Electronic Journal of Statistics
Vol. 13 (2019) 1978–2011
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1567

Estimation from nonlinear observations via convex programming with application to bilinear regression*

Sohail Bahmani

School of Electrical and Computer Engineering,
Georgia Institute of Technology,
Atlanta, GA 30332-0250
e-mail: sohail.bahmani@ece.gatech.edu

Abstract: We propose a computationally efficient estimator, formulated as a convex program, for a broad class of nonlinear regression problems that involve *difference of convex* (DC) nonlinearities. The proposed method can be viewed as a significant extension of the "anchored regression" method formulated and analyzed in [10] for regression with convex nonlinearities. Our main assumption, in addition to other mild statistical and computational assumptions, is availability of a certain approximation oracle for the average of the gradients of the observation functions at a ground truth. Under this assumption and using a PAC-Bayesian analysis we show that the proposed estimator produces an accurate estimate with high probability. As a concrete example, we study the proposed framework in the bilinear regression problem with Gaussian factors and quantify a sufficient sample complexity for exact recovery. Furthermore, we describe a computationally tractable scheme that provably produces the required approximation oracle in the considered bilinear regression problem.

MSC 2010 subject classifications: Primary 62F10, 90C25; secondary 62P30.
Keywords and phrases: Nonlinear regression, convex programming, PAC-Bayesian analysis, bilinear regression.

Received June 2018.

1. Introduction

Let $f_1^+, f_2^+, \dots, f_n^+$ be i.i.d. copies of a random *convex* function $f^+ \colon \mathbb{R}^d \to \mathbb{R}$. Similarly, let $f_1^-, f_2^-, \dots, f_n^-$ be i.i.d. copies of a random convex function $f^- \colon \mathbb{R}^d \to \mathbb{R}$. For simplicity, we also assume that the functions f^+ and f^- are differentiable.[1] We observe a parameter $\boldsymbol{x}_\star \in \mathbb{R}^d$ indirectly through the measurements

$$y_i = f_i^+(\boldsymbol{x}_\star) - f_i^-(\boldsymbol{x}_\star) + \xi_i \qquad , i = 1, \dots, n\,, \tag{1}$$

where ξ_is denote additive noise. Given the data

$$\left(f_i^+(\cdot)\,, f_i^-(\cdot)\,, y_i\right) \qquad i = 1, \dots, n\,,$$

*This work was supported in part by the Semiconductor Research Corporation (SRC) and DARPA.

[1] Religiously, we may add "almost everywhere almost surely."

the goal is to accurately estimate $\boldsymbol{x}_\star$, up to the possible inherent ambiguities, by a computationally tractable procedure.

One can immediately notice the *difference of convex* (DC) structure [2] in the observation model (1). Many parametric regression problems can be abstracted as (1) due to richness of the set of DC functions [24]; for instance, any smooth function can be expressed in the DC form using positive and negative semidefinite parts of its Hessian. While it is evident from the considered form of the observed data, we emphasize that the DC decomposition of the observation function is assumed to be known and our proposed estimator relies on such a DC decomposition.

In the context of the model (1), the standard estimators based on empirical risk minimization such as (nonlinear) least squares would lead to nonconvex optimization problems that are generally computationally hard. Thus, without making any assumption, our search for a computationally efficient estimator for (1) may be futile. Of course, statistical assumptions are also necessary to make the estimation meaningful; the observations must convey (enough) information about the ground truth parameter. All of the assumptions we make are discussed in more detail in Section 1.1.

Throughout we use the notation $\mathbb{E}_{\mathcal{D}}$ or $\mathbb{E}_{\mathcal{D}^n}$ to denote the expectation with respect to a single or multiple observations. Outer product of vectors is denoted by the binary operation $\otimes$. Furthermore, $\|\cdot\|$, $\|\cdot\|_{\mathrm{F}}$, and $\|\cdot\|_{\mathrm{op}}$ respectively denote the usual Euclidean norm, Frobenius norm, and operator norm.

1.1. Statistical and computational assumptions

In this section we describe the main statistical and computational assumptions we rely on in our analysis some of which were alluded to above. Stating some of these assumptions requires us to define certain parameters of the model for which we provide the motivations subsequently.

To avoid long expressions, for $i = 1, \dots, n$, we define

$$q_i(\boldsymbol{h}) \stackrel{\text{def}}{=} \frac{1}{2}\left|\langle \nabla f_i^+(\boldsymbol{x}_\star) - \nabla f_i^-(\boldsymbol{x}_\star), \boldsymbol{h}\rangle\right| , \tag{2}$$

which are clearly nonnegative and positive homogeneous. Therefore, by the triangle inequality, they also satisfy

$$|q_i(\boldsymbol{h}) - q_i(\boldsymbol{h}')| \le q_i(\boldsymbol{h} - \boldsymbol{h}'), \tag{3}$$

for every pair of $\boldsymbol{h}, \boldsymbol{h}' \in \mathbb{R}^d$.

As it becomes clear in the sequel, the central piece in our analysis is to establish a lower bound for the empirical process $\frac{1}{n}\sum_{i=1}^n q_i(\boldsymbol{h})$ uniformly for a set of vectors $\boldsymbol{h}$. A crucial point in our proof is that $\mathbb{E}_{\mathcal{D}}\left(q_i(\boldsymbol{z})\right)$ is linear in $\|\boldsymbol{z}\|$. If $\mathbb{E}_{\mathcal{D}}(q_i(\boldsymbol{z}))$ had a different modulus of continuity and did not admit a lower bound

[2] Sometimes this structure is referred to as *convex-concave*, indicating the decomposition into the sum of a convex function and a concave function.

with linear growth in $\|\boldsymbol{z}\|$, then a nontrivial lower bound for the mentioned empirical process that holds uniformly in an arbitrarily small neighborhood of the origin might not exist. The consequence would be an error bound that does not vanish by removing the additive noise. These circumstances are observed and well-understood, for instance, in the contexts of *ratio limit theorems* [18, 19], the issue of a nontrivial *version space* in learning problems [39, 40], and implicitly in specific applications such one-bit compressed sensing and its generalizations [44, 45].

We use the random function $q(\boldsymbol{h}) = \frac{1}{2}\left|\langle\nabla f^+(\boldsymbol{x}_\star) - \nabla f^-(\boldsymbol{x}_\star), \boldsymbol{h}\rangle\right|$, which has the same law as the functions q_i, to define a few important quantities below.

Conditioning: Let $\mathbb{S}^{d-1}$ denote the usual unit sphere in $\mathbb{R}^d$. Given $\mathcal{S} \subseteq \mathbb{S}^{d-1}$, we define $\lambda_{\mathcal{D}}$ and $\Lambda_{\mathcal{D}}$ as

$$\lambda_{\mathcal{D}} \stackrel{\text{def}}{=} \inf_{\boldsymbol{z}\in\mathcal{S}} \mathbb{E}_{\mathcal{D}}\left(q(\boldsymbol{z})\right), \tag{4}$$

and

$$\Lambda_{\mathcal{D}} \stackrel{\text{def}}{=} \sup_{\boldsymbol{z}\in\mathcal{S}} \mathbb{E}_{\mathcal{D}}\left(q(\boldsymbol{z})\right). \tag{5}$$

The dependence of $\lambda_{\mathcal{D}}$ and $\Lambda_{\mathcal{D}}$ on $\mathcal{S}$ will always be clear from the context, thus we do not make this dependence explicit merely to simplify the notation. Our results will depend on the *condition number* $\Lambda_{\mathcal{D}}/\lambda_{\mathcal{D}}$. In particular, it is important to have $\lambda_{\mathcal{D}} > 0$.

While generically $\mathcal{S}$ can be set to $\mathbb{S}^{d-1}$ in the definitions (4) and (5), in some applications we may choose $\mathcal{S}$ to be a proper subset of $\mathbb{S}^{d-1}$. This restriction helps us avoiding a *degeneracy* that leads to $\lambda_{\mathcal{D}} = 0$ and vacuous error bounds. An interesting example occurs in the *bilinear regression* problem discussed in Section 4.

Our proposed estimator, described in Section 2, can be viewed as an approximation to

$$\begin{aligned}\operatorname*{argmax}_{\boldsymbol{x}}\ \mathbb{E}\Big(&\frac{1}{2}\langle\nabla f^+(\boldsymbol{x}_\star) + \nabla f^-(\boldsymbol{x}_\star), \boldsymbol{x}-\boldsymbol{x}_\star\rangle - \\ &\max\{f^+(\boldsymbol{x}) - f^+(\boldsymbol{x}_\star), f^-(\boldsymbol{x}) - f^-(\boldsymbol{x}_\star)\}\Big),\end{aligned}$$

disregarding the additive constants in the objective function. The importance of $\lambda_{\mathcal{D}}$ can be explained by inspecting the uniqueness of the above "idealized" estimator. By convexity of $f^\pm(\cdot)$ we have

$$f^\pm(\boldsymbol{x}) - f^\pm(\boldsymbol{x}_\star) \geq \langle\nabla f^\pm(\boldsymbol{x}_\star), \boldsymbol{x}-\boldsymbol{x}_\star\rangle,$$

and thereby

$$\begin{aligned}&\max\{f^+(\boldsymbol{x}) - f^+(\boldsymbol{x}_\star), f^-(\boldsymbol{x}) - f^-(\boldsymbol{x}_\star)\} \\ &\geq \max\{\langle\nabla f^+(\boldsymbol{x}_\star), \boldsymbol{x}-\boldsymbol{x}_\star\rangle, \langle\nabla f^-(\boldsymbol{x}_\star), \boldsymbol{x}-\boldsymbol{x}_\star\rangle\}.\end{aligned}$$

Therefore, the objective function of the idealized estimator is dominated by

$$-\frac{1}{2}\mathbb{E}\left(\left|\langle \nabla f^{+}(\boldsymbol{x}_\star) - \nabla f^{-}(\boldsymbol{x}_\star), \boldsymbol{x} - \boldsymbol{x}_\star\rangle\right|\right)\,.$$

The points $\boldsymbol{x}$ for which $f^{+}(\boldsymbol{x}) - f^{-}(\boldsymbol{x}) = f^{+}(\boldsymbol{x}_\star) - f^{-}(\boldsymbol{x}_\star)$ almost surely are effectively equivalent to $\boldsymbol{x}_\star$. Thus, in view of (4), with $\mathcal{S} \subseteq \mathbb{S}^{d-1}$ being the complement of the directions from $\boldsymbol{x}_\star$ to its equivalents, having $\lambda_{\mathcal{D}} > 0$ guarantees that the idealized estimator can only be $\boldsymbol{x} = \boldsymbol{x}_\star$.

Regularity: For technical reasons we also need some regularity for the data distribution. To exclude pathologically heavy-tailed data distributions we make the mild {assumption} that the (directional) second moment of $\nabla f^{+}(\boldsymbol{x}_\star) - \nabla f^{-}(\boldsymbol{x}_\star)$ is bounded from above by its corresponding (directional) first moment. This assumption can be made precise in terms of $q(\boldsymbol{z})$ as follows. For some constant $\eta_{\mathcal{D}} > 1$ we assume that

$$\sqrt{\mathbb{E}_{\mathcal{D}}\left(q^2(\boldsymbol{z})\right)} \leq \eta_{\mathcal{D}}\mathbb{E}_{\mathcal{D}}\left(q(\boldsymbol{z})\right)\,, \tag{6}$$

holds for all $\boldsymbol{z} \in \mathbb{R}^d$. Furthermore, with $\boldsymbol{g}$ denoting a standard normal random variable, we define

$$\mathit{\Gamma}_{\mathcal{D}} \stackrel{\text{def}}{=} \sqrt{\mathbb{E}_{\boldsymbol{g}}\mathbb{E}_{\mathcal{D}}\left(q^2(\boldsymbol{g})\right)} = \frac{1}{2}\sqrt{\mathbb{E}_{\mathcal{D}}\left(\left\|\nabla f^{+}(\boldsymbol{x}_\star) - \nabla f^{-}(\boldsymbol{x}_\star)\right\|^2\right)}\,, \tag{7}$$

which is a measure of smoothness of the functions q_i near the origin. The main factor in the sample complexity we establish is $\mathit{\Gamma}_{\mathcal{D}}^2/\mathit{\Lambda}_{\mathcal{D}}^2$ that can be interpreted as the *effective dimension* of the problem since it is bounded by the ratio of the trace and the operator norm of the correlation matrix of $\nabla f^{+}(\boldsymbol{x}_\star) - \nabla f^{-}(\boldsymbol{x}_\star)$.

Approximation oracle: We assume an *approximation oracle* is available that provides a vector $\boldsymbol{a}_0 \in \mathbb{R}^d$ which, for some $\varepsilon \in (0, 1]$, obeys

$$\left\|\boldsymbol{a}_0 - \frac{1}{2n}\sum_{i=1}^{n}\nabla f_i^{+}(\boldsymbol{x}_\star) + \nabla f_i^{-}(\boldsymbol{x}_\star)\right\| \leq \frac{1-\varepsilon}{2}\lambda_{\mathcal{D}}\,. \tag{8}$$

Having access to the approximation oracle above is the strongest assumption we make. This assumption could be excessive for prediction tasks where the goal is merely accurate approximation of $f^{+}(\boldsymbol{x}_\star) - f^{-}(\boldsymbol{x}_\star)$ for the unseen data. However, in this paper we are analyzing an estimation task in which accurate estimation of $\boldsymbol{x}_\star$ is the goal rather than predicting $f^{+}(\boldsymbol{x}_\star) - f^{-}(\boldsymbol{x}_\star)$. A standard approach to such estimation problems is to optimize an *empirical risk* that quantifies the consistency of any candidate estimate with the observations. Because these risk functions are generally nonconvex, accuracy guarantees for iterative estimation procedures is often established assuming that they are *initialized* at a point, say $\boldsymbol{x}_0$, in a relatively small neighborhood of the ground truth $\boldsymbol{x}_\star$ (i.e., $\boldsymbol{x}_0 \approx \boldsymbol{x}_\star$). The imposed bound (8) can be derived from such initialization conditions; e.g., if $\nabla f^+(\cdot)+\nabla f^-(\cdot)$ is a sufficiently smooth mapping, then $\boldsymbol{x}_0 \approx \boldsymbol{x}_\star$ would imply $\boldsymbol{a}_0 = \frac{1}{2n}\sum_{i=1}^n \nabla f_i^+(\boldsymbol{x}_0)+\nabla f_i^-(\boldsymbol{x}_0) \approx \frac{1}{2n}\sum_{i=1}^n \nabla f_i^+(\boldsymbol{x}_\star)+\nabla f_i^-(\boldsymbol{x}_\star)$. Finally, if the vectors $\nabla f^+(\boldsymbol{x}_\star)+\nabla f^-(\boldsymbol{x}_\star)$ are sufficiently light-tailed in the sense of being bounded in a certain *Orlicz norm*[3], then we can simply require

$$\left\|\boldsymbol{a}_0 - \frac{1}{2}\mathbb{E}_{\mathcal{D}}\left(\nabla f^+(\boldsymbol{x}_\star)+\nabla f^-(\boldsymbol{x}_\star)\right)\right\| \le \frac{1-\varepsilon'}{2}\lambda_{\mathcal{D}}\,,$$

and then resort to a matrix concentration inequality such as the matrix Bernstein inequality [27, Theorem 2.7] or the matrix Rosenthal inequality [15, 25, 36] to recover the condition (8). We do not attempt to provide a general framework to address these details in this paper. However, in the context of the bilinear regression problem, following the idea of "spectral initialization" used in nonconvex methods (see, e.g., [12, 32, 33, 35, 42]) we provide an explicit example for an implementable approximation oracle in Section 4.

The three assumptions stated above are primarily related to the statistical model. We also make the following assumptions on the computational model in order to provide a tractable method.

Computational assumptions: As mentioned above, we emphasize that our approach requires the access to the DC decomposition of the observation function. Computing such a decomposition can be intractable in general (see, e.g., [2] and references therein). However, assuming access to an efficiently computable DC form is a reasonable compromise for creating a concrete computational framework. In many applications the DC decomposition is provided explicitly or is easy to compute. For instance, many statistical problems are concerned with observations of the form $y_i = \phi(\langle \boldsymbol{a}_i, \boldsymbol{x}_\star\rangle)$ for a certain nonlinear function $\phi : \mathbb{R} \to \mathbb{R}$ and data point $\boldsymbol{a}_i$. In some interesting instances, the desired DC decomposition is relatively easy to compute because it reduces to computing a DC decomposition of $\phi(\cdot)$ over $\mathbb{R}$. Of course, as a natural requirement for implementing optimization algorithms such as the first-order methods, we also need the components of the DC decomposition (and their gradients) to be computable.

2. The estimator and the main results

Given $\boldsymbol{a}_0$, the output of the approximation oracle which obeys (8), we formulate the estimator of $\boldsymbol{x}_\star$ as

$$\widehat{\boldsymbol{x}} \in \underset{\boldsymbol{x}}{\operatorname{argmax}}\ \langle \boldsymbol{a}_0, \boldsymbol{x}\rangle - \frac{1}{n}\sum_{i=1}^n \max\left\{f_i^+(\boldsymbol{x}) - y_i, f_i^-(\boldsymbol{x})\right\}\,, \tag{9}$$

which is a convex program that can be solved efficiently.

