The nonsmooth landscape of phase retrieval

Damek Davis[*] Dmitriy Drusvyatskiy[†] Courtney Paquette[‡]

Abstract

We consider a popular nonsmooth formulation of the real phase retrieval problem. We show that under standard statistical assumptions, a simple subgradient method converges linearly when initialized within a constant relative distance of an optimal solution. Seeking to understand the distribution of the stationary points of the problem, we complete the paper by proving that as the number of Gaussian measurements increases, the stationary points converge to a codimension two set, at a controlled rate. Experiments on image recovery problems illustrate the developed algorithm and theory.

Keywords: Phase retrieval, stationary points, subdifferential, variational principle, subgradient method, spectral functions, eigenvalues

1 Introduction

Phase retrieval is a common task in computational science, with numerous applications including imaging, X-ray crystallography, and speech processing. In this work, we consider a popular real counterpart of the problem. Given a set of tuples $\{(a_i, b_i)\}_{i=1}^m \subset \mathbb{R}^d \times \mathbb{R}$, the (real) phase retrieval problem seeks to determine a vector $x \in \mathbb{R}^d$ satisfying $(a^T x)^2 = b_i$ for each index $i = 1, \ldots, m$. Due to its combinatorial nature, this problem is known to be NP-hard [14]. One can model the real phase retrieval problem in a variety of ways. Here, we consider the following "robust formulation":

$$\min_x \; f_S(x) := \frac{1}{m}\sum_{i=1}^m |(a_i^T x)^2 - b_i|.$$

This model of the problem has gained some attention recently with the work of Duchi-Ruan [10] and Eldar-Mendelson [12]. Indeed, this model exhibits a number of desirable

[*] School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14850, USA; people.orie.cornell.edu/dsd95/.

[†] Department of Mathematics, U. Washington, Seattle, WA 98195; www.math.washington.edu/~ddrusv. Research of Drusvyatskiy was supported by the AFOSR YIP award FA9550-15-1-0237 and by the NSF DMS 1651851 and CCF 1740551 awards.

[‡] Industrial and Systems Engineering Department, Lehigh University, Bethlehem, PA 18015; sites.math.washington.edu/~yumiko88/.

properties, making it amenable to numerical methods. Namely, in contrast to other possible formulations, mild statistical assumptions imply that f_S is both *weakly convex* [10, Corollary 3.2] and *sharp* [12, Theorem 2.4], with high probability. That is, there exist numerical constants $\rho, \kappa > 0$ such that

$$\text{the assignment } x \mapsto f_S(x) + \frac{\eta}{2}\|x\|^2 \text{ is a convex function,}$$

and the inequality

$$f_S(x) - \inf f_S \geq \kappa \|x - \bar{x}\| \|x + \bar{x}\| \qquad \text{holds for all } x \in \mathbb{R}^d.$$

Here, $\pm\bar{x}$ are the true signals and $\|\cdot\|$ denotes the ℓ_2-norm. Weak convexity is a well studied concept in optimization literature [5,13,23,25], while sharpness and the closely related notion of error bounds [2, 7, 21] classically underly rapid local convergence guarantees in nonlinear programming. Building on these observations, Duchi and Ruan [10] showed that with proper initialization, the so-called *prox-linear algorithm* [7,8,10,11,20] quadratically converges to $\pm\bar{x}$ (even in presence of outliers). The only limitation of their approach is that the prox-linear method requires, at every iteration, invoking an iterative solver for a convex subproblem. For large-scale instances ($m \gg 1, d \gg 1$), the numerical resolution of such problems is nontrivial. In the current work, we analyze a lower-cost alternative when there are no errors in the measurements.

We will show that the robust phase retrieval objective favorably lends itself to classical subgradient methods. This is somewhat surprising because, until recently, convergence rates of subgradient methods in nonsmooth, nonconvex optimization have remained elusive; see the discussion in [6]. We will prove that under mild statistical assumptions and proper initialization, the standard Polyak subgradient method

$$x_{k+1} = x_k - \left(\frac{f_S(x_k) - \min f_S}{\|g_k\|^2}\right) g_k \qquad \text{with} \qquad g_k \in \partial f_S(x_k),$$

linearly converges to $\pm\bar{x}$, with high probability. We note that high quality initialization, in turn, is straightforward to obtain; see e.g. [10, Section 3.3] and [29]. The argument we present is appealingly simple, relying only on weak convexity and sharpness of the function.

Aside from the current work and that of [10], we are not aware of any other attempts to optimize the robust phase retrieval objective directly. Other works focus on different problem formulations. Notably, Candès et al. [3] and Sun et al. [27] optimize the smooth loss $\frac{1}{m}\sum_{i=1}^m(\langle a_i, x\rangle^2 - b_i)^2$ using a second-order trust region method and a gradient method, respectively. Wang et al. [29] instead minimize the highly nonsmooth function $\frac{1}{m}\sum_{i=1}^m(|\langle a_i, x\rangle| - \sqrt{b_i})^2$ by a gradient descent-like method. Another closely related recent work is that of Tan and Vershynin [28]. One can interpret their scheme as a stochastic subgradient method on the formulation $\frac{1}{m}\sum_{i=1}^m ||\langle a_i, x\rangle| - \sqrt{b_i}|$, though this is not explicitly stated in the paper. Under proper initialization and assuming that a_i are uniformly sampled from a sphere, they prove linear convergence. Their argument relies on sophisticated probabilistic tools. In contrast, we disentangle the probabilistic statements (weak convexity and sharpness) from the deterministic convergence of Algorithm 1. As a proof of concept, we illustrate the proposed subgradient method synthetic and large-scale real image recovery problems.

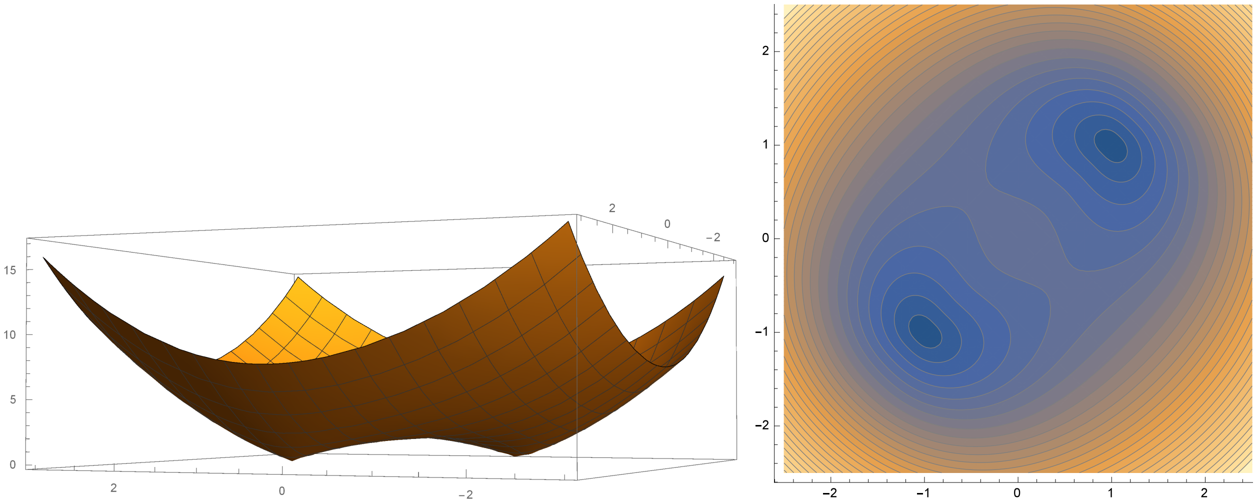

Figure 1: Depiction of the population objective f_P with $\bar{x} = (1, 1)$: graph (left), contours (right).

Weak convexity and sharpness, taken together, imply existence of a small neighborhood $\mathcal{X}$ of $\{\pm\bar{x}\}$ devoid of extraneous stationary points of f_S (see Lemma 3.1). On the other hand, it is intriguing to determine where the objective function f_S may have stationary points outside of this neighborhood. We complete the paper by proving that as the number of Gaussian measurements increases, the stationary points of the problem converge to a codimension two set, at a controlled rate. This suggests that there are much larger regions than the neighborhood $\mathcal{X}$, where the objective function has benign geometry.

We follow an intuitive and transparent strategy. Setting the groundwork, assume that a_i are i.i.d samples from a normal distribution $\mathsf{N}(0, I_{d\times d})$. Hence the problem $\min f_S$ is an empirical average approximation of the population objective

$$\min_x \; f_P(x) := \mathbb{E}_a[|(a^T x)^2 - (a^T \bar{x})^2|].$$

Seeking to determine the location of stationary points of f_S, we begin by first determining the stationary points of f_P. We base our analysis on the elementary observation that $f_P(x)$ depends on x only through the eigenvalues of the rank two matrix $X := xx^T - \bar{x}\bar{x}^T$. More precisely, equality holds:

$$f_P(x) = \frac{4}{\pi}\left[\mathrm{Tr}(X)\cdot \arctan\left(\sqrt{\left|\frac{\lambda_{\max}(X)}{\lambda_{\min}(X)}\right|}\right) + \sqrt{|\lambda_{\max}(X)\lambda_{\min}(X)|}\right] - \mathrm{Tr}(X).$$

See Figure 1 for a graphical illustration.

Using basic perturbation properties of eigenvalues, we will show that the stationary points of f_P are precisely

$$\{0\} \cup \{\pm\bar{x}\} \cup \{x \in \bar{x}^{\perp} : \|x\| = c \cdot \|\bar{x}\|\}, \tag{1.1}$$

where $c \approx 0.4416$ is a numerical constant. Intuitively, this region, excluding $\{\pm\bar{x}\}$, is where numerical methods may stagnate. In particular, f_P has no extraneous stationary points outside of the subspace $\bar{x}^{\perp}$. Along the way, we prove a number of results in matrix theory, which may be of independent interest. For example, we show that all stationary points of a composition of an orthogonally invariant gauge function with the map $x \mapsto xx^T - \bar{x}\bar{x}^T$ must be either perpendicular or collinear with $\bar{x}$.

Having located the stationary points of the population objective f_P, we turn to the stationary points of the subsampled function f_S. This is where the techniques commonly used for smooth formulations of the problem, such as those in [27], are no longer applicable; indeed, the subdifferential $\partial f_P(x)$ is usually a very poor approximation of $\partial f_S(x)$. Nonetheless, we show that the *graphs* of the subdifferentials ∂f_P and ∂f_S are close with high probability – a result closely related to the celebrated Attouch's convergence theorem [1]. The analysis of the stationary points of the subsampled objective flows from there. Namely, we show that there is a constant C such that whenever $m \geq Cd$, all stationary points x of f_S satisfy

$$\frac{\|x\|\|x-\bar{x}\|\|x+\bar{x}\|}{\|\bar{x}\|^3} \lesssim \sqrt[4]{\frac{d}{m}} \qquad \text{or} \qquad \left\{ \begin{array}{c} \left|\frac{\|x\|}{\|\bar{x}\|} - c\right| \lesssim \sqrt[4]{\frac{d}{m}} \cdot \left(1 + \frac{\|\bar{x}\|}{\|x\|}\right) \\ \frac{|\langle x, \bar{x}\rangle|}{\|x\|\|\bar{x}\|} \lesssim \sqrt[4]{\frac{d}{m}} \cdot \frac{\|\bar{x}\|}{\|x\|} \end{array} \right\},$$

with high probability; compare with (1.1). The argument we present is very general, relying only on weak convexity and concentration of f_S around its mean. Therefore, we believe that the technique may be of independent interest.

Finally, we comment on the structure of stationary points for the variant of the phase retrieval problem, in which the measurements b are corrupted by gross outliers. It is straightforward to obtain a full characterization of the stationary points of the population objective using the techniques developed in earlier sections.

The outline for the paper is as follows. Section 2 summarizes notation and basic results we will need. In Section 3, we analyze the linear convergence of the Polyak subgradient method for a class of nonsmooth, nonconvex functions, which includes the subsampled objective f_S. In Section 4, we perform a few proof-of-concept experiments, illustrating the performance of the Polyak subgradient method on synthetic and real large-scale image recovery problems. Section 5 is devoted to characterizing the nonsmooth landscape of the population objective f_P. In Section 6, we develop a concentration theorem for the subdifferential graphs of f_S and f_P, and briefly comment on robust extensions.

2 Notation

Throughout, we mostly follow standard notation. The symbol $\mathbb{R}$ will denote the real line, while $\mathbb{R}_+$ and $\mathbb{R}_{++}$ will denote nonnegative and strictly positive real numbers, respectively. We always endow $\mathbb{R}^d$ with the dot product $\langle x, y\rangle = x^T y$ and the induced norm $\|x\| := \sqrt{\langle x, x\rangle}$. The symbol $\mathbb{S}^{d-1}$ will denote the unit sphere in $\mathbb{R}^d$, while $B(x, r) := \{y : \|x - y\| < r\}$ will stand for the open ball around x of radius $r > 0$. For any set $Q \subset \mathbb{R}^d$, the distance function is defined by $\operatorname{dist}(x; Q) := \inf_{y \in Q} \|y - x\|$. The adjoint of a linear map $A \colon \mathbb{R}^d \to \mathbb{R}^m$ will be written as $A^* \colon \mathbb{R}^m \to \mathbb{R}^d$.

Since the main optimization problem we consider is nonsmooth, we will use some basic generalized derivative constructions. For a more detailed discussion, see for example the monographs of Mordukhovich [22] and Rockafellar-Wets [26].

Consider a function $f\colon \mathbb{R}^d \to \mathbb{R}$ and a point $\bar{x}$. The *Fréchet subdifferential* of f at $\bar{x}$, denoted $\hat{\partial} f(\bar{x})$, is the set of all vectors $v \in \mathbb{R}^d$ satisfying

$$f(x) \geq f(\bar{x}) + \langle v, x - \bar{x} \rangle + o(\|x - \bar{x}\|) \qquad \text{as } x \to \bar{x}.$$

Thus v lies in $\hat{\partial} f(\bar{x})$ if and only if the affine function $x \mapsto f(\bar{x}) + \langle v, x - \bar{x} \rangle$ minorities f near $\bar{x}$ up to first-order. Since the assignment $x \mapsto \hat{\partial} f(x)$ may have poor continuity properties, it is useful to extend the definition slightly. The *limiting subdifferential* of f at $\bar{x}$, denoted $\partial f(\bar{x})$, consists of all vectors $v \in \mathbb{R}^d$ such that there exist sequences x_i and $v_i \in \hat{\partial} f(x_i)$ satisfying $(x_i, f(x_i), v_i) \to (\bar{x}, f(\bar{x}), v)$. We say that $\bar{x}$ is *stationary* for f if the inclusion $0 \in \partial f(\bar{x})$ holds. The *graph* of ∂f is the set

$$\operatorname{gph} \partial f := \{(x, y) \in \mathbb{R}^d \times \mathbb{R}^d : y \in \partial f(x)\}.$$

For essentially all functions that we will encounter, the two subdifferentials, $\hat{\partial} f(\bar{x})$ and $\partial f(\bar{x})$, coincide. This is the case for C^1-smooth functions f, where $\hat{\partial} f(\bar{x})$ and $\partial f(\bar{x})$ consist only of the gradient $\nabla f(\bar{x})$. Similarly for convex function f, both subdifferentials reduce to the subdifferential in the sense of convex analysis:

$$v \in \partial f(\bar{x}) \qquad \Longleftrightarrow \qquad f(x) \geq f(\bar{x}) + \langle v, x - \bar{x} \rangle \qquad \text{for all } x \in \mathbb{R}^d.$$

Most of the nonsmooth functions we will encounter have a simple composite form:

$$F(x) := h(c(x)),$$

where $h\colon \mathbb{R}^m \to \mathbb{R}$ is a finite convex function and $c\colon \mathbb{R}^d \to \mathbb{R}^n$ is a C^1-smooth map. For such composite functions, the two subdifferentials coincide, and admit the intuitive chain rule [26, Theorem 10.6, Corollary 10.9]:

$$\partial F(x) = \nabla c(x)^* \partial h(c(x)) \qquad \text{for all } x \in \mathbb{R}^d.$$

A function $f\colon \mathbb{R}^d \to \mathbb{R}$ is called *ρ-weakly convex* if $f + \frac{\rho}{2}\|\cdot\|^2$ is a convex function. It follows immediately from [26, Theorem 12.17] that a lower-semicontinuous function f is ρ-weakly convex if and only if the inequality

$$f(y) \geq f(x) + \langle v, y - x \rangle - \frac{\rho}{2}\|y - x\|^2,$$

holds for all points $x, y \in \mathbb{R}^d$ and vectors $v \in \partial f(x)$.

Finally, we will often use implicitly the observation that the Lipschitz constant of any lower-semicontinuous function f on a convex open set U coincides with $\sup\{\|\zeta\| : x \in U,\ \zeta \in \partial f(x)\}$; see e.g. [26, Theorem 9.13].

3 Subgradient method

In this work, we consider the robust formulation of the (real) phase retrieval problem. Setting the stage, suppose we are given vectors $\{a_i\}_{i=1}^m$ in $\mathbb{R}^d$ and measurements $b := \langle a_i, \bar{x} \rangle^2$, for a fixed but unknown vector $\bar{x}$. The goal of the phase retrieval problem is to recover the vector $\bar{x} \in \mathbb{R}^d$, up to a sign flip. The formulation of the problem we consider in this work is:

$$\min_x \ f_S(x) := \frac{1}{m} \sum_{i=1}^{m} |\langle a_i^T, x \rangle^2 - b_i|.$$

The function f_S (in contrast to other possible formulations) has a number of desirable properties, which we will highlight as we continue.

In this section, we show that the landscape of the phase retrieval objective f_S favorably lends itself to classical subgradient methods. This is somewhat surprising because, until recently, convergence rates of subgradient methods in nonsmooth, nonconvex optimization have remained elusive; see the discussion in [6]. We show that, with proper initialization and under appropriate statistical assumptions, the standard Polyak subgradient method [24] linearly converges to $\pm x$.

3.1 Linear convergence of the subgradient method

The linear convergence guarantees that we present are mostly independent of the structure of f_S and instead rely only on a few general regularity properties, which f_S satisfies under mild statistical assumptions. Consequently, it will help the exposition in the current section to abstract away from f_S. Throughout the section, fix a function $g : \mathbb{R}^d \to \mathbb{R}$ for which there exist constants $\rho > 0$ and $\kappa > 0$ satisfying the following

1. **Weak Convexity.** The function $g + \frac{\rho}{2}\|\cdot\|^2$ is convex;

2. **Sharpness.** There exists $\bar{x} \in \mathbb{R}^d$ such that we have

$$g(x) - \min g \geq \kappa \|x - \bar{x}\| \|x + \bar{x}\| \qquad \text{for all } x \in \mathbb{R}^d.$$

3. **Minimizers.** The points $\pm\bar{x}$ minimize g.

Duchi and Ruan [10], following the work of Eldar-Mendelson [12], showed that the robust phase retrieval loss $f_S(\cdot)$ satisfies these three conditions with high-probability, under reasonable statistical assumptions. We will discuss these guarantees in Section 3.2, where we will instantiate the subgradient method on the robust phase retrieval objective. Consider now the standard Polyak subgradient method applied to g.

```
Algorithm 1: Polyak Subgradient Method
  Data: $x_0 \in \mathbb{R}^d$
  State k: $(k \geq 1)$
  Choose $\zeta_k \in \partial g(x_k)$.
  if $\zeta_k \neq 0$ then
      Set $x_{k+1} = x_k - \frac{g(x_k) - \min g}{\|\zeta_k\|^2} \zeta_k$.
  else
      Exit algorithm.
  end
```

We will shortly prove that the subgradient method converges linearly when initialized within a constant relative distance of $\pm\bar{x}$. To get some intuition for this guarantee, it is instructive to first observe that Properties 1, 2, and 3 imply existence of a neighborhood around $\pm\bar{x}$ of constant "relative size", where g has no extraneous stationary points. We will prove shortly that within this neighborhood, there is a domain of linear convergence of the subgradient method.

