Supplementary Material On Asymptotic Linear Convergence of Projected Gradient Descent for Constrained Least Squares

Trung Vu, *Graduate Student Member, IEEE*, Raviv Raich, *Senior Member, IEEE*

I. Related Works

In this section, we review existing approaches to convergence analysis of iterative first-order methods in optimization including projected gradient descent. We present several aspects of convergence, namely, convergence to a global versus a local optimum and speed of convergence. Finally, we clarify our contribution in this work with regard to previous works in the literature.

A. Convergence of Iterative First-Order Methods

Convergence properties of iterative algorithms such as PGD often involve two key aspects: the quality of convergent points and the speed of convergence. On the one hand, the quality of convergent points provides useful insights into when the algorithm converges, whether it converges to a stationary point or a set of stationary points of the problem, and how big is the gap between the objective function at the convergent point and the optimal objective value. On the other hand, the speed of convergence concerns the order of convergence, the rate of convergence, and the number of iterations required to obtain sufficiently small errors. Let $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ be the sequence of updates generated by a certain iterative first-order method (e.g., PGD). In order to prove the convergence of the algorithm, it is common [1]–[5] to consider the convergence of the following quantities to $\mathbf{0}$ as $k \to \infty$: *(i)* the norm of the generalized gradient ($\|\frac{1}{\eta}(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)})\|$), *(ii)* the gap between current objective function and the optimal value ($|f(\boldsymbol{x}^{(k)}) - f^*|$), and *(iii)* the distance to a convergent point ($\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|$). Here, we note that f^* and $\boldsymbol{x}^*$ are the limiting points of the objective function $f(\boldsymbol{x}^{(k)})$ and the parameter $\boldsymbol{x}^{(k)}$ as the number of iterations k goes to infinity, respectively. In (i), the convergence of the generalized gradient norm to 0 implies the stationarity condition of the constrained problem is satisfied. It follows that the algorithm converges to *a set of stationary points* of the problem. In (ii), the convergence on the function side is often obtained via the monotonicity of the objective-value sequence $\{f(\boldsymbol{x}^{(k)})\}_{k=0}^{\infty}$ (e.g., decreasing to a limiting value f^*). This in turn implies the sequence $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ converges to *a set of local optima* that yields the same objective function value f^*.[1] In (iii), the convergence of $\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|$ implies convergence to a unique point that is often an *isolated local optimum point* of the problem. Typically, convergence on the domain side is used in linear convergence proofs for strongly convex settings.

Trung Vu and Raviv Raich are with the School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR 97331, USA. (e-mails: vutru@oregonstate.edu and raich@oregonstate.edu).

B. Convergence to a Global Optimum

In general, a stationary point can be a saddle point, a local/global minimum, or local/global maximum of the problem. When both the objective function and the constraint set are convex, it is well-known that all stationary points are also global optima of the problem. Convergence analysis of iterative algorithm (e.g., PGD) in convex optimization therefore focus on providing a universal upper bound on the distance to the global solutions. Analysis on the domain side (iii) is usually used in the presence of *strong convexity* that guarantees the *uniqueness* of the global optimum [1]-Section 8.6. Without the strong convexity, one may resort to analysis on the function side (ii) in order to prove convergence to *a set* of global optima [6]-Section 10.4.3. When convexity is not guaranteed, due to a non-convex objective and/or a non-convex constraint set, convergence analysis has recourse to a set of stationary points by bounding the generalized gradient norm through iterations (i) [3]-Section 2.3.2. Notwithstanding, recent advances in structured non-convex optimization have shed light on convergence guarantees to global solutions of the problem. By exploiting the special structure of some classes of non-convex problems and using appropriate initialization, PGD can be shown to converge to a unique global optimum despite the non-convexity of these problems. Examples of such powerful results include sparse recovery with restricted isometry properties [7], matrix completion with incoherence properties [8], empirical risk minimization with restricted strong convexity and smoothness properties [9], and spherically constrained quadratic minimization with hidden convexity [10].

C. Convergence to a Local Optimum

In general non-convex settings, domain-side convergence analysis is restricted to the local region around the convergence point $\boldsymbol{x}^*$. Such points can be a saddle point, a

[1] An example for such scenario is minimizing a convex but not strongly convex function $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$ subject to $\boldsymbol{x} \in \mathbb{R}^n$ and $\|\boldsymbol{x}\|_2^2 = 1$. The $2n$ vectors $\{\boldsymbol{e}_i\}_{i=1}^n$ and $\{-\boldsymbol{e}_i\}_{i=1}^n$ are local minimizers that obtain the same objective function value. It is worthwhile mentioning that they are also the global solutions of the foregoing problem.

