Acceleration and Implicit Regularization in Gaussian Phase Retrieval

Tyler Maunu
Brandeis University

Martin Molina-Fructuoso
Brandeis University

Abstract

We study accelerated optimization methods in the Gaussian phase retrieval problem. In this setting, we prove that gradient methods with Polyak or Nesterov momentum have similar implicit regularization to gradient descent. This implicit regularization ensures that the algorithms remain in a nice region, where the cost function is strongly convex and smooth despite being nonconvex in general. This ensures that these accelerated methods achieve faster rates of convergence than gradient descent. Experimental evidence demonstrates that the accelerated methods converge faster than gradient descent in practice.

1 INTRODUCTION

While convex optimization is by now a mature field with an array of well-understood algorithms, the properties of nonconvex optimization algorithms have only recently begun to be illuminated. Understanding nonconvex optimization algorithms is of fundamental importance for progress in data science and machine learning since many optimization problems we wish to solve are nonconvex. Furthermore, it is important to understand how variants of these nonconvex optimization algorithms, such as stochastic or accelerated methods, perform in various settings.

A prototypical nonconvex optimization problem arises when one considers the Burer-Monteiro factorization of programs over positive semidefinite (PSD) matrices (Burer and Monteiro, 2003). Here, for a PSD matrix $\boldsymbol{S} \in \mathbb{R}^{n \times n}$ one uses the parametrization $\boldsymbol{S} = \boldsymbol{U}\boldsymbol{U}^T$, and then applies a standard algorithm like gradient

descent to optimize over $\boldsymbol{U}$ rather than $\boldsymbol{S}$. This factorization can be used in solving many problems, such as matrix completion, phase retrieval, covariance sketching, and more (Sanghavi et al., 2017; Li et al., 2019; Ma et al., 2020). The Burer-Monteiro factorization has been shown to work well in practice despite it typically leading to nonconvex optimization problems. Furthermore, it represents a concrete case where we can understand the properties of nonconvex optimization algorithms.

While many works have analyzed the optimization landscapes of factorized matrix problems, we draw our inspiration from the work of Ma et al. (2020). Here, the authors study how gradient descent interacts with a random observation model on the specific problems of phase retrieval, matrix completion, and blind deconvolution. In this work, we focus on one of the specific problems of Ma et al. (2020) for clarity, although these results can be generalized. The *phase retrieval* problem seeks a signal vector $\boldsymbol{x}_\star \in \mathbb{R}^n$ from observations

$$y_i = |\boldsymbol{a}_i^T \boldsymbol{x}_\star|^2 \tag{1}$$

collected from a set of known sensing vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m$. In the general case, one considers sensing vectors and signals $\boldsymbol{a}_i, \boldsymbol{x}_\star \in \mathbb{C}^n$, but here we focus on the real case for simplicity of exposition. The Gaussian phase retrieval problem refers to the case where the sensing vectors $\boldsymbol{a}_i$ are standard Gaussian (Candes et al., 2015).

In effect, they show that gradient descent with proper initialization interacts well with the random measurement model by developing an analysis of its *implicit regularization*. Broadly speaking, implicit regularization refers to a bias in an optimization algorithm that only exists implicitly rather than explicitly. In effect, the implicit regularization result of Ma et al. (2020) ensures that the optimization problem in question behaves much like a convex optimization problem despite the problem being nonconvex in general.

Implicit regularization can be used to show that certain optimization methods perform well on nonconvex problems, and the type of regularization varies from

setting to setting. In certain settings, it ensures that the algorithm finds a solution that minimizes some norm (Gunasekar et al., 2017; Li et al., 2018; Vaskevicius et al., 2019; Min et al., 2021; Zhao et al., 2022) or yields outputs that are projections with respect to a Bregman divergence or minimum entropy solutions (Gunasekar et al., 2018; Woodworth et al., 2020; Even et al., 2023). Alternatively, as we mentioned in the case of Ma et al. (2020), one can show that methods maintain some independence properties in random sensing models. Almost all implicit regularization results for nonconvex methods rely on finding a good starting point, either by spectral initialization or by some other cleverly chosen point.

While many results exist exploring the implicit biases of gradient descent and its stochastic counterpart, the implicit biases of accelerated algorithms are not well understood. It is essential to better understand their implicit biases since in practical nonconvex optimization such as training state-of-the-art deep learning methods, practitioners use some variant of momentum (such as in Kingma and Ba (2014)).Therefore, in this work, we set out to better understand the implicit regularization of gradient descent with momentum in the context of the specific nonconvex problem of phase retrieval.

1.1 Outline of Contributions

Here, we summarize the main contributions of our work. These findings are also presented in Figure 1: accelerated gradient descent (AGD) methods achieve faster convergence than gradient descent on the Gaussian phase retrieval problem in theory and practice.

1. We develop the first implicit regularization analysis for two accelerated gradient methods on the Gaussian phase retrieval problem. These methods utilize Polyak's momentum (Polyak, 1964) and Nesterov's momentum (Nesterov, 1983). Our analysis shows that with spectral initialization, these accelerated methods maintain incoherence and locality properties throughout their iterations. To accomplish this, we develop a novel leave-one-out analysis that works on pairs of iterates rather than the iterates themselves. To our knowledge, this is the first implicit regularization result for accelerated gradient methods on a matrix factorization problem.

2. As a result of this analysis, we are able to show that accelerated gradient methods converge faster than gradient descent on the Gaussian phase retrieval problem. These are the first results of guaranteed convergence for accelerated algorithms on the phase retrieval problem, even
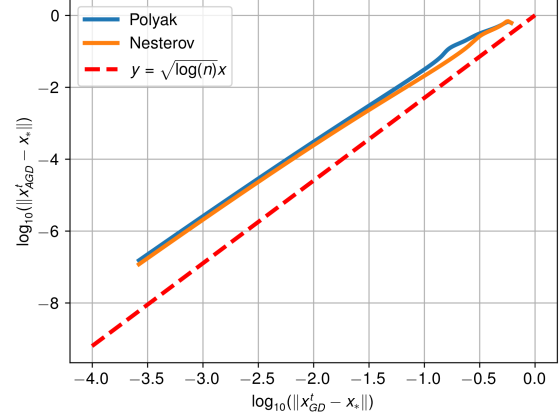

Figure 1: Accelerated gradient methods achieve faster convergence than gradient descent in both practice and in theory on the Gaussian phase retrieval problem. Here, Polyak and Nesterov momentum are used to accelerate gradient descent. The y-axis measures the log-error of the accelerated method, whereas the x-axis measures the log-error of the standard gradient method at each iteration. As we can see, after the same amount of iterations has passed for each method, accelerated gradient descent is approximately a factor of $\sqrt{\log(n)}$ closer to $\boldsymbol{x}_*$ on a log scale, which is supported by our theory.

though such methods were previously considered in works such as Xu et al. (2018); Wang and Abernethy (2020).

3. While past works experimentally verify the fast convergence of momentum methods, we also supplement our results to show that accelerated gradient methods achieve faster convergence than gradient descent in practice.

1.2 Review of Related Work

Many techniques exist to solve the phase retrieval problem (1), and so we review some of them here.

Other variants of the phase retrieval problem consider different types of sensing vectors. These include holographic phase retrieval (Barmherzig and Sun, 2022), ptychography (Thibault et al., 2008), near-field phase retrieval (Wang et al., 2020), and far-field phase retrieval (Zhuang et al., 2022). Here, we considered the problem of Gaussian phase retrieval (Candes et al., 2015).

One popular technique for recovering the vector $\boldsymbol{x}_*$ involves solving convex relaxations (Candes et al., 2013). However, solving the resulting semidefinite programs is typically expensive.

Many past works have studied first-order algorithms to recover $\boldsymbol{x}^\star$ from observations (1), and, as mentioned, these have shown efficient recovery. Candes et al. (2015) study the Wirtinger Flow algorithm, which has become a staple in the phase retrieval literature. This algorithm is gradient descent on the least squares objective that we consider. Here, the authors prove local linear convergence of the WF algorithm and show it needs $O(n \log(1/\epsilon))$ iterations to converge to an ϵ-accurate solution. Sun et al. (2018) show that the objective has no local minimizers, all global minimizers are equal up to a phase shift, and all saddle points are strict saddles. To take advantage of this structure of the problem, they advocate the use of a trust-region algorithm.

Later, Ma et al. (2020) improved the analysis of gradient descent with an implicit regularization style analysis to $O(\log(n) \log(1/\epsilon))$. Even more recently, these results have recently been improved to $O(\log n + \log(1/\epsilon))$ by Chen et al. (2019); Lee and Stöger (2023), whose results rely on random initialization. The work of Chen et al. (2019) focuses on a two-stage analysis of gradient descent, while the work of Lee and Stöger (2023) focuses on an alternating least squares approach.

Accelerated gradient methods for phase retrieval were considered in Xu et al. (2018); Wang and Abernethy (2020). Some preliminary results by Wang and Abernethy (2020) indicated that heavy ball for phase retrieval may converge to the benign region more quickly. However, proof of this fact for the finite sample setting has not yet been established. Other work has considered higher-order methods, such as sketched Newton methods (Luo et al., 2023).

Also, some works have studied phase retrieval in alternative settings. Sparse phase retrieval is a common paradigm where the underlying signal of interest is assumed to be sparse (Eldar et al., 2014; Shechtman et al., 2014; Wang et al., 2017). The works of Wu and Rebeschini (2020, 2021) consider implicit regularization for sparse phase retrieval, where the underlying vector is assumed to be sparse. Others considered phase retrieval with generative priors, such as the work of Hand et al. (2018). Finally, there are also works on robust phase retrieval, such as Candes et al. (2013); Duchi and Ruan (2019).

Geometric approaches to solving the phase retrieval problem exist (Huang et al., 2017; Hou et al., 2020; Maunu et al., 2023), and some methods also consider alternative energies (Duchi and Ruan, 2019; Wang et al., 2018; Maunu et al., 2023).

Finally, our work adds to the literature on implicit regularization of methods for nonconvex optimization.

Much past work has focused on gradient descent, but some works have looked at accelerated methods as well (Pagliana and Rosasco, 2019; Wang et al., 2023; Ghosh et al., 2023). None of these results has considered the implicit regularization of accelerated regularization in matrix factorization problems, though.

1.3 Notation

We use bold lowercase letters to denote vectors and bold uppercase letters to denote matrices. When applied to vectors, $\|\cdot\|_2$ represents the Euclidean norm, whereas when applied to matrices, it represents the spectral norm.

2 BACKGROUND

In this section, we give the necessary background to understand our contributions. First, we will outline simple approaches to solve the phase retrieval problem via first-order optimization algorithms. After this, we review related work on the use of accelerated first-order methods for the phase retrieval problem. Finally, we will review the general theory of first-order accelerated algorithms in convex optimization.

2.1 Phase Retrieval

The phase retrieval problem arises in settings such as X-ray crystallography (Candes et al., 2015), and it also has applications to optical problems, in particular, medical optical imaging (Shechtman et al., 2015), microscopy (Miao et al., 2008), and astronomical imaging (Shechtman et al., 2015).

While the practical phase retrieval problem considers various measurement operators $\mathcal{A}$, such as Fourier observations, we consider a specific mathematical setting. This is also referred to as the *Gaussian phase retrieval problem.*

Assumption 1. *The sensing vectors $\boldsymbol{a}_i$, $i = 1, \ldots, m$, are sampled i.i.d. from a standard Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.*

Other variants of the phase retrieval problem can be studied, where the sensing vectors take on other forms. See the examples in Barmherzig et al. (2019). In ptychography, the observations are Fourier measurements of an object modulated by a fixed illumination pattern focused over patches at a time (Thibault et al., 2008).

It is natural to consider recovering the underlying signal $\boldsymbol{x}_\star$ by solving an optimization problem. For example, one can consider the least squares problem

$$f(\boldsymbol{x}) = \frac{1}{4m} \sum_{i=1}^{m} ((\boldsymbol{a}_i^T \boldsymbol{x})^2 - y_i)^2. \qquad (2)$$

We note that this is equivalent to a Burer-Monteiro factorization (Burer and Monteiro, 2003) on the low-rank matrix sensing problem (Fazel et al., 2008). In fact, this type of problem has a further structure in that the sensing matrices are rank one, which has been considered in Chen et al. (2015); Cai and Zhang (2015).