Lemma 3.1 (Neighborhood with no stationary points)**.** *Aside from $\pm\bar{x}$, the function g has no stationary points x satisfying*

$$\min\left\{\frac{\|x-\bar{x}\|}{\|x+\bar{x}\|}, \frac{\|x+\bar{x}\|}{\|x-\bar{x}\|}\right\} < \frac{2\kappa}{\rho}. \tag{3.1}$$

Proof. Consider a stationary point x of g satisfying $\|x-\bar{x}\| < \frac{2\kappa}{\rho}\|x+\bar{x}\|$. Properties 1, 2, and 3 then imply

$$\kappa\|x-\bar{x}\|\|x+\bar{x}\| \leq g(x) - g(\bar{x}) \leq \frac{\rho}{2}\|x-\bar{x}\|^2.$$

Hence, we conclude $x = \bar{x}$. The symmetric case, $\|x+\bar{x}\| < \frac{2\kappa}{\rho}\|x-\bar{x}\|$, is analogous. □

We are now ready to establish the linear rate of convergence of the subgradient method. For simplicity, we only consider initializing the method at a point within a constant relative error of $\bar{x}$. The symmetric statement for initializing near $-\bar{x}$ is completely symmetric.

Theorem 3.2 (Linear rate)**.** *Fix a real $\gamma \in (0,1)$ and suppose that we initialize Algorithm 1 at some point x_0 within the constant relative error:*

$$\frac{\|x_0-\bar{x}\|}{\|\bar{x}\|} \leq \frac{2\kappa\gamma}{\rho+\kappa}. \tag{3.2}$$

Then every iterate x_k produced by Algorithm 1 continues to satisfy (3.2) *and the sequence $\{x_k\}$ converges Q-linearly to $\bar{x}$ at the rate:*

$$\|x_{k+1}-\bar{x}\|^2 \leq \left(1 - \frac{4(1-\gamma)\kappa^3(\rho+(1-\gamma)\kappa)\|\bar{x}\|^2}{(\rho+\kappa)^2 L_g^2}\right)\|x_k-\bar{x}\|^2, \tag{3.3}$$

where L_g is the Lipschitz constant of g on the ball $B\left(\bar{x}, \frac{2\kappa\gamma\|\bar{x}\|}{\rho+\kappa}\right)$.

Proof. Define the ball $B := B\left(\bar{x}, \frac{2\kappa\gamma\|\bar{x}\|}{\rho+\kappa}\right)$. Observe first since the point $\left(1-\frac{2\gamma\kappa}{\rho+\kappa}\right)\bar{x}$ lies in B, every point $x \in B$ satisfies the inequalities:

$$\begin{aligned}\|x+\bar{x}\| &\geq \min_{z\in B}\ \|z+\bar{x})\| = \left\|\left(1-\frac{2\gamma\kappa}{\rho+\kappa}\right)\bar{x}+\bar{x}\right\| \\ &= \frac{2(\rho+(1-\gamma)\kappa)}{\rho+\kappa}\|\bar{x}\| = \frac{2\rho}{\rho+\kappa}\|\bar{x}\| + \frac{2(1-\gamma)\kappa}{\rho+\kappa}\|\bar{x}\| \geq \frac{\rho}{\kappa}\|x-\bar{x}\| + \frac{2(1-\gamma)\kappa}{\rho+\kappa}\|\bar{x}\|.\end{aligned} \tag{3.4}$$

We now proceed by induction. Suppose that x_k lies in B. To complete the inductive step, we must show that x_{k+1} lies in B and the inequality (3.3) holds. To this end, using Properties 1 and 2, we deduce

$$\begin{aligned}
&\|x_{k+1} - \bar{x}\|^2 \\
&= \|x_k - \bar{x}\|^2 + 2\langle x_k - \bar{x}, x_{k+1} - x_k\rangle + \|x_{k+1} - x_k\|^2 \\
&= \|x_k - \bar{x}\|^2 + \frac{2(g(x_k) - g(\bar{x}))}{\|\zeta_k\|^2}\langle \bar{x} - x_k, \zeta_k\rangle + \frac{(g(x_k) - g(\bar{x}))^2}{\|\zeta_k\|^2} \\
&\leq \|x_k - \bar{x}\|^2 + \frac{2(g(x_k) - g(\bar{x}))}{\|\zeta_k\|^2}\left(g(\bar{x}) - g(x_k) + \frac{\rho}{2}\|x_k - \bar{x}\|^2\right) + \frac{(g(x_k) - g(\bar{x}))^2}{\|\zeta_k\|^2} \qquad \text{(Property 1)} \\
&= \|x_k - \bar{x}\|^2 + \frac{(g(x_k) - g(\bar{x}))}{\|\zeta_k\|^2}\left(\rho\|x_k - \bar{x}\|^2 - (g(x_k) - g(\bar{x}))\right) \\
&\leq \|x_k - \bar{x}\|^2 + \frac{(g(x_k) - g(\bar{x}))}{\|\zeta_k\|^2}\left(\rho\|x_k - \bar{x}\|^2 - \kappa\|x_k - \bar{x}\|\|x_k + \bar{x}\|\right) \qquad \text{(Property 2)} \\
&= \|x_k - \bar{x}\|^2 - \frac{\rho(g(x_k) - g(\bar{x}))}{\|\zeta_k\|^2}\left(\frac{\kappa}{\rho}\|x_k + \bar{x}\| - \|x_k - \bar{x}\|\right)\|x_k - \bar{x}\|.
\end{aligned}$$

Equation (3.4) shows the inequality $\|x_k - \bar{x}\| \leq \frac{\kappa}{\rho}\|x_k + \bar{x}\|$. Hence we may apply Property 2 again to lower bound $g(x_k) - g(\bar{x})$, yielding

$$\|x_{k+1} - \bar{x}\|^2 \leq \left(1 - \frac{\kappa\rho\|x_k + \bar{x}\|}{\|\zeta_k\|^2}\left(\frac{\kappa}{\rho}\|x_k + \bar{x}\| - \|x_k - \bar{x}\|\right)\right)\|x_k - \bar{x}\|^2.$$

Finally, using the inequalities, $\|x_k + \bar{x}\| \geq \frac{2(\rho + (1-\gamma)\kappa)}{\rho + \kappa}\|\bar{x}\|$ and $\frac{\kappa}{\rho}\|x_k + \bar{x}\| - \|x - \bar{x}\| \geq \frac{2(1-\gamma)\kappa^2}{\rho(\rho+\kappa)}\|\bar{x}\|$ in (3.4), yields

$$\|x_{k+1} - \bar{x}\|^2 \leq \left(1 - \frac{4(1-\gamma)\kappa^3(\rho + (1-\gamma)\kappa)\|\bar{x}\|^2}{(\rho + \kappa)^2 L_g^2}\right)\|x_k - \bar{x}\|^2.$$

To finish the inductive step, we only need to note $\|x_{k+1} - \bar{x}\| \leq \|x_k - \bar{x}\|$, and hence x_{k+1} lies in B. The result follows. □

3.2 Convergence for the phase retrieval objective

We now turn to an application of Theorem 3.2 to the phase retrieval loss f_S. In particular, to run the subgradient method, we must only compute a subgradient of f_S, which can be easily done using the chain rule:

$$\frac{1}{m}\sum_{i=1}^{m} 2\langle a_i, x\rangle \cdot \operatorname{sign}(\langle a_i, x\rangle^2 - b_i)a_i \in \partial f_S(x).$$

Each iteration of Algorithm 1 thus requires a single pass through the set of measurement vectors. Next, observe that Property 3 clearly holds for f_S. Thus for a successful application of Theorem 3.2, we must only address the following questions:

(i) Describe the statistical conditions on the data generating mechanism, which insure Properties 1 and 2 hold with high probability.

(ii) Estimate the Lipschitz constant of f_S on the ball $B := B\left(\bar{x}, \frac{2\kappa\rho\|\bar{x}\|}{\rho+\kappa}\right)$.

(iii) Describe a good initialization procedure for producing $x_0 \in B$.

Essentially all of these points follow from the work of Duchi and Ruan [10], Eldar-Mendelson [12], and Wang et al. [29]. We summarize them here for the sake of completeness. Henceforth, let us suppose that $a_i \in \mathbb{R}^d$ (for $i = 1, \ldots, m$) are independent realizations of a random vector $a \in \mathbb{R}^d$.

3.2.1 Sharpness

In order, to ensure sharpness (or in the language of [12], "stability"), we make the following assumption on the distribution of a.

Assumption A. There exist constants $\kappa^*_{\mathrm{st}}, p_0 > 0$ such that for all $u, v \in \mathbb{S}^{d-1}$, we have

$$\mathbb{P}\left(|\langle a, v\rangle\langle a, u\rangle| \geq \kappa^*_{\mathrm{st}}\right) \geq p_0,$$

Roughly speaking, this mild assumption simply says that the random vector a has sufficient support in all directions. In particular, the standard Gaussian $a \sim \mathsf{N}(0, I_d)$ satisfies Assumption A with $\kappa^*_{\mathrm{st}} = 0.365$ and $p_0 = 0.25$; see [10, Example 1]. The following is proved in [10, Corollary 3.1].

Theorem 3.3 (Sharpness). *Suppose that Assumption A holds. Then there exists a numerical constant $c < \infty$ such that if $mp_0^2 \geq cd$, we have*

$$\mathbb{P}\left(f_S(x) - f_S(\bar{x}) \geq \frac{1}{2}\kappa^*_{\mathrm{st}} p_0 \|x - \bar{x}\|\|x + \bar{x}\| \quad \text{for all } x \in \mathbb{R}^d\right) \geq 1 - 2\exp\left(-\frac{mp_0^2}{32}\right).$$

Thus Assumption A implies sharpness of the problem with high probability.

3.2.2 Weak convexity

We next look at weak convexity of the objective f_S. We will need the following definition.

Definition 3.4. A random vector $a \in \mathbb{R}^d$ is σ^2-*sub-Gaussian* if for all unit vectors $v \in \mathbb{S}^{d-1}$, we have

$$\mathbb{E}\left[\exp\left(\frac{\langle a, v\rangle^2}{\sigma^2}\right)\right] \leq e.$$

Assumption B. The random vector a is σ^2-sub-Gaussian.

The following is a direct consequence of [10, Corollary 3.2].

Theorem 3.5 (Weak convexity). *Suppose that Assumption B holds. Then there exists a numerical constant $c < \infty$ such that whenever $m \geq cd$, the function f_S is $4\sigma^2$-weakly convex, with probability at least $1 - \exp\left(-\frac{m}{c}\right)$.*

Proof. This follows almost immediately from [10, Corollary 3.2]. Define the separable function $h(z_1,\ldots,z_m) := \frac{1}{m}\sum_{i=1}^m |z_i|$ and the map $F\colon \mathbb{R}^d \to \mathbb{R}^m$ with the i'th coordinate given by $F_i(x) := (a_i^T x)^2 - b_i$. Observe the equality $f_S(x) = h(F(x))$. Corollary 3.2 in [10] shows that there exists a numerical constant $c < \infty$ such that whenever $m \geq cd$, with probability at least $1 - \exp\left(-\frac{m}{c}\right)$, we have

$$f_S(y) \geq h(F(x) + \nabla F(x)(y-x)) - 2\sigma^2\|y-x\|^2 \qquad \text{for all } x, y \in \mathbb{R}^d.$$

Since h is convex, for any vector $v \in \partial h(F(x))$ we have

$$h(F(x) + \nabla F(x)(y-x)) \geq h(F(x)) + \langle v, \nabla F(x)(y-x)\rangle = f_S(x) + \langle \nabla F(x)^* v, y-x\rangle.$$

Taking into account the equality $\partial f_S(x) = \nabla F(x)^* \partial h(F(x))$, we conclude that f_S is $4\sigma^2$-weakly convex. □

3.2.3 Lipschitz constant on a ball

Let us next estimate the Lipschitz constant of f_S on a ball of a fixed radius. To this end, observe the chain of inequalities

$$\begin{aligned} |f_S(x) - f_S(y)| &\leq \frac{1}{m}\sum_{i=1}^m \left| |\langle a_i, x\rangle^2 - \langle a_i, \bar{x}\rangle^2| - |\langle a_i, y\rangle^2 - \langle a_i, \bar{x}\rangle^2| \right| \\ &\leq \frac{1}{m}\sum_{i=1}^m |\langle a_i, x\rangle^2 - \langle a_i, y\rangle^2| \qquad (3.5) \\ &= \|x-y\|\|x+y\| \cdot \frac{1}{m}\sum_{i=1}^m |\langle a_i, v\rangle\langle a_i, w\rangle|, \end{aligned}$$

where we set $v := \frac{x-y}{\|x-y\|}$ and $w := \frac{x+y}{\|x+y\|}$. Thus we would like to upper-bound the term $\frac{1}{m}\sum_{i=1}^m |\langle a_i, v\rangle\langle a_i, w\rangle|$ by a numerical constant, with high probability. Intuitively, there are two key ingredients that would ensure this bound: the random vector $a \in \mathbb{R}^d$ should have light tails (sub-Gaussian) and a should not concentrate too much along any single direction. A standard way to model the latter is through an isotropy assumption.

Definition 3.6 (Isotropy)**.** A random vector $a \in \mathbb{R}^d$ is *isotropic* if $\mathbb{E}[aa^T] = I_d$.

Note that $a \in \mathbb{R}^d$ is isotropic if and only if $\mathbb{E}[\langle a, v\rangle^2] = 1$ for all unit vectors $v \in \mathbb{S}^{d-1}$.

Assumption C. The random vector a is isotropic.

Assumptions B and C imply that the term $\frac{1}{m}\sum_{i=1}^m |\langle a_i, v\rangle\langle a_i, w\rangle|$ cannot deviate too much from its mean, uniformly over all unit vectors $v, w \in \mathbb{R}^d$. Indeed, the following is a special case of [12, Theorem 2.8].

Theorem 3.7 (Concentration)**.** *Suppose that Assumptions B and C hold. Then there exist constants c_1, c_2, c_3 depending only on σ so that with probability at least $1-2\exp(-c_2c_1^2\min\{m, d^2\})$, the inequality holds:*

$$\sup_{v,w\in\mathbb{S}^{d-1}} \left| \frac{1}{m}\sum_{i=1}^m |\langle a_i, v\rangle\langle a_i, w\rangle| - \mathbb{E}_a[|\langle a, v\rangle\langle a, w\rangle|] \right| \leq c_1^3 c_3 \left(\sqrt{\frac{d}{m}} + \frac{d}{m}\right).$$

We can now establish Lipschitz behavior of f_S on bounded sets.

Corollary 3.8 (Lipschitz constant on a ball)**.** *Suppose that Assumptions B and C hold. Then there exist constants c_1, c_2, c_3 depending only on σ such that with probability at least*

$$1 - 2\exp(-c_2 c_1^2 \min\{m, d^2\}),$$

we have

$$|f_S(x) - f_S(y)| \le \left(1 + c_1^3 c_3 \left(\sqrt{\frac{d}{m}} + \frac{d}{m}\right)\right) \|x - y\| \|x + y\| \qquad \text{for all } x, y \in \mathbb{R}^d, \tag{3.6}$$

and consequently

$$\max_{\zeta \in \partial f_S(x)} \|\zeta\| \le 2\left(1 + c_1^3 c_3 \cdot \left(\sqrt{\frac{d}{m}} + \frac{d}{m}\right)\right) \|x\| \qquad \text{for all } x \in \mathbb{R}^d. \tag{3.7}$$

Proof. Combining inequalities (3.5) with Theorem 3.7, we deduce that there exist constants c_1, c_2, c_3 depending only on σ such that with probability

$$1 - 2\exp(-c_2 c_1^2 \min\{m, d^2\}),$$

all points $x, y \in \mathbb{R}^d$ satisfy

$$|f_S(x) - f_S(y)| \le \left(\mathbb{E}_a[|\langle a, v\rangle\langle a, w\rangle|] + c_1^3 c_3 \left(\sqrt{\frac{d}{m}} + \frac{d}{m}\right)\right) \|x - y\| \|x + y\|,$$

where we set $v := \frac{x-y}{\|x-y\|}$ and $w := \frac{x+y}{\|x+y\|}$. Isotropy, in turn, implies

$$\mathbb{E}_a[|\langle a, v\rangle\langle a, w\rangle|] \le \sqrt{\mathbb{E}_a[|\langle a, v\rangle|^2]} \cdot \sqrt{\mathbb{E}_a[|\langle a, w\rangle|^2]} = 1,$$

Equation (3.6) follows immediately. Consequently, notice

$$\operatorname*{limsup}_{x,y \to z} \frac{|f_S(x) - f_S(y)|}{\|x - y\|} \le 2\left(1 + c_1^3 c_3 \left(\sqrt{\frac{d}{m}} + \frac{d}{m}\right)\right) \|z\|.$$

Since the Lipschitz constant of f_S at x coincides with the value $\max_{\zeta \in \partial f(x)} \|\zeta\|$ (see e.g. [26, Theorem 9.13]), the estimate (3.7) follows. □

We now have all the ingredients in place to apply Theorem 3.2 to the robust phase retrieval objective. Namely, under Assumptions A, B, and C, we may set

$$\rho := 4\sigma^2; \quad \kappa := \frac{1}{2}\kappa_{\text{st}}^* p_0; \quad L_g := 2\left(1 + c_1^3 c_3 \cdot \left(\sqrt{\frac{d}{m}} + \frac{d}{m}\right)\right)\left(1 + \frac{2\kappa}{\rho + \kappa}\right) \|\bar{x}\|. \tag{3.8}$$

In particular, for the Gaussian distribution, all ρ and κ are dimension independent. Thus, we have proved the following convergence guarantee – the main result of this section. To simplify the formulas, we apply Theorem 3.2 only with $\gamma := 1/2$.

Corollary 3.9 (Linear convergence for phase retrieval)**.** *Suppose that Assumptions A, B, and C hold. Then there exists a numerical constant $c < \infty$ and constants $R < \infty$ depending only on κ^*_{st}, p_0, and σ such that the following is true. Whenever we are in the regime, $\frac{c}{p_0^2} \le \frac{m}{d} \le d$, and we initialize Algorithm 1 at x_0 satisfying*

$$\frac{\|x_0 - \bar{x}\|}{\|\bar{x}\|} \le R, \tag{3.9}$$

we can be sure with probability at least

$$1 - 6\exp\left(-m \cdot \min\left\{\tfrac{p_0^2}{32}, c^{-1}, \tilde{c}\right\}\right)$$

that the produced iterates $\{x_k\}$ converge linearly to $\bar{x}$ at the rate:

$$\|x_{k+1} - \bar{x}\|^2 \le \left(1 - \frac{\kappa^3(\rho + \kappa/2)}{2(\rho+\kappa)^2\left(1 + \hat{c}\cdot\left(\sqrt{\frac{p_0^2}{c}} + \frac{p_0^2}{c}\right)\right)^2\left(1 + \frac{2\kappa}{\rho+\kappa}\right)^2}\right)\|x_k - \bar{x}\|^2. \tag{3.10}$$

Here, ρ, κ are defined in (3.8) *and $\hat{c}, \tilde{c}$ are constants that depend only on σ. In particular, the linear rate depends only on κ^*_{st}, p_0, and σ.*

Thus under typical statistical assumptions, the subgradient method converges linearly to $\bar{x}$, as long as one can initialize the method at a point x_0 satisfying the relative error condition $\|x_0 - \bar{x}\| \le R\|\bar{x}\|$, where R is a constant. A number of authors have proposed initialization strategies that can achieve this guarantee using only a constant multiple of d measurements [3, 10, 28–30]. For completeness, we record the strategy that was proposed in [29], and rigorously justified in [10]. To simplify the exposition, we only state the guarantees of the initialization under Gaussian assumptions on the measurement vectors a_i.