[3]For a precise definition, interested readers are referred to [27, Appendix A.1] and the references therein.

Let us first demystify the formulation of the estimator by some intuitive explanations. Using the identity

$$\max\{u, v\} = \frac{u + v + |u - v|}{2}\,,$$

the objective function in (9) can be expressed as

$$\begin{aligned}&\langle \boldsymbol{a}_0, \boldsymbol{x}\rangle - \frac{1}{n}\sum_{i=1}^{n} \max\left\{f_i^+(\boldsymbol{x}) - y_i, f_i^-(\boldsymbol{x})\right\}\\ &= -\frac{1}{2n}\sum_{i=1}^{n}\left(f_i^+(\boldsymbol{x}) + f_i^-(\boldsymbol{x}) - \langle 2\boldsymbol{a}_0, \boldsymbol{x}\rangle - y_i\right) - \frac{1}{2n}\sum_{i=1}^{n}\left|f_i^+(\boldsymbol{x}) - f_i^-(\boldsymbol{x}) - y_i\right|.\end{aligned}$$

Suppose that, instead of (8), we have $\boldsymbol{a}_0 = \frac{1}{2n}\sum_{i=1}^{n}\nabla f_i^+(\boldsymbol{x}_\star) + \nabla f_i^-(\boldsymbol{x}_\star)$. Inspecting the first sum, it is evident that it is, up to additive constants, a Bregman divergence that admits $\boldsymbol{x}_\star$ as a minimizer. Furthermore, the second sum is expected to be minimized at a point close to $\boldsymbol{x}_\star$ because the observations obey $y_i \approx f_i^+(\boldsymbol{x}_\star) - f_i^-(\boldsymbol{x}_\star)$. Therefore, we can expect that the estimator $\widehat{\boldsymbol{x}}$ is not far from $\boldsymbol{x}_\star$. The Karush-Kuhn-Tucker (KKT) stationarity condition, explains that the approximation error of $\boldsymbol{a}_0$ in (8) is tolerable because the second sum contains nondifferentiable terms with (potentially) large subdifferentials. Our analysis makes these intuitive explanations rigorous in an implicit manner.

With the definitions and assumptions stated in Section 1.1, our main result is the following theorem that provides the sample complexity for accuracy (9) in a generic setting. This theorem is a simple consequence of Proposition 1 and Lemma 1 stated subsequently.

Theorem 1. *Given a set $\mathcal{S} \subseteq \mathbb{S}^{d-1}$ and parameter $\varepsilon \in (0, 1)$, suppose (4), (5), (6), (7), and (8) hold. Furthermore, for a solution $\widehat{\boldsymbol{x}}$ of (9), suppose that we have*

$$\widehat{\boldsymbol{x}} - \boldsymbol{x}_\star \in \|\widehat{\boldsymbol{x}} - \boldsymbol{x}_\star\|\,\mathcal{S}\,. \tag{10}$$

If the number of measurements obeys

$$n \geq C\max\left\{\eta_{\mathcal{D}}^2\log\frac{2}{\delta},\, \frac{\varGamma_{\mathcal{D}}^2}{\lambda_{\mathcal{D}}\varLambda_{\mathcal{D}}}\right\}\frac{\varLambda_{\mathcal{D}}^2}{\lambda_{\mathcal{D}}^2}\eta_{\mathcal{D}}^2\varepsilon^{-4}\,,$$

for a sufficiently large absolute constant $C > 0$, then with probability $\geq 1 - \delta$ we have

$$\|\widehat{\boldsymbol{x}} - \boldsymbol{x}_\star\| \leq \frac{\frac{1}{n}\sum_{i=1}^{n}|\xi_i|}{\frac{1}{2}\lambda_{\mathcal{D}}}\,.$$

Proof. Given the lower bound on n, Proposition 1 below together with (4) guarantee that, with probability $\geq 1 - \delta$, we have

$$\frac{1}{n}\sum_{i=1}^{n} q_i(\boldsymbol{h}) \geq (1 - \frac{1}{2}\varepsilon)\lambda_{\mathcal{D}}\,,$$

for all $\boldsymbol{h} \in \mathcal{S}$. The desired error bound follows immediately from Lemma 1 with $\boldsymbol{x}_0 = \boldsymbol{x}_\star$ and $\varepsilon_0 = 0$. $\square$

The condition (10) in the theorem may appear unnatural at first. Clearly, the condition holds if we choose $\mathcal{S} = \mathbb{S}^{d-1}$. However, the condition (10) is imposed to address the situations where a set of equivalent ground truth vectors $\boldsymbol{x}_\star$ exists and we only need to prove accuracy with respect to the closest point in this set. This relaxed accuracy requirement induces additional structure on $\widehat{\boldsymbol{x}} - \boldsymbol{x}_\star$ that should be considered to avoid the degeneracy at $\lambda_{\mathcal{D}} = 0$. The sole purpose of (10) is to capture the mentioned additional structures. The bilinear regression problem discussed below in Section 4 is an example where it is important to have a nontrivial set $\mathcal{S}$.

Furthermore, with $\kappa \stackrel{\text{def}}{=} \Lambda_{\mathcal{D}}/\lambda_{\mathcal{D}}$ and $d_{\text{eff}} \stackrel{\text{def}}{=} \Gamma_{\mathcal{D}}^2/\Lambda_{\mathcal{D}}^2$, the achievable sample complexity stated by the Theorem 1 can be rewritten as

$$n \geq C \max\{\eta_{\mathcal{D}}^2 \log \frac{2}{\delta}, \kappa\, d_{\text{eff}}\}\kappa^2 \eta_{\mathcal{D}}^2 \varepsilon^{-4}\,,$$

signifying the role of the effective dimension $d_{\text{eff}} \leq d$, and the conditioning κ of the problem.

Under the assumptions stated in Section 1.1, accuracy of the estimator (9) can be reduced to the existence of an appropriate uniform lower bound for the empirical process $\frac{1}{n}\sum_{i=1}^n q_i(\boldsymbol{h})$ as a function of $\boldsymbol{h}$. The following Lemma 1, proved in Section 3.2, provides the precise form of this reduction.

Lemma 1. *Let $\boldsymbol{x}_0$ be one of the possibly many vectors equivalent to $\boldsymbol{x}_\star$ meaning that*

$$f^+(\boldsymbol{x}_0) - f^-(\boldsymbol{x}_0) = f^+(\boldsymbol{x}_\star) - f^-(\boldsymbol{x}_\star)\,,$$

almost surely. Given a set $\mathcal{S} \subseteq \mathbb{S}^{d-1}$, recall the definition (4) and assume that an analog of the condition (8) with respect to $\boldsymbol{x}_0$ holds, namely,

$$\left\| \boldsymbol{a}_0 - \frac{1}{2n}\sum_{i=1}^n \nabla f_i^+(\boldsymbol{x}_0) + \nabla f_i^-(\boldsymbol{x}_0) \right\| \leq \frac{1-\varepsilon}{2}\lambda_{\mathcal{D}}\,, \tag{11}$$

for some constant parameter $\varepsilon \in (0,1)$. Furthermore, suppose that (10) holds and that for a certain absolute constant $\varepsilon_0 \in [0,1)$,

$$\frac{1}{2n}\sum_{i=1}^n \left|\langle \nabla f_i^+(\boldsymbol{x}_0) - \nabla f_i^-(\boldsymbol{x}_0), \boldsymbol{h}\rangle\right| \geq (1 - \frac{\varepsilon+\varepsilon_0}{2})\lambda_{\mathcal{D}}\,, \tag{12}$$

holds for every $\boldsymbol{h} \in \mathcal{S}$. Then the estimate $\widehat{\boldsymbol{x}}$ obeys

$$\|\widehat{\boldsymbol{x}} - \boldsymbol{x}_\star\| \leq \frac{\frac{1}{n}\sum_{i=1}^n |\xi_i|}{\frac{1-\varepsilon_0}{2}\lambda_{\mathcal{D}}}\,.$$

Again, in the generic case we choose $\mathcal{S} = \mathbb{S}^{d-1}$, $\boldsymbol{x}_0 = \boldsymbol{x}_\star$, and $\varepsilon_0 = 0$ in Lemma 1. For the structured problems mentioned above, however, with a nontrivial choice of $\mathcal{S}$ in (10), we may need to choose $\boldsymbol{x}_0 \neq \boldsymbol{x}_\star$ and an appropriate $\varepsilon_0 > 0$.

Lemma 1 provides an error bound that is proportional to $\frac{1}{n}\sum_{i=1}^n |\xi_i|$. This dependence is satisfactory for a deterministic noise model where we ought to consider the worst-case scenarios. However, we may obtain improved noise dependence for random noise models. In fact, simple modifications in the proof of Lemma 1 allow us to replace $\frac{1}{n}\sum_{i=1}^n |\xi_i|$ in the error bound by the maximum of the two expressions

$$\left|\sup_{\boldsymbol{x}} \frac{1}{n}\sum_{i=1}^n \xi_i \mathbb{1}\left(f_i^+(\boldsymbol{x}) - f_i^-(\boldsymbol{x}) > f_i^+(\boldsymbol{x}_\star) - f_i^-(\boldsymbol{x}_\star)\right)\right|$$

and

$$\left|\sup_{\boldsymbol{x}} \frac{1}{n}\sum_{i=1}^n -\xi_i \mathbb{1}\left(f_i^+(\boldsymbol{x}) - f_i^-(\boldsymbol{x}) > f_i^+(\boldsymbol{x}_\star) - f_i^-(\boldsymbol{x}_\star) + \xi_i\right)\right| .$$

These expressions may provide much tighter bounds when the noise is random with a well-behaved distribution. For instance, if $\xi_1, \ldots, \xi_n$ are i.i.d. zero-mean Gaussian random variables, the first expression reduces to *the Gaussian complexity* of the functions $\mathbb{1}\left(f_i^+(\boldsymbol{x}) - f_i^-(\boldsymbol{x}) > f_i^+(\boldsymbol{x}_\star) - f_i^-(\boldsymbol{x}_\star)\right)$ which may be of order $n^{-1/2}$. To keep the exposition simple, we focus on the deterministic noise model in this paper.

Clearly, to prove accuracy of (9) through Lemma 1, establishing an inequality of the form (12) is crucial. Proposition 1 below can provide a guarantee for such an inequality in the case $\boldsymbol{x}_0 = \boldsymbol{x}_\star$ and under the assumptions made in Section 1.

Proposition 1. *Let $\varepsilon \in (0,1)$ be a constant parameter. With the definitions (2), (4), (5), (6), and (7), for any $\delta \in (0,1]$, if for a sufficiently large absolute constant $C > 0$ we have*

$$n \geq C \max\left\{\eta_{\mathcal{D}}^2 \log\frac{2}{\delta},\ \frac{\varGamma_{\mathcal{D}}^2}{\lambda_{\mathcal{D}}\varLambda_{\mathcal{D}}}\right\} \frac{\varLambda_{\mathcal{D}}^2}{\lambda_{\mathcal{D}}^2}\eta_{\mathcal{D}}^2 \varepsilon^{-4} ,$$

then with probability $\geq 1-\delta$ the bound

$$\frac{1}{n}\sum_{i=1}^n q_i(\boldsymbol{h}) \geq (1-\varepsilon)\mathbb{E}_{\mathcal{D}}(q(\boldsymbol{h})) ,$$

holds for every $\boldsymbol{h} \in \mathcal{S}$.

The proof of this proposition is provided in Section 3.

2.1. Related work

In a prior work [10], we considered the "convex regression" model, a special case of (1) with purely convex nonlinearities (i.e., $f_i^- \equiv 0$ and $f_i^+ \equiv f_i$ for convex functions f_i). With a slightly weaker approximation oracle that produces an *anchor* $\boldsymbol{a}_0$ for which $\langle \boldsymbol{a}_0, \boldsymbol{x}_\star \rangle / \|\boldsymbol{a}_0\| \, \|\boldsymbol{x}_\star\|$ is nonvanishing, statistical accuracy of estimation via the convex program

$$\begin{aligned} &\operatorname*{argmax}_{\boldsymbol{x}} \ \langle \boldsymbol{a}_0, \boldsymbol{x} \rangle \\ &\text{subject to } \frac{1}{n} \sum_{i=1}^{n} \max \left\{ f_i(\boldsymbol{x}) - y_i, 0 \right\} \leq \text{average noise}\,, \end{aligned}$$

is studied in [10]. The effect of convex regularization (e.g., ℓ_1-regularization) in structured estimation (e.g., sparse estimation) is also considered and analyzed in [10]. Evidently, the solution of the convex program above is insensitive to (positive) scaling of the anchor $\boldsymbol{a}_0$. The estimator (9) is, however, sensitive to the scaling of $\boldsymbol{a}_0$ which is a main reason for the need for a slightly stronger approximation oracle in this paper. An interesting example where the described convex regression applies is the *phase retrieval* problem that was previously studied in [8, 9, 20, 21].

As will be seen in Section 4, bilinear regression can be modeled by (1) as well. Succinctly, the goal in a bilinear regression problem is to recover signal components $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$, up to the inevitable scaling ambiguity, from bilinear observations of the form $\langle \boldsymbol{a}_i^{(1)}, \boldsymbol{x}^{(1)} \rangle \langle \boldsymbol{a}_i^{(2)}, \boldsymbol{x}^{(2)} \rangle$ for $i = 1, \dots, n$. In the context of the closely related *blind deconvolution* problem, solving such a system of bilinear equations in the *lifted domain* through nuclear-norm minimization has been analyzed in [3] and [7]. Despite their accuracy guarantees, the nuclear-norm minimization methods are practically not scalable to large problem sizes which motivated the analysis of nonconvex techniques (see, e.g., [32, 33, 35]). Inspired by the results on the phase retrieval problem mentioned above, [1] proposed and analyzed a convex program for bilinear regression that operates in the natural space of the signals, thereby avoiding the prohibitive computational cost of the lifted convex formulations. Unlike the mentioned methods for phase retrieval that only require a (directional) approximation of the ground truth, the proposed estimator in [1] requires the exact knowledge of the signs of all of the multiplied linear forms $\langle \boldsymbol{a}_i^{(1)}, \boldsymbol{x}_\star^{(1)} \rangle$ (or $\langle \boldsymbol{a}_i^{(2)}, \boldsymbol{x}_\star^{(2)} \rangle$). This requirement is rather strong and may severely limit the applicability of the considered method. In Section 4, we look into the problem of bilinear regression as a special case of the general regression problem (1); under the common Gaussian model for the measurement vectors we derive the sample complexity of (9) and explain an efficient method to construct an admissible vector $\boldsymbol{a}_0$ only using the given observations.

3. Main proofs

There are various techniques under the umbrella of empirical process theory that can be employed to establish Proposition 1 and thereby Theorem 1. For instance, techniques relying on the concepts of VC dimension [47, 48] or Rademacher complexity [11, 26, 29] including the small-ball method [28, 39, 40] that are primarily developed in the field of statistical learning theory. However, some techniques, such as the small-ball method, are designed particularly to handle the type of heavy-tailed data we consider in our model. In this paper, we use another common technique, the PAC-Bayesian (or pseudo-Bayesian) method, that is suitable for heavy-tailed data. This method, proposed in [38], has been used previously for establishing various generalization bounds for classification [see e.g., 17, 30, 38] and accuracy in regression problems [4–6, 14, 43]. The bounds obtained using this technique appear in different forms; we refer the interested reader to the survey paper [37] and the monograph [13] for a broader view of the related results and techniques. Compared to the small-ball method, the PAC-Bayesian argument does not rely on the *symmetrization* [46, Lemma 2.3.1] and *Rademacher contraction* [31, Theorem 4.12] ideas and has a more elementary nature.

Our analysis below in Section 3.1 parallels that used in [43] which in turn was inspired by [6]. The technical tools we use can be found in the PAC-Bayesian literature; we provide the proofs to make the manuscript self-contained. We emphasize that the novelty of this work is the general regression model (1) and the computationally efficient estimator (9) rather than the methods of analysis.

