Regularization and the small-ball method I: Sparse recovery. preprint `arXiv:1601.05584 [math.ST]`, 2016.

[18] G. Lecué and S. Mendelson. Regularization and the small-ball method II: Complexity dependent error rates. preprint `arXiv:1608.07681 [math.ST]`, 2016.

[19] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and processes.* Springer Science & Business Media, 2013.

[20] S. Ling and T. Strohmer. Self-calibration and biconvex compressive sensing. *Inverse Problems*, 31:115002, 2015.

[21] P. Mccullagh and J. A. Nelder. *Generalized linear models*, volume 37 of *Monographs on statistics and applied probability.* Chapman and Hall/CRC, London ; New York, 2nd edition, 1989.

[22] C. McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141 (1):148–188, 1989.

[23] S. Mendelson. Learning without concentration. In *Proceedings of the 27th Conference on Learning Theory (COLT)*, volume 35 of *JMLR W&CP*, pages 25–39, 2014.

[24] S. Mendelson. Learning without concentration for general loss functions. preprint; `arXiv:1410.3192`, 2014.

[25] A. Nemirovski. *Topics in Non-parametric Statistics*, chapter 5, pages 183–206. Springer Berlin Heidelberg, Berlin, Heidelberg, 2000.

[26] S. Oymak, A. Jalali, M. Fazel, Y. Eldar, and B. Hassibi. Simultaneously structured models with application to sparse and low-rank matrices. *Information Theory, IEEE Transactions on*, 61(5):2886–2908, 2015.

[27] R. E. A. C. Paley and A. Zygmund. A note on analytic functions in the unit circle. *Mathematical Proceedings of the Cambridge Philosophical Society*, 28(3):266–272, 1932.

[28] Y. Plan and R. Vershynin. The generalized LASSO with non-linear observations. *IEEE Transactions on Information Theory*, 62(3):1528–1537, Mar. 2016.

[29] Y. Plan, R. Vershynin, and E. Yudovina. High-dimensional estimation with geometric constraints. *Information and Inference*, 2016.

[30] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms.* Cambridge University Press, New York, NY, USA, 2014.

[31] C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, volume 2 of *Probability Theory*, pages 583–602, Berkeley, Calif., 1972. University of California Press.

[32] J. A. Tropp. *Convex Recovery of a Structured Signal from Independent Random Linear Measurements*, pages 67–101. Springer International Publishing, Cham, 2015.

[33] A. W. van Der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer, 1996.

[34] Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2014.

[35] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006.