Theorem 3.10 ([10, Equation (15)])**.** *Assume that $a_i \sim \mathsf{N}(0, I_d)$ are i.i.d. standard Gaussian. Define the value $\hat{r}^2 := \frac{1}{m}\sum_{i=1}^m b_i$ and the index set $\mathcal{I}_{\mathrm{sel}} := \{i \in [m] \mid b_i \le \frac{1}{2}\hat{r}^2\}$. Set*

$$X^{\mathrm{init}} := \sum_{i \in \mathcal{I}_{\mathrm{sel}}} a_i a_i^T \qquad \text{and} \qquad \hat{d} := \operatorname*{argmin}_{d \in \mathbb{S}^{d-1}} d^T \mathrm{X}^{\mathrm{init}} d.$$

Then as soon as $\frac{m}{d} \gtrsim \varepsilon^{-2}$ the point $x_0 = \hat{r}\hat{d}$ satisfies

$$\frac{\min\{\|x_0 - \bar{x}\|, \|x_0 + \bar{x}\|\}}{\|\bar{x}\|} \lesssim \varepsilon \log\frac{1}{\varepsilon}$$

with probability at least $\ge 1 - 5\exp(-cm\varepsilon^2)$, where c is a numerical constant.

For more details and intuition underlying the initialization procedure, see [10, Section 3.3].

4 Numerical Illustration

In this section, as a proof of concept, we apply the subgradient method to medium and large-scale phase retrieval problems. All of our experiments were performed on a standard desktop: Intel(R) Core(TM) i7-4770 CPU3.40 GHz with 8.00 GB RAM.

We begin with simulated data. Set $d = 5000$. We generated a standard Gaussian random matrix $A \in \mathbb{R}^{m \times d}$ for each value $m \in \{11,000, 12225, 13500, 14750, 16000, 17250, 18500\}$; afterwards, we generated a Gaussian vector $\bar{x} \sim \mathsf{N}(0, I_d)$ and set $b = (A\bar{x})^2$. We then applied the initialization procedure, detailed in Theorem 3.10, followed by the subgradient method. Figure 4 plots the progress of the iterates produced by the subgradient method in each of the seven experiments. The top curve corresponds to $m = 11,000$, the bottom curve corresponds to $m = 18500$, while the curves for the other values of m interpolate in between. The iterates corresponding to $m = 11,000$ stagnate; evidently the number of measurements is too small. Indeed, the iterates do not even converge to a stationary point of the problem; this is in contrast to the prox-linear method in [10]. The iterates for the rest of the experiments converge to the true signal $\pm\bar{x}$ at an impressive linear rate.

In out second experiment, we use digit images from the MNIST data set [17]; these are relatively small so that the measurement matrices can be stored in memory. We illustrate the generic behavior of the algorithm on digit seven in Figure 2. The dimensions of the image we use are 32×32 (with 3 RGB channels). Hence, after vectorizing the dimension of the variable is $d = 3072$, while the number of Gaussian measurements is $m = 3d = 9216$. The initialization produced appears to be reasonable; the digit is visually discernible. The true image and the final image produced by the method are essentially identical. The convergence plot appears in Figure 3.

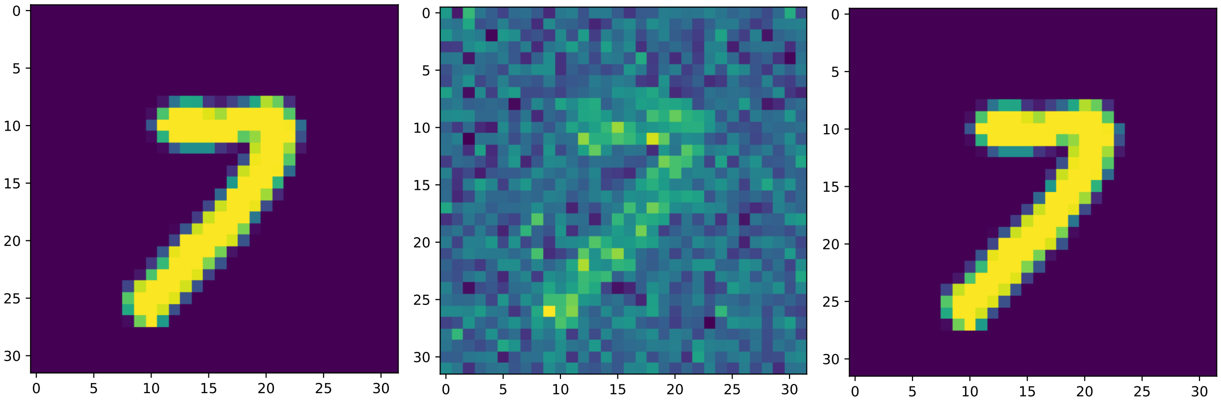

Figure 2: Digit recovery; left is the true digit, middle is the initial, right is the digit produced by the subgradient method. Dimension of the problem: $(n, d, m) = (32, 3072, 9216)$.

We next apply the subgradient method for recovering large-scale real images. To allow an easy comparison with previous work, we generate the data using the same process as in [10, Section 6.3]. We first describe how we generate the operator A. To this end, let $H \in \{-1, 1\}^{l \times l}/\sqrt{l}$ be a symmetric normalized Hadamard matrix. Consequently H satisfies the equation $H^2 = I_l$. Note that by the virtue of being Hadamard, matrix vector multiplication Hv requires time $l \log(l)$. For some integer k, we then generate k i.i.d. diagonal sign matrices

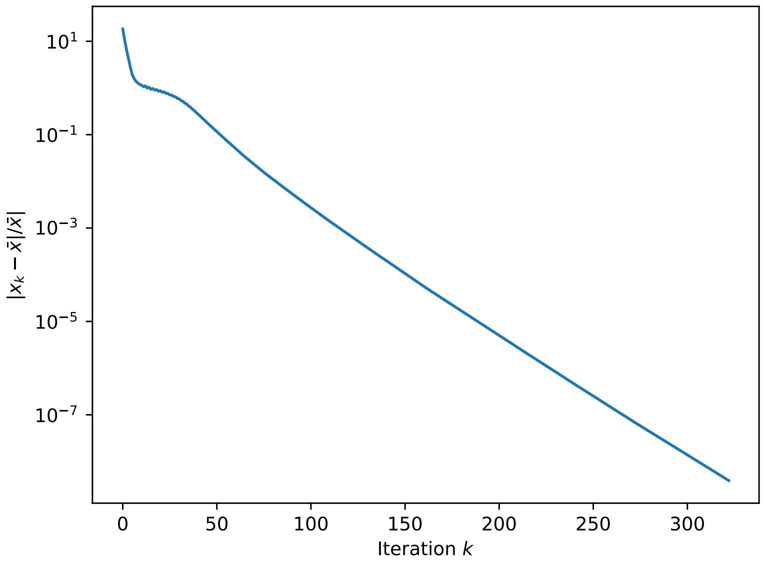

Figure 3: Convergence plot on MNIST digit (iterates vs. $\|x_k - \bar{x}\|/\|\bar{x}\|$).

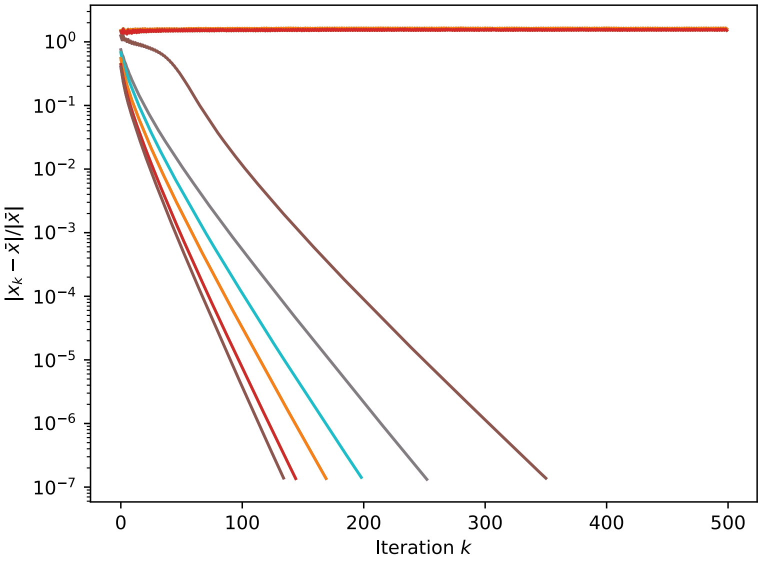

Figure 4: Convergence plot for the experiment on simulated data (iteration vs. $\|x_k - \bar{x}\|/\|\bar{x}\|$).

$S_1, \ldots, S_k \in \mathrm{diag}(\{-1,1\}^l)$ uniformly at random, and define $A = \begin{bmatrix} HS_1 & HS_2 & \ldots & HS_k \end{bmatrix}^T \in \mathbb{R}^{kl \times l}$.

We work with square colored images, represented as an array $\overline{X} \in \mathbb{R}^{n \times n \times 3}$. The number 3 appears because colored images have 3 RGB channels. We then stretch the matrix $\overline{X}$ into a $3n^2$-dimensional vector $\bar{x}$ and set the measurements $b_i := (A(i,\cdot)\bar{x})^2$, where $A(i,\cdot)$ denotes the i'th row of A. Thus if the image is $n \times n$, the number of variables in the problem formulation is $d := 3n^2$ and the number of measurements is $m := kd = 3kn^2$. We use the initialization procedure proposed in Theorem 3.10, with a standard power method (with a shift) to find the minimal eigenvalue of X^{init}. We complete the experiment by running the subgradient method (Algorithm 1), which requires no parameter tunning.

We perform a large scale experiment on two pictures taken by the Hubble telescope. Figure 5 describes the results of the experiment, while Figure 6 plots the iterate progress. The image on the left is 1024×1024 and we use $k = 3$ Hadamard matrices. Hence the dimensions of the problem are $d \approx 2^{22}$ and $m = 3d \approx 2^{24}$. The image on the right is 2048×2048 and we use $k = 3$ Hadamard matrices. Hence the dimensions of the problem are $d \approx 2^{24}$ and $m = 3d \approx 2^{25}$. For the image on the left, the entire experiment, including initialization and the subgradient method completed in 3 min. For the image on the right, it completed in 25.6 min. The vast majority of time was taken up by the initialization. Thus a more careful implementation and/or tunning of the initialization procedure could speed up the experiment.

5 Nonsmooth landscape of the robust phase retrieval

In this section, we pursue a finer analysis of the stationary points of the robust phase retrieval objective f_S. To motivate the discussion, recall that under Assumptions A and B, Lemma 5.1 shows that there are no extraneous stationary points x satisfying

$$\min\left\{\frac{\|x-\bar{x}\|}{\|x+\bar{x}\|}, \frac{\|x+\bar{x}\|}{\|x-\bar{x}\|}\right\} < \frac{\kappa_{\text{st}}^* p_0}{4\sigma^2}. \tag{5.1}$$

This result is uninformative when x is far away from $\bar{x}$ or when x is close to the origin. Therefore, it is intriguing to determine the location of *all* the stationary points of f_S. In this section, we will see that under a Gaussian observation model, the stationary points of f_S cluster around the codimension two set, $\{0, \pm\bar{x}\} \cup (\bar{x}^{\perp} \cap c \cdot \mathbb{S}^{d-1})$, where $c \approx 0.4416$ is a numerical constant.

5.1 A matrix analysis interlude

Before continuing, we introduce some basic matrix notation. We mostly follow [9, 18, 19]. The symbol $\mathcal{S}^d$ will denote the Euclidean space of real symmetric $d \times d$-matrices with the trace inner product $\langle X, Y\rangle := \operatorname{Tr}(XY)$. A function $f\colon \mathbb{R}^d \to \mathbb{R}$ is called *symmetric* if equality, $f(\sigma x) = f(x)$, holds for all coordinate permutations σ. For any symmetric function $f\colon \mathbb{R}^d \to \mathbb{R}$, we define the induced function on the symmetric matrices $f_\lambda\colon \mathcal{S}^d \to \mathbb{R}$ as the composition

$$f_\lambda(X) := f(\lambda(X)),$$

where $\lambda\colon \mathcal{S}^d \to \mathbb{R}^d$ assigns to each matrix $X \in \mathcal{S}^d$ its eigenvalues in nonincreasing order

$$\lambda_1(X) \geq \lambda_2(X) \geq \ldots \geq \lambda_n(X).$$

Note that f coincides with the restriction of f_λ to diagonal matrices, $f_\lambda(\operatorname{Diag}(x)) = f(x)$. Any function on $\mathcal{S}^d$ that has the form f_λ for some symmetric function f, is called *spectral.*

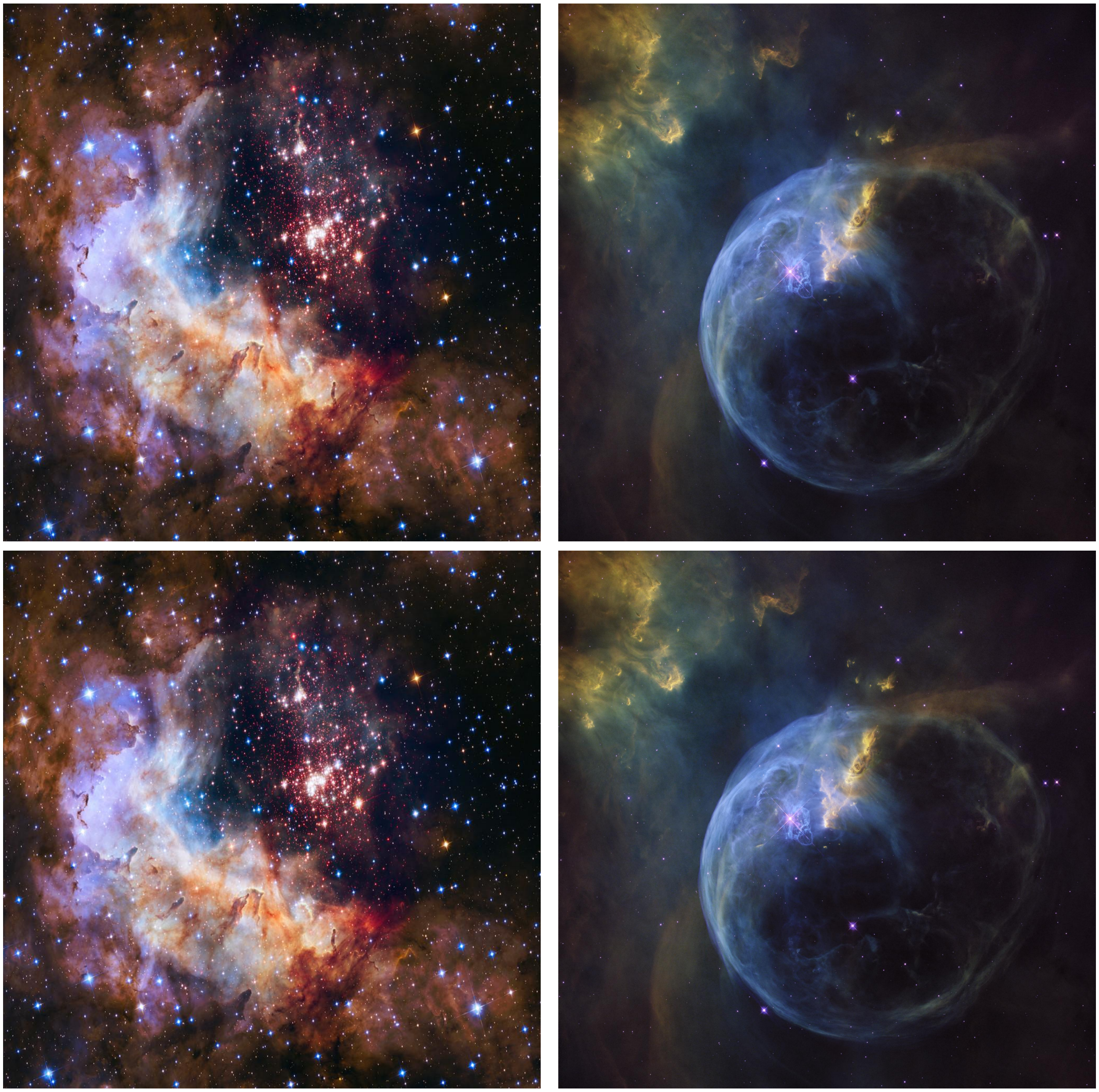

Figure 5: Image recovery; top row are the true images, bottom row are the images produced by the subgradient method. We do not record the images produced by the initialization as they were both completely black. Dimensions of the problem: $(n, k, d, m) \approx (1024, 3, 2^{22}, 2^{24})$ (left) and $(n, k, d, m) \approx (2048, 3, 2^{24}, 2^{25})$ (right).

Equivalently, spectral functions on $\mathcal{S}^d$ are precisely those that are invariant under conjugation by orthogonal matrices. Henceforth, let $\mathbb{O}^d$ be the set of real $d \times d$ orthogonal matrices.

Recall that two matrices $X, V \in \mathcal{S}^d$ commute if and if they can be simultaneously diagonalized. When describing variational properties of convex spectral functions, a stronger notion is needed. We say that X, V admit a *simultaneous ordered spectral decomposition* if

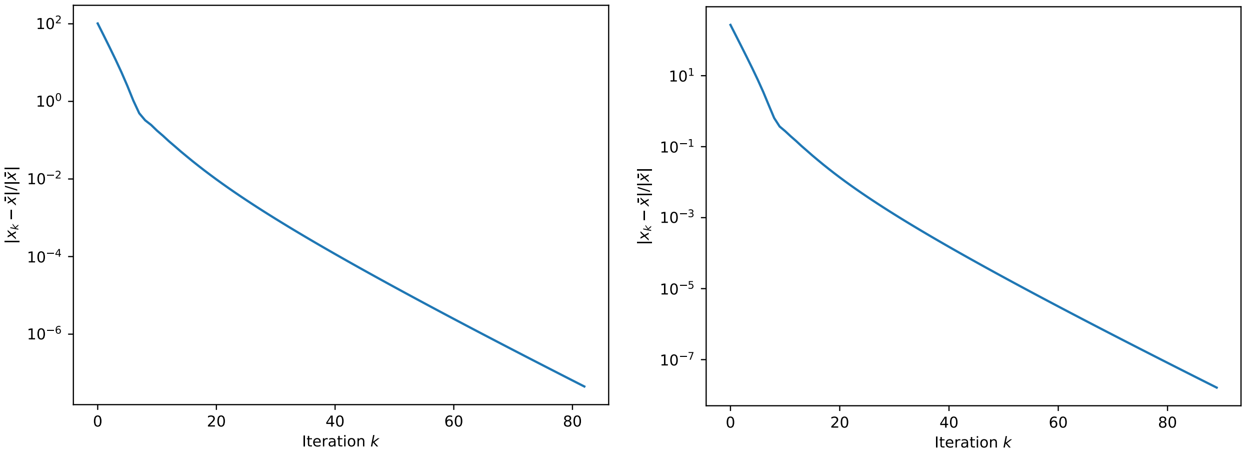

Figure 6: Convergence plot on the two Hubble images (iterates vs. $\|x_k - \bar{x}\|/\|\bar{x}\|$).

there exists a matrix $U \in \mathbb{O}^d$ satisfying

$$UVU^T = \mathrm{Diag}(\lambda(V)) \qquad \text{and} \qquad UXU^T = \mathrm{Diag}(\lambda(X)).$$

Thus the definition stipulates that X and V admit a simultaneous diagonalization, where the diagonals of the two diagonal matrices are simultaneously ordered.