The core idea in the PAC-Bayes theory is the variational inequality[4]

$$\mathbb{E}_{\boldsymbol{z}\sim\mu} R(\boldsymbol{z}) \le \log \mathbb{E}_{\boldsymbol{z}\sim\nu} \exp(R(\boldsymbol{z})) + D_{\mathrm{KL}}(\mu,\nu)\,, \tag{13}$$

where $D_{\mathrm{KL}}(\mu,\nu) = \mathbb{E}_{\boldsymbol{z}\sim\mu}\left(\log \frac{\mathrm{d}\mu(\boldsymbol{z})}{\mathrm{d}\nu(\boldsymbol{z})}\right)$ denotes the Kullback-Leibler divergence (or relative entropy) between probability measures μ and ν with $\mu \ll \nu$. In PAC-Bayesian analyses, the fact that this bound is deterministic and holds for any probability measure $\mu \ll \nu$ is leveraged to control the supremum of stochastic processes. In particular, for a stochastic process $R(\cdot)$ with the domain $\mathcal{X}$, we may *approximate* $\sup_{\boldsymbol{x}\in\mathcal{X}}\ R(\boldsymbol{x})$ by the supremum of $\sup_{\mu_{\boldsymbol{x}}:\boldsymbol{x}\in\mathcal{X}}\ \mathbb{E}_{\boldsymbol{z}\sim\mu_{\boldsymbol{x}}} R(\boldsymbol{z})$ with respect to a certain set of probability measures $\mu_{\boldsymbol{x}}$ indexed by the elements of $\mathcal{X}$ (e.g., $\mathbb{E}_{\boldsymbol{z}\sim\mu_{\boldsymbol{x}}}(\boldsymbol{z}) = \boldsymbol{x}$). Then, under some regularity conditions on the stochastic process, the approximate bound can be converted to an exact bound.

3.1. A PAC-Bayesian proof of Proposition 1

We use the PAC-Bayesian analysis to establish Proposition 1, the main ingredient in proving the accuracy of (9).

Proof of Proposition 1. For $i = 1, \ldots, n$, let

$$w_i(\boldsymbol{z}) \stackrel{\text{def}}{=} \log\left(1 - [\alpha q_i(\boldsymbol{z})]_{\le 1} + \frac{1}{2}\,[\alpha q_i(\boldsymbol{z})]_{\le 1}^2\right)\,, \tag{14}$$

[4] While this inequality is sometimes interpreted using the Fenchel–Legendre transform, it is simply a Jensen's inequality in disguise.

where $[u]_{\leq 1} \stackrel{\text{def}}{=} \min(u, 1)$ and $\alpha > 0$ is a normalizing factor to be specified later. As it becomes clear below, the function $w_i(\boldsymbol{z})$ should be viewed as an approximation for $\alpha q_i(\boldsymbol{z})$ that serves two purposes in the PAC-Bayesian argument. First, the use of the logarithm leads to *cumulant generating functions* that can be relatively easily approximated by $-\mathbb{E}_{\mathcal{D}}(\alpha q_i(\boldsymbol{z}))$. Second, the use of the truncation legitimizes the evaluation of moment generating functions and also allows us to have bounded the deviations caused by the parameter perturbation in the PAC-Bayesian argument.

Let $\gamma_{\boldsymbol{h}}$ denote the normal distribution with mean $\boldsymbol{h}$ and covariance $\sigma^2\boldsymbol{I}$ for a parameter σ. By (13), for every $\boldsymbol{h} \in \mathcal{S} \subseteq \mathbb{S}^{d-1}$, we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{z}\sim\gamma_{\boldsymbol{h}}}\left(\sum_{i=1}^n w_i(\boldsymbol{z}) - n\log\mathbb{E}_{\mathcal{D}}\exp\left(w_1(\boldsymbol{z})\right)\right)\\
&\quad\leq \log\mathbb{E}_{\boldsymbol{z}\sim\gamma_{\boldsymbol{0}}}\exp\left(\sum_{i=1}^n w_i(\boldsymbol{z}) - n\log\mathbb{E}_{\mathcal{D}}\exp\left(w_1(\boldsymbol{z})\right)\right) + \frac{\sigma^{-2}}{2}\,.
\end{aligned}\tag{15}$$

Furthermore, by Markov's inequality with probability $\geq 1-\delta/2$ we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{z}\sim\gamma_{\boldsymbol{0}}}\exp\left(\sum_{i=1}^n w_i(\boldsymbol{z}) - n\log\mathbb{E}_{\mathcal{D}}\exp\left(w_1(\boldsymbol{z})\right)\right)\\
&\leq \frac{2}{\delta}\mathbb{E}_{\mathcal{D}^n}\mathbb{E}_{\boldsymbol{z}\sim\gamma_{\boldsymbol{0}}}\exp\left(\sum_{i=1}^n w_i(\boldsymbol{z}) - n\log\mathbb{E}_{\mathcal{D}}\exp\left(w_1(\boldsymbol{z})\right)\right)\\
&= \frac{2}{\delta}\mathbb{E}_{\boldsymbol{z}\sim\gamma_{\boldsymbol{0}}}\mathbb{E}_{\mathcal{D}^n}\exp\left(\sum_{i=1}^n w_i(\boldsymbol{z}) - n\log\mathbb{E}_{\mathcal{D}}\exp\left(w_1(\boldsymbol{z})\right)\right)\\
&= \frac{2}{\delta}\,,
\end{aligned}$$

where the exchange of expectations on the second line is valid as the argument is a bounded function. Therefore, on the same event because of (15) for every $\boldsymbol{h} \in \mathcal{S}$ we have

$$\mathbb{E}_{\boldsymbol{z}\sim\gamma_{\boldsymbol{h}}}\left(\sum_{i=1}^n w_i(\boldsymbol{z}) - n\log\mathbb{E}_{\mathcal{D}}\exp\left(w_1(\boldsymbol{z})\right)\right) \leq \frac{\sigma^{-2}}{2} + \log\frac{2}{\delta}\,,$$

or equivalently

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\boldsymbol{z}\sim\gamma_{\boldsymbol{h}}} w_i(\boldsymbol{z}) \leq \mathbb{E}_{\boldsymbol{z}\sim\gamma_{\boldsymbol{h}}}\left(\log\mathbb{E}_{\mathcal{D}}\exp(w_1(\boldsymbol{z}))\right) + \frac{\sigma^{-2} + 2\log\frac{2}{\delta}}{2n}\,. \tag{16}$$

The definition (14) and the facts that $-u \leq \log(1-u+\frac{1}{2}u^2) \leq -u + \frac{1}{2}u^2$ and $1 - [u]_{\leq 1} + \frac{1}{2}[u]_{\leq 1}^2 \leq 1 - u + \frac{1}{2}u^2$, for all $u \geq 0$, imply the bounds

$$\mathbb{E}_{\boldsymbol{z}\sim\gamma_{\boldsymbol{h}}} w_i(\boldsymbol{z}) \geq -\mathbb{E}_{\boldsymbol{z}\sim\gamma_{\boldsymbol{h}}}\left([\alpha q_i(\boldsymbol{z})]_{\leq 1}\right)\,,$$

and

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{z}\sim\gamma_{\boldsymbol{h}}} \log \mathbb{E}_{\mathcal{D}} \exp(w_1(\boldsymbol{z}))\\
&= \mathbb{E}_{\boldsymbol{z}\sim\gamma_{\boldsymbol{h}}} \log \left(1 - \mathbb{E}_{\mathcal{D}}\left([\alpha q_1(\boldsymbol{z})]_{\leq 1}\right) + \frac{1}{2}\mathbb{E}_{\mathcal{D}}\left([\alpha q_1(\boldsymbol{z})]_{\leq 1}^2\right)\right)\\
&\leq -\mathbb{E}_{\boldsymbol{z}\sim\gamma_{\boldsymbol{h}}}\mathbb{E}_{\mathcal{D}}\left(\alpha q_1(\boldsymbol{z})\right) + \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim\gamma_{\boldsymbol{h}}}\mathbb{E}_{\mathcal{D}}\left(\alpha^2 q_1^2(\boldsymbol{z})\right) .
\end{aligned}$$

Using (3), (7), and the Cauchy-Schwarz inequality, we also have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{z}\sim\gamma_{\boldsymbol{h}}}\mathbb{E}_{\mathcal{D}}\left(q_1(\boldsymbol{z})\right) &\geq \mathbb{E}_{\boldsymbol{z}\sim\gamma_{\boldsymbol{h}}}\mathbb{E}_{\mathcal{D}}\left(q_1(\boldsymbol{h}) - q_1(\boldsymbol{h}-\boldsymbol{z})\right)\\
&= \mathbb{E}_{\mathcal{D}}\left(q_1(\boldsymbol{h})\right) - \mathbb{E}_{\boldsymbol{z}\sim\gamma_{\boldsymbol{0}}}\mathbb{E}_{\mathcal{D}}\left(q_1(\boldsymbol{z})\right)\\
&\geq \mathbb{E}_{\mathcal{D}}\left(q_1(\boldsymbol{h})\right) - \sigma\varGamma_{\mathcal{D}} .
\end{aligned}$$

Thus, it follows from (16) that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\alpha}\mathbb{E}_{\boldsymbol{z}\sim\gamma_{\boldsymbol{h}}}\left([\alpha q_i(\boldsymbol{z})]_{\leq 1}\right) \geq \mathbb{E}_{\mathcal{D}}\left(q_1(\boldsymbol{h})\right) - \sigma\varGamma_{\mathcal{D}}&\\
- \frac{\alpha}{2}\mathbb{E}_{\boldsymbol{z}\sim\gamma_{\boldsymbol{h}}}\mathbb{E}_{\mathcal{D}}\left(q_1^2(\boldsymbol{z})\right) - \frac{\sigma^{-2} + 2\log\frac{2}{\delta}}{2\alpha n} .&
\end{aligned} \tag{17}$$

Applying Lemmas 2 and 3, stated and proved in the appendix, to (17) shows that for all $\boldsymbol{h}\in\mathcal{S}$, with probability $\geq 1-\delta$, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} q_i(\boldsymbol{h}) + \sigma\varGamma_{\mathcal{D}} + \frac{1}{\alpha}\sqrt{\frac{\log\frac{2}{\delta}}{2n}}\\
&\geq \mathbb{E}_{\mathcal{D}}\left(q_1(\boldsymbol{h})\right) - \sigma\varGamma_{\mathcal{D}}\\
&\quad - \frac{\alpha}{2}\mathbb{E}_{\boldsymbol{z}\sim\gamma_{\boldsymbol{h}}}\mathbb{E}_{\mathcal{D}}\left(q_1^2(\boldsymbol{z})\right) - \frac{\sigma^{-2} + 2\log\frac{2}{\delta}}{2\alpha n}\\
&\geq \mathbb{E}_{\mathcal{D}}\left(q_1(\boldsymbol{h})\right) - \sigma\varGamma_{\mathcal{D}} - \frac{\alpha}{2}\left(\eta_{\mathcal{D}}\mathbb{E}_{\mathcal{D}}\left(q_1(\boldsymbol{h})\right) + \sigma\varGamma_{\mathcal{D}}\right)^2\\
&\quad - \frac{\sigma^{-2} + 2\log\frac{2}{\delta}}{2\alpha n}\\
&\geq \mathbb{E}_{\mathcal{D}}\left(q_1(\boldsymbol{h})\right) - \sigma\varGamma_{\mathcal{D}} - \alpha\left(\eta_{\mathcal{D}}^2\left(\mathbb{E}_{\mathcal{D}}\left(q_1(\boldsymbol{h})\right)\right)^2 + \sigma^2\varGamma_{\mathcal{D}}^2\right)\\
&\quad - \frac{\sigma^{-2} + 2\log\frac{2}{\delta}}{2\alpha n} ,
\end{aligned}$$

By rearranging the terms, we reach at

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} q_i(\boldsymbol{h}) \geq \mathbb{E}_{\mathcal{D}}\left(q_1(\boldsymbol{h})\right)&\\
- 2\sigma\varGamma_{\mathcal{D}} - \alpha\left(\eta_{\mathcal{D}}^2\left(\mathbb{E}_{\mathcal{D}}\left(q_1(\boldsymbol{h})\right)\right)^2 + \sigma^2\varGamma_{\mathcal{D}}^2\right)&\\
- \frac{1}{\alpha}\left(\sqrt{\frac{\log\frac{2}{\delta}}{2n}} + \frac{\sigma^{-2} + 2\log\frac{2}{\delta}}{2n}\right) .&
\end{aligned}$$

Recalling (4), (5), and (7), we can choose

$$\sigma = \frac{\varepsilon\lambda_{\mathcal{D}}}{4\mathit{\Gamma}_{\mathcal{D}}}\,,$$

and

$$\alpha = \lambda_{\mathcal{D}}\left(\sqrt{\frac{\log\frac{2}{\delta}}{2n}} + \frac{16\mathit{\Gamma}_{\mathcal{D}}^2\lambda_{\mathcal{D}}^{-2}\varepsilon^{-2} + 2\log\frac{2}{\delta}}{2n}\right)^{1/2}\left(\frac{\mathit{\Lambda}_{\mathcal{D}}}{\lambda_{\mathcal{D}}}\right)^{-1/2}\,,$$

to obtain

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} q_i(\boldsymbol{h}) &\geq \mathbb{E}_{\mathcal{D}}\left(q_1(\boldsymbol{h})\right) - \frac{\varepsilon\lambda_{\mathcal{D}}}{2} \\
&\quad - 2\lambda_{\mathcal{D}}\left(\sqrt{\frac{\log\frac{2}{\delta}}{2n}} + \frac{16\mathit{\Gamma}_{\mathcal{D}}^2\lambda_{\mathcal{D}}^{-2}\varepsilon^{-2} + 2\log\frac{2}{\delta}}{2n}\right)^{1/2}\left(\frac{\mathit{\Lambda}_{\mathcal{D}}}{\lambda_{\mathcal{D}}}\eta_{\mathcal{D}}^2 + \frac{\varepsilon^2}{6}\right)^{1/2} \\
&\geq \left(1 - \frac{\varepsilon}{2}\right)\mathbb{E}_{\mathcal{D}}(q_1(\boldsymbol{h})) \\
&\quad - 2\lambda_{\mathcal{D}}\left(\sqrt{\frac{\log\frac{2}{\delta}}{2n}} + \frac{16\mathit{\Gamma}_{\mathcal{D}}^2\lambda_{\mathcal{D}}^{-2}\varepsilon^{-2} + 2\log\frac{2}{\delta}}{2n}\right)^{1/2}\left(\frac{\mathit{\Lambda}_{\mathcal{D}}}{\lambda_{\mathcal{D}}}\eta_{\mathcal{D}}^2 + \frac{\varepsilon^2}{6}\right)^{1/2}.
\end{aligned}$$

It is then straightforward to deduce

$$\frac{1}{n}\sum_{i=1}^{n} q_i(\boldsymbol{h}) \geq (1-\varepsilon)\,\mathbb{E}_{\mathcal{D}}\left(q_1(\boldsymbol{h})\right)\,,$$

assuming

$$n \gtrsim \max\left\{\eta_{\mathcal{D}}^2\log\frac{2}{\delta},\ \frac{\mathit{\Gamma}_{\mathcal{D}}^2}{\lambda_{\mathcal{D}}\mathit{\Lambda}_{\mathcal{D}}}\right\}\frac{\mathit{\Lambda}_{\mathcal{D}}^2}{\lambda_{\mathcal{D}}^2}\eta_{\mathcal{D}}^2\varepsilon^{-4}\,,$$

with a sufficiently large hidden constant. □

3.2. Proof of Lemma 1

Below we provide a proof of Lemma 1.