local minimum, or a local maximum of the problem. The ROC associated with $\boldsymbol{x}^*$ is the neighborhood in which the algorithm (e.g., PGD) is guaranteed to converge to $\boldsymbol{x}^*$ when initialized inside this region. To a certain extent, the ROC in the aforementioned global convergence analysis is the entire feasible space. However, while global convergence analysis does not require the initialization to be close to the global solution, it often ignores the local structure near the solution needed for establishing sharp bounds on the speed of convergence. In particular, bounding techniques employed in global convergence analysis hold universally, including worst-case scenarios. Thus, in many problem-specific settings where the solution lies in a benign neighborhood, the global analysis could lead to conservative convergence rate bounds. As an illustration, in minimizing a smooth and strongly convex function f, gradient descent with a fixed step size achieves the rate of convergence at most $(\kappa-1)/(\kappa+1)$ [11], where κ is the (global) condition number of f. Recall that the condition number of a differentiable convex function is the ratio of its smoothness L to strong convexity μ [4]. For any quadratic function, this global bound is also an exact and attainable estimate thanks to the fact that the objective curvature is unchanged everywhere. For non-quadratic objectives, on the other hand, this global bound may be loose as κ takes into account the worst-case scenario, in which the objective function is most ill-conditioned. The asymptotic behavior of gradient descent near the solution indeed relies on the condition number of the local Hessian $\kappa(\boldsymbol{x}^*)$ of the objective function, defining as $\lambda_{\max}(\nabla^2 f(\boldsymbol{x}^*))/\lambda_{\min}(\nabla^2 f(\boldsymbol{x}^*))$. Generally, we have $\mu \le \lambda_{\min}(\nabla^2 f(\boldsymbol{x})) \le \lambda_{\max}(\nabla^2 f(\boldsymbol{x}^*)) \le L$, for any $\boldsymbol{x}$ in the domain of f, which implies $\kappa(\boldsymbol{x}^*) \le \kappa$. This local condition number $\kappa(\boldsymbol{x}^*)$ can be significantly smaller than the global condition number κ and hence, *a local convergence analysis can yield a tighter bound that reflects the actual convergence speed of the algorithm near the solution*. Similar situation also occurs for constrained least squares in which the Hessian restricted to the constrained set can depend on the local structure of the set.

D. Speed of Convergence

To illustrate the concept of convergence speed, let us consider the convergence on the domain side, i.e., the distance $\left\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^*\right\|$. Let μ be a number between 0 and 1. The convergence of $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ to $\boldsymbol{x}^*$ is said to be at rate $\mu \triangleq \mu(\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty})$ if $\mu = \inf_{\{\epsilon_k\}_{k=0}^{\infty}} \lim_{k\to\infty} \epsilon_{k+1}/\epsilon_k$, for any monotonically decreasing sequence $\{\epsilon_k\}_{k=0}^{\infty}$ satisfying $\left\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^*\right\| \le \epsilon_k$ for all index k. The asymptotic rate of convergence of gradient descent to $\boldsymbol{x}^*$, denoted by ρ, is defined by the worst-case rate of convergence among all possible sequences $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ that are generated by the algorithm and converge to $\boldsymbol{x}^*$, i.e., $\rho = \sup_{\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}} \mu(\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty})$. Depending on the value of ρ in the interval $[0,1]$, the convergence is said to be *sublinear* when $\rho = 1$, *linear* when $0 < \rho < 1$, or *superlinear* when $\rho = 0$. The lower the value of ρ is, the faster the speed of convergence is and the fewer the number of iterations needed is to obtain a close approximation of the solution. Thus, analytical estimation of the convergence rate plays a pivotal role in convergence analysis. We would like to note two distinct methods for linear convergence rate analysis dating back to the 1960s. The first approach was proposed by Polyak [2], based on his earlier study into nonlinear difference equations [12]. The author analyzed the asymptotic convergence of gradient descent for minimizing some objective function f. Assuming $\boldsymbol{x}^*$ is a non-singular local minimum of f, Polyak showed that for any $\delta > 0$, there exists $\epsilon > 0$ such that if $\left\|\boldsymbol{x}^{(0)}-\boldsymbol{x}^*\right\| < \epsilon$ then the sequence $\{\boldsymbol{x}^{(k)}\}_{k=0}^{\infty}$ generated by gradient descent satisfies

$$\left\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^*\right\| \le \left\|\boldsymbol{x}^{(0)}-\boldsymbol{x}^*\right\| (\rho+\delta)^k, \tag{1}$$