The following is a foundational theorem in the convex analysis of spectral functions, due to Lewis [18]. An extension to the nonconvex setting was proved in [19], while a much simplified argument was recently presented in [9].

Theorem 5.1 (Spectral convex analysis)**.** *Consider a symmetric function $f\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$. Then f is convex if and only if f_λ is convex. Moreover, if f is convex, then the subdifferential $\partial f_\lambda(X)$ consists of all matrices $V \in \mathcal{S}^d$ satisfying $\lambda(V) \in \partial f(\lambda(X))$ and such that X and V admit a simultaneous ordered spectral decomposition.*

5.2 Landscape of the population objective

Henceforth, we fix a point $0 \neq \bar{x} \in \mathbb{R}^d$ and assume that $a \in \mathbb{R}^d$ is a normally distributed random vector $a \sim \mathsf{N}(0, I_d)$. In this section, we will investigate the population objective of the robust phase retrieval problem:

$$f_P(x) := \mathbb{E}_a\left[|\langle a, x\rangle^2 - \langle a, \bar{x}\rangle^2|\right].$$

Our aim is to prove the following result; see Figure 7 for a graphical depiction.

Theorem 5.2 (Landscape of the population objective)**.**
The stationary points of the population objective f_P are precisely

$$\{0\} \cup \{\pm\bar{x}\} \cup \{x \in \bar{x}^\perp : \|x\| = c \cdot \|\bar{x}\|\}, \tag{5.2}$$

where $c > 0$ (approx. $c \approx 0.4416$) is the unique solution of the equation $\frac{\pi}{4} = \frac{c}{1+c^2} + \arctan(c)$.

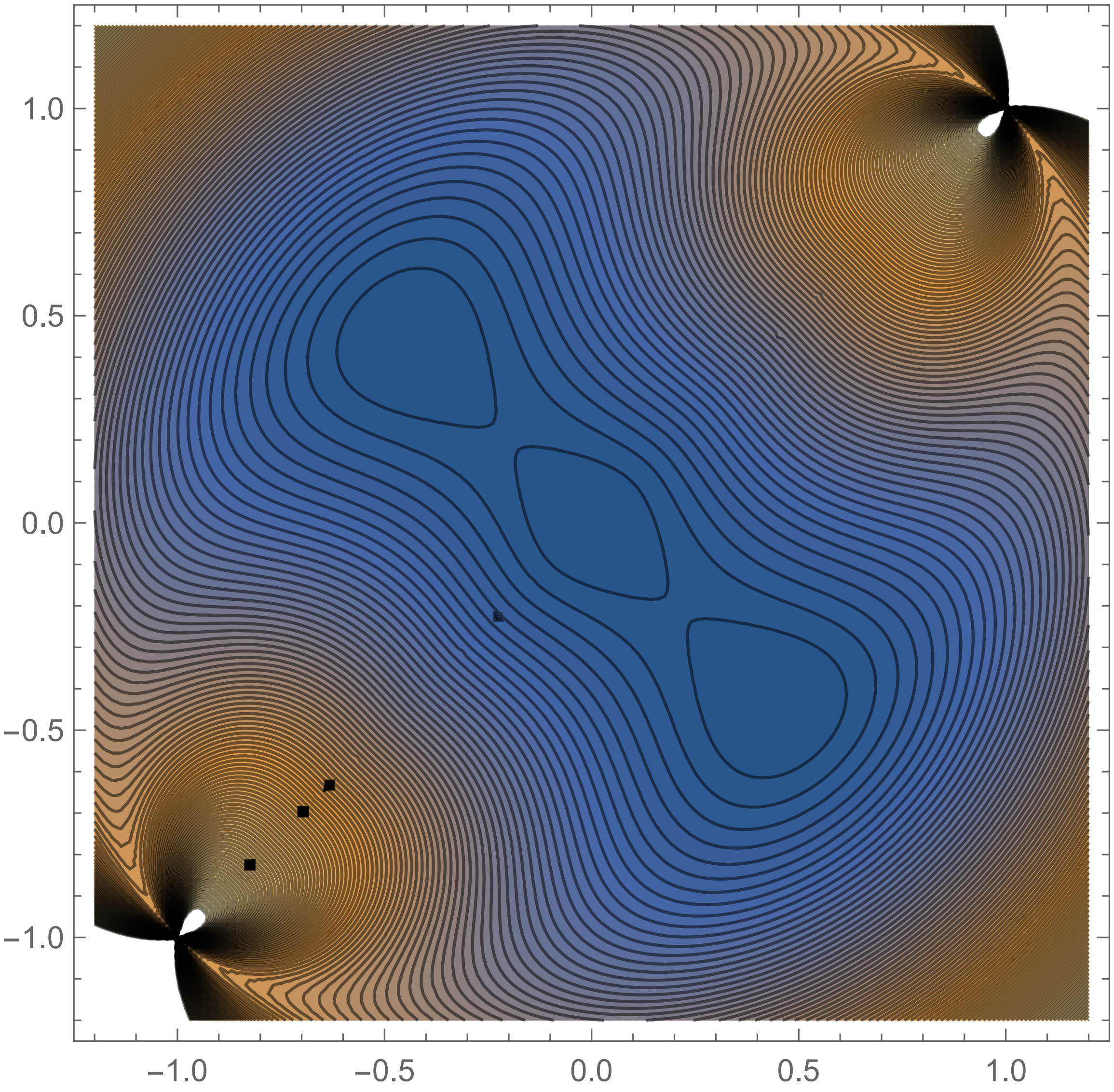

Figure 7: The contour plot of the function $x \mapsto \|\nabla f_P(x)\|$, where $\bar{x} = (1,1)$. The global minimizers of f_P are $\pm\bar{x}$, while the three extraneous stationary points are $(0,0)$ and $\pm c(-1,1)$, where $c \approx 0.4416$.

Theorem 5.2 provides an exact characterization of the stationary points of the population objective f_P. Looking ahead, when we will pass to the subsampled objective f_S in Section 6, we will show that every stationary point of f_S is *close* to an *approximately* stationary point of f_P. Therefore it will be useful to have an extension of Theorem 5.2 that locates approximately stationary points of f_P. This is the content of the following theorem.

Theorem 5.3 (Location of approximate stationary points)**.** *There exists a numerical constant $\gamma > 0$ such that the following holds. For any point $x \in \mathbb{R}^d$ with*

$$\varepsilon := \operatorname{dist}(0; \partial f_P(x)) \leq \gamma \|x\|,$$

it must be the case that $\|x\| \lesssim \|\bar{x}\|$ and x satisfies either

$$\|x\|\|x-\bar{x}\|\|x+\bar{x}\| \lesssim \varepsilon\|\bar{x}\|^2 \qquad \textit{or} \qquad \left\{ \begin{array}{c} |\|x\| - c\|\bar{x}\|| \lesssim \varepsilon\dfrac{\|\bar{x}\|}{\|x\|} \\ |\langle x, \bar{x}\rangle| \lesssim \varepsilon\|\bar{x}\| \end{array} \right\},$$

where $c > 0$ is the unique solution of the equation $\frac{\pi}{4} = \frac{c}{1+c^2} + \arctan(c)$.

We present the proofs of Theorem 5.2 in Section 5.3, and defer the proof of Theorem 5.3 to the Appendix (Section B), as the latter requires a much more delicate argument. At their core, the arguments rely on the observation that the population objective $f_P(x)$ depends on the input vector x only through the eigenvalues of the rank two matrix $xx^T - \bar{x}\bar{x}^T$. This observation was already implicitly used by Candès et al. [4]. Since this matrix will appear often in the arguments, we will use the symbol $X := xx^T - \bar{x}\bar{x}^T$ throughout. For ease of reference, we record the following simple observation: the matrix X is typically indefinite.

Lemma 5.4 (Eigenvalues of the rank two matrix)**.** *Suppose x and $\bar{x}$ are not collinear. Then X has exactly one strictly positive and one strictly negative eigenvalue.*

Proof. Suppose the claim is false. Then either X is positive semidefinite or negative semidefinite. Let us dispense with the first case. Observe $X \succeq 0$ if and only if $(x^Tv)^2 - (\bar{x}^Tv)^2 \geq 0$ for all v. Hence if X were positive semidefinite, we would deduce $x^\perp \subset \bar{x}^\perp$; that is, x and $\bar{x}$ are collinear, a contradiction. The case $X \preceq 0$ is analogous. □

The following lemma, as we alluded to above, shows that $f_P(x)$ depends on x only through the eigenvalues of the rank two matrix $X = xx^T - \bar{x}\bar{x}^T$.

Lemma 5.5 (Spectral representation of the population objective)**.**
For all points $x \in \mathbb{R}^d$, equality holds:

$$f_P(x) = \mathbb{E}_v\left[\Big|\langle \lambda(X), v\rangle\Big|\right], \tag{5.3}$$

where $v_i \in \mathbb{R}$ are i.i.d. chi-squared random variables $v_i \sim \chi_1^2$.

Proof. Observe the equalities:

$$\begin{aligned} f_P(x) = \mathbb{E}_a\left[|\langle a, x\rangle^2 - \langle a, \bar{x}\rangle^2|\right] &= \mathbb{E}_a[|\langle a, x-\bar{x}\rangle\langle a, x+\bar{x}\rangle|] \\ &= \mathbb{E}_a[|(x-\bar{x})^T aa^T(x+\bar{x})|] \\ &= \mathbb{E}_a\left[|\mathrm{Tr}\left(a^T(x+\bar{x})(x-\bar{x})^Ta\right)|\right]. \end{aligned}$$

Thus in terms of the matrix $M := (x+\bar{x})(x-\bar{x})^T$, we have $f_P(x) = \mathbb{E}_a\left[|\mathrm{Tr}\left(a^TMa\right)|\right]$. Taking into account the equalities $a^TMa = a^T\left(\frac{M+M^T}{2}\right)a = a^TXa$, we deduce

$$f_P(x) = \mathbb{E}_a\left[|\mathrm{Tr}\left(a^TXa\right)|\right].$$

Form now an eigenvalue decomposition $X = U\mathrm{Diag}(\lambda(X))U^T$, where $U \in \mathbb{R}^{d\times d}$ is an orthogonal matrix. Rotation invariance of the Gaussian distribution then implies

$$\mathbb{E}_a\left[|\mathrm{Tr}(a^TXa)|\right] = \mathbb{E}_a\left[|\mathrm{Tr}((Ua)^TX(Ua))|\right] = \mathbb{E}_u\left[\left|\sum_{i=1}^d \lambda_i(X)u_i^2\right|\right],$$

where u_i are i.i.d standard normals. The result follows. □

Thus Lemma 5.5 shows that the population objective f_P is a spectral function of X. Combined with Lemma 5.4, we deduce that there are two ways to rewrite the population objective in composite form:

$$f_P(x) = \varphi_\lambda(X) \qquad \text{and} \qquad f_P(x) = \zeta(\lambda_1(X), \lambda_d(X)),$$

where

$$\varphi(z) := \mathbb{E}_v\left[\left|\langle z, v\rangle\right|\right] \qquad \text{and} \qquad \zeta(y_1, y_2) := \mathbb{E}_{v_1,v_2}\left[|v_1 y_1 + v_2 y_2|\right]. \tag{5.4}$$

Notice that φ and ζ are norms on $\mathbb{R}^d$ and $\mathbb{R}^2$, respectively. It is instructive to compute ζ in closed form, yielding the following lemma. Since the proof is a straightforward computation, we have placed it in the appendix.

Lemma 5.6 (Explicit representation of the outer function)**.**
Let $v_1, v_2 \sim \chi_1^2$ be i.i.d. chi-squared. Then for all real $(y_1, y_2) \in \mathbb{R}_+ \times \mathbb{R}_-$, equality holds:

$$\mathbb{E}_{v_1,v_2}\left[|v_1 y_1 + v_2 y_2|\right] = \frac{4}{\pi}\left[(y_1 + y_2)\arctan\left(\sqrt{-\frac{y_1}{y_2}}\right) + \sqrt{-y_1 y_2}\right] - (y_1 + y_2).$$

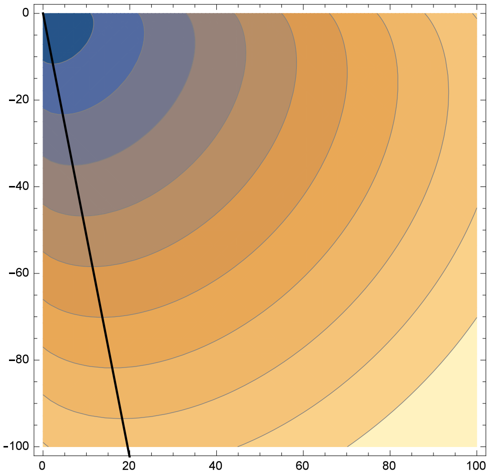

Figure 8: Contour plot of the function $\zeta(y_1, y_2) := \mathbb{E}_{v_1,v_2}\left[|v_1 y_1 + v_2 y_2|\right]$ on $\mathbb{R}_+ \times \mathbb{R}_-$. The black line depicts all points (y_1, y_2) with $\nabla_{y_1}\zeta(y_1, y_2) = 0$; for the explanation of the significance of this line, see Lemma 5.8.

Thus we have arrived at the following explicit representation of $f_P(x)$. Figure 1 in the introduction depicts the graph and the contours of the population objective.

Corollary 5.7 (Explicit representation of the population objective)**.**
The explicit representation holds:

$$f_P(x) = \frac{4}{\pi}\left[\mathrm{Tr}(X) \cdot \arctan\left(\sqrt{\left|\frac{\lambda_{\max}(X)}{\lambda_{\min}(X)}\right|}\right) + \sqrt{|\lambda_{\max}(X)\lambda_{\min}(X)|}\right] - \mathrm{Tr}(X).$$

5.3 Proof of Theorem 5.2

We next move on to the proof of Theorem 5.2. Let us first dispense with the easy implication, namely that every point in the set (5.2) is indeed stationary for f_P; in the process, we will see how the slope $c \approx 0.4416$ arises. Clearly $\pm\bar{x}$ are minimizers of f_P and are therefore stationary. The chain rule $\partial f_P(x) = \partial \varphi_\lambda(X)x$ implies that $x = 0$ is stationary as well. Fix now a point $x \in \bar{x}^\perp \setminus \{0\}$. Observe that the extremal eigenvalues of X are

$$\lambda_1(X) = \|x\|^2 \qquad \text{and} \qquad \lambda_d(X) = -\|\bar{x}\|^2,$$

with corresponding eigenvectors

$$e_1 := \frac{x}{\|x\|} \qquad \text{and} \qquad e_d := \frac{\bar{x}}{\|\bar{x}\|}.$$

Since $\lambda_1(X)$ and $\lambda_d(X)$ each have multiplicity one, the individual eigenvalue functions $\lambda_1(\cdot)$ and $\lambda_d(\cdot)$ are smooth at X with gradients

$$\nabla\lambda_1(X) = e_1 e_1^T \qquad \text{and} \qquad \nabla\lambda_d(X) = e_d e_d^T.$$

See for example [16, Theorem 5.11]. Setting $(y_1, y_2) := (\|x\|^2, -\|\bar{x}\|^2)$ and applying the chain rule to the decomposition $f_P(x) = \zeta(\lambda_1(X), \lambda_d(X))$ shows

$$\nabla f_P(x) = \left(\nabla_{y_1}\zeta(y_1,y_2)e_1e_1^T + \nabla_{y_2}\zeta(y_1,y_2)e_de_d^T\right)x = \nabla_{y_1}\zeta(y_1,y_2)x.$$

Thus a point $x \in \bar{x}^\perp \setminus \{0\}$ is stationary for f_P if and only if the partial derivative $\nabla_{y_1}\zeta(y_1,y_2)$ vanishes. The points (y_1,y_2) satisfying the equation $0 = \nabla_{y_1}\zeta(y_1,y_2)$ trace out exactly the line depicted in Figure 8.

Lemma 5.8. *The solutions of the equation* $0 = \nabla_{y_1}\zeta(y_1,y_2)$ *on* $\mathbb{R}_{++} \times \mathbb{R}_{--}$ *are precisely the tuples* $\{(c^2y, -y)\}_{y>0}$*, where* $c > 0$ *is the unique solution of the equation*

$$\frac{\pi}{4} = \frac{c}{1+c^2} + \arctan(c).$$

Note $c \approx 0.4416$*.*

Proof. Differentiating shows that $\omega(c) := \frac{c}{1+c^2} + \arctan(c)$ is a continuous strictly increasing function on $[0, +\infty)$ with $\omega(0) = 0$ and $\lim_{c\to+\infty}\omega(c) = \pi/2$. Hence the equation $\pi/4 = \omega(c)$ has a unique solution in the set $(0, \infty)$. A short computation yields the expression

$$\nabla_{y_1}\zeta(y_1,y_2) = \frac{4}{\pi}\left(\frac{y_1+y_2}{2\sqrt{-y_1/y_2}(y_1-y_2)} - \frac{y_2}{2\sqrt{-y_1y_2}} + \arctan\left(\sqrt{-\frac{y_1}{y_2}}\right)\right) - 1.$$

Set $y_1 = -c^2y_2$ for some $c > 0$ and $y_2 < 0$. Then plugging in this value of y_1, equality $0 = \nabla_{y_1}\zeta(y_1,y_2)$ holds if and only if

$$\pi/4 = \left(\frac{c}{1+c^2} + \arctan(c)\right).$$

This equation is independent of y_1 and its solution in c is exactly the value satisfying $\pi/4 = \omega(c)$. □

Thus we have proved the following.

Proposition 5.9. *Let $c > 0$ be the unique solution of the equation $\frac{\pi}{4} = \frac{c}{1+c^2} + \arctan(c)$. Then a point $x \in \bar{x}^\perp \setminus \{0\}$ is stationary for f_P if and only if equality $\|x\| = c\|\bar{x}\|$ holds.*

In particular, we have proved one implication in Theorem 5.2. To prove the converse, we must show that every stationary point of f_P lies in the set (5.2). Various approaches are possible based either on the decomposition $f_P(x) = \varphi_\lambda(X)$ or $f_P(x) = \zeta(\lambda_1(X), \lambda_d(X))$. We will focus on the former. We will prove a strong result about the location of stationary points of arbitrary convex spectral functions of X. Indeed, it will be more convenient to consider the more abstract setting as follows.

Throughout, we fix a symmetric convex function $f\colon \mathbb{R}^d \to \mathbb{R}$ and a point $0 \neq \bar{x} \in \mathbb{R}^d$, and define the function

$$g(x) := f_\lambda(xx^T - \bar{x}\bar{x}^T).$$

Note, the population objective f_P has this representation with $f = \varphi$. The chain rule directly implies

$$\partial g(x) = \partial f_\lambda(X)x.$$

Therefore, using Theorem 5.1 let us also fix a matrix $V \in \partial f_\lambda(X)$ and a matrix $U \in \mathbb{O}^d$ satisfying

$$\lambda(V) \in \partial f(\lambda(X)), \qquad V = U(\mathrm{Diag}(\lambda(V))U^T, \qquad \text{and} \qquad X = U\mathrm{Diag}(\lambda(X))U^T.$$

The following two elementary lemmas will form the core of the argument.

Lemma 5.10 (Eigenvalue correlation).
The following are true.