Proof of Lemma 1. By optimality of $\widehat{\boldsymbol{x}}$ in (9) we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\max\left\{f_i^+\left(\widehat{\boldsymbol{x}}\right) - y_i, f_i^-\left(\widehat{\boldsymbol{x}}\right)\right\} \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\max\left\{f_i^+\left(\boldsymbol{x}_0\right) - y_i, f_i^-\left(\boldsymbol{x}_0\right)\right\} + \langle \boldsymbol{a}_0, \widehat{\boldsymbol{x}} - \boldsymbol{x}_0\rangle
\end{aligned}$$

$$= \frac{1}{n}\sum_{i=1}^{n} f_i^-\left(\boldsymbol{x}_0\right) + \max\left\{-\xi_i, 0\right\} + \left\langle \boldsymbol{a}_0, \widehat{\boldsymbol{x}} - \boldsymbol{x}_0\right\rangle .$$

For $i = 1, \ldots, n$ let $y_{\star i} = y_i - \xi_i = f_i^+(\boldsymbol{x}_\star) - f_i^-(\boldsymbol{x}_\star) = f_i^+(\boldsymbol{x}_0) - f_i^-(\boldsymbol{x}_0)$ and observe that

$$\max\left\{f_i^+\left(\widehat{\boldsymbol{x}}\right) - y_{\star i}, f_i^-\left(\widehat{\boldsymbol{x}}\right)\right\} \le \max\left\{f_i^+\left(\widehat{\boldsymbol{x}}\right) - y_i, f_i^-\left(\widehat{\boldsymbol{x}}\right)\right\} + \max\left\{\xi_i, 0\right\} .$$

Therefore, we deduce that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\max\left\{f_i^+\left(\widehat{\boldsymbol{x}}\right) - y_{\star i}, f_i^-\left(\widehat{\boldsymbol{x}}\right)\right\}\\
&\le \frac{1}{n}\sum_{i=1}^{n} f_i^-\left(\boldsymbol{x}_0\right) + \max\left\{\xi_i, 0\right\} + \max\left\{-\xi_i, 0\right\} + \left\langle \boldsymbol{a}_0, \widehat{\boldsymbol{x}} - \boldsymbol{x}_0\right\rangle\\
&= \frac{1}{n}\sum_{i=1}^{n} f_i^-\left(\boldsymbol{x}_0\right) + |\xi_i| + \left\langle \boldsymbol{a}_0, \widehat{\boldsymbol{x}} - \boldsymbol{x}_0\right\rangle ,
\end{aligned}$$

or equivalently

$$\frac{1}{n}\sum_{i=1}^{n}\max\left\{f_i^+\left(\widehat{\boldsymbol{x}}\right) - f_i^+(\boldsymbol{x}_0), f_i^-\left(\widehat{\boldsymbol{x}}\right) - f_i^-\left(\boldsymbol{x}_0\right)\right\} \le \frac{1}{n}\sum_{i=1}^{n}|\xi_i| + \left\langle \boldsymbol{a}_0, \widehat{\boldsymbol{x}} - \boldsymbol{x}_0\right\rangle .$$

Invoking the assumption (11) and using Cauchy-Schwarz inequality we can write

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\max\left\{f_i^+\left(\widehat{\boldsymbol{x}}\right) - f_i^+(\boldsymbol{x}_0), f_i^-\left(\widehat{\boldsymbol{x}}\right) - f_i^-\left(\boldsymbol{x}_0\right)\right\}\\
&\le \frac{1}{n}\sum_{i=1}^{n}|\xi_i| + \langle \boldsymbol{a}_0 - \frac{1}{2n}\sum_{i=1}^{n}\nabla f_i^+(\boldsymbol{x}_0) + \nabla f_i^-(\boldsymbol{x}_0), \widehat{\boldsymbol{x}} - \boldsymbol{x}_0\rangle\\
&\quad + \langle\frac{1}{2n}\sum_{i=1}^{n}\nabla f_i^+(\boldsymbol{x}_0) + \nabla f_i^-(\boldsymbol{x}_0), \widehat{\boldsymbol{x}} - \boldsymbol{x}_0\rangle\\
&\le \frac{1}{n}\sum_{i=1}^{n}|\xi_i| + \frac{1-\varepsilon}{2}\lambda_{\mathcal{D}}\left\|\widehat{\boldsymbol{x}} - \boldsymbol{x}_0\right\| + \langle\frac{1}{2n}\sum_{i=1}^{n}\nabla f_i^+(\boldsymbol{x}_0) + \nabla f_i^-(\boldsymbol{x}_0), \widehat{\boldsymbol{x}} - \boldsymbol{x}_0\rangle .
\end{aligned}$$

Rearranging the terms gives the equivalent inequality

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\max\left\{f_i^+\left(\widehat{\boldsymbol{x}}\right) - f_i^+(\boldsymbol{x}_0), f_i^-\left(\widehat{\boldsymbol{x}}\right) - f_i^-\left(\boldsymbol{x}_0\right)\right\} - \frac{1}{2}\langle\nabla f_i^+(\boldsymbol{x}_0) + \nabla f_i^-(\boldsymbol{x}_0), \widehat{\boldsymbol{x}} - \boldsymbol{x}_0\rangle\\
&\le \frac{1}{n}\sum_{i=1}^{n}|\xi_i| + \frac{1-\varepsilon}{2}\lambda_{\mathcal{D}}\left\|\widehat{\boldsymbol{x}} - \boldsymbol{x}_0\right\| .
\end{aligned}$$

Observe that

$$
\begin{aligned}
&\max\left\{f_i^+(\widehat{\boldsymbol{x}})-f_i^+(\boldsymbol{x}_0),f_i^-(\widehat{\boldsymbol{x}})-f_i^-(\boldsymbol{x}_0)\right\}-\frac{1}{2}\langle\nabla f_i^+(\boldsymbol{x}_0)+\nabla f_i^-(\boldsymbol{x}_0),\widehat{\boldsymbol{x}}-\boldsymbol{x}_0\rangle\\
&\geq\frac{1}{2}\left|\langle\nabla f_i^+(\boldsymbol{x}_0)+\nabla f_i^-(\boldsymbol{x}_0),\widehat{\boldsymbol{x}}-\boldsymbol{x}_0\rangle\right|\,.
\end{aligned}
$$

Using the assumption that (12) holds, we obtain

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n\max\left\{f_i^+(\widehat{\boldsymbol{x}})-f_i^+(\boldsymbol{x}_0),f_i^-(\widehat{\boldsymbol{x}})-f_i^-(\boldsymbol{x}_0)\right\}-\frac{1}{2}\langle\nabla f_i^+(\boldsymbol{x}_0)+\nabla f_i^-(\boldsymbol{x}_0),\widehat{\boldsymbol{x}}-\boldsymbol{x}_0\rangle\\
&\geq\frac{1}{2n}\sum_{i=1}^n\left|\langle\nabla f_i^+(\boldsymbol{x}_0)+\nabla f_i^-(\boldsymbol{x}_0),\widehat{\boldsymbol{x}}-\boldsymbol{x}_0\rangle\right|\\
&\geq\left(1-\frac{\varepsilon+\varepsilon_0}{2}\right)\lambda_{\mathcal{D}}\left\|\widehat{\boldsymbol{x}}-\boldsymbol{x}_0\right\|\,.
\end{aligned}
$$

Therefore, we conclude that

$$
\left(1-\frac{\varepsilon+\varepsilon_0}{2}\right)\lambda_{\mathcal{D}}\left\|\widehat{\boldsymbol{x}}-\boldsymbol{x}_0\right\|\leq\frac{1}{n}\sum_{i=1}^n|\xi_i|+\frac{1-\varepsilon}{2}\lambda_{\mathcal{D}}\left\|\widehat{\boldsymbol{x}}-\boldsymbol{x}_0\right\|\,,
$$

which, since $\varepsilon_0\in[0,1)$, is equivalent to

$$
\left\|\widehat{\boldsymbol{x}}-\boldsymbol{x}_0\right\|\leq\frac{\frac{1}{n}\sum_{i=1}^n|\xi_i|}{\frac{1-\varepsilon_0}{2}\lambda_{\mathcal{D}}}\,. \qquad \square
$$

4. Application to bilinear regression

In this section we apply the general result above to the problem of bilinear regression. Suppose that the vectors $\boldsymbol{x}_\star^{(1)}$ and $\boldsymbol{x}_\star^{(2)}$ are observed through the bilinear measurements

$$
y_i=\langle\boldsymbol{a}_i^{(1)},\boldsymbol{x}_\star^{(1)}\rangle\langle\boldsymbol{x}_\star^{(2)},\boldsymbol{a}_i^{(2)}\rangle,\qquad\qquad i=1,\ldots,n\,, \tag{18}
$$

with known vector pairs $(\boldsymbol{a}_i^{(1)},\boldsymbol{a}_i^{(2)})$. In bilinear regression, the goals is to recover $\boldsymbol{x}_\star^{(1)}$ and $\boldsymbol{x}_\star^{(2)}$ (up to the inherent ambiguities) from the above measurements.

To apply our general framework, we introduce an equivalent formulation of the bilinear observations that is compatible with the DC observation model of (1). Let $\boldsymbol{x}_\star$ denote the concatenation of $\boldsymbol{x}_\star^{(1)}\in\mathbb{R}^{d_1}\backslash\{\boldsymbol{0}\}$ and $\boldsymbol{x}_\star^{(2)}\in\mathbb{R}^{d_2}\backslash\{\boldsymbol{0}\}$. Similarly, for $i=1,\ldots,n$ let $\boldsymbol{a}_i^{\pm}$ denote the concatenation of $\boldsymbol{a}_i^{(1)}$ and $\pm\boldsymbol{a}_i^{(2)}$. It is easy to verify that the bilinear measurements above can also be expressed in the form

$$
y_i=\frac{1}{4}\left|\langle\boldsymbol{a}_i^+,\boldsymbol{x}_\star\rangle\right|^2-\frac{1}{4}\left|\langle\boldsymbol{a}_i^-,\boldsymbol{x}_\star\rangle\right|^2\,,
$$

which is a special case of the DC observation model (1) with

$$f_i^+(\boldsymbol{x}) = \frac{1}{4}\left|\langle \boldsymbol{a}_i^+, \boldsymbol{x}\rangle\right|^2 \qquad \text{and} \qquad f_i^-(\boldsymbol{x}) = \frac{1}{4}\left|\langle \boldsymbol{a}_i^-, \boldsymbol{x}\rangle\right|^2 . \tag{19}$$

The problem setup and additional notations are as follows. Denote the $\ell \times \ell$ identity matrix by $\boldsymbol{I}_\ell$. For $k = 1, 2$, let $\boldsymbol{a}_1^{(k)}, \boldsymbol{a}_2^{(k)}, \ldots, \boldsymbol{a}_n^{(k)}$ be i.i.d. copies of $\boldsymbol{a}^{(k)} \sim \mathrm{Normal}(\boldsymbol{0}, \boldsymbol{I}_{d_k})$ with $\boldsymbol{a}_i^{(1)}$ and $\boldsymbol{a}_i^{(2)}$ also drawn independently for all $1 \le i \le n$. Similar to the definition of $\boldsymbol{a}_i^{\pm}$s above, we also denote the concatenation of $\boldsymbol{a}^{(1)}$ and $\pm\boldsymbol{a}^{(2)}$ by $\boldsymbol{a}^{\pm}$. The functions $f^{\pm}$ are defined analogous to $f_i^{\pm}$ with $\boldsymbol{a}^{\pm}$ replacing $\boldsymbol{a}_i^{\pm}$. For brevity, we set $d = d_1 + d_2$. Furthermore, some of the unspecified constants in the derivations below are overloaded and may take different values from line to line. For any vector $\boldsymbol{x} \in \mathbb{R}^d$ with partitions as $\boldsymbol{x}^{(1)} \in \mathbb{R}^{d_1}$ and $\boldsymbol{x}^{(2)} \in \mathbb{R}^{d_2}$, we use the notation $\boldsymbol{x}^-$ to denote the concatenation of $\boldsymbol{x}^{(1)}$ and $-\boldsymbol{x}^{(2)}$.

Evidently, any reciprocal scaling of $\boldsymbol{x}_\star^{(1)}$ and $\boldsymbol{x}_\star^{(2)}$ is also consistent with the bilinear measurements (18) and will be considered a valid solution. Throughout this section, we choose $\boldsymbol{x}_\star$ to be a "balanced" solution meaning that $\left\|\boldsymbol{x}_\star^{(1)}\right\| = \left\|\boldsymbol{x}_\star^{(2)}\right\|$. Also, without loss of generality, we may assume $\langle \boldsymbol{a}_0, \boldsymbol{x}_\star\rangle \ge 0$. The accuracy, however, is measured with respect to a closest consistent solution

$$\widehat{\boldsymbol{x}}_\star \in \operatorname*{argmin}_{\boldsymbol{x}} \left\{ \|\widehat{\boldsymbol{x}} - \boldsymbol{x}\| \,:\, \boldsymbol{x}^{(1)} = t\boldsymbol{x}_\star^{(1)}, \boldsymbol{x}^{(2)} = t^{-1}\boldsymbol{x}_\star^{(2)}, t \in \mathbb{R}\backslash\{0\} \right\} . \tag{20}$$

To state the accuracy guarantees for (9) in the described bilinear regression problem, we first bound the important quantities given by (4), (5), (6), and (7) for the restriction set

$$\mathcal{S} = \left\{ \boldsymbol{z} \in \mathbb{S}^{d-1} \,:\, \left|\langle \boldsymbol{z}, {\boldsymbol{x}_\star}^- \rangle\right| \le \frac{1}{2}\|\boldsymbol{x}_\star\| \right\} . \tag{21}$$

This choice of $\mathcal{S}$ allows us to find a nontrivial bound for $\lambda_{\mathcal{D}}$ and it is important in the proof of Theorem 2.

4.1. Quantifying $\lambda_{\mathcal{D}}$, $\varLambda_{\mathcal{D}}$, $\varGamma_{\mathcal{D}}$, and $\eta_{\mathcal{D}}$

Let $\boldsymbol{h} \in \mathcal{S}$ be a vector partitioned into $\boldsymbol{h}^{(1)} \in \mathbb{R}^{d_1}$ and $\boldsymbol{h}^{(2)} \in \mathbb{R}^{d_2}$. We can write

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\left|\langle \nabla f^+(\boldsymbol{x}_\star)) - \nabla f^-(\boldsymbol{x}_\star)), \boldsymbol{h}\rangle\right| &= \mathbb{E}_{\mathcal{D}}\left|\langle \boldsymbol{a}^{(1)} \otimes \boldsymbol{a}^{(2)}, \boldsymbol{h}^{(1)} \otimes \boldsymbol{x}_\star^{(2)} + \boldsymbol{x}_\star^{(1)} \otimes \boldsymbol{h}^{(2)}\rangle\right| \\
&= \sqrt{\frac{2}{\pi}}\,\mathbb{E}_{\mathcal{D}}\left\|\left(\boldsymbol{h}^{(1)} \otimes \boldsymbol{x}_\star^{(1)} + \boldsymbol{x}_\star^{(1)} \otimes \boldsymbol{h}^{(2)}\right)\boldsymbol{a}^{(2)}\right\| .
\end{aligned}$$

Using the Cauchy-Schwarz inequality and Lemma 7 in the appendix, respectively, we obtain

$$\mathbb{E}_{\mathcal{D}}\left|\langle \boldsymbol{a}^{(1)} \otimes \boldsymbol{a}^{(2)}, \boldsymbol{h}^{(1)} \otimes \boldsymbol{x}_\star^{(2)} + \boldsymbol{x}_\star^{(1)} \otimes \boldsymbol{h}^{(2)}\rangle\right| \le \sqrt{\frac{2}{\pi}}\left\|\boldsymbol{h}^{(1)} \otimes \boldsymbol{x}_\star^{(2)} + \boldsymbol{x}_\star^{(1)} \otimes \boldsymbol{h}^{(2)}\right\|_{\mathrm{F}} ,$$

and

$$\mathbb{E}_{\mathcal{D}}\left|\langle \boldsymbol{a}^{(1)}\otimes\boldsymbol{a}^{(2)},\boldsymbol{h}^{(1)}\otimes\boldsymbol{x}_\star^{(2)}+\boldsymbol{x}_\star^{(1)}\otimes\boldsymbol{h}^{(2)}\rangle\right| \geq \frac{2}{\pi}\left\|\boldsymbol{h}^{(1)}\otimes\boldsymbol{x}_\star^{(2)}+\boldsymbol{x}_\star^{(1)}\otimes\boldsymbol{h}^{(2)}\right\|_{\mathrm{F}}.$$

Observe that

$$\begin{aligned}
&\left\|\boldsymbol{h}^{(1)}\otimes\boldsymbol{x}_\star^{(2)}+\boldsymbol{x}_\star^{(1)}\otimes\boldsymbol{h}^{(2)}\right\|_{\mathrm{F}}^2\\
&=\left\|\boldsymbol{h}^{(1)}\otimes\boldsymbol{x}_\star^{(2)}\right\|_{\mathrm{F}}^2+\left\|\boldsymbol{x}_\star^{(1)}\otimes\boldsymbol{h}^{(2)}\right\|_{\mathrm{F}}^2+2\langle\boldsymbol{h}^{(1)}\otimes\boldsymbol{x}_\star^{(2)},\,\boldsymbol{x}_\star^{(1)}\otimes\boldsymbol{h}^{(2)}\rangle\\
&=\left\|\boldsymbol{h}^{(1)}\right\|^2\left\|\boldsymbol{x}_\star^{(2)}\right\|^2+\left\|\boldsymbol{h}^{(2)}\right\|^2\left\|\boldsymbol{x}_\star^{(1)}\right\|^2+2\langle\boldsymbol{h}^{(1)},\boldsymbol{x}_\star^{(1)}\rangle\langle\boldsymbol{h}^{(2)},\boldsymbol{x}_\star^{(2)}\rangle\\
&=\langle\boldsymbol{X}_\star,\boldsymbol{h}\otimes\boldsymbol{h}\rangle\,,
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{X}_\star &= \begin{bmatrix}\left\|\boldsymbol{x}_\star^{(2)}\right\|^2\boldsymbol{I}_{d_1} & \boldsymbol{x}_\star^{(1)}\otimes\boldsymbol{x}_\star^{(2)}\\ \boldsymbol{x}_\star^{(2)}\otimes\boldsymbol{x}_\star^{(1)} & \left\|\boldsymbol{x}_\star^{(1)}\right\|^2\boldsymbol{I}_{d_2}\end{bmatrix}\\
&=\frac{1}{2}\left\|\boldsymbol{x}_\star\right\|^2\boldsymbol{I}+\frac{1}{2}\boldsymbol{x}_\star\otimes\boldsymbol{x}_\star-\frac{1}{2}\boldsymbol{x}_\star^-\otimes\boldsymbol{x}_\star^-\,.
\end{aligned}$$

Thus, we have

$$\frac{1}{\pi}\sqrt{\langle\boldsymbol{X}_\star,\boldsymbol{h}\otimes\boldsymbol{h}\rangle}\leq\frac{1}{2}\mathbb{E}_{\mathcal{D}}\left|\langle\nabla f^+(\boldsymbol{x}_\star)-\nabla f^-(\boldsymbol{x}_\star),\boldsymbol{h}\rangle\right|\leq\frac{1}{\sqrt{2\pi}}\sqrt{\langle\boldsymbol{X}_\star,\boldsymbol{h}\otimes\boldsymbol{h}\rangle}\,. \tag{22}$$

Since $\boldsymbol{h}\in\mathcal{S}$, by definition $|\langle\boldsymbol{h},\boldsymbol{x}_\star^-\rangle|\leq\frac{1}{2}\left\|\boldsymbol{x}_\star\right\|$, and it is easy to verify that

$$\frac{3}{8}\left\|\boldsymbol{x}_\star\right\|^2\leq\langle\boldsymbol{X}_\star,\boldsymbol{h}\otimes\boldsymbol{h}\rangle\leq\left\|\boldsymbol{x}_\star\right\|^2\,.$$

Therefore, (22) implies that

$$\frac{\sqrt{6}}{4\pi}\left\|\boldsymbol{x}_\star\right\|\leq\frac{1}{2}\mathbb{E}_{\mathcal{D}}\left|\langle\nabla f^+(\boldsymbol{x}_\star)-\nabla f^-(\boldsymbol{x}_\star),\boldsymbol{h}\rangle\right|\leq\frac{1}{\sqrt{2\pi}}\left\|\boldsymbol{x}_\star\right\|\,,$$

which also means

$$\frac{\sqrt{6}}{4\pi}\left\|\boldsymbol{x}_\star\right\|\leq\lambda_{\mathcal{D}}\leq\varLambda_{\mathcal{D}}\leq\frac{1}{\sqrt{2\pi}}\left\|\boldsymbol{x}_\star\right\|\,. \tag{23}$$

Note that without the restriction of $\boldsymbol{h}$ to the prescribed set $\mathcal{S}$ in (21), we could have had $\lambda_{\mathcal{D}}=0$, which leads to vacuous bounds.