Due to the ambiguity between $\pm\boldsymbol{x}$ in the observations (1) as well as in the cost function (2), we can only hope to recover $\boldsymbol{x}_*$ up to sign. Therefore, we adopt the following notion of distance as in Ma et al. (2020).

$$\text{dist}(\boldsymbol{x}, \boldsymbol{x}_*) = \min\{\|\boldsymbol{x} - \boldsymbol{x}_*\|, \|\boldsymbol{x} + \boldsymbol{x}_*\|\}. \quad (3)$$

Candes et al. (2015) introduced a gradient descent algorithm for the phase retrieval problem. The gradient of the functional (2) is given by

$$\nabla f(\boldsymbol{x}) = \frac{1}{m} \sum_{i=1}^{m} ((\boldsymbol{a}_i^T \boldsymbol{x})^2 - y_i) \boldsymbol{a}_i \boldsymbol{a}_i^T \boldsymbol{x}. \quad (4)$$

The gradient descent algorithm follows the iterative updates

$$\boldsymbol{x}^{t+1} = \left(\boldsymbol{I} - \eta \frac{1}{m} \sum_{i=1}^{m} ((\boldsymbol{a}_i^T \boldsymbol{x})^2 - y_i) \boldsymbol{a}_i \boldsymbol{a}_i^T \right) \boldsymbol{x}^t \quad (5)$$

Ma et al. (2020) showed this algorithm exhibits an implicit regularization property that leads to linear convergence,

$$\|\boldsymbol{x}^t - \boldsymbol{x}_\star\| \leq C(1 - \eta \|\boldsymbol{x}_\star\|_2^2/2)^t \|\boldsymbol{x}_\star\|_2 \quad (6)$$

for some absolute constant C.

2.2 Accelerated Methods in Phase Retrieval

In this work, we consider accelerated first-order methods. These are not new and indeed have been considered in past works for phase retrieval, such as Xu et al. (2018); Ajayi et al. (2018); Xiong et al. (2018, 2020); Wang and Abernethy (2020). In Ajayi et al. (2018), the authors study an adaptation of Nesterov's accelerated method for the matrix sensing problem that exhibits linear convergence up to an error that depends on the momentum parameter. In Wang and Abernethy (2020), the authors study how a heavy ball performs with random initialization, although it lacks a correct proof. Xiong et al. (2018, 2020) show convergence of (P) and (N) using techniques from control theory. However, they do not demonstrate acceleration over standard gradient methods and require a generic regularity condition. In Xu et al. (2018), the authors demonstrate the advantages of accelerated Wirtinger flow but offer no proof of convergence. None of these

past works show convergence to the ground truth signal, and furthermore, no past works consider these accelerated gradient methods from the lens of implicit regularization. Therefore, our work answers the important outstanding question in the literature on the convergence of these momentum methods.

The two accelerated gradient methods considered in this paper are Polyak's Heavy Ball (Polyak, 1964) and Nesterov's Accelerated Gradient (Nesterov, 1983) methods. The iteration that defines gradient descent with Polyak momentum is

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \nabla f(\boldsymbol{x}^t) + \beta(\boldsymbol{x}^t - \boldsymbol{x}^{t-1}) \quad (P)$$

where $\boldsymbol{x}^1 = \boldsymbol{x}^0$. On the other hand, gradient descent with Nesterov momentum takes the form

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \nabla f(\boldsymbol{x}^t + \beta(\boldsymbol{x}^t - \boldsymbol{x}^{t-1})) + \beta(\boldsymbol{x}^t - \boldsymbol{x}^{t-1}) \quad (N)$$

where again $\boldsymbol{x}^1 = \boldsymbol{x}_0$.

We note that the primary difference between the two is that Nesterov momentum takes the gradient at an extrapolated point, whereas Polyak momentum uses the gradient at the current iterate. As we point out in our analysis, controlling the Nesterov momentum's implicit regularization is more challenging due to this extrapolation.

In some convex settings, the convergence of these accelerated methods is well understood. For a twice continuously differentiable function f, we say that f is μ-strongly convex and L-smooth if

$$\mu \boldsymbol{I} \preceq \nabla^2 f(\boldsymbol{x}) \preceq L\boldsymbol{I}, \quad (7)$$

for all $\boldsymbol{x}$. The condition number is the ratio $\kappa = L/\mu$. Under the assumption of strong convexity and smoothness, the convergence of (P) and (N) can be succinctly stated in the following theorem.

Theorem 1 ((Polyak, 1964; Nesterov et al., 2018)). *Suppose that we run gradient descent, (P) or (N) on a μ-strongly convex and L-smooth function f. Then, gradient descent with step size $\eta = 1/L$ achieves the convergence bound*

$$\|\boldsymbol{x}_t - \boldsymbol{x}_*\|_2 \lesssim \exp(-t/\kappa)\|\boldsymbol{x}_t - \boldsymbol{x}_0\|_2. \quad (8)$$

On the other hand, (P) with step size $\eta = \frac{4}{(\sqrt{\mu} + \sqrt{L})^2}$ and momentum parameter $\beta = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$ and (N) with step size $\eta = 1/L$ and momentum parameter $\beta = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$ achieve the convergence bound

$$\|\boldsymbol{x}_t - \boldsymbol{x}_*\|_2 \lesssim \exp(-c_1 t/\sqrt{\kappa})\|\boldsymbol{x}_t - \boldsymbol{x}_0\|_2, \quad (9)$$

Here, the constant $c_1 = 2$ for (P) and $c_1 = 1$ for (N).

As we can see, the accelerated methods achieve a faster rate, where the improvement is by a factor of $\sqrt{\kappa}$. This means that, to reach an error ϵ, GD needs $O(\kappa \log(1/\epsilon))$ iterations while the accelerated methods need $O(\sqrt{\kappa} \log(1/\epsilon))$ iterations. Notice that one needs to select the parameters for the momentum methods carefully. However, as we will see in the next section, in the Gaussian phase retrieval problem, we have bounds for μ and L that allow us to set the parameters in these methods.

3 CONVERGENCE FOR ACCELERATED GRADIENT METHODS

In this section, we give our main theorems proving convergence of (P) and (N) on the Gaussian phase retrieval problem. First, in Section 3.1, we review the ideas of implicit regularization from Ma et al. (2020) used to prove the convergence of gradient descent on this problem. Then, in Section 3.2, we give our main convergence theorems and a sketch of their proofs. Finally, Section 3.3 discusses these results and compares them with those for gradient descent.

3.1 Implicit Regularization

Ma et al. (2020) give implicit regularization results for low-rank matrix recovery problems. The main idea is to ensure that the iterates of gradient descent remain in a good region, which the authors call the *region of incoherence and contraction (RIC)*. In this region, one can show that the cost function (2) is strongly convex and smooth.

To see how this might be the case, consider the Hessian of the cost function,

$$\nabla^2 f(\boldsymbol{x}) = \frac{1}{m} \sum_{i=1}^{m} (3(\boldsymbol{a}_i^T \boldsymbol{x})^2 - y_i) \boldsymbol{a}_i \boldsymbol{a}_i^T. \qquad (10)$$

One can show that $\boldsymbol{I} \preceq \mathbb{E}_{\boldsymbol{a}_i} \nabla^2 f(\boldsymbol{x}) \preceq 10\boldsymbol{I}$ for $\boldsymbol{x}$ in a small ball around $\boldsymbol{x}_*$,

$$\|\boldsymbol{x}^t - \boldsymbol{x}_\star\|_2 \leq 2C_1 \|\boldsymbol{x}_\star\|_2 \qquad \text{(LOC)}$$

This is a locality property, which says the iterate needs to be sufficiently close to optimal for the cost function to be strongly convex and smooth. Thus, we might expect the function to be smooth and strongly convex for large samples. However, the concentration of the Hessian requires a suboptimal number of samples $m = \Omega(n^2)$ samples, while $m = \Omega(n)$ samples is sufficient to identify $\boldsymbol{x}_*$ (Chen et al., 2015).

To transfer these ideas to the small sample setting, one can show that the Hessian of f is sufficiently well-

behaved, provided that $\boldsymbol{x}$ obeys an additional condition besides locality. The extra condition that defines the RIC besides (LOC) is the incoherence condition,

$$\max_i |\boldsymbol{a}_i^T (\boldsymbol{x} - \boldsymbol{x}_\star)| \leq C_2 \log(n) \|\boldsymbol{x}_*\|. \qquad \text{(INC)}$$

This condition ensures that the vector $\boldsymbol{x} - \boldsymbol{x}_*$ is not too aligned with any specific vector $\boldsymbol{a}_i$.

Under the two assumptions (LOC) and (INC), Lemma 1 of Ma et al. (2020) states that with probability at least $1 - \mathcal{O}(mn^{-10})$,

$$\frac{1}{2} \boldsymbol{I} \preceq \nabla^2 f(\boldsymbol{x}) \preceq O(\log(n)) \boldsymbol{I}. \qquad (11)$$

The main idea is that the additional restriction of incoherence allows one to apply concentration bounds that hold for sample sizes $m = \Omega(n \log n)$ rather than $m = \Omega(n^2)$.

With this lemma in hand, one might expect to be able to show that (P) and (N) achieve faster convergence due to Theorem 1. However, there is an outstanding issue: this theorem assumes that the optimization landscape is globally strongly convex and smooth. However, this is not the case for the phase retrieval function, since it is only strongly convex and smooth inside the RIC.

To find an initial point in the RIC, we use a spectral initialization (Ma et al., 2020). Indeed, setting

$$\boldsymbol{x}^0 = \sqrt{\lambda_1(\boldsymbol{Y})/3} \widetilde{\boldsymbol{x}}^0,$$

where $\lambda_1(\boldsymbol{Y})$ and $\widetilde{\boldsymbol{x}}^0$ are the leading eigenvalue and eigenvector of $\boldsymbol{Y} = \frac{1}{m} \sum_{i=1}^{m} y_i \boldsymbol{a}_i \boldsymbol{a}_i^T$, guarantees that inequalities (LOC) and (INC) are satisfied for $\boldsymbol{x}^0$.

3.2 Convergence of (P) and (N)

The main theorem in this work is the same as the main theorem of Ma et al. (2020) for phase retrieval with the notable change that the convergence improves from κ to $\sqrt{\kappa}$, as in Theorem 1. In practical terms, with spectral initialization, (P) and (N) need $O(\sqrt{\log n} \log(1/\epsilon))$ iterations rather than the $O(\log n \log(1/\epsilon))$ to reach an error ϵ. As we demonstrate in experiments, this is backed up by real scenarios, where we see accelerated method demonstrating faster convergence. While we assume $\|\boldsymbol{x}_*\| = 1$ in the theorem to simplify our presentation, the proof can be extended to signals with arbitrary length without much difficulty.

Theorem 2 (Convergence of (P) and (N)). *Suppose that $\boldsymbol{x}_*$ is a fixed vector with $\|\boldsymbol{x}_*\| = 1$ and $\boldsymbol{a}_j$ follow Assumption 1. Provided that $0 \leq \eta \lesssim \frac{1}{c \log n \|\boldsymbol{x}_0\|^2}$, $m \gtrsim n \log n$, $\beta = \frac{\sqrt{c \log(n)} - \sqrt{1/2}}{\sqrt{c \log(n)} + \sqrt{1/2}}$ for some sufficiently large constant c, then with probability at least*

$1 - O(mn^{-9})$, (P) *and* (N) *with spectral initialization achieve the following contraction and incoherence:*

$$\text{dist}(\boldsymbol{x}^t, \boldsymbol{x}_*) \leq \epsilon \left(1 - \sqrt{\eta}\|\boldsymbol{x}_*\|_2^2/2\right)^t \|\boldsymbol{x}_*\|_2 \qquad (12)$$

$$\max_j |\boldsymbol{a}_j^T(\boldsymbol{x}^t - \boldsymbol{x}_*)| \leq c_2\sqrt{\log n}\|\boldsymbol{x}_*\|_2. \qquad (13)$$

To fix ideas, we present the main ideas in the proof for Polyak momentum. The proof for Nesterov momentum needs similar steps and estimations with more involved expressions. The fundamental ingredient of the proof of the rate of our algorithm is the fact that, for a strongly convex and smooth function, two consecutive iterations of (P) with the parameters of Theorem 2 achieve the contraction

$$\left\|\begin{bmatrix} \boldsymbol{x}^{t+1} - \boldsymbol{x}_* \\ \boldsymbol{x}^t - \boldsymbol{x}_* \end{bmatrix}\right\|_2 \leq \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}\right) \left\|\begin{bmatrix} \boldsymbol{x}^t - \boldsymbol{x}_* \\ \boldsymbol{x}^{t-1} - \boldsymbol{x}_* \end{bmatrix}\right\|_2. \qquad (14)$$

The repeated application of this relation, i.e.,

$$\left\|\begin{bmatrix} \boldsymbol{x}^{t+1} - \boldsymbol{x}_* \\ \boldsymbol{x}^t - \boldsymbol{x}_* \end{bmatrix}\right\|_2 \leq \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} - \sqrt{\mu}}\right)^t \left\|\begin{bmatrix} \boldsymbol{x}^1 - \boldsymbol{x}_* \\ \boldsymbol{x}^0 - \boldsymbol{x}_* \end{bmatrix}\right\|_2 \qquad (15)$$

would be enough to show the algorithm's convergence if f was strongly convex and smooth. However, because f is nonconvex in general, to use this reasoning directly, we need to make sure that all the iterations of gradient descent remain in the RIC, i.e., that they satisfy (LOC) and (INC).