1. ***Eigenvalues.*** *We have $\lambda_i(X) = \langle U_i, x\rangle^2 - \langle U_i, \bar{x}\rangle^2$ for $i \in \{1, d\}$, and consequently*

$$\begin{aligned} 0 \le \lambda_1(X) \le \langle U_1, x\rangle^2 \le \|x\|^2 \\ 0 \le -\lambda_d(X) \le \langle U_d, \bar{x}\rangle^2 \le \|\bar{x}\|^2. \end{aligned} \tag{5.5}$$

2. ***Anticorrelation.*** *Equality holds:*

$$\langle U_1, x\rangle\langle U_d, x\rangle = \langle U_1, \bar{x}\rangle\langle U_d, \bar{x}\rangle.$$

3. ***Correlation.*** *Provided $x \notin \{\pm\bar{x}\}$, we have span$\{x, \bar{x}\} \subset$ span$\{U_1, U_d\}$ and*

$$\langle x, \bar{x}\rangle = \langle U_1, x\rangle\langle U_1, \bar{x}\rangle + \langle U_d, x\rangle\langle U_d, \bar{x}\rangle.$$

Proof. From the eigenvalue decomposition, we obtain

$$\begin{aligned} \lambda_1(X) = U_1^T X U_1 = \langle U_1, x\rangle^2 - \langle U_1, \bar{x}\rangle^2 \\ \lambda_d(X) = U_d^T X U_d = \langle U_d, x\rangle^2 - \langle U_d, \bar{x}\rangle^2. \end{aligned}$$

Taking into account that always $\lambda_1(X) \geq 0$ and $\lambda_1(X) \leq 0$ (Lemma 5.4), we conclude $\lambda_1(X) \leq \langle U_1, x\rangle^2$ and $\lambda_d(X) \geq -\langle U_d, \bar{x}\rangle^2$. Claim 1 follows. For Claim 2, simply observe

$$0 = U_d^T X U_1 = \langle U_1, x\rangle\langle U_d, x\rangle - \langle U_1, \bar{x}\rangle\langle U_d, \bar{x}\rangle.$$

To see Claim 3, for each $i \in \{1, d\}$ notice

$$\langle U_i, x\rangle x - \langle U_i, \bar{x}\rangle\bar{x} = XU_i = \lambda_i(X)U_i.$$

Suppose $x \notin \{\pm\bar{x}\}$. Then if x and $\bar{x}$ are not collinear, we may divide through by $\lambda_i(X)$ and deduce, $\operatorname{span}\{U_1, U_d\} = \operatorname{span}\{x, \bar{x}\}$. On the other hand, if x and $\bar{x}$ are collinear, then exactly one λ_1 or λ_d is nonzero, and then x lies in the span of the corresponding column of U. In either case, we may write $x = \langle U_1, x\rangle U_1 + \langle U_d, x\rangle U_d$ and $\bar{x} = \langle U_1, \bar{x}\rangle U_1 + \langle U_d, \bar{x}\rangle U_d$ in terms of their orthogonal expansions. We deduce

$$\langle x, \bar{x}\rangle = \langle \langle U_1, x\rangle U_1 + \langle U_d, x\rangle U_d, \langle U_1, \bar{x}\rangle U_1 + \langle U_d, \bar{x}\rangle U_d\rangle = \langle U_1, x\rangle\langle U_1, \bar{x}\rangle + \langle U_d, x\rangle\langle U_d, \bar{x}\rangle,$$

as claimed. □

Lemma 5.11 (Spectral subdifferential)**.** *The following hold:*

$$\max\{|\lambda_1(V)\langle U_1, x\rangle|, |\lambda_d(V)\langle U_d, x\rangle|\} \leq \|Vx\|, \tag{5.6}$$

and

$$g(x) - g(\bar{x}) \leq \lambda_1(V)\lambda_1(X) + \lambda_d(V)\lambda_d(X). \tag{5.7}$$

Proof. To see (5.6), observe that for all unit vectors $z \in \mathbb{S}^{d-1}$, we have $\|Vx\| \geq \langle z, Vx\rangle$. Thus, testing against all $z \in \{\pm U_1, \pm U_d\}$ yields the lower bounds (5.6). To prove the final bound (5.7), we exploit the convexity of f_λ. The subgradient inequality implies

$$f_\lambda(X) - f_\lambda(0) \leq \langle V, X\rangle = \lambda_1(V)\lambda_1(X) + \lambda_d(V)\lambda_d(X).$$

The result follows. □

The following corollary follows quickly from the previous two lemmas.

Corollary 5.12 (Stationary point inclusion)**.** *Suppose that x is stationary for g, that is $Vx = 0$. Then one of the following conditions holds:*

1. $g(x) \leq g(\bar{x})$
2. $x = 0$
3. $\langle x, \bar{x}\rangle = 0$, $\lambda_1(V) = 0$.

Moreover, if $\bar{x}$ minimizes g, then a point x is stationary for g if and only if x satisfies 1, 2, or 3.

Proof. Suppose $Vx = 0$ and that the first two conditions fail, that is $x \neq 0$ and $g(x) > g(\bar{x})$. We will show that the third condition holds. To this end, inequalities (5.6) and (5.7), along with Lemma 5.10, directly imply the following:

$$0 < g(x) - g(\bar{x}) \leq \lambda_1(V)\lambda_1(X) + \lambda_d(V)\lambda_d(X), \tag{5.8}$$

$$\lambda_1(V)\langle U_1, x\rangle = \lambda_d(V)\langle U_d, x\rangle = 0, \tag{5.9}$$

$$x = \langle U_1, x\rangle U_1 + \langle U_d, x\rangle U_d. \tag{5.10}$$

Aiming towards a contradiction, suppose $\lambda_1(V) \neq 0$. Then (5.9) and (5.10) imply $\langle U_1, x\rangle = 0$ and $\langle U_d, x\rangle \neq 0$. The second equation in (5.9), in turn, yields $\lambda_d(V) = 0$. Appealing to Lemma 5.10, we moreover deduce

$$0 \leq \lambda_1(X) = \langle U_1, x\rangle^2 - \langle U_1, \bar{x}\rangle^2 \leq 0.$$

Thus $\lambda_1(X) = 0$ and therefore the right-hand-side of (5.8) is zero, a contradiction. We have shown the equality $\lambda_1(V) = 0$, as claimed.

Inequality (5.8) implies $\lambda_d(V) \neq 0$ and $\lambda_d(X) \neq 0$, and hence by Inequality (5.9), we have $\langle U_d, x\rangle = 0$. Combining the latter equality with Lemma 5.10, we conclude $0 = \langle U_1, x\rangle\langle U_d, x\rangle = \langle U_1, \bar{x}\rangle\langle U_d, \bar{x}\rangle$. Note $\langle U_d, \bar{x}\rangle \neq 0$, since otherwise we would get $\lambda_d(X) = 0$ by (5.5). We conclude $\langle U_1, \bar{x}\rangle = 0$. Finally, Lemma 5.10 then yields

$$\langle x, \bar{x}\rangle = \langle U_1, x\rangle\langle U_1, \bar{x}\rangle + \langle U_d, x\rangle\langle U_d, \bar{x}\rangle = 0,$$

thereby completing the proof.

Now suppose that $\bar{x}$ minimizes g. Clearly $\pm\bar{x}$ is a stationary point of g. In addition, 0 is a stationary point of g because $V \cdot 0 = 0$. Thus, it remains to show that all points satisfying 3 are stationary. Thus suppose x satisfies 3 and $x \neq 0$. Then the eigenvalues of X are precisely $\|x\|^2$ and $-\|\bar{x}\|^2$ with eigenvectors $U_1 = \pm\frac{x}{\|x\|}$ and $U_d = \pm\frac{\bar{x}}{\|\bar{x}\|}$, respectively. Thus, we have $U^TVx = \mathrm{Diag}(\lambda(V))U^Tx = (\lambda_1(V)\langle U_1, x\rangle, 0, \ldots, 0, \lambda_d(V)\langle U_d, x\rangle)^T = 0$. We conclude $Vx = 0$, as required. □

The proof of Theorem 5.2 is now immediate.

Proof of Theorem 5.2. We have already proved that every point in the set (5.2) is stationary for f_P (Proposition 5.9). Thus we focus on the converse. In light of Proposition 5.9, it is sufficient to show that every stationary point x of f_P lies in the set $\{0, \pm\bar{x}\} \cup x^{\perp}$. This is immediate from Corollary 5.12 under the identification $f_P(x) = g(x) = \varphi_\lambda(X)$. □

6 Concentration and stability

Having determined the stationary points of the population objective f_P, we next turn to the stationary points of f_S. Our strategy is to show that with high probability, every stationary point of f_S is close to some stationary point of f_P. The difficulty is that it is not true that $\partial f_S(x)$ concentrates around $\partial f_P(x)$. Instead, we will see that the graphs of the two subdifferentials ∂f_S and ∂f_P concentrate, which is sufficient for our purposes. Our argument will rely on two basic properties, namely (1) the subsampled objective f_S concentrates well around f_P, and (2) the function f_S is weakly convex.

6.1 Concentration of subdifferential graphs

Armed with the concentration (Theorem 3.7) and the weak convexity (Theorem 3.5) guarantees, we can show that the graphs of ∂f_P and ∂f_S are close. The following theorem will be our main technical tool, and is of interest in its own right. In essence, the result is a quantitative extension of the celebrated Attouch's convergence theorem [1] in convex analysis. Henceforth, for any function $l\colon \mathbb{R}^d \to \overline{\mathbb{R}}$ and a point $\bar{x} \in \mathbb{R}^d$, with $f(\bar{x})$ finite, we define the local Lipschitz constant

$$\operatorname{lip}(l;\bar{x}) := \limsup_{x\to\bar{x}} \frac{|l(x)-l(\bar{x})|}{\|x-\bar{x}\|}.$$

Theorem 6.1 (Comparison)**.** *Consider four lsc functions $f, g, l, u\colon \mathbb{R}^d \to \overline{\mathbb{R}}$ and a pair $(x, v) \in \operatorname{gph}\partial g$. Suppose that l is locally Lipschitz continuous and that the following conditions*

$$\left\{\begin{aligned} l(y) \leq & f(y) - g(y) \leq u(y) \\ g(y) \geq g(x) &+ \langle v, y-x\rangle - \frac{\rho}{2}\|y-x\|^2 \end{aligned}\right\} \qquad \textit{hold for all points } y \in \mathbb{R}^d.$$

Then for any $\gamma > 0$, there exists a point $\hat{x}$ satisfying

$$\|\hat{x}-x\| \leq 2\gamma \qquad \textit{and} \qquad \operatorname{dist}(v;\partial f(\hat{x})) \leq 2\rho\gamma + \frac{u(x)-l(x)}{\gamma} + \operatorname{lip}(l;\hat{x}).$$

In particular, if $l(\cdot)$ is constant, we have the estimate

$$\operatorname{dist}\Big((x,v), \operatorname{gph}\partial f\Big) \leq \sqrt{4(\rho + \sqrt{2+\rho^2})} \cdot \sqrt{u(x)-l(x)}. \tag{6.1}$$

Proof. From the two assumptions, for any point $y \in \mathbb{R}^d$ we have

$$f(y) \geq g(y) + l(y) \geq g(x) + l(y) + \langle v, y-x\rangle - \frac{\rho}{2}\|y-x\|^2.$$

Define the function

$$\zeta(y) := f(y) - \langle v, y-x\rangle + \frac{\rho}{2}\|y-x\|^2 - l(y).$$

Clearly then we have

$$\zeta(x) - \inf\zeta \leq f(x) - l(x) - g(x) \leq u(x) - l(x). \tag{6.2}$$

Choose now any minimizer

$$\hat{x} \in \operatorname*{argmin}_y \left\{\zeta(y) + \frac{u(x)-l(x)}{4\gamma^2}\cdot\|y-x\|^2\right\}.$$

First order optimality conditions and the sum rule [26, Exercise 10.10] immediately imply

$$\frac{u(x)-l(x)}{2\gamma^2}\cdot(x-\hat{x}) \in \partial\zeta(\hat{x}) \subset \partial f(\hat{x}) - v + \rho(\hat{x}-x) + \operatorname{lip}(l;\hat{x})B(0,1),$$

and hence

$$\operatorname{dist}(v;\partial f(\hat{x})) \leq \frac{u(x)-l(x)}{2\gamma^2}\cdot\|\hat{x}-x\| + \rho\|\hat{x}-x\| + \operatorname{lip}(l;\hat{x}). \tag{6.3}$$

Next, we estimate the distance $\|\hat{x}-x\|$. To this end, observe from the definition of $\hat{x}$, we have

$$\zeta(\hat{x}) + \frac{u(x)-l(x)}{4\gamma^2}\cdot\|\hat{x}-x\|^2 \leq \zeta(x)$$

and hence

$$\frac{u(x)-l(x)}{4\gamma^2}\cdot\|\hat{x}-x\|^2 \leq \zeta(x)-\zeta(\hat{x}) \leq u(x)-l(x), \tag{6.4}$$

where the last inequality follows from (6.2). In the case $u(x) = l(x)$, we deduce $\zeta(x) = \zeta(\hat{x})$. Thus we equally well could have set $\hat{x} = x$, and the theorem follows immediately from (6.3). On the other hand, in the setting $u(x) > l(x)$, the inequality (6.4) immediately yields $\|\hat{x}-x\| \leq 2\gamma$, as claimed. Combining this inequality with (6.3) then gives the desired guarantee

$$\operatorname{dist}(v;\partial f(\hat{x})) \leq 2\rho\gamma + \frac{u(x)-l(x)}{\gamma} + \operatorname{lip}(l;\hat{x}).$$

Supposing l is a constant, we have the estimate

$$\operatorname{dist}\Big((x,v),\operatorname{gph}\partial f\Big) \leq \sqrt{4\gamma^2 + \left(2\rho\gamma + \frac{u(x)-l(x)}{\gamma}\right)^2}.$$

Minimizing the right-hand-side in γ yields the choice $\gamma = \frac{\sqrt{u(x)-l(x)}}{(8+4\rho^2)^{1/4}}$. With this value of γ, a quick computation yields the claimed guarantee (6.1). □

Let us now specialize the theorem to the setting where the lower and upper bounds $l(\cdot), u(\cdot)$ are functions of the product $\|x-\bar{x}\|\cdot\|x+\bar{x}\|$, as in phase retrieval.

Corollary 6.2. *Fix two functions $f, g\colon \mathbb{R}^d \to \mathbb{R}$. Suppose that g is ρ-weakly convex and that there is a point $\bar{x}$ and a real $\delta > 0$ such that the inequality*

$$|f(x)-g(x)| \leq \delta\|x-\bar{x}\|\cdot\|x+\bar{x}\| \qquad \textit{holds for all } x \in \mathbb{R}^d.$$

Then for any stationary point x of g, there exists a point $\hat{x}$ satisfying

$$\left\{\begin{array}{c} \|x-\hat{x}\| \leq \sqrt{\frac{4\delta}{\rho+2\delta}}\cdot\sqrt{\|x-\bar{x}\|\|x+\bar{x}\|}, \\ \operatorname{dist}(0;\partial f(\hat{x})) \leq (\delta + 2\sqrt{\delta(\rho+2\delta)})\cdot(\|x-\bar{x}\|+\|x+\bar{x}\|) \end{array}\right\}.$$

Proof. Set $u(x) := \delta\|x-\bar{x}\|\cdot\|x-\bar{x}\|$ and $l(x) := -\delta\|x-\bar{x}\|\cdot\|x-\bar{x}\|$ and observe $\operatorname{lip}(l;x) \leq \delta(\|x-\bar{x}\|+\|x+\bar{x}\|)$. Applying Theorem 6.1, we deduce that for any $\gamma > 0$, there exists a point $\hat{x}$ satisfying

$$\|\hat{x}-x\| \leq 2\gamma \qquad \text{and} \qquad \operatorname{dist}(0;\partial f(\hat{x})) \leq 2\rho\gamma + \frac{2\delta\|x-\bar{x}\|\|x+\bar{x}\|}{\gamma} + \delta(\|\hat{x}-\bar{x}\|+\|\hat{x}+\bar{x}\|).$$

The triangle inequality implies

$$\|\hat{x}-\bar{x}\| \leq 2\gamma + \|x-\bar{x}\| \qquad \text{and} \qquad \|\hat{x}+\bar{x}\| \leq 2\gamma + \|x+\bar{x}\|,$$

and therefore

$$\operatorname{dist}(0;\partial f(\hat{x})) \leq 2(\rho+2\delta)\gamma + \frac{2\delta\|x-\bar{x}\|\|x+\bar{x}\|}{\gamma} + \delta(\|x-\bar{x}\|+\|x+\bar{x}\|)$$

Minimizing this expression in $\gamma > 0$ yields the choice $\gamma := \sqrt{\frac{\delta\|x-\bar{x}\|\|x+\bar{x}\|}{\rho+2\delta}}$. Plugging in this value of γ and applying the AM-GM inequality then implies

$$\begin{aligned}\operatorname{dist}(0;\partial f(\hat{x})) &\leq 4\sqrt{\delta(\rho+2\delta)\|x-\bar{x}\|\|x+\bar{x}\|} + \delta(\|x-\bar{x}\|+\|x+\bar{x}\|)\\ &\leq (\delta + 2\sqrt{\delta(\rho+2\delta)})(\|x-\bar{x}\|+\|x+\bar{x}\|).\end{aligned}$$

The result follows. □

We now arrive at the main result of the section.

Corollary 6.3 (Subsampled stationary points)**.** *Consider the robust phase retrieval objective $f_S(\cdot)$ generated from i.i.d standard Gaussian vectors. There exist numerical constants $c_1, c_2 > 0$ such that whenever $m \geq c_1 d$, then with probability at least $1-2\exp(-\min\{m/c_1, c_2 m, d^2\})$, every stationary point x of f_S satisfies $\|x\| \lesssim \|\bar{x}\|$ and one of the two conditions:*

$$\frac{\|x\|\|x-\bar{x}\|\|x+\bar{x}\|}{\|\bar{x}\|^3} \lesssim \sqrt[4]{\frac{d}{m}} \qquad or \qquad \left\{\begin{array}{c} \left|\frac{\|x\|}{\|\bar{x}\|} - c\right| \lesssim \sqrt[4]{\frac{d}{m}} \cdot \left(1 + \frac{\|\bar{x}\|}{\|x\|}\right) \\ \frac{|\langle x, \bar{x}\rangle|}{\|x\|\|\bar{x}\|} \lesssim \sqrt[4]{\frac{d}{m}} \cdot \frac{\|\bar{x}\|}{\|x\|} \end{array}\right\},$$

where $c > 0$ is the unique solution of the equation $\frac{\pi}{4} = \frac{c}{1+c^2} + \arctan(c)$.