We can also evaluate $\Gamma_{\mathcal{D}}$ as

$$\begin{aligned}
\Gamma_{\mathcal{D}} &= \frac{1}{2}\sqrt{\mathbb{E}_{\mathcal{D}}\left(\|\nabla f^{+}(\boldsymbol{x}_\star) - \nabla f^{-}(\boldsymbol{x}_\star)\|^2\right)} \\
&= \frac{1}{2}\sqrt{\mathbb{E}_{\mathcal{D}}\left(\left\|\boldsymbol{a}^{(1)}\langle \boldsymbol{a}^{(2)}, \boldsymbol{x}_\star^{(2)}\rangle\right\|^2 + \left\|\boldsymbol{a}^{(2)}\langle \boldsymbol{a}^{(1)}, \boldsymbol{x}_\star^{(1)}\rangle\right\|^2\right)} \\
&= \frac{1}{2}\sqrt{\mathbb{E}_{\mathcal{D}}\left(d_1\left|\langle \boldsymbol{a}^{(2)}, \boldsymbol{x}_\star^{(2)}\rangle\right|^2 + d_2\left|\langle \boldsymbol{a}^{(1)}, \boldsymbol{x}_\star^{(1)}\rangle\right|^2\right)} \\
&= \frac{1}{2}\sqrt{d_1\left\|\boldsymbol{x}_\star^{(2)}\right\|^2 + d_2\left\|\boldsymbol{x}_\star^{(1)}\right\|^2} \\
&= \frac{1}{4}\sqrt{d}\,\|\boldsymbol{x}_\star\|\ .
\end{aligned} \tag{24}$$

Furthermore, using the lower bound in (22), we can write

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left(\left|\langle \nabla f^{+}(\boldsymbol{x}_\star) - \nabla f^{-}(\boldsymbol{x}_\star), \boldsymbol{h}\rangle\right|^2\right) \\
&= \mathbb{E}_{\mathcal{D}}\left(\left|\langle \boldsymbol{h}^{(1)}, \boldsymbol{a}^{(1)}\rangle\langle \boldsymbol{a}^{(2)}, \boldsymbol{x}_\star^{(2)}\rangle + \langle \boldsymbol{x}_\star^{(1)}, \boldsymbol{a}^{(1)}\rangle\langle \boldsymbol{a}^{(2)}, \boldsymbol{h}^{(2)}\rangle\right|^2\right) \\
&= \left\|\boldsymbol{h}^{(1)}\right\|^2\left\|\boldsymbol{x}_\star^{(2)}\right\|^2 + 2\langle \boldsymbol{h}^{(1)}, \boldsymbol{x}_\star^{(1)}\rangle\langle \boldsymbol{h}^{(2)}, \boldsymbol{x}_\star^{(2)}\rangle + \left\|\boldsymbol{h}^{(2)}\right\|^2\left\|\boldsymbol{x}_\star^{(1)}\right\|^2 \\
&= \left\|\boldsymbol{h}^{(1)} \otimes \boldsymbol{x}_\star^{(2)} + \boldsymbol{x}_\star^{(1)} \otimes \boldsymbol{h}^{(2)}\right\|_{\mathrm{F}}^2 \\
&= \langle \boldsymbol{X}_\star, \boldsymbol{h} \otimes \boldsymbol{h}\rangle \\
&\le \frac{\pi^2}{4}\left(\mathbb{E}_{\mathcal{D}}\left(\left|\langle \nabla f^{+}(\boldsymbol{x}_\star) - \nabla f^{-}(\boldsymbol{x}_\star), \boldsymbol{h}\rangle\right|\right)\right)^2\ .
\end{aligned}$$

Thus, we are guaranteed to have

$$\eta_{\mathcal{D}} \le \frac{\pi}{2}\,. \tag{25}$$

4.2. Accuracy guarantee

To prove accuracy of (9) in the considered bilinear regression problem, we need to apply Lemma 1 with $\widehat{\boldsymbol{x}}_\star$ given by (20) as the reference ground truth. Therefore, we also use Proposition 1 with a nontrivial restriction set $\mathcal{S}$ in our analysis to establish an inequality of the form (12). Because $\widehat{\boldsymbol{x}}_\star$ depends on the observations, however, it cannot be used as the reference ground truth in Proposition 1. Lemma 4 in the appendix shows that the bound obtained using Proposition 1, with the balanced ground truth (i.e., $\boldsymbol{x}_\star$) as the reference point and the restrictions set (21), can be extended to the cases where other equivalent solutions are considered as the reference ground truth.

For any $t \in \mathbb{R}\backslash 0$, let $D_t : \mathbb{R}^d \to \mathbb{R}^d$ be the reciprocal scaling operator described by

$$D_t(\boldsymbol{x}) = \begin{bmatrix} t\boldsymbol{x}^{(1)} \\ t^{-1}\boldsymbol{x}^{(2)} \end{bmatrix},$$

where $\boldsymbol{x}$ is the concatenation of $\boldsymbol{x}^{(1)} \in \mathbb{R}^{d_1}$ and $\boldsymbol{x}^{(2)} \in \mathbb{R}^{d_2}$. Furthermore, for $\theta \in [0,1]$, we define the cone $\mathcal{K}_{t,\theta}$ as

$$\mathcal{K}_{t,\theta} \stackrel{\text{def}}{=} \left\{\boldsymbol{h} \,:\, \left|\langle D_t(\boldsymbol{x}_\star^-), \boldsymbol{h}\rangle\right| \leq \theta \left\|D_t(\boldsymbol{x}_\star^-)\right\| \|\boldsymbol{h}\|\right\}. \tag{26}$$

This specific choice of the cone $\mathcal{K}_{t,\theta}$ is important for the following reason: If, for some $t_{\text{opt}} \in \mathbb{R}\backslash\{0\}$, $\widehat{\boldsymbol{x}}_\star = D_{t_{\text{opt}}}(\boldsymbol{x}_\star)$ is the solution described by (20), then elementary calculus shows that

$$\langle D_{t_{\text{opt}}}(\boldsymbol{x}_\star^-), \widehat{\boldsymbol{x}} - \widehat{\boldsymbol{x}}_\star\rangle = 0,$$

which means that $\widehat{\boldsymbol{x}} - \widehat{\boldsymbol{x}}_\star \in \mathcal{K}_{t_{\text{opt}},0}$. Leveraging this property we can show that $D_{t_{\text{opt}}^{-1}}(\widehat{\boldsymbol{x}} - \widehat{\boldsymbol{x}}_\star) \in \mathcal{K}_{1,\frac{1}{2}}$ which allows us to invoke Lemma 4.

The following theorem establishes the sample complexity of (9) for exact recovery in the noiseless bilinear regression problem.

Theorem 2 (bilinear regression)**.** *We observe n noiseless bilinear measurements (18) corresponding to the functions $f_i^{\pm}$ described by (19). Suppose that (8) holds for some $\varepsilon \in [7/8, 1)$. If the number of measurements obeys*

$$n \gtrsim \varepsilon^{-4} \max\left\{d, \log\frac{8}{\delta}\right\}, \tag{27}$$

with a sufficiently large hidden constant, then with probability $\geq 1-\delta$, the solution to (9) coincides with $\widehat{\boldsymbol{x}}_\star$ given by (20).

Proof. Because of (27), we may assume

$$C \max\left\{\sqrt{\frac{d}{n}} + \sqrt{\frac{\log\frac{8}{\delta}}{n}}, \left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log\frac{8}{\delta}}{n}}\right)^2\right\} \leq \frac{1}{9}, \tag{28}$$

where $C > 0$ is the constant in Lemma 6. With $\varepsilon' = 6\varepsilon - 5$, it follows from (40) in Lemma 6 that

$$\begin{aligned}\left\|\boldsymbol{a}_0 - \frac{1}{2n}\sum_{i=1}^n \nabla f_i^+(\widehat{\boldsymbol{x}}_\star) + \nabla f_i^-(\widehat{\boldsymbol{x}}_\star)\right\| &\leq 6\left\|\boldsymbol{a}_0 - \frac{1}{2n}\sum_{i=1}^n \nabla f_i^+(\boldsymbol{x}_\star) + \nabla f_i^-(\boldsymbol{x}_\star)\right\| \\ &\leq 3(1-\varepsilon)\lambda_{\mathcal{D}} = \frac{1-\varepsilon'}{2}\lambda_{\mathcal{D}},\end{aligned} \tag{29}$$

holds with probability $\geq 1 - \delta/2$.

Furthermore, the approximations (23), (24), and (25) show that because of (27), Proposition 1, with $\boldsymbol{x}_\star$ taken as the reference ground truth, ensures

$$\frac{1}{n}\sum_{i=1}^{n}\left|\langle\nabla f_i^+(\boldsymbol{x}_\star)-\nabla f_i^-(\boldsymbol{x}_\star),\boldsymbol{h}\rangle\right| \geq (1-\frac{1}{2}\varepsilon)\mathbb{E}_{\mathcal{D}}\left(\left|\langle\nabla f^+(\boldsymbol{x}_\star)-\nabla f^-(\boldsymbol{x}_\star),\boldsymbol{h}\rangle\right|\right)\,,$$

to hold for all $\boldsymbol{h}\in\mathcal{S}=\mathbb{S}^{d-1}\cap\mathcal{K}_{1,\frac{1}{2}}$ with probability $\geq 1-\delta/2$. On the same event, if t_{opt}, defined as above through $\widehat{\boldsymbol{x}}_\star = D_{t_{\mathrm{opt}}}(\boldsymbol{x}_\star)$, obeys $\sqrt{2/3}\leq|t_{\mathrm{opt}}|\leq\sqrt{3/2}$, then Lemma 4 implies that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left|\langle\nabla f_i^+(\widehat{\boldsymbol{x}}_\star)-\nabla f_i^-(\widehat{\boldsymbol{x}}_\star),\boldsymbol{h}\rangle\right| \\
&\geq (1-\frac{1}{2}\varepsilon)\mathbb{E}_{\mathcal{D}}\left(\left|\langle\nabla f^+(\widehat{\boldsymbol{x}}_\star)-\nabla f^-(\widehat{\boldsymbol{x}}_\star),\boldsymbol{h}\rangle\right|\right) \\
&= (1-\frac{1}{2}\varepsilon)\mathbb{E}_{\mathcal{D}}\left(\left|\langle\nabla f^+(\boldsymbol{x}_\star)-\nabla f^-(\boldsymbol{x}_\star),D_{t_{\mathrm{opt}}^{-1}}(\boldsymbol{h})\rangle\right|\right)\,,
\end{aligned}$$

for all $\boldsymbol{h}\in\mathbb{S}^{d-1}\cap\mathcal{K}_{t_{\mathrm{opt}},0}$. Note that the expectations on the right-hand side are only with respect to $f^{\pm}$; the vector $\widehat{\boldsymbol{x}}_\star$ and the scalar t_{opt} should be treated as deterministic variables. Using (4) we obtain

$$\frac{1}{2n}\sum_{i=1}^{n}\left|\langle\nabla f_i^+(\widehat{\boldsymbol{x}}_\star)-\nabla f_i^-(\widehat{\boldsymbol{x}}_\star),\boldsymbol{h}\rangle\right| \geq (1-\frac{1}{2}\varepsilon)\lambda_{\mathcal{D}}\left\|D_{t_{\mathrm{opt}}^{-1}}(\boldsymbol{h})\right\|\,.$$

Therefore, the bound

$$\left\|D_{t_{\mathrm{opt}}^{-1}}(\boldsymbol{h})\right\|^2 = t_{\mathrm{opt}}^{-2}\left\|\boldsymbol{h}^{(1)}\right\|^2 + t_{\mathrm{opt}}^2\left\|\boldsymbol{h}^{(2)}\right\|^2 \geq \min\left\{t_{\mathrm{opt}}^{-2},t_{\mathrm{opt}}^2\right\}\|\boldsymbol{h}\|^2$$

and the choice of $\varepsilon'=6\varepsilon-5$ made above yield

$$\frac{1}{2n}\sum_{i=1}^{n}\left|\langle\nabla f_i^+(\widehat{\boldsymbol{x}}_\star)-\nabla f_i^-(\widehat{\boldsymbol{x}}_\star),\boldsymbol{h}\rangle\right| \geq (1-\frac{5+\varepsilon'}{12})\min\left\{\left|t_{\mathrm{opt}}^{-1}\right|,|t_{\mathrm{opt}}|\right\}\lambda_{\mathcal{D}}\|\boldsymbol{h}\|\,. \tag{30}$$

It only remains to bound $|t_{\mathrm{opt}}|$ appropriately, not only to approximate the term $\min\left\{\left|t_{\mathrm{opt}}^{-1}\right|,|t_{\mathrm{opt}}|\right\}$, but also to satisfy the condition $\sqrt{2/3}\leq|t_{\mathrm{opt}}|\leq\sqrt{3/2}$ used previously. First, we show that $\|\widehat{\boldsymbol{x}}_\star-\boldsymbol{x}_\star\|$ is small through Lemma 6. Note that the previous application of Lemma 6, in which we had (28), also guarantees

$$\|\widehat{\boldsymbol{x}}_\star-\boldsymbol{x}_\star\| \leq \frac{48}{5}\left\|\boldsymbol{a}_0-\frac{1}{2n}\sum_{i=1}^{n}\nabla f_i^+(\boldsymbol{x}_\star)+\nabla f_i^-(\boldsymbol{x}_\star)\right\| \leq \frac{4}{5}(1-\varepsilon')\lambda_{\mathcal{D}}\,.$$

Therefore, using the upper bound in (23), we get

$$\|\widehat{\boldsymbol{x}}_\star-\boldsymbol{x}_\star\| \leq \frac{8}{25}(1-\varepsilon')\|\boldsymbol{x}_\star\|$$

Because $\widehat{\boldsymbol{x}}_\star = D_{t_{\mathrm{opt}}}(\boldsymbol{x}_\star)$, and $\boldsymbol{x}_\star$ is balanced, we also have

$$\begin{aligned}\|\widehat{\boldsymbol{x}}_\star - \boldsymbol{x}_\star\|^2 &= |t_{\mathrm{opt}} - 1|^2 \left\|\boldsymbol{x}_\star^{(1)}\right\|^2 + \left|t_{\mathrm{opt}}^{-1} - 1\right|^2 \left\|\boldsymbol{x}_\star^{(2)}\right\|^2 \\ &\geq \frac{1}{2} \max\left\{\left|t_{\mathrm{opt}}^{-1} - 1\right|^2, |t_{\mathrm{opt}} - 1|^2\right\} \|\boldsymbol{x}_\star\|^2 ,\end{aligned}$$

which together with the previous inequality imply

$$\max\left\{\left|t_{\mathrm{opt}}^{-1} - 1\right|, |t_{\mathrm{opt}} - 1|\right\} \leq \frac{1}{2}(1 - \varepsilon') .$$

Therefore, we obtain

$$\begin{aligned}\min\left\{\left|t_{\mathrm{opt}}^{-1}\right|, |t_{\mathrm{opt}}|\right\} &= \left(\max\left\{\left|t_{\mathrm{opt}}^{-1}\right|, |t_{\mathrm{opt}}|\right\}\right)^{-1} \\ &\geq \left(1 + \max\left\{\left|t_{\mathrm{opt}}^{-1} - 1\right|, |t_{\mathrm{opt}} - 1|\right\}\right)^{-1} \\ &\geq \left(1 + \frac{1}{2}(1 - \varepsilon')\right)^{-1} \\ &\geq \left(1 - \frac{5 + \varepsilon'}{12}\right)^{-1} (1 - \varepsilon') ,\end{aligned}$$

where the fourth line holds since $1/4 \leq \varepsilon' = 6\varepsilon - 5 \leq 1$. Using the derived bound in (30) yields

$$\frac{1}{2n} \sum_{i=1}^{n} \left|\langle \nabla f_i^+(\widehat{\boldsymbol{x}}_\star) - \nabla f_i^-(\widehat{\boldsymbol{x}}_\star), \boldsymbol{h}\rangle\right| \geq (1 - \varepsilon')\lambda_{\mathcal{D}} \|\boldsymbol{h}\| .$$

Hence, in view of (11), we may invoke Lemma 1 with $\boldsymbol{x}_0 = \widehat{\boldsymbol{x}}_\star$, ε' in place of ε, and $\varepsilon_0 = \varepsilon'$, and prove the exact recovery (i.e., $\widehat{\boldsymbol{x}} = \widehat{\boldsymbol{x}}_\star$), which occurs with probability $\geq 1 - \delta$. □

4.3. Approximation oracle

We provide a computationally tractable procedure that can serve as the approximation oracle discussed in Section 1.1 and requires no information other than the given measurements (18). This approach basically follows the idea of "spectral initialization" used for the nonconvex phase retrieval and blind deconvolution methods [12, 32, 42]; refinements of this approach can be found in [16, 34, 41] and references therein. We use the measurements to find an approximation $\boldsymbol{a}_0$ of $\boldsymbol{x}_\star/2$ and show, by Lemma 5, that $\boldsymbol{x}_\star/2$ itself is an approximation for $\frac{1}{2n}\sum_{i=1}^{n} \nabla f_i^+(\boldsymbol{x}_\star) + \nabla f_i^-(\boldsymbol{x}_\star)$.