While the detailed proof of Theorem 2 is given in the supplementary material, we sketch the induction we use here.

1. First, we notice that for points in the RIC (namely close enough to $\boldsymbol{x}_*$ (LOC) and in the incoherence region (INC)), the sampled function (2) is strongly convex and smooth with high probability.

2. Second, we prove a contraction-like result for iterations $\boldsymbol{x}_t$ in the RIC of the form

$$\left\|\begin{bmatrix} \boldsymbol{x}^{t+1} - \boldsymbol{x}_* \\ \boldsymbol{x}^t - \boldsymbol{x}_* \end{bmatrix}\right\|_2 \leq M \left\|\begin{bmatrix} \boldsymbol{x}^t - \boldsymbol{x}_* \\ \boldsymbol{x}^{t-1} - \boldsymbol{x}_* \end{bmatrix}\right\|_2, \qquad (16)$$

where M depends only on the constants of strong convexity and smoothness of f. This result follows from the facts that the segment $[\boldsymbol{x}_*, \boldsymbol{x}^t]$ is in the RIC for (P) and that f can be considered strongly convex and smooth in the RIC (11). In a similar way, we can show that for (N), the segment $[\boldsymbol{x}_*, \boldsymbol{x}^t + \beta(\boldsymbol{x}^t - \boldsymbol{x}^{t-1})]$ is in the RIC.

3. We then proceed by induction to show that all iterations for a starting point $\boldsymbol{x}_0$ in the RIC remain in the RIC with high probability using a leave-one-out argument.

Our proof strategy is based on a leave-one-out approach, where a new function is constructed

$$f^{(\ell)}(\boldsymbol{x}) = \frac{1}{4m} \sum_{i:i\neq\ell} ((\boldsymbol{a}_i^T\boldsymbol{x})^2 - y_i)^2, \qquad (17)$$

for $\ell = 1, \ldots, m$. The leave one out sequences $\boldsymbol{x}^{t,(\ell)}$ are defined by running one of the optimization methods (P) or (N) on (17). The following lemma ensures that the accelerated methods remain close to the leave-one-out sequences throughout their iterations.

Lemma 1 (Leave-one-out proximity). *Suppose that* $\boldsymbol{a}_j \overset{i.i.d.}{\sim} N(\boldsymbol{0}, \boldsymbol{I})$. *Then, for the sequences generated by* (P) *or* (4), *with probability at least* $1 - \mathcal{O}(mn^{-10})$,

$$\max_{1\leq\ell\leq m} \left\|\begin{bmatrix} \boldsymbol{x}^{t+1} - \boldsymbol{x}^{t+1,(\ell)} \\ \boldsymbol{x}^t - \boldsymbol{x}^{t,(\ell)} \end{bmatrix}\right\|_2 \leq C_3\sqrt{\frac{\log(n)}{n}}.$$

In the past analysis of gradient descent, proximity to leave-one-out sequences was essential for proving that the iterates remain in the RIC. Therefore, our analysis can be seen as an extension of these results to the accelerated case, where we show that the accelerated gradient sequences also remain close to the leave-one-out sequences.

3.3 Discussion

It is important to point out that we only actually need $O(\log n)$ iterations to reach a region where (LOC) implies (INC). From that point on, the rates of convergence provided by the methods we studied are not probabilistic. This fact is also used in Ma et al. (2020), although they use a conservative bound on the number of iterations to achieve this fact.

One of the main challenges in the proof of implicit regularization and convergence of the accelerated gradient methods is that these methods do not generally define monotonic sequences for common Lyapunov functions. For example, we do not expect the sequences $(\|\boldsymbol{x}^t - \boldsymbol{x}_*\|)_{t\in\mathbb{N}}$ or $(f(\boldsymbol{x}^t) - f(\boldsymbol{x}_*))_{t\in\mathbb{N}}$ to be monotonic. This then makes it hard to guarantee that the sequences of iterates $(\boldsymbol{x}^t)$ remain within the RIC. Fortunately, we are able to prove an analogous implicit regularization result to that observed for gradient descent in Ma et al. (2020) by looking at sequences that consider pairs of consecutive iterates. That is, we are able to prove a contraction result, as well as leave-one-out proximity for the sequence $\left(\begin{bmatrix} \boldsymbol{x}^{t+1} \\ \boldsymbol{x}^t \end{bmatrix}\right)_{t\in\mathbb{N}}$.

Since the bounds require that pairs of iterates maintain properties rather than individual iterates, the constants in our estimates are worse than those required for GD by a constant absolute factor. This essentially means that our sample complexity is a constant factor worse than that required for GD. On the other hand, our convergence rate is arbitrarily better than that of GD, since we get a better convergence rate by a factor of $\sqrt{\log n}$.

There are a few limitations of our methodology. First, current results on random initialization and alternating projections separate the dimension from the $\log(1/\epsilon)$ in the rate, while our rate is $\sqrt{\log n}\log(1/\epsilon)$. Thus, while our results are better than GD with spectral initialization (Ma et al., 2020), they are not better than GD with random initialization in the random model (Chen et al., 2019). Proving convergence of accelerated methods with random initialization remains an open question (Wang and Abernethy, 2020). On the other hand, our work opens the door to accelerating in the random setting by showing how to apply leave-one-out analysis to accelerated gradient methods.

4   EXPERIMENTS

While our results are primarily theoretical in nature, we give a few simulations showing that accelerated gradient methods are well-behaved, and actually perform better than GD on simulated phase retrieval tasks.

Our first experiment is displayed in Figure 2. Here, we test GD against the two momentum based methods (P) and (N). Here, the sensing vectors satisfy Assumption 1, $\boldsymbol{x}_*$ is drawn uniformly from S^{n-1}, $\eta = 0.2/(\log n)$, and $\beta = \frac{\sqrt{10\log n}-\sqrt{2}}{\sqrt{10\log n}+\sqrt{2}}$. We vary $n = 10, 50, 100$ and $m = 200, 500, 1000$. As we can see, where gradient descent converges, the accelerated methods converge faster.

In our second experiment, we test GD against the two momentum-based methods (P) and (N) in the same setup as before, but now the initial vector is randomly chosen from S^{n-1}. The results are displayed in Figure 3. As we can see, the accelerated methods still converge faster than GD when it converges. Therefore, we expect that our theoretical results can be extended to this setting as well.

In our third experiment in Figure 4, we repeat the experiment in Figure 1 across a range of dimensions. For each dimension, we compute the slope of the line given by plotting $\log \|\boldsymbol{x}_{AGD}^t - \boldsymbol{x}_*\|$ against $\log \|\boldsymbol{x}_{GD}^t - \boldsymbol{x}_*\|$. This slope can be interpreted as the speedup of the accelerated method. With the spectral initialization, we expect the speedup to be proportional to $\sqrt{\log n}$,
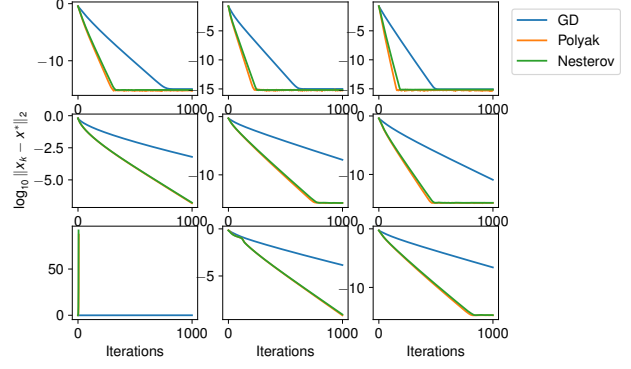

Figure 2: Experiment demonstrating the faster convergence of accelerated methods when compared to GD. Across the rows we vary $m = 200, 500, 1000$, and down the columns we vary $n = 10, 50, 100$. As we can see, (P) and (N) have identical performance.
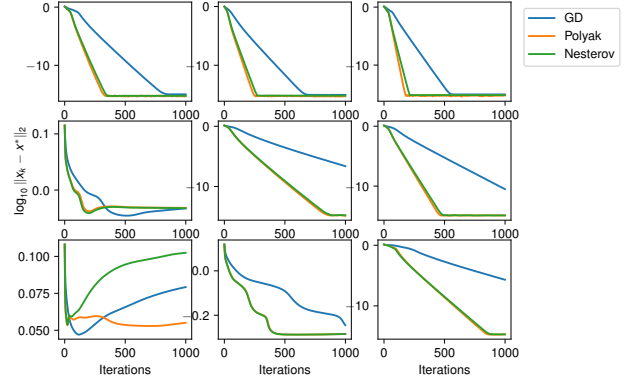

Figure 3: Experiment demonstrating the faster convergence of accelerated methods when compared to GD with random initialization. Across the rows we vary $m = 200, 500, 1000$, and down the columns we vary $n = 10, 50, 100$. As we can see, (P) and (N) have identical performance.

which we observe in practice.

Finally, we compare gradient descent to the accelerated methods on the coded diffraction experiment of Section IV of Candes et al. (2015). Here, the vector $\boldsymbol{x}_*$ is the Stanford main quad image, and $\boldsymbol{a}_j$ are Fourier measurements with random modulation patterns; see the supplementary material or Section IV or Candes et al. (2015) for details. As we can see, the accelerated methods recover the image much faster than Wirtinger Flow (WF), which is just gradient descent.

5   CONCLUSIONS

This work presents the first analysis of implicit regularization of accelerated gradient methods on a ma-
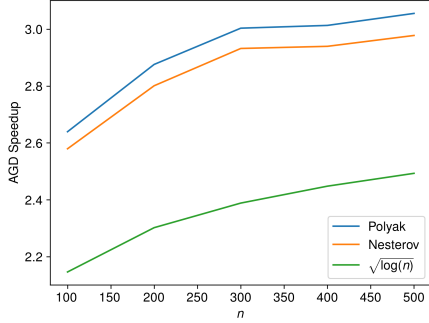
Figure 4: Plot of speedup of accelerated gradient descent versus momentum. Here, the y-axis is the slope of the linear fit of $\log \|\boldsymbol{x}_{AGD}^t - \boldsymbol{x}_*\|$ versus $\log \|\boldsymbol{x}_{GD}^t - \boldsymbol{x}_*\|$, and the x-axis is n. As we can see, these are well-approximated by the curve $\sqrt{\log n}$.
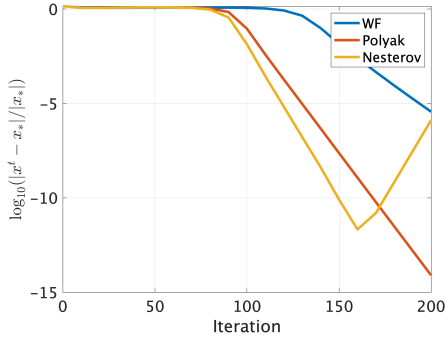

Figure 5: Errors versus iteration for Wirtinger flow as well as Nesterov/Polyak accelerated Wirtinger flow. All methods have the same per-iteration complexity. As we see, the accelerated methods recover the true image much faster.

trix factorization problem. We consider the problem of Gaussian phase retrieval, where the measurement vectors follow a standard Gaussian distribution. In this case, we are able to show that accelerated gradient iterates remain well-behaved in the sense that they remain within the region of incoherence and contraction. We believe that our results can be generalized to other settings, such as matrix completion and blind deconvolution.

Many interesting future directions can be pursued. For example, one can consider extending the results to show that accelerated methods converge more quickly than GD for random initialization. Furthermore, one can consider accelerated methods for sparse phase retrieval or generative phase retrieval. This could be done by applying ideas from nonsmooth optimization (Beck and Teboulle, 2009) or accelerating smooth formulations with other implicit regularization (Wu and Rebeschini, 2021).