Proof. Theorem 3.7 shows that there exist constants $c_1, c_2 > 0$ such with probability at least $1-2\exp(-c_1\min\{m, d^2\})$, we have

$$|f_S(x) - f_P(x)| \leq \frac{c_2}{2}\left(\sqrt{\frac{d}{m}} + \frac{d}{m}\right)\|x-\bar{x}\|\|x+\bar{x}\| \qquad \text{for all } x \in \mathbb{R}^d. \tag{6.5}$$

Lemma 3.5, in turn, shows that there exist numerical constants $c_3, \rho > 0$ such that provided $m \geq c_3 d$, the function f_S is ρ-weakly convex, with probability at least $1 - \exp\left(-\frac{m}{c_3}\right)$. Let us now try to apply Corollary 6.2. To simplify notation, define $\Delta := \sqrt{\frac{d}{m}}$ and set $\delta := c_2\Delta$. Notice $\delta \geq \frac{c_2}{2}(\Delta + \Delta^2)$ and hence we may apply Corollary 6.2. We deduce that with high probability, for any stationary point x of f_S there exists a point $\hat{x} \in \mathbb{R}^d$ satisfying

$$\left\{\begin{array}{c} \|x-\hat{x}\| \leq \sqrt{\frac{4c_2\Delta}{\rho+2c_2\Delta}} \cdot \sqrt{\|x-\bar{x}\|\|x+\bar{x}\|}, \\ \operatorname{dist}(0;\partial f_P(\hat{x})) \leq (c_2\Delta + 2\sqrt{c_2\Delta(\rho+2c_2\Delta)}) \cdot (\|x-\bar{x}\|+\|x+\bar{x}\|) \end{array}\right\}. \tag{6.6}$$

Notice $\sqrt{\frac{4c_2\Delta}{\rho+2c_2\Delta}} \leq \sqrt{\Delta}\cdot\sqrt{\frac{4c_2}{\rho}} \leq 2C'\sqrt{\Delta}$ and $(c_2\Delta + 2\sqrt{c_2\Delta(\rho+2c_2\Delta)}) \leq C'\sqrt{\Delta}$ for some numerical constant C'. For notational convenience, set $D_x := \|x-\bar{x}\|+\|x+\bar{x}\|$. Thus, by the AM-GM inequality, the inclusion $\hat{x} \in B(x, C'\sqrt{\Delta}D_x)$ holds.

Claim 1. There exist constants $C'', \tau > 0$ such that with high probability, for all $\Delta < C''$, the inequality $\|x\| \leq \tau \|\bar{x}\|$ holds for any stationary point x of f_S.

Proof. We may assume that $\|\bar{x}\| \leq \|x\|$ since otherwise the result is trivial. Next, observe that $\|x\|$ and $\|\hat{x}\|$ have comparable norms:

$$\|\hat{x}\| \leq \|x\| + C'\sqrt{\Delta} D_x \leq (1 + 4C'\sqrt{\Delta})\|x\|,$$
$$\|\hat{x}\| \geq \|x\| - C'\sqrt{\Delta} D_x \geq (1 - 4C'\sqrt{\Delta})\|x\|,$$

where we have used the bound $D_x \leq 4\|x\|$ twice. To make the last bound meaningful, we may set $C'' < (\frac{1}{8C'})^2$, thereby ensuring $1 - 4C'\sqrt{\Delta} \geq 1/2$. Because the norms are comparable, we deduce

$$\operatorname{dist}(0; \partial f_P(\hat{x})) \leq C'\sqrt{\Delta} D_x \leq 4C'\sqrt{\Delta}\|x\| \leq \frac{4C'\sqrt{\Delta}}{(1 - 4C'\sqrt{\Delta})}\|\hat{x}\|. \tag{6.7}$$

Let us now decrease C'' if necessary to have $C'' < \min\{\left(\frac{1}{8C'}\right)^2, \left(\frac{\gamma}{8C'}\right)^2\}$, where γ is the fixed constant from Theorem 5.3. Then for all $\Delta < C''$, we have $1 - 4C'\sqrt{\Delta} \geq \frac{1}{2}$ and $\frac{4C'\sqrt{\Delta}}{1-4C'\sqrt{\Delta}} \leq 8C'\sqrt{\Delta} \leq \gamma$. Now we can apply Theorem 5.3 to $\hat{x}$, which guarantees that $\|\hat{x}\| \lesssim \|\bar{x}\|$. Thus because the norms of $\|x\|$ and $\|\hat{x}\|$ are comparable, we obtain the desired result. □

Provided $\Delta \leq \min\{\left(\frac{1}{8C'}\right)^2, \left(\frac{\gamma}{8C'}\right)^2\}$, we obtain from (6.7) and Claim 1 the estimate

$$\operatorname{dist}(0; \partial f_P(\hat{x})) \leq \varepsilon := C'\sqrt{\Delta} D_x \leq 8C'\sqrt{\Delta}\|\hat{x}\| \leq \gamma\|\hat{x}\|.$$

Applying Theorem 5.3 we find that the point $\hat{x} \in B(x, C'\sqrt{\Delta} D_x)$ satisfies either

$$\|\hat{x}\|\|\hat{x} - \bar{x}\|\|\hat{x} + \bar{x}\| \lesssim \sqrt{\Delta} D_x \|\bar{x}\|^2 \qquad \text{or} \qquad \left\{ \begin{array}{c} |\|\hat{x}\| - c\|\bar{x}\|| \lesssim \sqrt{\Delta} D_x \dfrac{\|\bar{x}\|}{\|\hat{x}\|} \\ |\langle \hat{x}, \bar{x} \rangle| \lesssim \sqrt{\Delta} D_x \|\bar{x}\| \end{array} \right\}. \tag{6.8}$$

Applying the triangle inequality and the bound $D_x \leq (2 + 2\tau)\|\bar{x}\|$, the claimed inequalities all follow (see Appendix A for a detailed explanation). □

6.2 Comments on Robustness

We have thus far assumed that the measurement vector $b = (Ax)^2$ has not been corrupted by errant noise. In this section, we record a few straightforward extensions of earlier results, which hold if the measurements b are noisy.

Assumption D. *Let $b_1, \ldots, b_m$ be m i.i.d. copies of*

$$\hat{b} = (a^T x)^2 + \delta \cdot \xi,$$

where $\delta \in \{0, 1\}$, $\xi \in \mathbb{R}$, and $a \in \mathbb{R}^d$ are independent random variables satisfying (1) $p_{\text{fail}} := P(\delta \neq 0) < 1$, (2) $\mathbb{E}[|\xi|] < \infty$, and (3) $a \sim \mathsf{N}(0, I_d)$.

Under corruption by $\delta \cdot \xi$, we define new population and subsampled objectives

$$\hat{f}_P(x) := \mathbb{E}_{a,\xi,\delta}[|(a^T x)^2 - (a^T \bar{x})^2 - \delta \cdot \xi|];$$
$$\hat{f}_S(x) := \frac{1}{m} \sum_{i=1}^{m} |(a_i^T x)^2 - b_i|.$$

Then by following the outline of the proof of Lemma 5.5, we arrive at a similar characterization of $\hat{f}_P$ as a spectral function.

Lemma 6.4 (Spectral representation of the population objective)**.** *For all points $x \in \mathbb{R}^d$, equality holds:*

$$\hat{f}_P(x) = \mathbb{E}_{v,\xi,\delta}\left[\Big|\langle \lambda(X), v\rangle - \delta \cdot \xi_i\Big|\right], \tag{6.9}$$

where $v_i \in \mathbb{R}$ are i.i.d. chi-squared random variables $v_i \sim \chi_1^2$.

Thus, we may write

$$\hat{f}_P(x) = \hat{\varphi}(\lambda(X)),$$

where $\hat{\varphi}$ is the convex symmetric function

$$\hat{\varphi}(z) := \mathbb{E}_{v,\xi,\delta}\left[\Big|\langle z, v\rangle - \delta \cdot \xi_i\Big|\right].$$

Moreover, provided that $\bar{x}$ is a minimizer of $\hat{f}_P$, the complete set of stationary points of $\hat{f}_p$ may be determined from Corollary 5.12. We prove this now.

Lemma 6.5. *For all $x \in \mathbb{R}^d$, the following inequality holds:*

$$\hat{f}_P(x) - \hat{f}_P(\pm\bar{x}) \geq (1 - 2p_{\text{fail}}) f_P(x).$$

Consequently, if $p_{\text{fail}} < 1/2$, the points $\pm\bar{x}$ are the only minimizers of $\hat{f}_p$, and there exists a numerical constant κ such that

$$\hat{f}_P(x) - \hat{f}_P(\pm\bar{x}) \geq \kappa(1 - 2p_{\text{fail}})\|x - \bar{x}\|\|x + \bar{x}\|.$$

Proof. By expanding the difference, we find that

$$\begin{aligned}\hat{f}_P(x) - \hat{f}_P(\pm\bar{x}) &= (1 - p_{\text{fail}})(f_P(x) - f_P(\bar{x})) + p_{\text{fail}}\mathbb{E}_{a,\xi}\left[|(a^T x)^2 - (a^T\bar{x})^2 - \xi| - |\xi|\right] \\ &\geq (1 - p_{\text{fail}}) f_P(x) - p_{\text{fail}}\mathbb{E}_a\left[|(a^T x)^2 - (a^T\bar{x})^2|\right] \\ &\geq (1 - 2p_{\text{fail}}) f_P(x).\end{aligned}$$

Only the sharpness inequality is left to prove, but this is simply a consequence of the sharpness of f_P, which was proved in [12, Corollary 3.7]. □

Therefore, by Corollary 5.12 we arrive at the complete characterization of the stationary points of $\hat{f}_P$.

Theorem 6.6. *The set of stationary points of $\hat{f}_P$ are precisely*

$$\{\pm x\} \cup \{0\} \cup \{x \mid \langle x, \bar{x}\rangle = 0, \textit{and } \exists \zeta \in \partial\hat{\varphi}(\lambda(X)), \max_i\{\zeta_i\} = 0\}.$$

The exact location of those stationary points orthogonal to $\bar{x}$ depends on the structure of the convex function $\hat{\varphi}$, which in turn depends the distribution of the noise $\delta \cdot \xi$. We will not attempt to characterize such $\hat{\varphi}$.

By the sharpness of $\hat{f}_P$, a quantitative version of Theorem 6.6 immediately follows from Theorem B.3. When coupled together with a concentration inequality like that in Theorem 3.7, such a theorem would imply concentration of the subdifferential graphs of $\hat{f}_S$ and $\hat{f}_P$. We omit these straightforward details.

Acknowledgments We thank Peng Zheng (University of Washington, Seattle) for invaluable help in developing the numerical illustrations.

References

[1] H. Attouch, J. L. Ndoutoume, and M. Théra. Epigraphical convergence of functions and convergence of their derivatives in Banach spaces. *Sém. Anal. Convexe*, 20:Exp. No. 9, 45, 1990.

[2] J.V. Burke and M.C. Ferris. Weak sharp minima in mathematical programming. *SIAM J. Control Optim.*, 31(5):1340–1359, 1993.

[3] Emmanuel J. Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via Wirtinger flow: theory and algorithms. *IEEE Trans. Inform. Theory*, 61(4):1985–2007, 2015.

[4] Emmanuel J. Candès, Thomas Strohmer, and Vladislav Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

[5] F. H. Clarke, R. J. Stern, and P. R. Wolenski. Proximal smoothness and the lower-C^2 property. *J. Convex Anal.*, 2(1-2):117–144, 1995.

[6] D. Davis and B. Grimmer. Proximally guided stochastic method for nonsmooth, nonconvex problems. *Preprint arXiv:1707.03505*, 2017.

[7] D. Drusvyatskiy and A.S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *To appear in Math. Oper. Res., arXiv:1602.06661*, 2016.

[8] D. Drusvyatskiy and C. Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Preprint arXiv:1605.00125*, 2016.

[9] D. Drusvyatskiy and C. Paquette. Variational analysis of spectral functions simplified. *J. Convex Anal.,*, 25(1), 2018.

[10] J.C. Duchi and F. Ruan. Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *Preprint arXiv:1705.02356*, 2017.

[11] J.C. Duchi and F. Ruan. Stochastic methods for composite optimization problems. *Preprint arXiv:1703.08570*, 2017.

[12] Yonina C. Eldar and Shahar Mendelson. Phase retrieval: stability and recovery guarantees. *Appl. Comput. Harmon. Anal.*, 36(3):473–494, 2014.

[13] H. Federer. Curvature measures. *Trans. Amer. Math. Soc.*, 93(3):418–491, 1959.

[14] Matthew Fickus, Dustin G. Mixon, Aaron A. Nelson, and Yang Wang. Phase retrieval from very few measurements. *Linear Algebra Appl.*, 449:475–499, 2014.

[15] Roger A. Horn and Charles R. Johnson. *Matrix analysis.* Cambridge University Press, Cambridge, second edition, 2013.

[16] Tosio Kato. *A short introduction to perturbation theory for linear operators.* Springer-Verlag, New York-Berlin, 1982.

[17] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998.

[18] A.S. Lewis. Convex analysis on the Hermitian matrices. *SIAM J. Optim.*, 6(1):164–177, 1996.

[19] A.S. Lewis. Nonsmooth analysis of eigenvalues. *Math. Program.*, 84(1, Ser. A):1–24, 1999.

[20] A.S. Lewis and S.J. Wright. A proximal method for composite minimization. *Math. Program.*, pages 1–46, 2015.

[21] Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, Mar 1993.

[22] B.S. Mordukhovich. *Variational Analysis and Generalized Differentiation I: Basic Theory.* Grundlehren der mathematischen Wissenschaften, Vol 330, Springer, Berlin, 2006.

[23] R.A. Poliquin and R.T. Rockafellar. Prox-regular functions in variational analysis. *Trans. Amer. Math. Soc.*, 348:1805–1838, 1996.

[24] B. T. Poljak. A general method for solving extremal problems. *Dokl. Akad. Nauk SSSR*, 174:33–36, 1967.

[25] R. Tyrrell Rockafellar. Favorable classes of Lipschitz-continuous functions in subgradient optimization. In *Progress in nondifferentiable optimization*, volume 8 of *IIASA Collaborative Proc. Ser. CP-82*, pages 125–143. Internat. Inst. Appl. Systems Anal., Laxenburg, 1982.

[26] R.T. Rockafellar and R.J-B. Wets. *Variational Analysis*. Grundlehren der mathematischen Wissenschaften, Vol 317, Springer, Berlin, 1998.

[27] J. Sun, Q. Qu, and J. Wright. A geometric analysis of phase retrieval. *To appear in Found. Comp. Math., arXiv:1602.06664*, 2017.

[28] Y.S Tan and R. Vershynin. Phase retreival via randomized kaczmarz: Theoretical guarantees. *arXiv:1605.08285*, 2017.

[29] G. Wang, G.B. Giannakis, and Y.C. Eldar. Solving systems of random quadratic equations via a truncated amplitude flow. *arXiv:1605.08285*, 2016.

[30] Huishuai Zhang, Yuejie Chi, and Yingbin Liang. Provable non-convex phase retrieval with outliers: Median truncated wirtinger flow. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 1022–1031. JMLR.org, 2016.

Appendices

A Auxiliary computations

Proof of Lemma 5.6. We let $\sigma_1 = y_1$ and $\sigma_2 = -y_2$. We may write

$$\begin{aligned}\mathbb{E}_v\left[|\sigma_1 v_1^2 - \sigma_2 v_2^2|\right] &= \frac{1}{2\pi}\int_{\mathbb{R}^2} |\sigma_1 v_1^2 - \sigma_2 v_2^2| \exp\left(-\left(\frac{v_1^2+v_2^2}{2}\right)\right) dv_1 dv_2 \\ &= \frac{1}{2\pi}\int_{R_1} \left(\sigma_1 v_1^2 - \sigma_2 v_2^2\right) \exp\left(-\left(\frac{v_1^2+v_2^2}{2}\right)\right) dv_1 dv_2 \\ &\quad + \frac{1}{2\pi}\int_{R_2} \left(\sigma_2 v_2^2 - \sigma_1 v_1^2\right) \exp\left(-\left(\frac{v_1^2+v_2^2}{2}\right)\right) dv_1 dv_2\end{aligned}$$

where

$$\begin{aligned}R_1 &= \{(v_1, v_2) \,:\, \sqrt{\sigma_1}|v_1| \geq \sqrt{\sigma_2}|v_2|\} \\ R_2 &= \{(v_1, v_2) \,:\, \sqrt{\sigma_2}|v_2| \geq \sqrt{\sigma_1}|v_1|\}.\end{aligned}$$

Using the convention $\arctan(\theta) \in \left[\frac{-\pi}{2}, \frac{\pi}{2}\right]$, we define the angle $\theta_1 := \arctan\left(\sqrt{\frac{\sigma_1}{\sigma_2}}\right)$. Passing to the polar coordinates, we deduce

$$\begin{aligned}&\frac{1}{2\pi}\int_{R_1} (\sigma_1 v_1^2 - \sigma_2 v_2^2) \exp\left(-\left(\frac{v_1^2+v_2^2}{2}\right)\right) dv_1 dv_2 \\ &\qquad\qquad = \frac{1}{2\pi}\int_{R_1} r^3(\sigma_1 \cos^2(\theta) - \sigma_2 \sin^2(\theta))\, e^{-r^2/2}\, dr d\theta.\end{aligned}$$

We break up the region R_1 into three wedges corresponding to the angles $[0, \theta_1]$, $[2\pi, 2\pi - \theta_1]$, and $[\pi + \theta_1, \pi - \theta_1]$. We will compute the integral over one of the regions. The rest will follow analogously. To this end, we successively deduce

$$\begin{aligned}
&\frac{1}{2\pi}\int_0^{\theta_1}\int_0^\infty r^3\Big(\sigma_1\cos^2(\theta) - \sigma_2\sin^2(\theta)\Big)e^{-r^2/2}\,drd\theta\\
&\qquad= \frac{1}{2\pi}\int_0^{\theta_1}\sigma_1\,(1+\cos(2\theta)) - \sigma_2\,(1-\cos(2\theta))\,d\theta\\
&\qquad= \frac{1}{2\pi}\big((\sigma_1-\sigma_2)\theta + (\sigma_1+\sigma_2)\sin(\theta)\cos(\theta)\big)\Big|_0^{\theta_1}\\
&\qquad= \frac{1}{2\pi}\big((\sigma_1-\sigma_2)\theta_1 + (\sigma_1+\sigma_2)\sin(\theta_1)\cos(\theta_1)\big)\\
&\frac{1}{2\pi}\int_{2\pi-\theta_1}^{2\pi}\int_0^\infty r^3\Big(\sigma_1\cos^2(\theta) - \sigma_2\sin^2(\theta)\Big)e^{-r^2/2}\,drd\theta\\
&\qquad= \frac{1}{2\pi}\big((\sigma_1-\sigma_2)\theta_1 + (\sigma_1+\sigma_2)\sin(\theta_1)\cos(\theta_1)\big)\\
&\frac{1}{2\pi}\int_{\pi-\theta_1}^{\pi+\theta_1}\int_0^\infty r^3\Big(\sigma_1\cos^2(\theta) - \sigma_2\sin^2(\theta)\Big)e^{-r^2/2}\,drd\theta\\
&\qquad= \frac{1}{2\pi}\big(2(\sigma_1-\sigma_2)\theta_1 + 2(\sigma_1+\sigma_2)\sin(\theta_1)\cos(\theta_1)\big).
\end{aligned}$$

Similarly, we see that for the region R_2, we have

$$\begin{aligned}
&\frac{1}{2\pi}\int_{R_2}(\sigma_2 v_2^2 - \sigma_1 v_1^2)\exp\left(-\left(\frac{v_1^2+v_2^2}{2}\right)\right)dv_1dv_2\\
&\qquad= \frac{1}{2\pi}\int_{R_2} r^3(\sigma_2\sin^2(\theta) - \sigma_1\cos^2(\theta))\,e^{-r^2/2}\,drd\theta.
\end{aligned}$$