Let $\lambda_{\max}$ and $\boldsymbol{v}_{\max}$ be respectively the leading eigenvalue and eigenvector of

$$\boldsymbol{S}_n \stackrel{\mathrm{def}}{=} \frac{1}{2n} \sum_{i=1}^{n} y_i \left(\boldsymbol{a}_i^+ \otimes \boldsymbol{a}_i^+ - \boldsymbol{a}_i^- \otimes \boldsymbol{a}_i^-\right) .$$

The fact that $\boldsymbol{S}_n$ has an all-zero diagonal and is symmetric ensures that $\lambda_{\max} \geq 0$. We show that

$$\boldsymbol{a}_0 = \left(\frac{\lambda_{\max}}{2}\right)^{1/2} \boldsymbol{v}_{\max} \tag{31}$$

meets the required condition (8) with high probability. To this end, first we show that $\boldsymbol{S}_n$ is well-concentrated around its expectation. Observe that $\langle \boldsymbol{a}_i^-, \boldsymbol{x}_\star \rangle = \langle \boldsymbol{a}_i^+, \boldsymbol{x}_\star^- \rangle$ and similarly $\langle \boldsymbol{a}_i^-, \boldsymbol{x}_\star^- \rangle = \langle \boldsymbol{a}_i^+, \boldsymbol{x}_\star \rangle$. Thus, we obtain

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}} \boldsymbol{S}_n &= \mathbb{E}_{\mathcal{D}} \left(\frac{1}{8} \left(\left|\langle \boldsymbol{a}_i^+, \boldsymbol{x}_\star \rangle\right|^2 - \left|\langle \boldsymbol{a}_i^-, \boldsymbol{x}_\star \rangle\right|^2 \right) \left(\boldsymbol{a}_i^+ \otimes \boldsymbol{a}_i^+ - \boldsymbol{a}_i^- \otimes \boldsymbol{a}_i^- \right) \right) \\
&= \frac{1}{8} \mathbb{E}_{\mathcal{D}} \left(\left|\langle \boldsymbol{a}_i^+, \boldsymbol{x}_\star \rangle\right|^2 \boldsymbol{a}_i^+ \otimes \boldsymbol{a}_i^+ \right) + \frac{1}{8} \mathbb{E}_{\mathcal{D}} \left(\left|\langle \boldsymbol{a}_i^-, \boldsymbol{x}_\star \rangle\right|^2 \boldsymbol{a}_i^- \otimes \boldsymbol{a}_i^- \right) \\
&\quad - \frac{1}{8} \mathbb{E}_{\mathcal{D}} \left(\left|\langle \boldsymbol{a}_i^+, \boldsymbol{x}_\star^- \rangle\right|^2 \boldsymbol{a}_i^+ \otimes \boldsymbol{a}_i^+ \right) - \frac{1}{8} \mathbb{E}_{\mathcal{D}} \left(\left|\langle \boldsymbol{a}_i^-, \boldsymbol{x}_\star^- \rangle\right|^2 \boldsymbol{a}_i^- \otimes \boldsymbol{a}_i^- \right) \\
&= \frac{1}{4} \left(2\boldsymbol{x}_\star \otimes \boldsymbol{x}_\star + \|\boldsymbol{x}_\star\|^2 \boldsymbol{I} \right) - \frac{1}{4} \left(2\boldsymbol{x}_\star^- \otimes \boldsymbol{x}_\star^- + \left\|\boldsymbol{x}_\star^-\right\|^2 \boldsymbol{I} \right) \\
&= \frac{1}{2} \left(\boldsymbol{x}_\star \otimes \boldsymbol{x}_\star - \boldsymbol{x}_\star^- \otimes \boldsymbol{x}_\star^- \right) .
\end{aligned}$$

By the triangle inequality we can write

$$\begin{aligned}
\|\boldsymbol{S}_n - \mathbb{E}_{\mathcal{D}} \boldsymbol{S}_n\|_{\mathrm{op}} \leq \frac{1}{8} &\left\| \frac{1}{n} \sum_{i=1}^n \left|\langle \boldsymbol{a}_i^+, \boldsymbol{x}_\star \rangle\right|^2 \boldsymbol{a}_i^+ \otimes \boldsymbol{a}_i^+ - 2\boldsymbol{x}_\star \otimes \boldsymbol{x}_\star - \|\boldsymbol{x}_\star\|^2 \boldsymbol{I} \right\|_{\mathrm{op}} \\
&+ \frac{1}{8} \left\| \frac{1}{n} \sum_{i=1}^n \left|\langle \boldsymbol{a}_i^-, \boldsymbol{x}_\star \rangle\right|^2 \boldsymbol{a}_i^- \otimes \boldsymbol{a}_i^- - 2\boldsymbol{x}_\star \otimes \boldsymbol{x}_\star - \|\boldsymbol{x}_\star\|^2 \boldsymbol{I} \right\|_{\mathrm{op}} \\
&+ \frac{1}{8} \left\| \frac{1}{n} \sum_{i=1}^n \left|\langle \boldsymbol{a}_i^+, \boldsymbol{x}_\star^- \rangle\right|^2 \boldsymbol{a}_i^+ \otimes \boldsymbol{a}_i^+ - 2\boldsymbol{x}_\star^- \otimes \boldsymbol{x}_\star^- - \left\|\boldsymbol{x}_\star^-\right\|^2 \boldsymbol{I} \right\|_{\mathrm{op}} \\
&+ \frac{1}{8} \left\| \frac{1}{n} \sum_{i=1}^n \left|\langle \boldsymbol{a}_i^-, \boldsymbol{x}_\star^- \rangle\right|^2 \boldsymbol{a}_i^- \otimes \boldsymbol{a}_i^- - 2\boldsymbol{x}_\star^- \otimes \boldsymbol{x}_\star^- - \left\|\boldsymbol{x}_\star^-\right\|^2 \boldsymbol{I} \right\|_{\mathrm{op}} .
\end{aligned}$$

Each of the summands on the right-hand side is small for a sufficiently large n. For example, as shown in [12, Lemma 7.4], if $n \geq C_\tau d \log d$ for a sufficiently large constant C_τ that depends only on $\tau \in (0,1)$, then

$$\left\| \frac{1}{n} \sum_{i=1}^n \left|\langle \boldsymbol{a}_i^+, \boldsymbol{x}_\star \rangle\right|^2 \boldsymbol{a}_i^+ \otimes \boldsymbol{a}_i^+ - 2\boldsymbol{x}_\star \otimes \boldsymbol{x}_\star - \|\boldsymbol{x}_\star\|^2 \boldsymbol{I} \right\|_{\mathrm{op}} \leq \tau \|\boldsymbol{x}_\star\|^2 ,$$

with probability $\geq 1 - 5\exp(-4\tau d) - 4d^{-2}$. Clearly, we can write similar inequalities for the other three summands and by a simple union bound conclude that

$$\|\boldsymbol{S}_n - \mathbb{E}_{\mathcal{D}} \boldsymbol{S}_n\|_{\mathrm{op}} \leq \frac{\tau}{4} \left(\|\boldsymbol{x}_\star\|^2 + \left\|\boldsymbol{x}_\star^-\right\|^2 \right) = \frac{\tau}{2} \|\boldsymbol{x}_\star\|^2 , \tag{32}$$

holds with probability $\geq 1 - c_\tau d^{-2}$ for some absolute constant c_τ depending only on τ. Recall that $\mathbb{E}_\mathcal{D} \boldsymbol{S}_n = \left(\boldsymbol{x}_\star \otimes \boldsymbol{x}_\star - \boldsymbol{x}_\star^- \otimes \boldsymbol{x}_\star^-\right)/2$. Because we chose $\left\|\boldsymbol{x}_\star^{(1)}\right\| = \left\|\boldsymbol{x}_\star^{(2)}\right\|$ and by the construction of $\boldsymbol{x}_\star$ and $\boldsymbol{x}_\star^-$ we have $\langle \boldsymbol{x}_\star, \boldsymbol{x}_\star^- \rangle = 0$. Thus $\boldsymbol{x}_\star$ and $\boldsymbol{x}_\star^-$ are eigenvectors of $\mathbb{E}_\mathcal{D} \boldsymbol{S}_n$. We may assume that $\langle \boldsymbol{v}_{\max}, \boldsymbol{x}_\star \rangle \geq 0$; otherwise we can simply use $-\boldsymbol{x}_\star$ as the target. Then, on the event (32), a variant of the Davis-Kahan theorem [50, Corollary 3] ensures

$$\left\|\boldsymbol{v}_{\max} - \frac{\boldsymbol{x}_\star}{\|\boldsymbol{x}_\star\|}\right\| \leq \frac{2^{1/2}\tau \|\boldsymbol{x}_\star\|^2}{\|\boldsymbol{x}_\star\|^2/2} = 2^{3/2}\tau\,.$$

Since $\boldsymbol{a}_0$ is defined by (31), we equivalently obtain

$$\left\|\boldsymbol{a}_0 - \left(\frac{\lambda_{\max}}{2}\right)^{1/2} \frac{\boldsymbol{x}_\star}{\|\boldsymbol{x}_\star\|}\right\| \leq 2\tau\lambda_{\max}^{1/2}\,.$$

Using (32), it is also easy to show that

$$\frac{1-\tau}{2}\|\boldsymbol{x}_\star\|^2 \leq \lambda_{\max} \leq \frac{1+\tau}{2}\|\boldsymbol{x}_\star\|^2\,.$$

Therefore, we deduce that

$$\begin{aligned}\left\|\boldsymbol{a}_0 - \frac{1}{2}\boldsymbol{x}_\star\right\| &\leq \left\|\boldsymbol{a}_0 - \left(\frac{\lambda_{\max}}{2}\right)^{1/2} \frac{\boldsymbol{x}_\star}{\|\boldsymbol{x}_\star\|}\right\| + \left\|\left(\frac{\lambda_{\max}}{2}\right)^{1/2} \frac{\boldsymbol{x}_\star}{\|\boldsymbol{x}_\star\|} - \frac{1}{2}\boldsymbol{x}_\star\right\| \\ &\leq 2^{1/2}\tau\left(1+\frac{\tau}{2}\right)\|\boldsymbol{x}_\star\| + \frac{1}{2}\tau\|\boldsymbol{x}_\star\|\,. \end{aligned} \tag{33}$$

It follows from Lemma 5 for $\boldsymbol{x} = \boldsymbol{x}_\star$, (33), and (23), that if $n \overset{\tau}{\gtrsim} \left(d + \log\frac{4}{\delta}\right)\log d$, then

$$\left\|\frac{1}{2n}\sum_{i=1}^n \nabla f_i^+(\boldsymbol{x}_\star) + \nabla f_i^-(\boldsymbol{x}_\star) - \boldsymbol{a}_0\right\| \leq C_\tau \lambda_\mathcal{D}\,,$$

with probability $\geq 1 - c_\tau d^{-2}$. Choosing an appropriate value for τ in terms of ε, the constant C_τ can also be made smaller than $(1-\varepsilon)/2$, thereby guaranteeing (8).

4.4. Numerical experiments

To evaluate the proposed method numerically, we ran 100 trials with the standard Gaussian measurements for each pair of $d_1 = d_2 = d/2 \in \{50, 100, 150\}$ and $n/d \in \{5, 6, 7, 8, 9\}$. The signal pairs $\boldsymbol{x}_\star^{(1)}$ and $\boldsymbol{x}_\star^{(2)}$ are drawn independently and uniformly from the $d/2$-dimensional unit sphere in each trial. We solved an

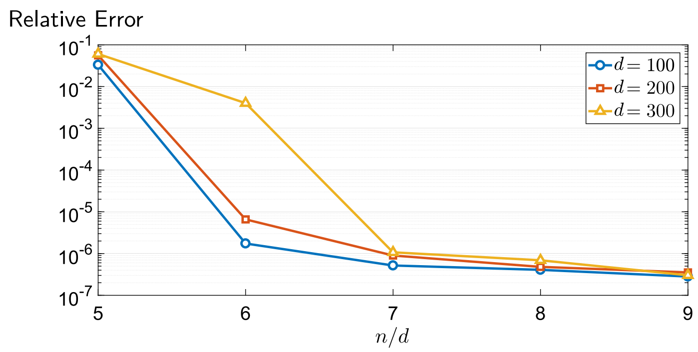

FIG 1. *Relative error of the estimate* (9) *versus the oversampling ratio* n/d

equivalent form of (9) which is the quadratically-constrained linear maximization

$$\begin{aligned} \max_{\boldsymbol{x}\in\mathbb{R}^{d_1+d_2},\boldsymbol{w}\in\mathbb{R}^n} \quad & \langle \boldsymbol{a}_0,\boldsymbol{x}\rangle - \frac{1}{n}\langle \mathbf{1}_n,\boldsymbol{w}\rangle \\ \text{subject to } & \frac{1}{4}\left|\langle \boldsymbol{a}_i^+,\boldsymbol{x}\rangle\right|^2 - y_i \le w_i, \qquad i=1,\dots,n \\ & \frac{1}{4}\left|\langle \boldsymbol{a}_i^-,\boldsymbol{x}\rangle\right|^2 \le w_i, \qquad\quad i=1,\dots,n\,, \end{aligned} \tag{34}$$

where $\mathbf{1}_n$ denotes the n-dimensional all-one vector, using the Gurobi solver [23] through the CVX package [22]. This solver relies on an *interior point method* for solving the second order cone program (SOCP) corresponding to (34). For better scalability, first order methods including stochastic and incremental methods can be used to solve (9) directly. We did not intend in this paper to find the best convex optimization method for solving (9).

Figure 1 shows the median of the relative error computed as

$$\text{Relative Error} = \sqrt{\frac{\left\|\sqrt{\frac{\|\widehat{\boldsymbol{x}}^{(2)}\|}{\|\widehat{\boldsymbol{x}}^{(1)}\|}}\,\widehat{\boldsymbol{x}}^{(1)} - \boldsymbol{x}_\star^{(1)}\right\|^2 + \left\|\sqrt{\frac{\|\widehat{\boldsymbol{x}}^{(1)}\|}{\|\widehat{\boldsymbol{x}}^{(2)}\|}}\,\widehat{\boldsymbol{x}}^{(2)} - \boldsymbol{x}_\star^{(2)}\right\|^2}{\left\|\boldsymbol{x}_\star^{(1)}\right\|^2 + \left\|\boldsymbol{x}_\star^{(2)}\right\|^2}}\,.$$

The experiment suggests that the proposed method succeeds when the oversampling ratio is around eight (i.e., $n \approx 8(d_1+d_2) = 8d$).