Figure 6: Recovered images after 140 iterations for the accelerated as well as Wirtinger flow algorithm on the coded diffraction test. As we can see, the accelerated methods have already recovered an accurate image before the gradient method even begins to converge. Note that the WF image is black because most pixels are much smaller in value than the maximum.

Acknowledgements

References

Tayo Ajayi, David Mildebrath, Anastasios Kyrillidis, Shashanka Ubaru, Georgios Kollias, and Kristofer Bouchard. Provably convergent acceleration in factored gradient descent with applications in matrix sensing. *arXiv preprint arXiv:1806.00534*, 2018.

David A Barmherzig and Ju Sun. Towards practical holographic coherent diffraction imaging via maximum likelihood estimation. *Optics Express*, 30(5): 6886–6906, 2022.

David A Barmherzig, Ju Sun, Po-Nan Li, Thomas Joseph Lane, and Emmanuel J Candes. Holographic phase retrieval and reference design. *Inverse Problems*, 35(9):094001, 2019.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1): 183–202, 2009.

Samuel Burer and Renato DC Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical programming*, 95(2):329–357, 2003.

T Tony Cai and Anru Zhang. ROP: Matrix recovery by rank one projects. *The Annals of Statistics*, 43 (1):102–138, 2015.

Emmanuel J Candes, 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.

Emmanuel J Candes, Yonina C Eldar, Thomas Strohmer, and Vladislav Voroninski. Phase retrieval via matrix completion. *SIAM review*, 57(2):225–251, 2015.

Yuxin Chen, Yuejie Chi, and Andrea J Goldsmith. Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Transactions on Information Theory*, 61(7):4034–4059, 2015.

Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176:5–37, 2019.

John C Duchi and Feng Ruan. Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *Information and Inference: A Journal of the IMA*, 8(3):471–529, 2019.

Yonina C Eldar, Pavel Sidorenko, Dustin G Mixon, Shaby Barel, and Oren Cohen. Sparse phase retrieval from short-time fourier measurements. *IEEE Signal Processing Letters*, 22(5):638–642, 2014.

Mathieu Even, Scott Pesme, Suriya Gunasekar, and Nicolas Flammarion. (s) gd over diagonal linear networks: Implicit regularisation, large stepsizes and edge of stability. *arXiv preprint arXiv:2302.08982*, 2023.

Maryam Fazel, Emmanuel Candes, Benjamin Recht, and Pablo Parrilo. Compressed sensing and robust recovery of low rank matrices. In *2008 42nd Asilomar Conference on Signals, Systems and Computers*, pages 1043–1047. IEEE, 2008.

Avrajit Ghosh, He Lyu, Xitong Zhang, and Rongrong Wang. Implicit regularization in heavy-ball momentum accelerated stochastic gradient descent. In *ICLR*, 2023.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. *Advances in neural information processing systems*, 30, 2017.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841. PMLR, 2018.

Paul Hand, Oscar Leong, and Vlad Voroninski. Phase retrieval under a generative prior. *Advances in Neural Information Processing Systems*, 31, 2018.

Thomas Y Hou, Zhenzhen Li, and Ziyun Zhang. Fast global convergence for low-rank matrix recovery via riemannian gradient descent with random initialization. *arXiv preprint arXiv:2012.15467*, 2020.

Wen Huang, Kyle A Gallivan, and Xiangxiong Zhang. Solving phaselift by low-rank riemannian optimization methods for complex semidefinite constraints. *SIAM Journal on Scientific Computing*, 39(5): B840–B859, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kiryung Lee and Dominik Stöger. Randomly initialized alternating least squares: Fast convergence for matrix sensing. *SIAM Journal on Mathematics of Data Science*, 5(3):774–799, 2023.

Yuanxin Li, Cong Ma, Yuxin Chen, and Yuejie Chi. Nonconvex matrix factorization from rank-one measurements. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1496–1505. PMLR, 2019.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47. PMLR, 2018.

Yuetian Luo, Wen Huang, Xudong Li, and Anru Zhang. Recursive importance sketching for rank constrained least squares: Algorithms and high-order convergence. *Operations Research*, 2023.

Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Foundations of Computational Mathematics*, 20:451–632, 2020.

Tyler Maunu, Thibaut Le Gouic, and Philippe Rigollet. Bures-wasserstein barycenters and low-rank matrix recovery. *AISTATS*, 2023.

Jianwei Miao, Tetsuya Ishikawa, Qun Shen, and Thomas Earnest. Extending x-ray crystallography to allow the imaging of noncrystalline materials, cells, and single protein complexes. *Annu. Rev. Phys. Chem.*, 59:387–410, 2008.

Hancheng Min, Salma Tarmoun, René Vidal, and Enrique Mallada. On the explicit role of initialization on the convergence and implicit bias of overparametrized linear networks. In *International Conference on Machine Learning*, pages 7760–7768. PMLR, 2021.

Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence

o (1/k2). In *Dokl. Akad. Nauk. SSSR*, volume 269, page 543, 1983.

Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

Nicolò Pagliana and Lorenzo Rosasco. Implicit regularization of accelerated methods in hilbert spaces. *Advances in Neural Information Processing Systems*, 32, 2019.

Boris Polyak. Introduction to optimization. optimization software. *Inc., Publications Division, New York*, 1, 1987.

Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

Sujay Sanghavi, Rachel Ward, and Chris D White. The local convexity of solving systems of quadratic equations. *Results in Mathematics*, 71:569–608, 2017.

Yoav Shechtman, Amir Beck, and Yonina C Eldar. Gespar: Efficient phase retrieval of sparse signals. *IEEE transactions on signal processing*, 62(4):928–938, 2014.

Yoav Shechtman, Yonina C Eldar, Oren Cohen, Henry Nicholas Chapman, Jianwei Miao, and Mordechai Segev. Phase retrieval with application to optical imaging: a contemporary overview. *IEEE signal processing magazine*, 32(3):87–109, 2015.

Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18:1131–1198, 2018.

Pierre Thibault, Martin Dierolf, Andreas Menzel, Oliver Bunk, Christian David, and Franz Pfeiffer. High-resolution scanning x-ray diffraction microscopy. *Science*, 321(5887):379–382, 2008.

Tomas Vaskevicius, Varun Kanade, and Patrick Rebeschini. Implicit regularization for optimal sparse recovery. *Advances in Neural Information Processing Systems*, 32, 2019.

Fei Wang, Yaoming Bian, Haichao Wang, Meng Lyu, Giancarlo Pedrini, Wolfgang Osten, George Barbastathis, and Guohai Situ. Phase imaging with an untrained neural network. *Light: Science & Applications*, 9(1):77, 2020.

Gang Wang, Liang Zhang, Georgios B Giannakis, Mehmet Akçakaya, and Jie Chen. Sparse phase retrieval via truncated amplitude flow. *IEEE Transactions on Signal Processing*, 66(2):479–491, 2017.

Gang Wang, Georgios B Giannakis, Yousef Saad, and Jie Chen. Phase retrieval via reweighted amplitude flow. *IEEE Transactions on Signal Processing*, 66 (11):2818–2833, 2018.

Jun-Kun Wang and Jacob Abernethy. Quickly finding a benign region via heavy ball momentum in non-convex optimization. *arXiv preprint arXiv:2010.01449*, 2020.

Li Wang, Zhiguo Fu, Yingcong Zhou, and Zili Yan. The implicit regularization of momentum gradient descent in overparametrized models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10149–10156, 2023.

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673. PMLR, 2020.

Fan Wu and Patrick Rebeschini. A continuous-time mirror descent approach to sparse phase retrieval. *Advances in Neural Information Processing Systems*, 33:20192–20203, 2020.

Fan Wu and Patrick Rebeschini. Hadamard wirtinger flow for sparse phase retrieval. In *International Conference on Artificial Intelligence and Statistics*, pages 982–990. PMLR, 2021.

Huaqing Xiong, Yuejie Chi, Bin Hu, and Wei Zhang. Convergence analysis of accelerated first-order methods for phase retrieval. *MTS*, 2018.

Huaqing Xiong, Yuejie Chi, Bin Hu, and Wei Zhang. Analytical convergence regions of accelerated gradient descent in nonconvex optimization under regularity condition. *Automatica*, 113:108715, 2020.

Rui Xu, Mahdi Soltanolkotabi, Justin P Haldar, Walter Unglaub, Joshua Zusman, Anthony FJ Levi, and Richard M Leahy. Accelerated wirtinger flow: A fast algorithm for ptychography. *arXiv preprint arXiv:1806.05546*, 2018.

Peng Zhao, Yun Yang, and Qiao-Chu He. High-dimensional linear regression via implicit regularization. *Biometrika*, 109(4):1033–1046, 2022.

Zhong Zhuang, David Yang, Felix Hofmann, David Barmherzig, and Ju Sun. Practical phase retrieval using double deep image priors. *arXiv preprint arXiv:2211.00799*, 2022.

Checklist

1. For all models and algorithms presented, check if you include:

   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes – see Section 2.1 as well as Section 2.2.]

   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm.

[Yes - the complexity of the method is the same as gradient descent, and we prove better iteration complexity. See Theorem 2.]

(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No - we do not provide source code, but can provide it upon request and will make a public github repository upon release of our paper.]

2. For any theoretical claim, check if you include:

(a) Statements of the full set of assumptions of all theoretical results. [Yes - we include all assumptions in the statement of Theorem 2.]

(b) Complete proofs of all theoretical results. [Yes - all proofs are sketched in Section 3.2 and full details are given in the appendix.]

(c) Clear explanations of any assumptions. [Yes - the main assumption is the Gaussian assumption in Assumption 1.]

3. For all figures and tables that present empirical results, check if you include:

(a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [No - as mentioned above, full source code and instructions will be put on github with the release of our paper.]

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes - see the description of the experiments in Section 4]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes - see Appendix.]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator If your work uses existing assets. [Yes - the only data we use is from Candes et al. (2015) with the Stanford quad picture in the coded diffraction test.]

(b) The license information of the assets, if applicable. [Not Applicable]

(c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

(d) Information about consent from data providers/curators. [Not Applicable]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A SUPPLEMENTARY MATERIAL

Recall that, by Assumption 1, the sensing vectors $\{\boldsymbol{a}_i\}$, $i = 1, \ldots, m$, are sampled i.i.d from a normal distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. In the proofs in this section, we use the following results that come from standard concentration inequalities for normal distributions, see Ma et al. (2020, Appendix A):

1. With probability at least $1 - \mathcal{O}(me^{-3n/2})$:

$$\max_{1 \leq i \leq m} \|\boldsymbol{a}_i\|_2 \leq \sqrt{6n} \tag{18}$$

2. With probability at least $1 - \mathcal{O}(mn^{-10})$, for any $\boldsymbol{y} \in \mathbb{R}^n$:

$$\max_{1 \leq i \leq m} |\boldsymbol{a}_i^T \boldsymbol{y}| \leq 5\sqrt{\log(n)} \|\boldsymbol{y}\|_2 \tag{19}$$

A.1 Proof of Theorem 2 for Polyak's Heavy Ball method

Proof of Theorem 2 for Polyak's Heavy Ball Method (P). By repeated application of Lemma 4, for any $t \leq n$, we have that $\boldsymbol{x}^t$ is in the RIC with probability at least $(1 - \mathcal{O}(mn^{-10}))^n \geq 1 - \mathcal{O}(mn^{-9})$. Moreover, by Lemma 3 we also have that

$$\left\| \begin{bmatrix} \boldsymbol{x}^{t+1} - \boldsymbol{x}_* \\ \boldsymbol{x}^t - \boldsymbol{x}_* \end{bmatrix} \right\|_2 \leq \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \right)^t \left\| \begin{bmatrix} \boldsymbol{x}^1 - \boldsymbol{x}_* \\ \boldsymbol{x}^0 - \boldsymbol{x}_* \end{bmatrix} \right\|_2$$

$$\leq \left(1 - \frac{2\sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \right)^t C_1. \tag{20}$$

In particular, for $T_0 = n$, we obtain

$$\left\| \begin{bmatrix} \boldsymbol{x}^{T_0} - \boldsymbol{x}_* \\ \boldsymbol{x}^{T_0-1} - \boldsymbol{x}_* \end{bmatrix} \right\|_2 \leq \left(1 - \frac{2\sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \right)^{T_0} C_1 \tag{21}$$

$$\leq \frac{1}{n} C_1, \tag{22}$$

as long as $n \geq \dfrac{-\log(n)}{\log\left(1 - \frac{2\sqrt{1/2}}{\sqrt{C\log(n)} + \sqrt{1/2}} \right)}$. Furthermore, the choice $t \geq n$ then implies that $\left(1 - \dfrac{2\sqrt{1/2}}{\sqrt{C\log(n)} + \sqrt{1/2}} \right)^t \leq \dfrac{1}{n}$. Therefore, as long as the iterates remain in the RIC for the first T_0 iterations, we get contraction to a neighborhood where $\operatorname{dist}(\boldsymbol{x}^{T_0}, \boldsymbol{x}_*) \leq C_1/n = O(1/n)$.