We break up the region R_2 into two wedges where the angles range from $[\theta_1, \pi - \theta_1]$ and $[\pi + \theta_1, 2\pi - \theta_1]$ as we did in R_1. We will show the explicit computation for one of these terms and note the rest following using similar computations:

$$\begin{aligned}
&\frac{1}{2\pi}\int_{\theta_1}^{\pi-\theta_1}\int_0^\infty r^3\Big(\sigma_2\sin^2(\theta) - \sigma_1\cos^2(\theta)\Big)e^{-r^2/2}\,drd\theta\\
&\qquad= \frac{1}{2\pi}\int_{\theta_1}^{\pi-\theta_1}\sigma_2\,(1-\cos(2\theta)) - \sigma_1\,(1+\cos(2\theta))\,d\theta\\
&\qquad= \frac{1}{2\pi}\big((\sigma_2-\sigma_1)\theta - (\sigma_1+\sigma_2)\sin(\theta)\cos(\theta)\big)\Big|_{\theta_1}^{\pi-\theta_1}\\
&\qquad= \frac{1}{2\pi}\big((\sigma_2-\sigma_1)(\pi-2\theta_1) + 2(\sigma_1+\sigma_2)\sin(\theta_1)\cos(\theta_1)\big)\\
&\frac{1}{2\pi}\int_{\pi+\theta_1}^{2\pi-\theta_1}\int_0^\infty r^3\Big(\sigma_2\sin^2(\theta) - \sigma_1\cos^2(\theta)\Big)e^{-r^2/2}\,drd\theta\\
&\qquad= \frac{1}{2\pi}\big((\sigma_2-\sigma_1)(\pi-2\theta_1) + 2(\sigma_1+\sigma_2)\sin(\theta_1)\cos(\theta_1)\big).
\end{aligned}$$

By combining the computed integrals, we arrive at the full answer

$$\mathbb{E}_v\left[|\sigma_1 v_1^2 - \sigma_2 v_2^2|\right] = \frac{4}{\pi}\left[(\sigma_1 - \sigma_2)\arctan\left(\sqrt{\frac{\sigma_1}{\sigma_2}}\right) + \sqrt{\sigma_1\sigma_2}\right] - (\sigma_1 - \sigma_2),$$

as claimed. □

Proof showing Equation (6.8) *implies Corollary 6.3.* We observe that $D_x \leq 2\|x\| + 2\|\bar{x}\|$, which by Claim 1 gives $D_x \leq (2\tau + 2)\|\bar{x}\|$. First by applying the triangle inequality with $\|\hat{x} - x\| \leq C'\sqrt{\Delta}D_x$ and (6.8), we obtain

$$|\|x\| - c\|\bar{x}\|| \leq |\|x - \hat{x}\| + \|\hat{x}\| - c\|\bar{x}\|| \lesssim C'\sqrt{\Delta}D_x + \sqrt{\Delta}D_x\frac{\|\bar{x}\|}{\|\hat{x}\|}.$$

Using the bound on D_x gives the desired inequality. Next, we conclude

$$\begin{aligned} |\langle x, \bar{x}\rangle| &\leq \|\bar{x}\|\|x - \hat{x}\| + |\langle \hat{x}, \bar{x}\rangle| \\ &\lesssim C'\sqrt{\Delta}D_x\|\bar{x}\| + \sqrt{\Delta}D_x\|\bar{x}\|. \end{aligned}$$

Applying the bound on D_x, the result is shown. Lastly, using $\|x\| \leq \tau\|\bar{x}\|$ and $\|\hat{x}\| \lesssim \|\bar{x}\|$, we conclude

$$\begin{aligned} \|x\|\|x - \bar{x}\|\|x + \bar{x}\| &\leq (\|x - \hat{x}\| + \|\hat{x}\|)(\|x - \bar{x}\|\|x + \bar{x}\|) \\ &\leq D_x^2\|x - \hat{x}\| + \|\hat{x}\|\|x - \bar{x}\|\|x + \bar{x}\| \\ &\lesssim \|\bar{x}\|^2 D_x\sqrt{\Delta} + \|\hat{x}\|(\|\hat{x} - \bar{x}\| + \|\hat{x} - x\|)\|x + \bar{x}\| \\ &\lesssim \|\bar{x}\|^3\sqrt{\Delta} + \|\bar{x}\|^2\sqrt{\Delta}D_x + \|\hat{x}\|\|\hat{x} - \bar{x}\|\|x + \bar{x}\| \\ &\lesssim \|\bar{x}\|^3\sqrt{\Delta} + \|\hat{x}\|\|\hat{x} - \bar{x}\|(\|x - \hat{x}\| + \|\hat{x} + \bar{x}\|) \\ &\lesssim \|\bar{x}\|^3\sqrt{\Delta} + \|\bar{x}\|^2 D_x\sqrt{\Delta} + \|\hat{x}\|\|\hat{x} - \bar{x}\|\|\hat{x} + \bar{x}\| \\ &\lesssim \|\bar{x}\|^3\sqrt{\Delta} + \sqrt{\Delta}D_x\|\bar{x}\|^2. \end{aligned}$$

Dividing through by $\|\bar{x}\|^3$, finishes the proof. □

B Proof of Theorem 5.3

In this section, we will prove Theorem 5.3. Contrasting with Theorem 5.2, the proof of Theorem 5.3 is much more delicate, in large part relying on perturbation bounds on eigenvalues; e.g. Gershgorin theorem [15, Corollary 6.1.3]. We continue using the notation of Section 5.3. Namely, fix a symmetric convex function $f\colon \mathbb{R}^d \to \mathbb{R}$ and a point $\bar{x} \in \mathbb{R}^d \setminus \{0\}$, and define the function

$$g(x) := f_\lambda(xx^T - \bar{x}\bar{x}^T).$$

The chain rule directly implies

$$\partial g(x) = \partial f_\lambda(X)x.$$

Therefore, using Theorem 5.1 let us also fix a matrix $V \in \partial f_\lambda(X)$ and a matrix $U \in \mathbb{O}^d$ satisfying

$$\lambda(V) \in \partial f(\lambda(X)), \qquad V = U(\mathrm{Diag}(\lambda(V))U^T, \qquad \text{and} \qquad X = U\mathrm{Diag}(\lambda(X))U^T.$$

We begin with two technical lemmas.

Lemma B.1. *Suppose that there exists $\kappa > 0$ such that the inequality*

$$g(x) - g(\bar{x}) \geq \kappa\|x - \bar{x}\|\|x + \bar{x}\| \qquad \textit{holds for all } x \in \mathbb{R}^d.$$

Then for any $x \notin \{\pm\bar{x}\}$, we have $\max\{|\lambda_1(V)|, |\lambda_d(V)|\} \geq \kappa/2$.

Proof. Using Lemma 5.10, for $i \in \{1, d\}$ we obtain

$$|\lambda_i(X)| = |\langle U_i, x\rangle^2 - \langle U_i, \bar{x}\rangle^2| = |\langle U_i, x - \bar{x}\rangle\langle U_i, x + \bar{x}\rangle| \leq \|x - \bar{x}\|\|x + \bar{x}\|.$$

Taking into account (5.7), yields

$$\begin{aligned}\kappa\|x - \bar{x}\|\|x + \bar{x}\| \leq g(x) - g(\bar{x}) &\leq \lambda_1(V)\lambda_1(X) + \lambda_d(V)\lambda_d(X) \\ &\leq 2\max\{|\lambda_1(V)|, |\lambda_d(V)|\}\|x - \bar{x}\|\|x + \bar{x}\|,\end{aligned}$$

as desired. □

Lemma B.2. *Suppose that there exists $\kappa > 0$ such that the inequality*

$$g(x) - g(\bar{x}) \geq \kappa\|x - \bar{x}\|\|x + \bar{x}\| \qquad \textit{holds for all } x \in \mathbb{R}^d. \tag{B.1}$$

Then any point $x \in \mathbb{R}^d \setminus \{0\}$ satisfies

$$\frac{\kappa\|x - \bar{x}\|\|x + \bar{x}\|}{\|x\|} - \frac{(|\lambda_1(V)| + |\lambda_d(V)|)\|\bar{x}\|^2}{\|x\|} \leq \mathrm{dist}(0; \partial g(x)).$$

Proof. First, note that for $x \in \{\pm\bar{x}\}$, the result holds trivially, so we may assume $x \notin \{\pm\bar{x}\}$. Recall the equality $\partial g(x) = \partial f_\lambda(X)x$. Fix now a vector $V \in \partial f_\lambda(X)$ satisfying $\mathrm{dist}(0; \partial g(x)) = \|Vx\|$. Using convexity, we deduce

$$g(x) - g(\bar{x}) = f_\lambda(xx^T - \bar{x}\bar{x}^T) - f_\lambda(0) \leq \langle V, xx^T - \bar{x}\bar{x}^T\rangle \leq \|x\|\mathrm{dist}(0, \partial g(x)) + |\bar{x}^T V\bar{x}|. \tag{B.2}$$

We next upper bound the term $|\bar{x}^T V\bar{x}|$. To this end, fix a matrix $U \in \mathbb{O}^d$ satisfying $V = U\mathrm{Diag}(\lambda(V))U^T$ and $X = U\mathrm{Diag}(\lambda(X))U^T$, and such that the inclusion $\lambda(V) \in \partial f(\lambda(X))$ holds. Taking into account $\bar{x} \in \mathrm{span}\{U_1, U_d\}$ (Lemma 5.10), we deduce

$$|\bar{x}^T V\bar{x}| = |\lambda_1(V)\langle U_1, \bar{x}\rangle^2 + \lambda_d(V)\langle U_d, \bar{x}\rangle^2| \leq (|\lambda_1(V)| + |\lambda_d(V)|)\|\bar{x}\|^2.$$

Combining this estimate with (B.2) and (B.1) completes the proof. □

We next prove a quantitative version of Corollary 5.12. The argument follows a similar outline.

Theorem B.3 (Quantitative Version of Corollary 5.12)**.** *Suppose that there exists a constant $\kappa > 0$ such that the inequality*

$$g(y) - g(\bar{x}) \geq \kappa\|y - \bar{x}\|\|y + \bar{x}\| \qquad \textit{holds for all } y \in \mathbb{R}^d.$$

Suppose $|\lambda_1(V)|, |\lambda_d(V)|$ are both upper bounded by a numerical constant[1] *and set $\varepsilon := \|Vx\|$. Then there exists a numerical constant $\gamma > 0$, such that whenever $\varepsilon \leq \gamma \cdot \|x\|$, we have that $\|x\| \lesssim \|\bar{x}\|$ and x satisfies either*

$$\|x\|\|x - \bar{x}\|\|x + \bar{x}\| \lesssim \varepsilon\|\bar{x}\|^2 \qquad \textit{or} \qquad \begin{Bmatrix} |\lambda_1(V)| \lesssim \varepsilon/\|x\| \\ |\langle x, \bar{x}\rangle| \lesssim \varepsilon\|\bar{x}\| \end{Bmatrix}.$$

Proof. Clearly, we may suppose $x \notin \{0, \pm\bar{x}\}$ and $\varepsilon \neq 0$, since otherwise the theorem would hold vacuously. We will prove the following precise bound, which immediately implies the statement of the theorem: there exists a numerical constant $\gamma > 0$, such that whenever $\varepsilon \leq \gamma\|x\|$, the inequalities $\|x\| \leq \delta\|\bar{x}\|$ and

$$\min\left\{\frac{\|x - \bar{x}\|\|x + \bar{x}\|}{\frac{2}{\kappa}\max\left\{\left(\frac{\|x\|}{\sqrt{2}} + \frac{\sqrt{2}\|\bar{x}\|^2}{\|x\|}\right), \frac{\|x\|(\kappa\sqrt{2} + 2|\lambda_d(V)|)}{\kappa}\right\}}, \max\left\{\frac{|\lambda_1(V)|\|x\|}{\sqrt{2}}, \frac{\kappa|\langle x, \bar{x}\rangle|}{2\sqrt{2}\delta\|x\| + 2\|\bar{x}\|}\right\}\right\} \leq \|Vx\|, \tag{B.3}$$

hold, where we define the numerical constant

$$\delta := \sqrt{\frac{2(|\lambda_1(V)| + |\lambda_d(V)|)}{\kappa}} + 1.$$

As a first step, we show that $\|x\|$ is within a numerical constant of $\|\bar{x}\|$.

Claim 2. Provided $\gamma < \frac{\kappa(1-1/\delta)^2}{2}$, the inequality, $\|x\| \leq \delta\|\bar{x}\|$, holds.

Proof. Assume for sake of contradiction $\frac{\|x\|}{\|\bar{x}\|} > \delta := \sqrt{\frac{2(|\lambda_1(V)| + |\lambda_d(V)|)}{\kappa}} + 1$. Lemma B.1 shows $\max\{|\lambda_1(V)|, |\lambda_d(V)|\} \geq \frac{\kappa}{2}$, and therefore $\delta > 1$. Using the bound $\operatorname{dist}(0; \partial g(x)) \leq \|Vx\| = \varepsilon$ and Lemma B.2, we deduce:

$$\frac{\kappa\|x - \bar{x}\|\|x + \bar{x}\|}{\|x\|^2} - \frac{\varepsilon}{\|x\|} \leq \frac{(|\lambda_1(V)| + |\lambda_d(V)|)\|\bar{x}\|^2}{\|x\|^2}.$$

Clearly, we have

$$\frac{\kappa\|x - \bar{x}\|\|x + \bar{x}\|}{\|x\|^2} \geq \frac{\kappa(\|x\| - \|\bar{x}\|)^2}{\|x\|^2} \geq \kappa(1 - 1/\delta)^2.$$

Let us now choose $\gamma < \frac{\kappa(1-1/\delta)^2}{2}$, thereby guaranteeing $\frac{\varepsilon}{\|x\|} \leq \frac{\kappa(1-1/\delta)^2}{2}$. Hence, we obtain

$$\frac{\kappa(1 - 1/\delta)^2}{2(|\lambda_1(V)| + |\lambda_d(V)|)} \leq \frac{1}{|\lambda_1(V)| + |\lambda_d(V)|}\left(\frac{\kappa\|x - \bar{x}\|\|x + \bar{x}\|}{\|x\|^2} - \frac{\varepsilon}{\|x\|}\right) \leq \frac{\|\bar{x}\|^2}{\|x\|^2} < \frac{1}{\delta^2}.$$

[1] This holds whenever $(t, s) \mapsto f(t, s, 0, \ldots, 0)$ is Lipschitz continuous.

Rearranging yields

$$\frac{\kappa}{2(|\lambda_1(V)| + |\lambda_d(V)|)} < \frac{1}{(\delta-1)^2},$$

a contradiction. □

Looking back at the expression, define the values:

$$\rho_1 = \frac{\|x\|}{\sqrt{2}} \quad \text{and} \quad \rho_3 = \frac{2}{\kappa}\max\left\{\left(\frac{\|x\|}{\sqrt{2}} + \frac{\sqrt{2}\|\bar{x}\|^2}{\|x\|}\right), \frac{\|x\|(\kappa\sqrt{2} + 2|\lambda_d(V)|)}{\kappa}\right\}.$$

Notice that the inequality, $\varepsilon\rho_3 \geq \|x-\bar{x}\|\|x+\bar{x}\|$, would immediately imply the validity of the theorem. Thus, we assume $\varepsilon\rho_3 < \|x-\bar{x}\|\|x+\bar{x}\|$ throughout. It suffices now to show

$$|\lambda_1(V)| \leq \varepsilon/\rho_1 \qquad \text{and} \qquad |\langle x, \bar{x}\rangle| \leq \frac{\varepsilon}{\kappa}\left(2\sqrt{2}\delta\|x\| + \|\bar{x}\|\right).$$

We do so in order. We begin by observing that the inequality (5.6) guarantees

$$\max\{|\lambda_1(V)\langle U_1, x\rangle|, |\lambda_d(V)\langle U_d, x\rangle|\} \leq \varepsilon. \tag{B.4}$$

Claim 3. The inequality $|\lambda_1(V)| < \varepsilon/\rho_1$ holds.

Proof. Let us assume the contrary, $|\lambda_1(V)| \geq \varepsilon/\rho_1$. Inequality (B.4) then implies $|\langle U_1, x\rangle| \leq \rho_1$, while Lemma 5.10 in turn guarantees

$$0 \leq \lambda_1(X) = \langle U_1, x\rangle^2 - \langle U_1, \bar{x}\rangle^2 \leq \rho_1^2.$$

Taking into account $\langle U_1, x\rangle^2 + \langle U_d, x\rangle^2 = \|x\|^2$ (Lemma 5.10, correlation), we deduce $\langle U_d, x\rangle^2 \geq \|x\|^2 - \rho_1^2$. Combining this with (B.4), we deduce

$$|\lambda_d(V)| \leq \frac{\varepsilon}{|\langle U_d, x\rangle|} \leq \frac{\varepsilon}{\sqrt{\|x\|^2 - \rho_1^2}}.$$

Therefore, using the correlation inequality (5.7), we find

$$\begin{aligned}
\varepsilon\rho_3\kappa < \kappa\|x-\bar{x}\|\|x+\bar{x}\| \leq g(x) - g(\bar{x}) &\leq \lambda_1(V)\lambda_1(X) + \lambda_d(V)\lambda_d(X) \\
&\leq |\lambda_1(V)|(\langle U_1, x\rangle^2 - \langle U_1, \bar{x}\rangle^2) + \frac{\varepsilon}{\sqrt{\|x\|^2 - \rho_1^2}}\left(\langle U_d, \bar{x}\rangle^2 - \langle U_d, x\rangle^2\right) \\
&\leq \varepsilon|\langle U_1, x\rangle| + \frac{\varepsilon\langle U_d, \bar{x}\rangle^2}{\sqrt{\|x\|^2 - \rho_1^2}} \\
&\leq \varepsilon\left(\rho_1 + \frac{\|\bar{x}\|^2}{\sqrt{\|x\|^2 - \rho_1^2}}\right).
\end{aligned}$$

Dividing through by ε and plugging in the value of ρ_1 yields

$$\rho_3\kappa < \frac{\|x\|}{\sqrt{2}} + \frac{\sqrt{2}\|\bar{x}\|^2}{\|x\|},$$

which contradicts the definition of ρ_3. □

Let us now decrease $\gamma > 0$ further by ensuring $\gamma < \min\{\frac{\kappa(1-1/\delta)^2}{2}, \frac{\kappa}{2\sqrt{2}}\}$. Thus, from Claim 3 and our standing assumption $\|Vx\| \leq \frac{\kappa\|x\|}{2\sqrt{2}}$, we conclude

$$|\lambda_1(V)| < \frac{\sqrt{2}\varepsilon}{\|x\|} < \frac{\kappa}{2}.$$

Lemma B.1 guarantees, $\max\{|\lambda_1(V)|, |\lambda_d(V)|\} \geq \kappa/2$; thus, we deduce $|\lambda_d(V)| \geq \kappa/2$. Applying (B.4), we find that

$$|\langle U_d, x\rangle| \leq \frac{\varepsilon}{|\lambda_d(V)|} \leq \frac{2\varepsilon}{\kappa}. \tag{B.5}$$

Thus, by Lemma 5.10, we have

$$|\langle U_1, \bar{x}\rangle\langle U_d, \bar{x}\rangle| = |\langle U_1, x\rangle\langle U_d, x\rangle| \leq \frac{2\|x\|\varepsilon}{\kappa}. \tag{B.6}$$

Claim 4. The inequality $|\langle U_d, \bar{x}\rangle| > |\langle U_1, \bar{x}\rangle|$ holds.