Appendix A: Technical lemmas

A.1. Lemmas used in Section 2

Lemma 2. *For any* $\alpha > 0$*, with probability* $\ge 1-\delta/2$ *we have*

$$\frac{1}{n}\sum_{i=1}^n \frac{1}{\alpha}\mathbb{E}_{\boldsymbol{z}\sim\gamma_{\boldsymbol{h}}}\left(\left[\alpha q_i(\boldsymbol{z})\right]_{\le 1}\right) \le \frac{1}{n}\sum_{i=1}^n q_i(\boldsymbol{h}) + \sigma\mathit{\Gamma}_{\mathcal{D}} + \frac{1}{\alpha}\sqrt{\frac{\log\frac{2}{\delta}}{2n}}\,,$$

for all $\boldsymbol{h}$.

Proof. The triangle inequality and subadditivity of $u \mapsto [u]_{\leq 1}$ over the nonnegative real numbers yields

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\alpha}\mathbb{E}_{\boldsymbol{z}\sim\gamma_{\boldsymbol{h}}}\left([\alpha q_i(\boldsymbol{z})]_{\leq 1}\right)\\
&\leq \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\alpha}\mathbb{E}_{\boldsymbol{z}\sim\gamma_{\boldsymbol{h}}}\left([\alpha q_i(\boldsymbol{h}) + |\alpha q_i(\boldsymbol{z}) - \alpha q_i(\boldsymbol{h})|]_{\leq 1}\right)\\
&\leq \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\alpha}\mathbb{E}_{\boldsymbol{z}\sim\gamma_{\boldsymbol{h}}}\left([\alpha q_i\left(\boldsymbol{h}\right)]_{\leq 1}\right) + \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\alpha}\mathbb{E}_{\boldsymbol{z}\sim\gamma_{\boldsymbol{h}}}\left([|\alpha q_i(\boldsymbol{z}) - \alpha q_i(\boldsymbol{h})|]_{\leq 1}\right)\,.
\end{aligned}$$

Clearly, $[\alpha q_i(\boldsymbol{h})]_{\leq 1} \leq \alpha q_i(\boldsymbol{h})$. Thus, we only need to bound the second term in the above inequality. Using (3) followed by the Hoeffding's inequality shows that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{z}\sim\gamma_{\boldsymbol{h}}}\left([\alpha\,|q_i(\boldsymbol{z}) - q_i(\boldsymbol{h})|]_{\leq 1}\right)\\
&\leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{z}\sim\gamma_{\boldsymbol{h}}}\left([\alpha\,|q_i(\boldsymbol{z}-\boldsymbol{h})|]_{\leq 1}\right)\\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{z}\sim\gamma_{\boldsymbol{0}}}\left([\alpha\,|q_i(\boldsymbol{z})|]_{\leq 1}\right)\\
&\leq \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\boldsymbol{z}\sim\gamma_{\boldsymbol{0}}}\left([\alpha\,|q_i(\boldsymbol{z})|]_{\leq 1}\right) + \sqrt{\frac{\log\frac{2}{\delta}}{2n}}\,,
\end{aligned}$$

holds with probability $\geq 1 - \delta/2$ for all $\boldsymbol{h}$. Therefore, on this event we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\alpha}\mathbb{E}_{\boldsymbol{z}\sim\gamma_{\boldsymbol{h}}}\left([\alpha q_i(\boldsymbol{z})]_{\leq 1}\right)\\
&\leq \frac{1}{n}\sum_{i=1}^{n}q_i(\boldsymbol{h}) + \frac{1}{\alpha}\left[\mathbb{E}_{\mathcal{D}}\mathbb{E}_{\boldsymbol{z}\sim\gamma_{\boldsymbol{0}}}\left(\alpha q_i(\boldsymbol{z})\right)\right]_{\leq 1} + \frac{1}{\alpha}\sqrt{\frac{\log\frac{2}{\delta}}{2n}}\\
&\leq \frac{1}{n}\sum_{i=1}^{n}q_i(\boldsymbol{h}) + \sigma\mathit{\Gamma}_{\mathcal{D}} + \frac{1}{\alpha}\sqrt{\frac{\log\frac{2}{\delta}}{2n}}
\end{aligned}$$

where concavity of $u \mapsto [u]_{\leq 1}$ is used in the first inequality, and the second inequality follows from the fact that $[u]_{\leq 1} \leq u$, the Cauchy-Schwarz inequality, and the definition (7). □

Lemma 3. *For all* $\boldsymbol{h}$*, we have*

$$\sqrt{\mathbb{E}_{\boldsymbol{z}\sim\gamma_{\boldsymbol{h}}}\mathbb{E}_{\mathcal{D}}q_1^2(\boldsymbol{z})} \leq \eta_{\mathcal{D}}\mathbb{E}_{\mathcal{D}}\left(q_1(\boldsymbol{h})\right) + \sigma\mathit{\Gamma}_{\mathcal{D}}\,.$$

Proof. It immediately follows from the triangle inequality, (3), and the equivalence of $\boldsymbol{z} \sim \gamma_{\boldsymbol{h}}$ and $\boldsymbol{z} - \boldsymbol{h} \sim \gamma_{\boldsymbol{0}}$, that

$$\begin{aligned}\sqrt{\mathbb{E}_{\boldsymbol{z}\sim\gamma_{\boldsymbol{h}}}\mathbb{E}_{\mathcal{D}} q_1^2(\boldsymbol{z})} &\le \sqrt{\mathbb{E}_{\mathcal{D}}\left(q_1^2(\boldsymbol{h})\right)} + \sqrt{\mathbb{E}_{\boldsymbol{z}\sim\gamma_{\boldsymbol{h}}}\mathbb{E}_{\mathcal{D}}\left(\left(q_1(\boldsymbol{z}) - q_1(\boldsymbol{h})\right)^2\right)} \\ &\le \sqrt{\mathbb{E}_{\mathcal{D}}\left(q_1^2(\boldsymbol{h})\right)} + \sqrt{\mathbb{E}_{\boldsymbol{z}\sim\gamma_{\boldsymbol{h}}}\mathbb{E}_{\mathcal{D}}\left(q_1^2(\boldsymbol{z}-\boldsymbol{h})\right)} \\ &\le \sqrt{\mathbb{E}_{\mathcal{D}}\left(q_1^2(\boldsymbol{h})\right)} + \sqrt{\mathbb{E}_{\boldsymbol{z}\sim\gamma_0}\mathbb{E}_{\mathcal{D}}\left(q_1^2(\boldsymbol{z})\right)},\end{aligned}$$

which by the assumption (6) and definition (7) yields is the desired bound. □

A.2. Lemmas used in Section 4

Note that in the following lemmas the functions f_i^+ and f_i^- are defined as in (19).

Lemma 4. *With $\mathcal{K}_{t,\theta}$ defined by (26), suppose that*

$$\frac{1}{n}\sum_{i=1}^{n}\left|\langle \nabla f_i^+(\boldsymbol{x}_\star) - \nabla f_i^-(\boldsymbol{x}_\star), \boldsymbol{h}\rangle\right| \ge (1-\varepsilon)\mathbb{E}_{\mathcal{D}}\left(\left|\langle \nabla f^+(\boldsymbol{x}_\star) - \nabla f^-(\boldsymbol{x}_\star), \boldsymbol{h}\rangle\right|\right),$$

for all vectors $\boldsymbol{h} \in \mathcal{K}_{1,\frac{1}{2}}$. Then, for all $t \in \mathbb{R}\backslash\{0\}$ with $\sqrt{2/3} \le |t| \le \sqrt{3/2}$, and all vectors $\boldsymbol{h} \in \mathcal{K}_{t,0}$, we have

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\left|\langle \nabla f_i^+(D_t(\boldsymbol{x}_\star)) - \nabla f_i^-(D_t(\boldsymbol{x}_\star)), \boldsymbol{h}\rangle\right| \\ &\ge (1-\varepsilon)\,\mathbb{E}_{\mathcal{D}}\left(\left|\langle \nabla f^+(\boldsymbol{x}_\star) - \nabla f^-(\boldsymbol{x}_\star), D_{t^{-1}}(\boldsymbol{h})\rangle\right|\right) \\ &= (1-\varepsilon)\,\mathbb{E}_{\mathcal{D}}\left(\left|\langle \nabla f^+(D_t(\boldsymbol{x}_\star)) - \nabla f^-(D_t(\boldsymbol{x}_\star)), \boldsymbol{h}\rangle\right|\right).\end{aligned}$$

Proof. We have the identity

$$\begin{aligned}\langle \nabla f_i^+(D_t(\boldsymbol{x}_\star)) - \nabla f_i^-(D_t(\boldsymbol{x}_\star)), \boldsymbol{h}\rangle &= \langle \boldsymbol{a}_i^{(1)} \otimes \boldsymbol{a}_i^{(2)}, t\boldsymbol{x}_\star^{(1)} \otimes \boldsymbol{h}^{(2)} + \boldsymbol{h}^{(1)} \otimes t^{-1}\boldsymbol{x}_\star^{(2)}\rangle \\ &= \langle \nabla f_i^+(\boldsymbol{x}_\star) - \nabla f_i^-(\boldsymbol{x}_\star), D_{t^{-1}}(\boldsymbol{h})\rangle, \qquad (35)\end{aligned}$$

for every $\boldsymbol{h}$ and $t \in \mathbb{R}\backslash\{0\}$. Furthermore, because $\boldsymbol{h} \in \mathcal{K}_{t,0}$, by definition $\langle D_t(\boldsymbol{x}_\star^-), \boldsymbol{h}\rangle = 0$, thereby we have the following

$$\begin{aligned}\left|\langle D_{t^{-1}}(\boldsymbol{x}_\star^-), \boldsymbol{h}\rangle\right| &= \left|\langle D_{t^{-1}}(\boldsymbol{x}_\star^-) - D_t(\boldsymbol{x}_\star^-), \boldsymbol{h}\rangle\right| \\ &= \left|t - t^{-1}\right|\left|\langle \boldsymbol{x}_\star, \boldsymbol{h}\rangle\right| \\ &\le \max\left\{\left|t^2-1\right|, \left|t^{-2}-1\right|\right\}\|\boldsymbol{x}_\star\|\,\|D_{t^{-1}}(\boldsymbol{h})\|.\end{aligned}$$

Applying the bound $\sqrt{2/3} \le |t| \le \sqrt{3/2}$, we obtain

$$\left|\langle \boldsymbol{x}_\star^-, D_{t^{-1}}(\boldsymbol{h})\rangle\right| = \left|\langle D_{t^{-1}}(\boldsymbol{x}_\star^-), \boldsymbol{h}\rangle\right| \le \frac{1}{2}\|D_{t^{-1}}(\boldsymbol{x}_\star)\|\,\|\boldsymbol{h}\|,$$

which means that $D_{t^{-1}}(\boldsymbol{h}) \in \mathcal{K}_{1,\frac{1}{2}}$. Therefore, it follows from the assumption of the lemma that

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\left|\langle \nabla f_i^+(\boldsymbol{x}_\star) - \nabla f_i^-(\boldsymbol{x}_\star), D_{t^{-1}}\left(\boldsymbol{h}\right)\rangle\right| \\ &\geq (1-\varepsilon)\mathbb{E}_{\mathcal{D}}\left(\left|\langle \nabla f^+(\boldsymbol{x}_\star) - \nabla f^-(\boldsymbol{x}_\star), D_{t^{-1}}\left(\boldsymbol{h}\right)\rangle\right|\right)\,,\end{aligned}$$

which, using (35), implies

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\left|\langle \nabla f_i^+(D_t(\boldsymbol{x}_\star)) - \nabla f_i^-(D_t(\boldsymbol{x}_\star)), \boldsymbol{h}\rangle\right| \\ &\geq (1-\varepsilon)\,\mathbb{E}_{\mathcal{D}}\left(\left|\langle \nabla f^+(D_t(\boldsymbol{x}_\star)) - \nabla f^-(D_t(\boldsymbol{x}_\star)), \boldsymbol{h}\rangle\right|\right)\,,\end{aligned}$$

as desired. □

We use standard matrix concentration inequalities to establish Lemmas 5 and 6 below. We can upper bound $\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{a}_i^+ \otimes \boldsymbol{a}_i^+ - \boldsymbol{I}\right\|_{\mathrm{op}}$ by a standard covering argument as in [49, Theorem 5.39] which guarantees

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{a}_i^+ \otimes \boldsymbol{a}_i^+ - \boldsymbol{I}\right\|_{\mathrm{op}} \leq C\max\left\{\sqrt{\frac{d}{n}}+\sqrt{\frac{\log\frac{4}{\delta}}{n}},\left(\sqrt{\frac{d}{n}}+\sqrt{\frac{\log\frac{4}{\delta}}{n}}\right)^2\right\}\,, \tag{36}$$

with probability $\geq 1-\delta/2$ for a sufficiently large absolute constant $C>0$. Similarly, we have

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{a}_i^- \otimes \boldsymbol{a}_i^- - \boldsymbol{I}\right\|_{\mathrm{op}} \leq C\max\left\{\sqrt{\frac{d}{n}}+\sqrt{\frac{\log\frac{4}{\delta}}{n}},\left(\sqrt{\frac{d}{n}}+\sqrt{\frac{\log\frac{4}{\delta}}{n}}\right)^2\right\}\,, \tag{37}$$

with probability $\geq 1-\delta/2$.

The first lemma below is an immediate consequence of the matrix concentration inequalities above and is stated merely for reference.

Lemma 5. *On the event that* (36) *and* (37) *hold, we have*

$$\begin{aligned}&\left\|\frac{1}{2n}\sum_{i=1}^{n}\nabla f_i^+(\boldsymbol{x}) + \nabla f_i^-(\boldsymbol{x}) - \frac{1}{2}\boldsymbol{x}\right\| \\ &\leq C\max\left\{\sqrt{\frac{d}{n}}+\sqrt{\frac{\log\frac{4}{\delta}}{n}},\left(\sqrt{\frac{d}{n}}+\sqrt{\frac{\log\frac{4}{\delta}}{n}}\right)^2\right\}\|\boldsymbol{x}\|\,,\end{aligned} \tag{38}$$

for every $\boldsymbol{x}\in\mathbb{R}^d$.

Proof. By definition

$$\nabla f_i^+(\boldsymbol{x}) + \nabla f_i^-(\boldsymbol{x}) = \frac{1}{2}\left(\boldsymbol{a}_i^+ \otimes \boldsymbol{a}_i^+\right)\boldsymbol{x} + \frac{1}{2}\left(\boldsymbol{a}_i^- \otimes \boldsymbol{a}_i^-\right)\boldsymbol{x}\,.$$

for any $\boldsymbol{x}$. A simple application of triangle inequality yields

$$\begin{aligned}&\left\|\frac{1}{2n}\sum_{i=1}^{n}\nabla f_i^+(\boldsymbol{x}) + \nabla f_i^-(\boldsymbol{x}) - \frac{1}{2}\boldsymbol{x}\right\| \\ &\le \frac{1}{4}\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{a}_i^+ \otimes \boldsymbol{a}_i^+ - \boldsymbol{I}\right\|_{\mathrm{op}}\|\boldsymbol{x}\| + \frac{1}{4}\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{a}_i^- \otimes \boldsymbol{a}_i^- - \boldsymbol{I}\right\|_{\mathrm{op}}\|\boldsymbol{x}\|\,.\end{aligned}$$

The result follows immediately using the matrix concentration inequalities (36) and (37). ☐

Lemma 6. *There exists an absolute constant $C > 0$ such that*

$$\begin{aligned}&\left\|\boldsymbol{a}_0 - \frac{1}{2n}\sum_{i=1}^{n}\nabla f_i^+(\boldsymbol{x}_\star) + \nabla f_i^-(\boldsymbol{x}_\star)\right\| \\ &\ge \frac{1}{8}\left(1 - C\max\left\{\sqrt{\frac{d}{n}} + \sqrt{\frac{\log\frac{4}{\delta}}{n}}, \left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log\frac{4}{\delta}}{n}}\right)^2\right\}\right)\|\widehat{\boldsymbol{x}}_\star - \boldsymbol{x}_\star\|\,,\end{aligned} \tag{39}$$

holds with probability $\ge 1-\delta$. Furthermore, for a sufficiently large n, on the same event we have

$$\begin{aligned}&\left\|\boldsymbol{a}_0 - \frac{1}{2n}\sum_{i=1}^{n}\nabla f_i^+(\widehat{\boldsymbol{x}}_\star) + \nabla f_i^-(\widehat{\boldsymbol{x}}_\star)\right\| \\ &\le \frac{5 + 3C\max\left\{\sqrt{\frac{d}{n}} + \sqrt{\frac{\log\frac{4}{\delta}}{n}}, \left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log\frac{4}{\delta}}{n}}\right)^2\right\}}{1 - C\max\left\{\sqrt{\frac{d}{n}} + \sqrt{\frac{\log\frac{4}{\delta}}{n}}, \left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log\frac{4}{\delta}}{n}}\right)^2\right\}}\times \\ &\qquad\left\|\boldsymbol{a}_0 - \frac{1}{2n}\sum_{i=1}^{n}\nabla f_i^+(\boldsymbol{x}_\star) + \nabla f_i^-(\boldsymbol{x}_\star\right\|.\end{aligned} \tag{40}$$