where $\rho = \max\{|1-\eta\lambda_{\max}|, |1-\eta\lambda_{\min}|\}$ and $\lambda_{\max}$ and $\lambda_{\min}$ are the largest and smallest eigenvalues of $\nabla^2 f(\boldsymbol{x}^*)$, respectively. Here we emphasize that f does not need to be smooth and strongly convex everywhere but only so around $\boldsymbol{x}^*$. By setting $\eta_{opt} = 2/(\lambda_{\max}+\lambda_{\min})$, the optimal rate of convergence is given by $\rho_{opt} = (\kappa^*-1)/(\kappa^*+1)$, where $\kappa^* = \lambda_{\max}/\lambda_{\min}$ is the condition number of the local Hessian $\nabla^2 f(\boldsymbol{x}^*)$. When f is a strongly convex quadratic, the local result coincides with the aforementioned global result in [11] ($\kappa^* = \kappa$). The expression of ρ in (1) is called the **asymptotic convergence rate** of gradient descent with fixed step size η.[2] The second approach was developed by Daniel [13] in 1967, while studying gradient descent with *exact line search*, i.e., choosing η that minimizes the objective at each iteration. Utilizing the Kantorovich inequality [14], the author proved that if $\boldsymbol{x}^{(0)}$ is sufficiently close to $\boldsymbol{x}^*$, there exist a constant ϵ and a sequence $\{q_k\}_{k=0}^{\infty}$ such that

$$\left\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^*\right\| \le \epsilon \prod_{i=0}^{k} q_i, \qquad \lim_{k\to\infty} q_k = (\kappa^*-1)/(\kappa^*+1).$$

Note that here the characteristics of convergence are also exploited through the Hessian $\nabla^2 f(\boldsymbol{x}^*)$. This result was then extended to study the asymptotic convergence of projected gradient descent for constrained optimization [15]–[17].

II. Proof of Example 1

Our goal in this proof is to establish the Lipschitz differentiability of the projection operator onto the unit sphere $\mathcal{C} = \{\boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x}\| = 1\}$. We start by establishing the Lipschitz differentiability at a point on $\mathcal{C}$ and then extend it to any nonzero point in $\mathbb{R}^n$. For the Lipschitz differentiability on $\mathcal{C}$, we introduce the following lemma:

Lemma 1. *For any $\boldsymbol{x}^* \in \mathcal{C}$, we have*

$$\sup_{\boldsymbol{y}\in\Pi_{\mathcal{C}}(\boldsymbol{x}^*+\boldsymbol{\delta})} \|\boldsymbol{y}-\boldsymbol{x}^*-\big(\boldsymbol{I}-\boldsymbol{x}^*(\boldsymbol{x}^*)^\top\big)\boldsymbol{\delta}\| \le 2\|\boldsymbol{\delta}\|^2. \tag{2}$$

Proof. We consider two cases:
Case 1: If $\boldsymbol{x}^*+\boldsymbol{\delta} = \boldsymbol{0}$, then $\Pi_{\mathcal{C}}(\boldsymbol{0}) = \mathcal{C}$ and $\|\boldsymbol{\delta}\| = \|\boldsymbol{x}^*\| = 1$. For any $\boldsymbol{y} \in \mathcal{C}$, substituting $\boldsymbol{\delta} = -\boldsymbol{x}^*$ and then using the fact

[2]It is worthwhile to mention that using a similar technique, Nesterov [4] proved that the asymptotic rate is at most $\hat{\rho} = (\kappa^*+1)/(\kappa^*+3)$. While this bound also exploits the local information of the optimization problem, we note that it is not as tight as the bound in (1).

that $\boldsymbol{I}-\boldsymbol{x}^*(\boldsymbol{x}^*)^\top$ is the projection onto the null space of $\boldsymbol{x}^*$, we have

$$\begin{aligned}\boldsymbol{y}-\boldsymbol{x}^*-\big(\boldsymbol{I}-\boldsymbol{x}^*(\boldsymbol{x}^*)^\top\big)\boldsymbol{\delta}&=\boldsymbol{y}-\boldsymbol{x}^*+\big(\boldsymbol{I}-\boldsymbol{x}^*(\boldsymbol{x}^*)^\top\big)\boldsymbol{x}^*\\&=\boldsymbol{y}-\boldsymbol{x}^*.\end{aligned}$$

Next, taking the norm and using the triangle inequality yield

$$\begin{aligned}\|\boldsymbol{y}-\boldsymbol{x}^*-\big(\boldsymbol{I}-\boldsymbol{x}^*(\boldsymbol{x}^*)^\top\big)\boldsymbol{\delta}\|&=\|\boldsymbol{y}-\boldsymbol{x}^*\|\\&\le\|\boldsymbol{y}\|+\|\boldsymbol{x}^*\|=2\|\boldsymbol{\delta}\|^2,\end{aligned}$$

where the last step stems from $\|\boldsymbol{y}\|=\|\boldsymbol{x}^*\|=\|\boldsymbol{\delta}\|=1$. Thus, (2) holds in this case.