After this, for $t \geq T_0$, locality implies incoherence, i.e., condition (LOC) implies (INC). Indeed, for $t = T_0$, by Cauchy-Schwartz inequality, concentration inequality (18), and inequality (21) we obtain:

$$\max_{1 \leq \ell \leq m} |\boldsymbol{a}_\ell^T (\boldsymbol{x}^{T_0} - \boldsymbol{x}_*)|$$
$$\leq \sqrt{6n} \|\boldsymbol{x}^{T_0} - \boldsymbol{x}_*\| \leq \sqrt{6n} \frac{1}{n} C_1$$
$$\lesssim C_2 \sqrt{\log(n)}. \tag{23}$$

Now, because $\boldsymbol{x}^{T_0}$ is in the RIC, we can use Lemma 3 to get that

$$\left\| \begin{bmatrix} \boldsymbol{x}^{T_0+1} - \boldsymbol{x}_* \\ \boldsymbol{x}^{T_0} - \boldsymbol{x}_* \end{bmatrix} \right\|_2 \leq \left(1 - \frac{2\sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \right) \left\| \begin{bmatrix} \boldsymbol{x}^{T_0} - \boldsymbol{x}_* \\ \boldsymbol{x}^{T_0-1} - \boldsymbol{x}_* \end{bmatrix} \right\|_2$$

$$\leq \frac{1}{n} C_1 \tag{24}$$

Following the same argument that led to (23) and (24), we can see that locality implies incoherence for iterates with $t \geq T_0$. We have just shown that the assumptions of Lemma 3 are satisfied for all $t \geq T_0 = n$, which implies that

$$\left\| \begin{bmatrix} x^{t+1} - x_* \\ x^t - x_* \end{bmatrix} \right\|_2 \leq \left(1 - \frac{2\sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \right)^t C_1 \tag{25}$$

$\square$

The previous proofs rely on the following three lemmas. We state the first without proof, since it was shown in Ma et al. (2020). This lemma gives high probability bounds on the Hessian seen in (11) provided that the point x is in the RIC.

Lemma 2 (Lemma 1 of Ma et al. (2020)). *For sufficiently small C_1, sufficiently large C_2, and $m \geq c_0 n \log n$ with sufficiently large c_0, with probability at least $1 - \mathcal{O}(mn^{-10})$, one has*

$$\nabla^2 f(x) \preceq (5C_2(10 + C_2) \log(n)) I_n \tag{26}$$

for all $x \in \mathbb{R}^n$ for which $\|x - x_*\|_2 \leq 2C_1$, and

$$\nabla^2 f(x) \succeq \frac{1}{2} I_n \tag{27}$$

for all $x \in \mathbb{R}^n$ such that

$$\|x - x_*\|_2 \leq 2C_1 \tag{28}$$

$$\max_i |a_i^T (x - x_*)| \leq C_2 \sqrt{\log(n)}$$

In the following proof of Theorem 2, as a short hand, we use $\mu = 1/2$ and $L = c \log n$ as lower and upper bounds on the Hessian by Lemma 2.

Next, using the previous lemma, we show that if the current pair of consecutive iterates lie within the RIC, then they get closer to x_*.

Lemma 3 (Contraction). *If $a_j \overset{i.i.d.}{\sim} N(0, I)$, then with probability at least $1 - \mathcal{O}(mn^{-10})$, then the iterates* (P)

$$\left\| \begin{bmatrix} x^{t+1} - x_* \\ x^t - x_* \end{bmatrix} \right\|_2 \leq \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \right) \left\| \begin{bmatrix} x^t - x_* \\ x^{t-1} - x_* \end{bmatrix} \right\|_2 \tag{29}$$

when x^t, x^{t-1} satisfy conditions (28), *$\eta \lesssim \frac{4}{(\sqrt{1/2} + \sqrt{\log(n)})^2}$, and $\beta = \frac{\sqrt{c \log(n)} - \sqrt{1/2}}{\sqrt{c \log(n)} + \sqrt{1/2}}$.*

Proof of Lemma 3. We can write one step of the heavy ball method as

$$\begin{bmatrix} x^{t+1} - x_* \\ x^t - x_* \end{bmatrix} = \begin{bmatrix} x^t - x_* - \eta \nabla f(x^t) + \beta(x^t - x^{t-1}) \\ x^t - x_* \end{bmatrix}$$

$$= M(t) \begin{bmatrix} x^t - x_* \\ x^{t-1} - x_* \end{bmatrix} \tag{30}$$

where $\xi^t \in [x_*, x^t]$, the segment joining x_* and x^t. Here we have made

$$M(t) = \begin{bmatrix} (1 + \beta)I - \nabla^2 f(\xi^t) & -\beta I \\ I & 0 \end{bmatrix} \tag{31}$$

A direct analysis shows that ξ^t is in the region of RIC. Indeed, if $\xi^t = (1 - \tau)x_* + \tau x^t$ for some $\tau \in (0, 1)$ then

$$\|\xi^t - x_*\|_2 = \tau \|x^t - x_*\|_2 \leq 2C_1$$

and

$$\max_i |a_i^T (\xi - x_*)| = \tau \max_i |a_i^t (x^t - x_*)| \leq C_2 \sqrt{\log n}.$$

Therefore, as f is strongly convex and smooth, we can bound the norm of the matrix factor (31) as is usually done in the heavy ball method, by bounding its eigenvalues (see for instance Polyak (1964, 1987)).

Bounding the eigenvalues in this way yields the contraction

$$\left\| \begin{bmatrix} \boldsymbol{x}^{t+1} - \boldsymbol{x}_* \\ \boldsymbol{x}^t - \boldsymbol{x}_* \end{bmatrix} \right\|_2 \leq \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \right) \left\| \begin{bmatrix} \boldsymbol{x}^t - x_* \\ \boldsymbol{x}^{t-1} - x_* \end{bmatrix} \right\|_2 \tag{32}$$

$\square$

We now put the previous two lemmas together to prove the main induction. This lemma ensures that pairs of consecutive iterates maintain the contraction and incoherence properties necessary to remain in the RIC. Note that the proof of this lemma relies on a further Lemma 5, where we show that the iterates (P) stay close to leave-one-out sequences.

Lemma 4 (Induction). *Suppose that we initialize* (P) *with* $\boldsymbol{x}^0 = \sqrt{\frac{\lambda_1(\boldsymbol{Y})}{3}} \widetilde{\boldsymbol{x}}^0$, *where* $\lambda_1(\boldsymbol{Y})$ *and* $\boldsymbol{x}^0$ *are the leading eigenvalue and eigenvector of* $\boldsymbol{Y} = \frac{1}{m} \sum_{i=1}^m y_i \boldsymbol{a}_i \boldsymbol{a}_i^T$. *Suppose further that* $\boldsymbol{a}_j \overset{i.i.d.}{\sim} N(\boldsymbol{0}, \boldsymbol{I})$. *Then, for* n *sufficiently large, with probability at least* $1 - \mathcal{O}(mn^{-10})$,

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}_*\|_2^2 + \|\boldsymbol{x}^t - \boldsymbol{x}_*\|_2^2 \leq C_1^2 \tag{33}$$

$$\max_{1 \leq i \leq m} |\boldsymbol{a}_i^T(\boldsymbol{x}^{t+1} - \boldsymbol{x}_*)| \leq C_2\sqrt{\log(n)} \tag{34}$$

Proof. The proof of all these relations relies on an induction argument. According to Lemma 5 in Ma et al. (2020), for a given $\delta > 0$ and m large enough one has $\|\boldsymbol{x}^0 - \boldsymbol{x}_*\| \leq 2\delta$. Choosing $\delta = \frac{C_1}{8}$ implies that $\|\boldsymbol{x}^1 - \boldsymbol{x}_*\|^2 + \|\boldsymbol{x}^0 - \boldsymbol{x}_*\|^2 \leq C_1$. Furthermore, by Lemma 3, for $\mathcal{C} = \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \right) < 1$,

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}_*\|_2^2 + \|\boldsymbol{x}^t - \boldsymbol{x}_*\|_2^2 \tag{35}$$
$$\leq \mathcal{C}^2(\|\boldsymbol{x}^t - \boldsymbol{x}_*\|_2^2 + \|\boldsymbol{x}^{t-1} - \boldsymbol{x}_*\|_2^2) \leq C_1^2$$

Then, we prove incoherence using induction as well:

$$\max_{1 \leq \ell \leq m} |\boldsymbol{a}_\ell^T(\boldsymbol{x}^{t+1} - \boldsymbol{x}_*)| \tag{36}$$

$$\leq |\boldsymbol{a}_\ell^T(\boldsymbol{x}^{t+1} - \boldsymbol{x}^{t+1,(\ell)}) + \boldsymbol{a}_\ell^T(\boldsymbol{x}^{t+1,(\ell)} - \boldsymbol{x}_*)| \tag{37}$$

$$\leq \|\boldsymbol{a}_\ell^T\| \|\boldsymbol{x}^{t+1} - \boldsymbol{x}^{t+1,(\ell)}\| + 5\sqrt{\log(n)} \|\boldsymbol{x}^{t+1,(\ell)} - \boldsymbol{x}_*\| \tag{38}$$

$$\leq \|\boldsymbol{a}_\ell^T\| C_3 \sqrt{\frac{\log(n)}{n}} \tag{39}$$

$$+ 5\sqrt{\log(n)}(\|\boldsymbol{x}^{t+1,(\ell)} - \boldsymbol{x}^{t+1}\| + \|\boldsymbol{x}^{t+1} - \boldsymbol{x}_*\|) \tag{40}$$

$$\leq \sqrt{6n} C_3 \sqrt{\frac{\log(n)}{n}} + 5\sqrt{\log(n)} \left(C_3 \sqrt{\frac{\log(n)}{n}} + C_1 \right) \tag{41}$$

$$\leq C_2\sqrt{\log(n)} \tag{42}$$

by taking n large enough and using the fact that C_2 can be taken larger than C_1 and C_3. Here, the second to last inequality was obtained using (33) and Lemma 5.

$\square$

To prove the upper bound on the incoherence in the previous Lemma, we must use leave-one-out sequences. Define the leave-one-out function

$$f^{(\ell)}(\boldsymbol{x}) = \frac{1}{4m} \sum_{i:i \neq \ell} ((\boldsymbol{a}_i^T \boldsymbol{x})^2 - y_i)^2$$

We let $\boldsymbol{x}^{t,(\ell)}$ be sequence defined by running (P) on this function with the same initialization, $\boldsymbol{x}^{0,(\ell)} = \boldsymbol{x}^0$.

Lemma 5 (Leave one out). *Suppose that $\boldsymbol{a}_j \overset{i.i.d.}{\sim} N(\boldsymbol{0}, \boldsymbol{I})$. Then, with probability at least $1 - \mathcal{O}(mn^{-10})$, for n sufficiently large,*

$$\max_{1 \leq \ell \leq m} \left\| \begin{bmatrix} \boldsymbol{x}^{t+1} - \boldsymbol{x}^{t+1,(\ell)} \\ \boldsymbol{x}^t - \boldsymbol{x}^{t,(\ell)} \end{bmatrix} \right\|_2 \leq C_3 \sqrt{\frac{\log(n)}{n}}.$$

Proof. According to 6 in Ma et al. (2020) we have that

$$\max_{1 \leq \ell \leq m} \| \boldsymbol{x}^0 - \boldsymbol{x}^{0,(\ell)} \|_2 \leq C_3 \sqrt{\frac{\log(n)}{n}} \tag{43}$$

for any constant C_3 (for m large enough so that $\frac{n \log(n)}{m} \leq C_3$). Our base case is satisfied if C_3 is chosen appropriately and we take into account that $\boldsymbol{x}^1 = \boldsymbol{x}^0, \boldsymbol{x}^{1,(\ell)} = \boldsymbol{x}^{0,(\ell)}$ (see Lemma 6 in Ma et al. (2020)).