Proof. First we prove (39). By optimality of $\widehat{\boldsymbol{x}}$ in (9), we can write

$$\begin{aligned}\langle\boldsymbol{a}_0, \widehat{\boldsymbol{x}} - \boldsymbol{x}_\star\rangle &\ge \frac{1}{n}\sum_{i=1}^{n}\max\left(f_i^+(\widehat{\boldsymbol{x}}) - f_i^+(\boldsymbol{x}_\star), f_i^-(\widehat{\boldsymbol{x}}) - f_i^-(\boldsymbol{x}_\star)\right) \\ &\ge \frac{1}{2n}\sum_{i=1}^{n}f_i^+(\widehat{\boldsymbol{x}}) + f_i^-(\widehat{\boldsymbol{x}}) - f_i^+(\boldsymbol{x}_\star) - f_i^-(\boldsymbol{x}_\star).\end{aligned}$$

Then, subtracting $\langle \frac{1}{2n}\sum_{i=1}^n \nabla f_i^+(\boldsymbol{x}_\star) + \nabla f_i^-(\boldsymbol{x}_\star), \widehat{\boldsymbol{x}} - \boldsymbol{x}_\star\rangle$ yields

$$\begin{aligned}
&\langle \boldsymbol{a}_0 - \frac{1}{2n}\sum_{i=1}^n \nabla f_i^+(\boldsymbol{x}_\star) + \nabla f_i^-(\boldsymbol{x}_\star), \widehat{\boldsymbol{x}} - \boldsymbol{x}_\star\rangle \\
&\geq \frac{1}{2n}\sum_{i=1}^n f_i^+(\widehat{\boldsymbol{x}}) + f_i^-(\widehat{\boldsymbol{x}}) - f_i^+(\boldsymbol{x}_\star) - f_i^-(\boldsymbol{x}_\star) - \langle \nabla f_i^+(\boldsymbol{x}_\star) + \nabla f_i^-(\boldsymbol{x}_\star), \widehat{\boldsymbol{x}} - \boldsymbol{x}_\star\rangle \\
&= \frac{1}{2n}\sum_{i=1}^n \frac{1}{4}\left|\langle \boldsymbol{a}_i^+, \widehat{\boldsymbol{x}} - \boldsymbol{x}_\star\rangle\right|^2 + \frac{1}{4}\left|\langle \boldsymbol{a}_i^-, \widehat{\boldsymbol{x}} - \boldsymbol{x}_\star\rangle\right|^2 . \qquad (41)
\end{aligned}$$

Applying the Cauchy-Schwarz inequality to the first line, and the standard matrix concentration inequalities (36) and (37) in the third line, we obtain with probability $\geq 1-\delta$ that

$$\begin{aligned}
&\left\|\boldsymbol{a}_0 - \frac{1}{2n}\sum_{i=1}^n \nabla f_i^+(\boldsymbol{x}_\star) + \nabla f_i^-(\boldsymbol{x}_\star)\right\| \|\widehat{\boldsymbol{x}} - \boldsymbol{x}_\star\| \\
&\geq \frac{1}{4}\left(1 - C\max\left\{\sqrt{\frac{d}{n}} + \sqrt{\frac{\log\frac{4}{\delta}}{n}}, \left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log\frac{4}{\delta}}{n}}\right)^2\right\}\right)\|\widehat{\boldsymbol{x}} - \boldsymbol{x}_\star\|^2 ,
\end{aligned}$$

and thereby

$$\begin{aligned}
&\left\|\boldsymbol{a}_0 - \frac{1}{2n}\sum_{i=1}^n \nabla f_i^+(\boldsymbol{x}_\star) + \nabla f_i^-(\boldsymbol{x}_\star)\right\| \\
&\geq \frac{1}{4}\left(1 - C\max\left\{\sqrt{\frac{d}{n}} + \sqrt{\frac{\log\frac{4}{\delta}}{n}}, \left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log\frac{4}{\delta}}{n}}\right)^2\right\}\right)\|\widehat{\boldsymbol{x}} - \boldsymbol{x}_\star\| .
\end{aligned}$$

Finally, it follows from the triangle inequality and the definition of $\widehat{\boldsymbol{x}}_\star$ in (20) that

$$\|\widehat{\boldsymbol{x}}_\star - \boldsymbol{x}_\star\| \leq \|\widehat{\boldsymbol{x}} - \widehat{\boldsymbol{x}}_\star\| + \|\widehat{\boldsymbol{x}} - \boldsymbol{x}_\star\| \leq 2\|\widehat{\boldsymbol{x}} - \boldsymbol{x}_\star\| ,$$

which together with the previous bound guarantees (39).

Next, we prove (40). By the triangle inequality and the fact that $\nabla f_i^\pm(\boldsymbol{x})$ is linear in $\boldsymbol{x}$, we have

$$\begin{aligned}
&\left\|\boldsymbol{a}_0 - \frac{1}{2n}\sum_{i=1}^n \nabla f_i^+(\widehat{\boldsymbol{x}}_\star) + \nabla f_i^-(\widehat{\boldsymbol{x}}_\star)\right\| \\
&\leq \left\|\frac{1}{2n}\sum_{i=1}^n \nabla f_i^+(\boldsymbol{x}_\star) + \nabla f_i^-(\boldsymbol{x}_\star) - \nabla f_i^+(\widehat{\boldsymbol{x}}_\star) - \nabla f_i^-(\widehat{\boldsymbol{x}}_\star)\right\| \\
&\quad + \left\|\boldsymbol{a}_0 - \frac{1}{2n}\sum_{i=1}^n \nabla f_i^+(\boldsymbol{x}_\star) + \nabla f_i^-(\boldsymbol{x}_\star)\right\|
\end{aligned}$$

$$= \left\| \frac{1}{2n} \sum_{i=1}^{n} \nabla f_i^{+}(\boldsymbol{x}_\star - \widehat{\boldsymbol{x}}_\star) + \nabla f_i^{-}(\boldsymbol{x}_\star - \widehat{\boldsymbol{x}}_\star) \right\| + \left\| \boldsymbol{a}_0 - \frac{1}{2n} \sum_{i=1}^{n} \nabla f_i^{+}(\boldsymbol{x}_\star) + \nabla f_i^{-}(\boldsymbol{x}_\star) \right\| .$$

Recall, from the first part of the proof, that (36) and (37) hold with probability $\geq 1 - \delta$. Then, on the same event, Lemma 5 implies that

$$\begin{aligned} &\left\| \boldsymbol{a}_0 - \frac{1}{2n} \sum_{i=1}^{n} \nabla f_i^{+}(\widehat{\boldsymbol{x}}_\star) + \nabla f_i^{-}(\widehat{\boldsymbol{x}}_\star) \right\| \\ &\leq \frac{1}{2} \left(1 + C \max \left\{ \sqrt{\frac{d}{n}} + \sqrt{\frac{\log \frac{4}{\delta}}{n}}, \left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log \frac{4}{\delta}}{n}} \right)^2 \right\} \right) \|\widehat{\boldsymbol{x}}_\star - \boldsymbol{x}_\star\| \\ &\quad + \left\| \boldsymbol{a}_0 - \frac{1}{2n} \sum_{i=1}^{n} \nabla f_i^{+}(\boldsymbol{x}_\star) + \nabla f_i^{-}(\boldsymbol{x}_\star) \right\| . \end{aligned}$$

Therefore, if n is sufficiently large to ensure the right-hand side of (39) is non-negative, we deduce that (40) holds as well. ☐

Lemma 7. *For any matrix $\boldsymbol{A}$ and standard normal random vector $\boldsymbol{z}$ (of appropriate dimension) we have*

$$\mathbb{E} \|\boldsymbol{A}\boldsymbol{z}\| \geq \sqrt{\frac{2}{\pi}} \|\boldsymbol{A}\|_{\mathrm{F}} .$$

Proof. The Euclidean and Frobenius norms as well as the standard normal distribution are rotationally invariant. Thus, the claim can be reduced to the case where $\boldsymbol{A}$ is diagonal with nonzero diagonal entries $s_1, s_2, \ldots, s_r$ and $\|\boldsymbol{A}\boldsymbol{z}\| = \sqrt{\sum_{i=1}^{r} s_i^2 z_i^2}$. By concavity of $u \mapsto \sqrt{u}$ and Jensen's inequality we have

$$\sqrt{\left(\sum_{i=1}^{r} s_i^2 \right)^{-1} \sum_{i=1}^{r} s_i^2 z_i^2} \geq \left(\sum_{i=1}^{r} s_i^2 \right)^{-1} \sum_{i=1}^{r} s_i^2 |z_i| .$$

Therefore, taking expectation with respect to $\boldsymbol{z}$ we can conclude

$$\mathbb{E} \sqrt{\sum_{i=1}^{r} s_i^2 z_i^2} \geq \sqrt{\left(\sum_{i=1}^{r} s_i^2 \right)^{-1}} \sum_{i=1}^{r} s_i^2 \mathbb{E} |z_i| = \sqrt{\frac{2}{\pi} \sum_{i=1}^{r} s_i^2} = \sqrt{\frac{2}{\pi}} \|\boldsymbol{A}\|_{\mathrm{F}} . \qquad ☐$$

References

[1] A. Aghasi, A. Ahmed, and P. Hand. Branchhull: Convex bilinear inversion from the entrywise product of signals with known signs. preprint `arXiv:1702.04342 [cs.IT]`, 2017.

[2] A. A. Ahmadi and G. Hall. DC decomposition of nonconvex polynomials with algebraic techniques. *Mathematical Programming - Series B*, 2017. MR3785671

[3] A. Ahmed, B. Recht, and J. Romberg. Blind deconvolution using convex programming. *IEEE Transactions on Information Theory*, 60(3):1711–1732, March 2014. MR3168432

[4] P. Alquier. PAC-Bayesian bounds for randomized empirical risk minimizers. *Mathematical Methods of Statistics*, 17(4):279–304, Dec 2008. MR2483458

[5] P. Alquier and K. Lounici. PAC-Bayesian bounds for sparse regression estimation with exponential weights. *Electronic Journal of Statistics*, 5:127–145, 2011. MR2786484

[6] J.-Y. Audibert and O. Catoni. Robust linear least squares regression. *the Annals of Statistics*, 39(5):2766–2794, 10 2011. MR2906886

[7] S. Bahmani and J. Romberg. Lifting for blind deconvolution in random mask imaging: Identifiability and convex relaxation. *SIAM Journal on Imaging Sciences*, 8(4):2203–2238, 2015. MR3404681

[8] 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.

[9] S. Bahmani and J. Romberg. A flexible convex relaxation for phase retrieval. *Electronic Journal of Statistics*, 11(2):5254–5281, 2017. MR3738211

[10] S. Bahmani and J. Romberg. Solving equations of random convex functions via anchored regression. *J. Found. Comp. Math.*, 2018. In press; preprint `arXiv:1702.05327 [cs.LG]`.

[11] O. Bousquet, V. Koltchinskii, and D. Panchenko. Some local measures of complexity of convex hulls and generalization bounds. In J. Kivinen and R. H. Sloan, editors, *Computational Learning Theory*, pages 59–73, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg. MR2040405

[12] E. J. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *Information Theory, IEEE Transactions on*, 61(4):1985–2007, Apr. 2015. MR3332993

[13] O. Catoni. *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*, volume 56 of *Lecture Notes–Monograph Series*. Institute of Mathematical Statistics, Beachwood, OH, USA, 2007. MR2483528

[14] O. Catoni and I. Giulini. Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression. *arXiv preprint*, Dec. 2017. `arXiv:1712.02747 [math, stat]`.

[15] R. Y. Chen, A. Gittens, and J. A. Tropp. The masked sample covariance estimator: An analysis using matrix concentration inequalities. *Information and Inference: A Journal of the IMA*, 1(1):2–20, 2012. MR3311439

[16] Y. Chen and E. Candés. Solving random quadratic systems of equations is nearly as easy as solving linear systems. In *Advances in Neural Information Processing Systems 28*, pages 739–747. Curran Associates, Inc., 2015. MR3628877

[17] P. Germain, A. Lacasse, F. Laviolette, and M. Marchand. PAC-Bayesian learning of linear classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 353–360, New York, NY, USA, 2009. ACM.

[18] E. Giné and V. Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *Annals of Probability*, 34(3):1143–1216, May 2006. MR2243881

[19] E. Giné, V. Koltchinskii, and J. A. Wellner. Ratio limit theorems for empirical processes. In E. Giné, C. Houdré, and D. Nualart, editors, *Stochastic Inequalities and Applications*, pages 249–278, Basel, 2003. Birkhäuser Basel. MR2073436

[20] T. Goldstein and C. Studer. Convex phase retrieval without lifting via PhaseMax. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1273–1281, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[21] T. Goldstein and C. Studer. Phasemax: Convex phase retrieval via basis pursuit. *IEEE Transactions on Information Theory*, 64(4):2675–2689, April 2018. MR3782281

[22] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 2.1, Mar. 2014. URL `http://cvxr.com/cvx`. MR3318084

[23] Gurobi Optimization, Inc. Gurobi optimizer reference manual, 2016. URL `http://www.gurobi.com`.

[24] P. Hartman. On functions representable as a difference of convex functions. *Pacific J. Math.*, 9(3):707–713, 1959. MR0110773

[25] M. Junge and Q. Zeng. Noncommutative Bennett and Rosenthal inequalities. *Annals of Probability*, 41(6):4287–4316, Nov. 2013. MR3161475

[26] V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001. MR1842526

[27] V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Lecture Notes in Mathematics: École d'Été de Probabilités de Saint-Flour XXXVIII-2008. Springer-Verlag Berlin Heidelberg, 2011. MR2829871

[28] 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. MR3431642

[29] V. Koltchinskii and D. Panchenko. Rademacher processes and bounding the risk of function learning. In E. Giné, D. M. Mason, and J. A. Wellner, editors, *High Dimensional Probability II*, pages 443–457, Boston, MA, 2000. Birkhäuser Boston. MR1857339

[30] J. Langford and J. Shawe-Taylor. PAC-Bayes & margins. In *Advances in Neural Information Processing Systems*, pages 439–446, 2003.

[31] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and processes*. Springer Science & Business Media, 2013. MR2814399

[32] X. Li, S. Ling, T. Strohmer, and K. Wei. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *Applied and Computational Harmonic Analysis*, 2018. . in press.

[33] S. Ling and T. Strohmer. Regularized gradient descent: A non-convex recipe for fast joint blind deconvolution and demixing. *Information and Inference: A Journal of the IMA*, 2018. . in press. MR3922402

[34] W. Luo, W. Alghamdi, and Y. M. Lu. Optimal spectral initialization for signal recovery with applications to phase retrieval. Preprint `arXiv:1811.04420 [cs.IT]`, 2018.

[35] C. Ma, K. Wang, Y. Chi, and Y. Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion and blind deconvolution. Preprint `arXiv:1711.10467 [cs.LG]`, 2017.

[36] L. Mackey, M. I. Jordan, R. Y. Chen, B. Farrell, and J. A. Tropp. Matrix concentration inequalities via the method of exchangeable pairs. *Annals of Probability*, 42(3):906–945, May 2014. MR3189061

[37] D. McAllester and T. Akinbiyi. *PAC-Bayesian Theory*, pages 95–103. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. MR3236859

[38] D. A. McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, Dec 1999.

[39] 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. MR3367000

[40] S. Mendelson. Learning without concentration. *Journal of the ACM*, 62(3):21:1–21:25, June 2015. ISSN 0004-5411. MR3367000

[41] M. Mondelli and A. Montanari. Fundamental limits of weak recovery with applications to phase retrieval. In *Proceedings of the 31st Conference On Learning Theory (COLT)*, volume 75 of *Proceedings of Machine Learning Research*, pages 1445–1450. PMLR, 2018.

[42] P. Netrapalli, P. Jain, and S. Sanghavi. Phase retrieval using alternating minimization. In *Advances in Neural Information Processing Systems 26*, pages 2796–2804. Curran Associates, Inc., 2013. MR3385838

[43] R. I. Oliveira. The lower tail of random quadratic forms with applications to ordinary least squares. *Probability Theory and Related Fields*, 166(3):1175–1194, Dec 2016. MR3568047

[44] Y. Plan and R. Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482–494, Jan 2013. MR3008160

[45] Y. Plan and R. Vershynin. The generalized lasso with non-linear observations. *IEEE Transactions on Information Theory*, 62(3):1528–1537, March 2016. MR3472264

[46] A. W. van Der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer, 1996. MR1385671

[47] V. N. Vapnik. *Statistical learning theory*. Wiley, 1998. MR1641250

[48] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability &*

Its Applications, 16(2):264–280, 1971. MR0288823

[49] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In G. Kutyniok and Y. Eldar, editors, *Compressed Sensing, Theory and Applications*, pages 210–268. Cambridge University Press, 2012. MR2963170

[50] Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the Davis-Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015. MR3371006