Proof. Let us assume the contrary $|\langle U_d, \bar{x}\rangle| \leq |\langle U_1, \bar{x}\rangle|$. Then from (B.6) we obtain[2] $\langle U_d, \bar{x}\rangle^2 < \frac{2\|x\|\varepsilon}{\kappa}$. Hence from Lemma 5.10, we find that $|\lambda_d(X)| \leq \langle U_d, \bar{x}\rangle^2 \leq \frac{2\|x\|\varepsilon}{\kappa}$. Putting these facts together with the correlation inequality (5.7), we successively deduce

$$\begin{aligned} \varepsilon\rho_3\kappa < \kappa\|x-\bar{x}\|\|x+\bar{x}\| \leq g(x) - g(\bar{x}) &\leq \lambda_1(V)\lambda_1(X) + |\lambda_d(V)| \cdot |\lambda_d(X)| \\ &\leq \frac{\sqrt{2}\varepsilon}{\|x\|} \cdot \lambda_1(X) + |\lambda_d(V)| \cdot \frac{2\|x\|\varepsilon}{\kappa} \\ &\leq \frac{\kappa\sqrt{2}\varepsilon\|x\|}{\kappa} + \frac{2|\lambda_d(V)|\varepsilon\|x\|}{\kappa}, \end{aligned}$$

where the last inequality uses the bound $\lambda_1(X) \leq \|x\|^2$. Therefore, we have reached a contradiction to the definition of ρ_3. □

Combining Claim 4 with the expression $\langle U_1, \bar{x}\rangle^2 + \langle U_d, \bar{x}\rangle^2 = \|\bar{x}\|^2$, we conclude $\langle U_d, \bar{x}\rangle^2 \geq \frac{\|\bar{x}\|^2}{2}$. Therefore, (B.6) and Claim 2 imply the strong result:

$$|\langle U_1, \bar{x}\rangle| \leq \frac{2\sqrt{2}\varepsilon\|x\|}{\kappa\|\bar{x}\|} \leq \frac{2\sqrt{2}\varepsilon\delta}{\kappa}. \tag{B.7}$$

Thus combining Claim 2, Lemma 5.10, and (B.5) we conclude

$$\begin{aligned} |\langle x, \bar{x}\rangle| = |\langle U_1, x\rangle\langle U_1, \bar{x}\rangle + \langle U_d, x\rangle\langle U_d, \bar{x}\rangle| &\leq |\langle U_1, \bar{x}\rangle| \cdot \|x\| + |\langle U_d, x\rangle| \cdot \|\bar{x}\| \\ &\leq \frac{\varepsilon}{\kappa}\left(2\sqrt{2}\delta\|x\| + 2\|\bar{x}\|\right). \end{aligned}$$

The proof is complete. □

[2] If $ab < \delta$, then $\min\{a, b\}^2 < \delta$.

In order to interpret the conclusion of Theorem B.3 on the phase retrieval objective f_P, we must show that the condition

$$\left\{ \begin{array}{c} |\lambda_1(V)| \lesssim \varepsilon/\|x\| \\ |\langle x, \bar{x}\rangle| \lesssim \varepsilon\|\bar{x}\| \end{array} \right\}$$

guarantees that the equation $\|x\| = c \cdot \|\bar{x}\|$ almost holds, where c is defined in Theorem 5.2. This is the content of the following two lemmas. Note that it is easy to verify the equality $\lambda_1(V) = \nabla_{y_1}\zeta(y_1, y_2)$, where we set $(y_1, y_2) := (\lambda_1(X), \lambda_d(X))$.

Lemma B.4 (Extension of Lemma 5.8)**.** *Fix a real constant $0 \le \varepsilon < 1$. The solutions of the inequality $|\nabla_{y_1}\zeta(y_1, y_2)| \le \varepsilon$ on $\mathbb{R}_{++} \times \mathbb{R}_{--}$ are precisely the elements of the open cone*

$$\{(c^2 y, -y) : 0 < y, 0 < c_1 \le c \le c_2\},$$

where c_1, c_2 are the unique solutions of the equations

$$\frac{\pi}{4}(1+\varepsilon) = \frac{c_2}{1+c_2^2} + \arctan\left(c_2\right),$$

and

$$\frac{\pi}{4}(1-\varepsilon) = \frac{c_1}{1+c_1^2} + \arctan\left(c_1\right).$$

Moreover, considering c_1 and c_2 as functions of ε, we have $c_2(\varepsilon) - c_1(\varepsilon) \le 5\pi\varepsilon$ whenever $0 < \varepsilon < 1/2$.

Proof. The proof is completely analogous to that of Lemma 5.8. We leave the details to the reader. The only point worth commenting is the inequality $c_2(\varepsilon) - c_1(\varepsilon) \le 5\pi\varepsilon$ whenever $0 < \varepsilon < 1/2$. To get this bound, observe that $0 < c_2(\varepsilon) \le c_2(.5) \le .83$ for all $\varepsilon \le 1/2$ as c_2 is a increasing function of ε. Therefore,

$$\begin{aligned} \frac{\pi}{2}\varepsilon = \frac{c_2}{1+c_2^2} - \frac{c_1}{1+c_1^2} + \arctan(c_2) - \arctan(c_1) \ge \frac{c_2}{1+c_2^2} - \frac{c_1}{1+c_1^2} = \frac{1-c_1c_2}{(1+c_1^2)(1+c_2^2)}(c_2 - c_1) \\ \ge \frac{1-c_2^2}{(1+c_1^2)(1+c_2^2)}(c_2 - c_1) \ge \frac{1-c_2^2}{(1+c_2^2)^2}(c_2 - c_1). \end{aligned} \tag{B.8}$$

Thus, we have

$$c_2(\varepsilon) - c_1(\varepsilon) \le \frac{\pi\varepsilon}{2}\frac{(1+c_2^2(\varepsilon))^2}{1-c_2^2(\varepsilon)} \le \frac{\pi\varepsilon}{2}\frac{(1+.83^2)^2}{1-.83^2} \le 5\pi\varepsilon,$$

as claimed. □

Lemma B.5. *Fix a real constant $0 \le \varepsilon < \frac{1}{3}$ and vectors $x, \bar{x} \in \mathbb{R}^d \setminus \{0\}$. Suppose $\lambda_1(X) = -c^2\lambda_d(X)$ for some real constant $c > 0$ and $|\langle x, \bar{x}\rangle| \le \varepsilon\|\bar{x}\|\|x\|$. Then we have*

$$1 - (1+c^2)(\varepsilon + \varepsilon^2) \le c^2\frac{\|\bar{x}\|^2}{\|x\|^2} \le 1 + (1+c^2)(\varepsilon + \varepsilon^2).$$

Proof. Fix a decomposition $x = \frac{\langle x,\bar{x}\rangle}{\|\bar{x}\|^2}\bar{x} + v$, where $v \in \bar{x}^{\perp}$. Note inequality $|\langle x,\bar{x}\rangle| \leq \varepsilon\|x\|\|\bar{x}\|$ implies that $\bar{x}$ and x are not collinear, and therefore $\|v\| > 0$. Define the constant $\alpha = \frac{\langle x,\bar{x}\rangle}{\|\bar{x}\|^2}$. Then a quick computation shows the following decomposition:

$$X = \begin{bmatrix} \frac{\bar{x}}{\|\bar{x}\|} & \frac{v}{\|v\|} \end{bmatrix} \begin{bmatrix} (\alpha^2-1)\|\bar{x}\|^2 & \alpha\|\bar{x}\|\|v\| \\ \alpha\|\bar{x}\|\|v\| & \|v\|^2 \end{bmatrix} \begin{bmatrix} \frac{\bar{x}}{\|\bar{x}\|} & \frac{v}{\|v\|} \end{bmatrix}^T.$$

Notice that the above 2×2-matrix is invertible, and therefore its eigenvalues must be $\lambda_1(X)$ and $\lambda_d(X)$. By the Gershgorin theorem [15, Corollary 6.1.3] applied to the 2×2 matrix, we know that $\lambda_1(X)$ and $\lambda_d(X)$ must lie in the union of the intervals

$$\bar{D}_1 = \{z : |z - \|v\|^2| \leq |\alpha|\|\bar{x}\|\|v\|\} \quad \text{and} \quad \bar{D}_2 = \{z \,:\, |z - (\alpha^2-1)\|\bar{x}\|^2| \leq |\alpha|\|\bar{x}\|\|v\|\}.$$

We next prove the following claim.

Claim 5. The intervals $\bar{D}_1$ and $\bar{D}_2$ are contained in the following intervals around $\|x\|^2$ and $-\|\bar{x}\|^2$, respectively:

$$\begin{aligned} \bar{D}_1 \subset D_1 &:= \{z \,:\, |z - \|x\|^2| \leq (\varepsilon^2+\varepsilon)\|x\|^2\}, \\ \bar{D}_2 \subset D_2 &:= \{z \,:\, |z + \|\bar{x}\|^2| \leq (\varepsilon^2+\varepsilon)\|x\|^2\}. \end{aligned}$$

Moreover, we have $D_1 \cap D_2 = \emptyset$ and $D_1 \subset \mathbb{R}_{++}$.

Proof. Consider the interval $\bar{D}_1$. A routine computation shows

$$|\alpha| \leq \frac{\varepsilon\|x\|}{\|\bar{x}\|}, \quad 0 \leq \alpha^2 \leq \frac{\varepsilon^2\|x\|^2}{\|\bar{x}\|^2}, \quad \text{and} \quad 0 \leq \alpha\langle x,\bar{x}\rangle \leq \varepsilon^2\|x\|^2.$$

Using $\|x\| \geq \|v\|$ and $\|v\|^2 = \|x\|^2 - 2\alpha\langle x,\bar{x}\rangle + \alpha^2\|\bar{x}\|^2$, we successively deduce for any $z \in \bar{D}_1$, the inequalities

$$\begin{array}{ccccc} -|\alpha|\|\bar{x}\|\|x\| & \leq & z - \|v\|^2 & \leq & |\alpha|\|\bar{x}\|\|x\| \\ -\varepsilon\|x\|^2 & \leq & z - \|x\|^2 + 2\alpha\langle x,\bar{x}\rangle - \alpha^2\|\bar{x}\|^2 & \leq & \varepsilon\|x\|^2 \\ -\varepsilon\|x\|^2 + \alpha^2\|\bar{x}\|^2 - 2\alpha\langle x,\bar{x}\rangle & \leq & z - \|x\|^2 & \leq & \varepsilon\|x\|^2 + \alpha^2\|\bar{x}\|^2 - 2\alpha\langle x,\bar{x}\rangle \\ -\varepsilon\|x\|^2 - \varepsilon^2\|x\|^2 & \leq & z - \|x\|^2 & \leq & \varepsilon\|x\|^2 + \varepsilon^2\|x\|^2. \end{array}$$

Thus we have shown $\bar{D}_1 \subset D_1$. Similarly, for all $z \in \bar{D}_2$, we compute

$$\begin{array}{ccccc} -|\alpha|\|\bar{x}\|\|v\| & \leq & z - (\alpha^2-1)\|\bar{x}\|^2 & \leq & |\alpha|\|\bar{x}\|\|v\| \\ -\varepsilon\|x\|^2 + \alpha^2\|\bar{x}\|^2 & \leq & z + \|\bar{x}\|^2 & \leq & \varepsilon\|x\|^2 + \alpha^2\|\bar{x}\|^2 \\ -\varepsilon\|x\|^2 & \leq & z + \|\bar{x}\|^2 & \leq & \varepsilon\|x\|^2 + \varepsilon^2\|x\|^2. \end{array}$$

We conclude $\bar{D}_2 \subset D_2$. Provided $\|x\| \neq 0$ and $\varepsilon^2 + \varepsilon < 1$, it is clear $D_1 \subset \mathbb{R}_{++}$. It remains to show that $D_2 \cap D_1 = \emptyset$. Clearly it is sufficient to guarantee that the sum of the radii of D_2 and D_1 is strictly smaller than the distance between the centers:

$$(\varepsilon^2+\varepsilon)\|x\|^2 + (\varepsilon^2+\varepsilon)\|x\|^2 < \|x\|^2 - (-\|\bar{x}\|^2).$$

Rearranging, we must guarantee $2(\varepsilon^2+\varepsilon) - 1 < \frac{\|\bar{x}\|^2}{\|x\|^2}$. Clearly this is the case as soon as $\varepsilon < 1/3$. The result follows. □

Thus we have proved $D_1 \cap D_2 = \emptyset$ and $D_1 \subset \mathbb{R}_{++}$. Since $\bar{D}_1$ and $\bar{D}_2$, each contains at least one eigenvalue, it must be the case that $\lambda_d(X)$ lies in $\bar{D}_2$ and $\lambda_1(X)$ lies in $\bar{D}_1$. We thus conclude

$$\begin{aligned} \left|\lambda_1(X) - \|x\|^2\right| &\leq (\varepsilon^2 + \varepsilon)\|x\|^2 \\ \left|\lambda_d(X) + \|\bar{x}\|^2\right| &\leq (\varepsilon^2 + \varepsilon)\|x\|^2. \end{aligned}$$

Writing $\lambda_1(X) = -c^2\lambda_d(X)$, we obtain

$$\left|-c^2\lambda_d(X) - c^2\|\bar{x}\|^2 + c^2\|\bar{x}\|^2 - \|x\|^2\right| \leq (\varepsilon^2 + \varepsilon)\|x\|^2,$$

and hence

$$\left|\|x\|^2 - c^2\|\bar{x}\|^2\right| \leq (1 + c^2)(\varepsilon^2 + \varepsilon)\|x\|^2.$$

The result follows. □

Combining Lemmas B.4 and B.5, we arrive at the following.

Corollary B.6 (Small $\lambda_1(V)$ and near orthogonality)**.** *Fix a real constant $0 \leq \varepsilon \leq \frac{1}{8}$ and consider a point $x \in \mathbb{R}^d \setminus \{0\}$ satisfying $|\nabla_{y_1}\zeta(\lambda_1(X), \lambda_d(X))| \leq \varepsilon$ and $|\langle x, \bar{x}\rangle| \leq \varepsilon\|x\|\|\bar{x}\|$. Then x satisfies*

$$\left|\|x\| - c\|\bar{x}\|\right| \leq 26\varepsilon\|\bar{x}\|,$$

where c is the solution of the equation $\frac{\pi}{4} = \frac{c}{1+c^2} + \arctan(c)$.

Proof. Define the quantities $c_1(\varepsilon)$ and $c_2(\varepsilon)$ to be the solutions of the equations

$$\begin{aligned} \frac{\pi}{4}(1 - \varepsilon) &= \frac{c_1}{1 + c_1^2} + \arctan(c_1), \\ \frac{\pi}{4}(1 + \varepsilon) &= \frac{c_2}{1 + c_2^2} + \arctan(c_2), \end{aligned}$$

respectively. First, since $c_2(\cdot)$ is an increasing function, it is easy to verify $c_2(\varepsilon) < 1$ whenever $0 < \varepsilon \leq \frac{1}{8}$; thus we have $2\varepsilon(1 + c_2^2(\varepsilon)) < \frac{1}{2}$. By Lemma B.4, we know that whenever $|\nabla_{y_1}\zeta(\lambda_1(X), \lambda_d(X))| \leq \varepsilon$, there exists $\hat{c}$ satisfying $\lambda_1(X) = -\hat{c}^2\lambda_d(X)$ and $0 < c_1(\varepsilon) \leq \hat{c} \leq c_2(\varepsilon)$. Lemma B.5, in turn, implies

$$(1 - 2\varepsilon(1 + \hat{c}^2))\|x\|^2 \leq \hat{c}^2\|\bar{x}\|^2 \leq (1 + 2\varepsilon(1 + \hat{c}^2))\|x\|^2.$$

Looking at the right-hand-side, we deduce

$$c_1^2(\varepsilon)\|\bar{x}\|^2 \leq \hat{c}^2\|\bar{x}\|^2 \leq \left(1 + 2\varepsilon(1 + c_2^2(\varepsilon))\right)\|x\|^2,$$

while looking at the left-hand-side yields

$$(1 - 2\varepsilon(1 + c_2^2(\varepsilon)))\|x\|^2 \leq \hat{c}^2\|\bar{x}\|^2 \leq c_2^2(\varepsilon)\|\bar{x}\|^2.$$

Isolating $\|x\|^2$ and taking square roots we obtain

$$\frac{c_1(\varepsilon)}{\sqrt{1 + 2\varepsilon(1 + c_2^2(\varepsilon))}}\|\bar{x}\| \leq \|x\| \leq \frac{c_2(\varepsilon)}{\sqrt{1 - 2\varepsilon(1 + c_2^2(\varepsilon))}}\|\bar{x}\|. \tag{B.9}$$

Applying Lemma B.4 and the inequality $c_2(\varepsilon) < 1$, we upper bound the right-hand-side:

$$\frac{c_2(\varepsilon)}{\sqrt{1-2\varepsilon(1+c_2^2(\varepsilon))}} \le \frac{5\pi\varepsilon + c}{\sqrt{1-4\varepsilon}}$$
$$= c\left(1 + \frac{5\pi\varepsilon/c + 1 - \sqrt{1-4\varepsilon}}{\sqrt{1-4\varepsilon}}\right) \le c\left(1 + \frac{5\pi\varepsilon/c + 4\varepsilon}{\sqrt{1/2}}\right) \le c(1+57\varepsilon).$$

Exactly the same reasoning shows

$$\frac{c_1(\varepsilon)}{\sqrt{1+2\varepsilon(1+c_2^2(\varepsilon))}} \ge c(1-57\varepsilon).$$

Thus the inequality $\big|\|x\| - c\|\bar{x}\|\big| \le 57c\varepsilon\|\bar{x}\| \le 26\varepsilon\|\bar{x}\|$ holds, as claimed. □

We are now ready to prove the inexact extension of Theorem 5.3.

Proof of Theorem 5.3. We use the decomposition $g = f_P(X)$ and $f = \varphi$. Let us verify that we may apply Theorem B.3. To this end, observe that the population objective satisfies

$$f_P(x) - f_P(\bar{x}) = \mathbb{E}_a\left[\left\langle a, \frac{x-\bar{x}}{\|x-\bar{x}\|}\right\rangle \left\langle a, \frac{x+\bar{x}}{\|x+\bar{x}\|}\right\rangle\right] \|x-\bar{x}\|\|x+\bar{x}\| \ge \kappa\|x-\bar{x}\|\|x+\bar{x}\|$$

for the numerical constant κ [12, Corollary 3.7]. Moreover, clearly ζ is globally Lipschitz (being a norm), and therefore $|\lambda_1(V)|$ and $|\lambda_d(V)|$ are bounded by a numerical constant. Thus provided $\varepsilon := \|Vx\|$ satisfies $\varepsilon \le \gamma \cdot \|x\|$ for the numerical constant γ, we can be sure that x satisfies either

$$\|x\|\|x-\bar{x}\|\|x+\bar{x}\| \lesssim \varepsilon\|\bar{x}\|^2 \qquad \text{or} \qquad \left\{\begin{array}{c} |\lambda_1(V)| \lesssim \varepsilon/\|x\| \\ |\langle x, \bar{x}\rangle| \lesssim \varepsilon\|\bar{x}\| \end{array}\right\}.$$

Now suppose the latter is the case, and let C be a numerical constant satisfying $|\lambda_1(V)| \le C\varepsilon/\|x\|$ and $|\langle x, \bar{x}\rangle| \le C\varepsilon\|\bar{x}\|$. We aim to apply Corollary B.6. To do so, we must ensure

$$|\lambda_1(V)| \le \frac{C\varepsilon}{\|x\|} \le \frac{1}{8} \quad \text{and} \quad \left|\left\langle \frac{x}{\|x\|}, \frac{\bar{x}}{\|\bar{x}\|}\right\rangle\right| \le C\varepsilon \cdot \frac{\|\bar{x}\|}{\|x\|\|\bar{x}\|} = \frac{C\varepsilon}{\|x\|} \le \frac{1}{8}.$$

Adjusting γ if necessary, we can be sure that $\varepsilon/\|x\|$ is below $\frac{1}{8C}$. Applying Corollary B.6, with $\frac{C\varepsilon}{\|x\|}$ in place of ε, we conclude $|\|x\| - c\|\bar{x}\|| \lesssim \varepsilon\frac{\|\bar{x}\|}{\|x\|}$, as claimed. □