Recall that

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \nabla f(\boldsymbol{x}^t) + \beta(\boldsymbol{x}^t - \boldsymbol{x}^{t-1}).$$

For the leave one out iteration we have:

$$\boldsymbol{x}^{t+1} - \boldsymbol{x}^{t+1,(\ell)} = \boldsymbol{x}^t - \eta \nabla f(\boldsymbol{x}^t) + \beta(\boldsymbol{x}^t - \boldsymbol{x}^{t-1}) \tag{44}$$

$$- [\boldsymbol{x}^{t,(\ell)} - \eta \nabla f^{(\ell)}(\boldsymbol{x}^{t,\ell})] - \beta(\boldsymbol{x}^{t,(\ell)} - \boldsymbol{x}^{t-1,(\ell)}), \tag{45}$$

that can be rewritten as (by adding and subtracting the term $\eta \nabla f(\boldsymbol{x}^{t,(\ell)})$):

$$\begin{bmatrix} \boldsymbol{x}^{t+1} - \boldsymbol{x}^{t+1,(\ell)} \\ \boldsymbol{x}^t - \boldsymbol{x}^{t,(\ell)} \end{bmatrix} = \boldsymbol{M}(t) \begin{bmatrix} \boldsymbol{x}^t - \boldsymbol{x}^{t,(\ell)} \\ \boldsymbol{x}^{t-1} - \boldsymbol{x}^{t-1,(\ell)} \end{bmatrix} \tag{46}$$

$$- \eta \begin{bmatrix} \nabla f(\boldsymbol{x}^{t,(\ell)}) - \nabla f^{(\ell)}(\boldsymbol{x}^{t.(\ell)}) \\ \boldsymbol{0} \end{bmatrix}, \tag{47}$$

where

$$\boldsymbol{M}(t) = \begin{bmatrix} \boldsymbol{I} - \eta \nabla^2 f(\boldsymbol{\xi}^t) + \beta \boldsymbol{I} & -\beta \boldsymbol{I} \\ \boldsymbol{I} & \boldsymbol{0} \end{bmatrix} \tag{48}$$

for some $\boldsymbol{\xi}^t \in [\boldsymbol{x}^t, \boldsymbol{x}^{t,(\ell)}]$. As in the proof of Lemma 3, $\boldsymbol{\xi}^t$ is in the RIC. If $\boldsymbol{\xi}^t = (1-\tau)\boldsymbol{x}^t + \tau \boldsymbol{x}^{t,(\ell)}$, then

$$\|\boldsymbol{\xi}^t - \boldsymbol{x}_*\|_2 \leq (1-\tau)\|\boldsymbol{x}^t - \boldsymbol{x}_*\| + \tau \|\boldsymbol{x}^{t,(\ell)} - \boldsymbol{x}_*\| \tag{49}$$

$$\leq (1-\tau)C_1 + \tau(\|\boldsymbol{x}^{t,(\ell)} - \boldsymbol{x}^t\| + \|\boldsymbol{x}^t - \boldsymbol{x}_*\|) \tag{50}$$

$$\leq (1-\tau)C_1 + \tau \left(C_3 \sqrt{\frac{\log(n)}{n}} + C_1 \right) \tag{51}$$

$$\leq 2C_1 \tag{52}$$

for large enough n. We can also show that $\boldsymbol{\xi}^t$ satisfies the incoherence condition using the induction hypothesis

$$|\boldsymbol{a}_\ell^T(\boldsymbol{\xi}^t - \boldsymbol{x}_*)| \tag{53}$$

$$\leq |(1-\tau)\boldsymbol{a}_\ell^T(\boldsymbol{x}^t - \boldsymbol{x}_*) + \tau \boldsymbol{a}_\ell^T(\boldsymbol{x}^{t,(\ell)} - \boldsymbol{x}_*)| \tag{54}$$

$$\leq (1-\tau)C_2 \sqrt{\log(n)} \tag{55}$$

$$+ \tau 5\sqrt{\log(n)}\|\boldsymbol{x}^{t,(\ell)} - \boldsymbol{x}^t + \boldsymbol{x}^t - \boldsymbol{x}_*\| \tag{56}$$

$$\leq (1-\tau)C_2 \sqrt{\log(n)} \tag{57}$$

$$+ \tau 5\sqrt{\log(n)} \left(C_3 \sqrt{\frac{\log(n)}{n}} + C_1 \right) \tag{58}$$

$$\leq C_2 \sqrt{\log(n)} \tag{59}$$

Thus, by bounding $\boldsymbol{M}(t)$ and then using induction hypothesis, we obtain

$$\left\| \boldsymbol{M}(t) \begin{bmatrix} \boldsymbol{x}^t - \boldsymbol{x}^{t,(\ell)} \\ \boldsymbol{x}^{t-1} - \boldsymbol{x}^{t-1,(\ell)} \end{bmatrix} \right\|_2 \le \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \right) \left\| \begin{bmatrix} \boldsymbol{x}^t - \boldsymbol{x}^{t,(\ell)} \\ \boldsymbol{x}^{t-1} - \boldsymbol{x}^{t-1,(\ell)} \end{bmatrix} \right\|_2 \tag{60}$$

$$\le \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \right) C_3 \sqrt{\frac{\log(n)}{n}}.$$

On the other hand, by the definition of $f^{(\ell)}$ and (4),

$$\nabla f(x^{t,(\ell)}) - \nabla f^{(\ell)}(\boldsymbol{x}^{t,(\ell)}) = \frac{1}{m}[(\boldsymbol{a}_\ell^T \boldsymbol{x}^{t,(\ell)})^2 - (\boldsymbol{a}_\ell^T \boldsymbol{x}_*)^2]\boldsymbol{a}_\ell \boldsymbol{a}_\ell^T \boldsymbol{x}^{t,(\ell)}. \tag{61}$$

so that

$$\left\| \begin{bmatrix} \nabla f(\boldsymbol{x}^{t,(\ell)}) - \nabla f^{(\ell)}(\boldsymbol{x}^{t.(\ell)}) \\ \boldsymbol{0} \end{bmatrix} \right\|_2 \le \frac{1}{m}\|\boldsymbol{a}_\ell\|\|\boldsymbol{a}_\ell^T \boldsymbol{x}^{t,(\ell)}\||(\boldsymbol{a}_\ell^T \boldsymbol{x}^{t,(\ell)})^2 - (\boldsymbol{a}_\ell^T \boldsymbol{x}_*)^2|. \tag{62}$$

We proceed to bound each term in (62). First,

$$\|\boldsymbol{a}_\ell\| \le \sqrt{6n} \tag{63}$$

with probability at least $1 - \mathcal{O}(me^{-1.5n})$, by the concentration inequality (18).

For the second factor, using the induction hypothesis and (33) yields

$$|\boldsymbol{a}_\ell^T \boldsymbol{x}^{t,(\ell)}| \le |\boldsymbol{a}_\ell^T||(\boldsymbol{x}^{t,(\ell)} - \boldsymbol{x}_*)| + |\boldsymbol{a}_\ell^T \boldsymbol{x}_*| \le 5\sqrt{\log(n)}(\|\boldsymbol{x}^{t,(\ell)} - \boldsymbol{x}^t\| + \|\boldsymbol{x}^t - \boldsymbol{x}_*\|) + 5\sqrt{\log(n)} \tag{64}$$

$$\le 5\sqrt{\log(n)}\left(C_3\sqrt{\frac{\log(n)}{n}} + C_1 \right) + 5\sqrt{\log(n)} \tag{65}$$

$$\le (C_4 + 5)\sqrt{\log(n)} \tag{66}$$

for n large enough so that $\sqrt{\frac{\log(n)}{n}}$ is as small as needed. We have also used concentration inequality (19).

For the third factor

$$|(\boldsymbol{a}_\ell^T \boldsymbol{x}^{t,(\ell)})^2 - (\boldsymbol{a}_\ell^T \boldsymbol{x}_*)^2| \le |(\boldsymbol{a}_\ell^T \boldsymbol{x}^{t,(\ell)} - \boldsymbol{a}_\ell^T \boldsymbol{x}_*)^2 + 2\boldsymbol{a}_\ell^T(\boldsymbol{x}^{t,(\ell)} - \boldsymbol{x}_*)\boldsymbol{a}_\ell^T \boldsymbol{x}_*| \tag{67}$$

$$\le |\boldsymbol{a}_\ell^T(\boldsymbol{x}^{t,(\ell)} - \boldsymbol{x}_*)||\boldsymbol{a}_\ell^T(\boldsymbol{x}^{t,(\ell)} - \boldsymbol{x}_*) + 2\boldsymbol{a}_\ell^T \boldsymbol{x}_*| \tag{68}$$

$$\le C_2\sqrt{\log(n)}(C_2\sqrt{\log(n)} + 10\sqrt{\log(n)}) \tag{69}$$

This follows from the fact that $\boldsymbol{x}^{t,(\ell)}$ is independent of $\boldsymbol{a}_\ell$, and by Gaussian concentration. The combination of all these bounds gives

$$\left\| \begin{bmatrix} \nabla f(\boldsymbol{x}^{t,(\ell)}) - \nabla f^{(\ell)}(\boldsymbol{x}^{t.(\ell)}) \\ \boldsymbol{0} \end{bmatrix} \right\|_2 \le \sqrt{6n}(C_4 + 5)\sqrt{\log(n)}C_2(C_2 + 10)\log(n)$$

$$\le C_4 + 5)C_4(C_4 + 10)\frac{n\log(n)}{m}\sqrt{\frac{\log(n)}{n}} \tag{70}$$

$$\le c\sqrt{\frac{\log(n)}{n}} \tag{71}$$

for any c provided m is sufficiently large, $m \ge n\log(n)$. Finally the combination of bounds (60) and (70) yields

$$\left\| \begin{bmatrix} \boldsymbol{x}^{t+1} - \boldsymbol{x}^{t+1,(\ell)} \\ \boldsymbol{x}^t - \boldsymbol{x}^{t,(\ell)} \end{bmatrix} \right\|_2 \le \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \right) C_3\sqrt{\frac{\log(n)}{n}} + c\eta\sqrt{\frac{\log(n)}{n}}, \tag{72}$$

and as c can be made as small as desired and because of the relation between $\left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)$ and η then

$$\left\| \begin{bmatrix} \boldsymbol{x}^{t+1} - \boldsymbol{x}^{t+1,(\ell)} \\ \boldsymbol{x}^t - \boldsymbol{x}^{t,(\ell)} \end{bmatrix} \right\|_2 \le C_3\sqrt{\frac{\log(n)}{n}} \tag{73}$$

$$\square$$

A.2 Proof of Theorem 2 for Nesterov's Accelerated Gradient method

Proof of Theorem 2 for Nesterov's method N. Again, in the following, as a short hand, we use $\mu = 1/2$ and $L = c \log n$ as lower and upper bounds on the Hessian by Lemma 2.

The proof follows the same pattern as for Polyak's method with the difference that the contraction constant for Nesterov's method is $\left(1 - \frac{\sqrt{\mu}}{\sqrt{L}}\right)$ instead of $\left(1 - \frac{2\sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}\right)$ in equation (20).

For Nesterov's method it is also sufficient to choose $T_0 = n$ because for n large enough

$$n \geq \frac{-\log(n)}{\log\left(1 - \frac{\sqrt{\mu}}{\sqrt{L}}\right)} \geq \frac{-\log(n)}{\log\left(1 - \frac{\sqrt{1/2}}{\sqrt{C \log(n)}}\right)}$$

which implies that for $t \geq T_0 = n$

$$\left(1 - \frac{\sqrt{1/2}}{\sqrt{C \log(n)}}\right)^t \leq \frac{1}{n} \tag{74}$$

The rest of the proof stays the same as the proof of Theorem 2 for (P) given in Appendix A.1. □

We provide the lemmas as before for completeness. Note that there are small differences in the proof and these lemmas in order to deal with the extrapolation in the gradient computation for (N). In particular, it takes more work to ensure that the extrapolated point $\boldsymbol{x}^t + \beta(\boldsymbol{x}^t - \boldsymbol{x}^{t-1})$ remains in the RIC. To start, however, we show that as long as $\boldsymbol{x}^t, \boldsymbol{x}^{t-1}$ points are sufficiently far inside the RIC, then the pair has a contraction property.

Lemma 6 (Contraction for Nesterov). *If $\boldsymbol{a}_j \overset{i.i.d.}{\sim} N(\boldsymbol{0}, \boldsymbol{I})$, then with probability at least $1 - \mathcal{O}(mn^{-10})$, the iterates (N) satisfy*

$$\left\| \begin{bmatrix} \boldsymbol{x}^{t+1} - \boldsymbol{x}_* \\ \boldsymbol{x}^t - \boldsymbol{x}_* \end{bmatrix} \right\|_2 = \left(1 - \frac{\sqrt{\mu}}{\sqrt{L}}\right) \left\| \begin{bmatrix} \boldsymbol{x}^t - \boldsymbol{x}_* \\ \boldsymbol{x}^{t-1} - \boldsymbol{x}_* \end{bmatrix} \right\|_2 \tag{75}$$

when $\boldsymbol{x}^t, \boldsymbol{x}^{t-1}$ satisfies conditions (28), $\eta = \frac{1}{c \log n}$, and $\beta = \frac{\sqrt{c \log(n)} - \sqrt{1/2}}{\sqrt{c \log(n)} + \sqrt{1/2}}$.

Proof. Let's look at the recursion for $\boldsymbol{x}^t$ in Nesterov's method. In this case, we have

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \nabla f(\boldsymbol{x}^t + \beta(\boldsymbol{x}^t - \boldsymbol{x}^{t-1})) + \beta(\boldsymbol{x}^t - \boldsymbol{x}^{t-1}) \tag{76}$$

We then have

$$\begin{aligned} \boldsymbol{x}^{t+1} - \boldsymbol{x}_* = \boldsymbol{x}^t - \boldsymbol{x}_* \\ - \eta \nabla^2 f(\boldsymbol{\xi}^t)(\boldsymbol{x}^t + \beta(\boldsymbol{x}^t - \boldsymbol{x}^{t-1}) - \boldsymbol{x}_*) \\ - \beta(\boldsymbol{x}^t - \boldsymbol{x}_*) \\ + \beta(\boldsymbol{x}^{t-1} - \boldsymbol{x}_*) \end{aligned}$$

that can be rewritten as

$$\begin{bmatrix} \boldsymbol{x}^{t+1} - \boldsymbol{x}_* \\ \boldsymbol{x}^t - \boldsymbol{x}_* \end{bmatrix} = \boldsymbol{M}(t) \begin{bmatrix} \boldsymbol{x}^t - \boldsymbol{x}_* \\ \boldsymbol{x}^{t-1} - \boldsymbol{x}_* \end{bmatrix} \tag{77}$$

where

$$\boldsymbol{M}(t) = \begin{bmatrix} (1+\beta)(\boldsymbol{I} - \eta \nabla^2 f(\boldsymbol{\xi}^t)) & -\beta(\boldsymbol{I} - \nabla^2 f(\boldsymbol{\xi}^t)) \\ \boldsymbol{I} & \boldsymbol{0} \end{bmatrix} \tag{78}$$

and $\boldsymbol{\xi}^t \in [\boldsymbol{x}_*, \boldsymbol{x}^t + \beta(\boldsymbol{x}^t - \boldsymbol{x}^{t-1})]$, the segment joining $\boldsymbol{x}_*$ and $\boldsymbol{x}^t + \beta(\boldsymbol{x}^t - \boldsymbol{x}^{t-1})$.

We can show directly that $\boldsymbol{\xi}^t$ is in the region of RIC. For some $\tau \in (0, 1)$ we have

$$\begin{aligned} \|\boldsymbol{\xi}^t - \boldsymbol{x}_*\|_2 &= \|(1-\tau)\boldsymbol{x}_* + \tau(\boldsymbol{x}^t + \beta(\boldsymbol{x}^t - \boldsymbol{x}^{t-1})) - \boldsymbol{x}_*\| \\ &= \tau \|\boldsymbol{x}^t - \boldsymbol{x}_* + \beta(\boldsymbol{x}^t - \boldsymbol{x}_*) + \beta(\boldsymbol{x}_* - \boldsymbol{x}^{t-1})\| \\ &\leq \frac{C_1}{3} + 2\beta \frac{C_1}{3} \leq 2C_1. \end{aligned} \tag{79}$$

and for incoherence we get

$$|\boldsymbol{a}_\ell^T(\boldsymbol{\xi}^t - \boldsymbol{x}_*)| \tag{80}$$

$$\leq |\boldsymbol{a}_\ell^T((1-\tau)\boldsymbol{x}_* + \tau(\boldsymbol{x}^t + \beta(\boldsymbol{x}^t - \boldsymbol{x}^{t-1})) - \boldsymbol{x}_*)| \tag{81}$$

$$= \tau|\boldsymbol{a}_\ell^T(\boldsymbol{x}^t - \boldsymbol{x}_* + \beta(\boldsymbol{x}^t - \boldsymbol{x}^{t-1}))| \tag{82}$$

$$\leq |(1+\beta)\boldsymbol{a}_\ell^T(\boldsymbol{x}^t - \boldsymbol{x}_*) + \beta\boldsymbol{a}_\ell^T(\boldsymbol{x}_* - \boldsymbol{x}^t)| \tag{83}$$

$$\leq (1+2\beta)\frac{C_2}{3}\sqrt{\log(n)} \leq C_2\sqrt{\log(n)} \tag{84}$$

Therefore, as f can be considered strongly convex and smooth, we can bound the eigenvalues of the matrix $M(t)$ to obtain

$$\left\| \begin{bmatrix} \boldsymbol{x}^{t+1} - \boldsymbol{x}_* \\ \boldsymbol{x}^t - \boldsymbol{x}_* \end{bmatrix} \right\|_2 = \left(1 - \frac{\sqrt{\mu}}{\sqrt{L}}\right) \left\| \begin{bmatrix} \boldsymbol{x}^t - \boldsymbol{x}_* \\ \boldsymbol{x}^{t-1} - \boldsymbol{x}_* \end{bmatrix} \right\|_2 \tag{85}$$

$\square$

Lemma 7 (induction for Nesterov). *For n sufficiently large, with probability at least $1 - \mathcal{O}(mn^{-10})$,*

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}_*\|_2^2 + \|\boldsymbol{x}^t - \boldsymbol{x}_*\|_2^2 \leq (C_1/3)^2 \tag{86}$$

$$\max_{1 \leq i \leq m} |\boldsymbol{a}_i^T(\boldsymbol{x}^{t+1} - \boldsymbol{x}_*)| \leq \frac{C_2}{3}\sqrt{\log(n)} \tag{87}$$

Proof. In exactly the same way as for heavy ball, Lemma 5 in Ma et al. (2020) implies that the conditions are satisfied for the base case. Now, by Lemma 6, for $\mathcal{C} = \left(1 - \frac{\sqrt{\mu}}{\sqrt{L}}\right) < 1$,

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}_*\|_2^2 + \|\boldsymbol{x}^t - \boldsymbol{x}_*\|_2^2 \tag{88}$$

$$\leq \mathcal{C}^2(\|\boldsymbol{x}^t - \boldsymbol{x}_*\|_2^2 + \|\boldsymbol{x}^{t-1} - \boldsymbol{x}_*\|_2^2) \leq \left(\frac{C_1}{3}\right)^2$$

We can also use induction to prove incoherence:

$$\max_{1 \leq \ell \leq m} |\boldsymbol{a}_\ell^T(\boldsymbol{x}^{t+1} - \boldsymbol{x}_*)| \tag{89}$$

$$\leq |\boldsymbol{a}_\ell^T(\boldsymbol{x}^{t+1} - \boldsymbol{x}^{t+1,(\ell)}) + \boldsymbol{a}_\ell^T(\boldsymbol{x}^{t+1,(\ell)} - \boldsymbol{x}_*)| \tag{90}$$

$$\leq \|\boldsymbol{a}_\ell^T\|\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^{t+1,(\ell)}\| + 5\sqrt{\log(n)}\|\boldsymbol{x}^{t+1,(\ell)} - \boldsymbol{x}_*\| \tag{91}$$

$$\leq \|\boldsymbol{a}_\ell^T\|C_3\sqrt{\frac{\log(n)}{n}} \tag{92}$$

$$+ 5\sqrt{\log(n)}(\|\boldsymbol{x}^{t+1,(\ell)} - \boldsymbol{x}^{t+1}\| + \|\boldsymbol{x}^{t+1} - \boldsymbol{x}_*\|) \tag{93}$$

$$\leq \sqrt{6n}C_3\sqrt{\frac{\log(n)}{n}} + 5\sqrt{\log(n)}\left(C_3\sqrt{\frac{\log(n)}{n}} + \frac{C_1}{3}\right) \tag{94}$$

$$\leq \frac{C_2}{3}\sqrt{\log(n)} \tag{95}$$

by taking n large enough, considering that C_2 can be chosen larger than C_1 and that C_3 can be chosen as small as necessary by taking a large enough c_0 in $m \geq c_0 n \log(n)$. $\square$

Here, as in the proof of the corresponding lemma for heavy ball, to prove the upper bound on the incoherence in the previous Lemma, we must use leave-one-out sequences. As before, the leave-one-out function is defined as

$$f^{(\ell)}(\boldsymbol{x}) = \frac{1}{4m}\sum_{i:i\neq\ell}((\boldsymbol{a}_i^T\boldsymbol{x})^2 - y_i)^2$$

and we let $\boldsymbol{x}^{t,(\ell)}$ be the sequence defined by running (N) on this function with the same initialization, $\boldsymbol{x}^{0,(\ell)} = \boldsymbol{x}^0$.

Lemma 8 (Leave one out for Nesterov)**.** *For n sufficiently large, with probability at least $1 - \mathcal{O}(mn^{-10})$,*

$$\max_{1 \le \ell \le m} \left\| \begin{bmatrix} \boldsymbol{x}^{t+1} - \boldsymbol{x}^{t+1,(\ell)} \\ \boldsymbol{x}^t - \boldsymbol{x}^{t,(\ell)} \end{bmatrix} \right\|_2 \le C_3 \sqrt{\frac{\log(n)}{n}}.$$

Proof. In an analogous way to the heavy ball case, we can write one iteration of Nesterov acceleration as

$$
\begin{aligned}
\begin{bmatrix} \boldsymbol{x}^{t+1} - \boldsymbol{x}^{t+1,(\ell)} \\ \boldsymbol{x}^t - \boldsymbol{x}^{t,(\ell)} \end{bmatrix} &= \boldsymbol{M}_2(t) \begin{bmatrix} \boldsymbol{x}^t - \boldsymbol{x}^{t,(\ell)} \\ \boldsymbol{x}^{t-1} - \boldsymbol{x}^{t-1,(\ell)} \end{bmatrix} \\
&\quad - \eta \begin{bmatrix} \nabla f(\boldsymbol{x}^{t,(\ell)} + \beta(\boldsymbol{x}^{t,(\ell)} - \boldsymbol{x}^{t-1,(\ell)})) \\ \boldsymbol{0} \end{bmatrix} \\
&\quad + \eta \begin{bmatrix} \nabla f^{(\ell)}(\boldsymbol{x}^{t,(\ell)} + \beta(\boldsymbol{x}^{t,(\ell)} - \boldsymbol{x}^{t-1,(\ell)})) \\ \boldsymbol{0} \end{bmatrix},
\end{aligned}
\tag{96}
$$

where

$$\boldsymbol{M}_2(t) = \begin{bmatrix} (1+\beta)(\boldsymbol{I} - \eta \nabla^2 f(\boldsymbol{\xi}^t)) & -\beta(\boldsymbol{I} - \eta \nabla^2 f(\boldsymbol{\xi}^t)) \\ \boldsymbol{I} & \boldsymbol{0} \end{bmatrix}. \tag{97}$$

with $\boldsymbol{\xi}^t = (1-\tau)\boldsymbol{z}^{t,(\ell)} + \tau\boldsymbol{z}^t$, where $\tau \in (0,1)$ and

$$\boldsymbol{z}^{t,(\ell)} = \boldsymbol{x}^{t,(\ell)} + \beta(\boldsymbol{x}^{t,(\ell)} - \boldsymbol{x}^{t-1,(\ell)}) \tag{98}$$

$$\boldsymbol{z}^t = \boldsymbol{x}^t + \beta(\boldsymbol{x}^t - \boldsymbol{x}^{t-1}). \tag{99}$$

We can show directly that $\boldsymbol{\xi}^t$ is in the region of RIC. First, we have that

$$\|\boldsymbol{\xi}^t - \boldsymbol{x}_*\| = (1-\tau)\|\boldsymbol{z}^{t,(\ell)} - \boldsymbol{x}_*\| + \tau\|\boldsymbol{z}^t - \boldsymbol{x}_*\| \tag{100}$$

$$\le (1-\tau)(\|\boldsymbol{z}^{t,(\ell)} - \boldsymbol{z}^t\| + \|\boldsymbol{z}^t - \boldsymbol{x}_*\|) + \tau\|\boldsymbol{z}^t - \boldsymbol{x}_*\| \tag{101}$$

$$\le (1-\tau)\|\boldsymbol{z}^{t,(\ell)} - \boldsymbol{z}^t\| + \|\boldsymbol{z}^t - \boldsymbol{x}_*\|. \tag{102}$$

We can bound $\|\boldsymbol{\xi}^t - \boldsymbol{x}_*\|$ in two parts:

$$\|\boldsymbol{z}^t - \boldsymbol{x}_*\| \le (1+\beta)\|\boldsymbol{x}^t - \boldsymbol{x}_*\| + \beta\|\boldsymbol{x}^{t-1} - \boldsymbol{x}_*\| \tag{103}$$

$$\le (1+2\beta)\frac{C_1}{3} \tag{104}$$

and

$$\|\boldsymbol{z}^t - \boldsymbol{z}^{t,(\ell)}\| \le (1+\beta)\|\boldsymbol{x}^t - \boldsymbol{x}^{t,(\ell)}\| + \beta\|\boldsymbol{x}^{t-1} - \boldsymbol{x}^{t-1,(\ell)}\| \tag{105}$$

$$\le (1+2\beta)C_3 \sqrt{\frac{\log n}{n}} \tag{106}$$

so that

$$\|\boldsymbol{z}^t - \boldsymbol{z}^{t,(\ell)}\| \le (1-\tau)(1+2\beta)C_3 \sqrt{\frac{\log n}{n}} + (1+2\beta)\frac{C_1}{3} \tag{107}$$

$$\le 2C_1 \tag{108}$$

by taking n large enough.

Now for the incoherence condition we have

$$|\boldsymbol{a}_\ell^T((1-\tau)\boldsymbol{z}^{t,(\ell)} - \boldsymbol{z}^t)| \leq (1-\tau)5\sqrt{\log(n)}\|\boldsymbol{z}^{t,(\ell)} - \boldsymbol{x}_*\| \tag{109}$$

$$+ \tau|\boldsymbol{a}_\ell^T(\boldsymbol{x}^t - \boldsymbol{x}_* + \beta(\boldsymbol{x}^t - \boldsymbol{x}_* + \boldsymbol{x}_* - \boldsymbol{x}^{t-1}))| \tag{110}$$

$$\leq (1-\tau)5\sqrt{\log(n)}\|\boldsymbol{z}^{t,(\ell)} - \boldsymbol{z}^t + \boldsymbol{z}^t - \boldsymbol{x}_*\| \tag{111}$$

$$+ \tau|\boldsymbol{a}_\ell^T((1+\beta)(\boldsymbol{x}^t - \boldsymbol{x}_*) + (\boldsymbol{x}_* - \boldsymbol{x}^{t-1}))| \tag{112}$$

$$\leq (1-\tau)5\sqrt{\log(n)}(\|\boldsymbol{x}^{t,(\ell)} - \boldsymbol{x}^t\|) \tag{113}$$

$$+ (1-\tau)5\sqrt{\log(n)}(\beta\|(\boldsymbol{x}^{t,(\ell)} - \boldsymbol{x}^t + \boldsymbol{x}^{t-1} - \boldsymbol{x}^{t-1,(\ell)}\|) \tag{114}$$

$$+ (1-\tau)5\sqrt{\log(n)}(1+2\beta)C_1 \tag{115}$$

$$+ \tau((1+\beta)C_2\sqrt{\log(n)} + C_2\sqrt{\log(n)}) \tag{116}$$

$$\leq (1-\tau)5\sqrt{\log(n)}\left((1+2\beta)C_3\sqrt{\frac{\log(n)}{n}} + (1+2\beta)C_1\right) \tag{117}$$

$$+ \tau(2+\beta)C_2\sqrt{\log(n)} \tag{118}$$

$$\leq \frac{C_2}{3}\sqrt{\log(n)} \tag{119}$$

for large enough n.

As $\boldsymbol{\xi}^t$ is in the RIC, for the choice $\eta = \frac{1}{L}$, $\beta = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$ the eigenvalues of $\boldsymbol{M}_2(\cdot)$ are bounded by $1 - \sqrt{\mu}/\sqrt{L}$. After using our induction hypothesis, we get

$$\|\boldsymbol{M}_2(t)\|_2 \left\| \begin{bmatrix} \boldsymbol{x}^t - \boldsymbol{x}^{t,(\ell)} \\ \boldsymbol{x}^{t-1} - \boldsymbol{x}^{t-1,(\ell)} \end{bmatrix} \right\|_2 \leq \left(1 - \frac{\sqrt{\mu}}{\sqrt{L}}\right) \left\| \begin{bmatrix} \boldsymbol{x}^t - \boldsymbol{x}^{t,(\ell)} \\ \boldsymbol{x}^{t-1} - \boldsymbol{x}^{t-1,(\ell)} \end{bmatrix} \right\|_2 \tag{120}$$

$$\leq \left(1 - \frac{\sqrt{\mu}}{\sqrt{L}}\right)C_3\sqrt{\frac{\log(n)}{n}} \tag{121}$$

On the other hand, if we make $\boldsymbol{z}^{t,(\ell)} = \boldsymbol{x}^{t,(\ell)} + \beta(\boldsymbol{x}^{t,(\ell)} - \boldsymbol{x}^{t-1,(\ell)})$, by the definition of $f^{(\ell)}$ we have:

$$\nabla f(\boldsymbol{z}^{t,(\ell)}) - \nabla f^{(\ell)}(\boldsymbol{z}^{t,(\ell)}) \tag{122}$$

$$= \frac{1}{m}((\boldsymbol{a}_\ell^T\boldsymbol{z}^{t,(\ell)})^2 - (\boldsymbol{a}_\ell^T\boldsymbol{x}_*)^2)\boldsymbol{a}_\ell\boldsymbol{a}_\ell^T\boldsymbol{z}^{t,(\ell)} \tag{123}$$

$$\tag{124}$$

so that

$$\left\| \begin{bmatrix} \nabla f(\boldsymbol{z}^{t,(\ell)}) - \nabla f^{(\ell)}(\boldsymbol{z}^{t,(\ell)}) \\ \boldsymbol{0} \end{bmatrix} \right\|_2 \tag{125}$$

$$\leq \frac{1}{m}\|\boldsymbol{a}_\ell\|\|\boldsymbol{a}_\ell^T\boldsymbol{z}^{t,(\ell)}\|(\boldsymbol{a}_\ell^T\boldsymbol{z}^{t,(\ell)})^2 - (\boldsymbol{a}_\ell^T\boldsymbol{x}_*)^2| \tag{126}$$

$$\square$$

We now can bound each term in (126). We have that

$$\|\boldsymbol{a}_\ell\| \leq \sqrt{6n} \tag{127}$$

with probability at least $1 - \mathcal{O}(me^{-1.5n})$, by concentration inequality (18). For the second factor we have

$$|\boldsymbol{a}_\ell^T\boldsymbol{z}^{t,(\ell)}| \leq |\boldsymbol{a}_\ell(\boldsymbol{x}^{t,(\ell)} - \boldsymbol{x}^t)| + |\boldsymbol{a}_\ell(\boldsymbol{x}^t - \boldsymbol{x}_*)|$$

$$+ \beta|\boldsymbol{a}_\ell(\boldsymbol{x}^{t,(\ell)} - \boldsymbol{x}^t)| + \beta|\boldsymbol{a}_\ell^T(\boldsymbol{x}^t - \boldsymbol{x}_*)|$$

$$+ \beta|\boldsymbol{a}_\ell(\boldsymbol{x}_* - \boldsymbol{x}^{t-1})| + \beta|\boldsymbol{a}_\ell^T(\boldsymbol{x}^{t-1} - \boldsymbol{x}^{t-1,(\ell)})|$$

$$+ |\boldsymbol{a}_\ell^T\boldsymbol{x}_*| \tag{128}$$

Using the induction hypothesis and locality we obtain

$$|\boldsymbol{a}_\ell^T \boldsymbol{z}^{t,(\ell)}| \le 5\sqrt{\log(n)}\left(C_3\sqrt{\frac{\log(n)}{n}} + C_1 \right. \tag{129}$$

$$\left. + C_3\sqrt{\frac{\log(n)}{n}} + 2\beta C_1 + \beta C_3\sqrt{\frac{\log(n)}{n}} \right) \tag{130}$$

$$+ 5\sqrt{\log(n)} \tag{131}$$

so that

$$|\boldsymbol{a}_\ell^T \boldsymbol{z}^{t,(\ell)}| \le 5\sqrt{\log(n)}\left((1+2\beta)C_3\sqrt{\frac{\log(n)}{n}} + (1+2\beta)C_1 \right) \tag{132}$$

$$+ 5\sqrt{\log(n)} \tag{133}$$

$$\le (C_4+5)\sqrt{\log(n)} \tag{134}$$

for sufficiently large n so that $\sqrt{\frac{\log(n)}{n}}$ is as small as needed. For the third factor, by making $\boldsymbol{z}^{t,(\ell)} = \boldsymbol{x}^{t,(\ell)} + \beta(\boldsymbol{x}^{t,(\ell)} - \boldsymbol{x}^{t-1,(\ell)})$,

$$|(\boldsymbol{a}_\ell^T \boldsymbol{z}^{t,(\ell)})^2 - (\boldsymbol{a}_\ell^T \boldsymbol{x}_*)^2| \le |(\boldsymbol{a}_\ell^T \boldsymbol{z}^{t,(\ell)} - \boldsymbol{a}_\ell^T \boldsymbol{x}_*)^2 - 2\boldsymbol{a}_\ell^T(\boldsymbol{z}^{t,(\ell)} - \boldsymbol{x}_*)\boldsymbol{a}_\ell^T \boldsymbol{x}_*| \tag{135}$$

$$\le |\boldsymbol{a}^T(\boldsymbol{z}^{t,(\ell)} - \boldsymbol{x}_*)||\boldsymbol{a}_\ell^T(\boldsymbol{z}^{t,(\ell)} - \boldsymbol{x}_*) - 2\boldsymbol{a}_\ell^T \boldsymbol{x}_*|$$

$$\le (C_2\sqrt{\log(n)} + 2\beta C_2\sqrt{\log(n)}) \tag{136}$$

$$\cdot (C_2\sqrt{\log(n)} + 2\beta C_2\sqrt{\log(n)} + 10\sqrt{\log(n)}) \tag{137}$$

$$= C_2(1+2\beta)\sqrt{\log(n)}((1+2\beta)C_2\sqrt{\log(n)} + 10\sqrt{\log(n)}) \tag{138}$$

Combining all these bounds gives

$$\frac{1}{m}\|\nabla f(\boldsymbol{z}^{t,(\ell)}) - \nabla f^{(\ell)}(\boldsymbol{z}^{t,(\ell)})\| \le \frac{1}{m}\sqrt{6n}(C_4+5)\sqrt{\log(n)}C_2(C_2+10)\log(n) \tag{139}$$

$$\le (C_4+5)C_2(C_2+10)\frac{n\log(n)}{m}\sqrt{\frac{\log(n)}{n}} \tag{140}$$

$$\le c\sqrt{\frac{\log(n)}{n}} \tag{141}$$

where c can be made as small as needed by taking m large. Finally, the combination of bounds (121) and (141) with equation (96) yields

$$\begin{bmatrix} \boldsymbol{x}^{t+1} - \boldsymbol{x}^{t+1,(\ell)} \\ \boldsymbol{x}^t - \boldsymbol{x}^{t,(\ell)} \end{bmatrix} \le \left[\left(1 - \frac{\sqrt{\mu}}{\sqrt{L}}\right) + c\eta \right] C_3\sqrt{\frac{\log(n)}{n}} \tag{142}$$

$$\le C_3\sqrt{\frac{\log(n)}{n}} \tag{143}$$

because c can be made as small as needed and $\eta = \frac{1}{L}$.

A.3 Additional Details for Experiments

All experiments were conducted in Python run on a 2021 Macbook Pro with 16 GB RAM and a 2 GHz Quad-Core Intel Core i5 processor.