Case 2: If $\boldsymbol{x}^*+\boldsymbol{\delta}\neq\boldsymbol{0}$, then $\Pi_{\mathcal{C}}(\boldsymbol{x}^*+\boldsymbol{\delta})$ is singleton containing the unique projection

$$\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}^*+\boldsymbol{\delta})=\frac{\boldsymbol{x}^*+\boldsymbol{\delta}}{\|\boldsymbol{x}^*+\boldsymbol{\delta}\|}.$$

Hence, (2) is equivalent to

$$\left\|\frac{\boldsymbol{x}^*+\boldsymbol{\delta}}{\|\boldsymbol{x}^*+\boldsymbol{\delta}\|}-\boldsymbol{x}^*-\big(\boldsymbol{I}-\boldsymbol{x}^*(\boldsymbol{x}^*)^\top\big)\boldsymbol{\delta}\right\|\le 2\|\boldsymbol{\delta}\|^2. \tag{3}$$

We prove (3) by *(i)* showing that for any scalars $u>0$ and $(1-u)^2\le v\le(1+u)^2$:

$$(17u-2)v^2-2u(1-u)^2v+(1-u)^4(u+2)\ge 0, \tag{4}$$

and *(ii)* showing that (4) is equivalent to (3) with $u=\|\boldsymbol{x}^*+\boldsymbol{\delta}\|>0$ and $v=\|\boldsymbol{\delta}\|^2\ge 0$.

(i) To prove (4), let us consider the following cases:

1) If $0<u\le 2/17$, then for $v\le(1+u)^2$, we have

$$\begin{aligned}&(17u-2)v^2-2u(1-u)^2v+(1-u)^4(u+2)\\&\quad\ge(17u-2)(1+u)^4-2u(1-u)^2(1+u)^2\\&\qquad+(1-u)^4(u+2)\\&\quad=16u^2(u+2)(u^2+2u+2)\ge 0.\end{aligned}$$

2) If $2/17<u\le 1/2$, then for $(1-u)^2\le v\le(1+u)^2$, the following holds

$$\begin{aligned}&(17u-2)v^2-2u(1-u)^2v+(1-u)^4(u+2)\\&\quad\ge(17u-2)(1-u)^4-2u(1-u)^2(1+u)^2\\&\qquad+(1-u)^4(u+2)\\&\quad=8u(1-u)^2(2-u)(1-2u)\ge 0.\end{aligned}$$

3) If $u>1/2$, using the quadratic vertex at $v=u(1-u)^2/(17u-2)$ as the minimum point, we obtain

$$\begin{aligned}&(17u-2)v^2-2u(1-u)^2v+(1-u)^4(u+2)\\&\quad\ge\frac{4(1-u)^4(4u^2+8u-1)}{17u-2}\ge 0.\end{aligned}$$

(ii) Now for $u=\|\boldsymbol{x}^*+\boldsymbol{\delta}\|>0$ and $v=\|\boldsymbol{\delta}\|^2\ge 0$, we have $(\boldsymbol{x}^*)^\top\boldsymbol{\delta}=(u^2-v-1)/2$ and

$$\begin{aligned}(3)\ &\Leftrightarrow\ \left\|\frac{\boldsymbol{x}^*+\boldsymbol{\delta}}{\|\boldsymbol{x}^*+\boldsymbol{\delta}\|}-\boldsymbol{x}^*-\big(\boldsymbol{I}-\boldsymbol{x}^*(\boldsymbol{x}^*)^\top\big)\boldsymbol{\delta}\right\|\le 2\|\boldsymbol{\delta}\|^2\\&\Leftrightarrow\ \|\boldsymbol{x}^*+\boldsymbol{\delta}-\|\boldsymbol{x}^*+\boldsymbol{\delta}\|(\boldsymbol{x}^*+\boldsymbol{\delta}-\boldsymbol{x}^*(\boldsymbol{x}^*)^\top\boldsymbol{\delta})\|^2\\&\qquad\le 4\|\boldsymbol{x}^*+\boldsymbol{\delta}\|^2\|\boldsymbol{\delta}\|^4\\&\Leftrightarrow\ \|(1-u)(\boldsymbol{x}^*+\boldsymbol{\delta})+u((\boldsymbol{x}^*)^\top\boldsymbol{\delta})\boldsymbol{x}^*\|_2^2\le 4u^2v^2\\&\Leftrightarrow\ (1-u)^2u^2+u^2\Big(\frac{u^2-v-1}{2}\Big)^2\\&\qquad+2u(1-u)\frac{u^2-v-1}{2}\frac{u^2-v+1}{2}\le 4u^2v^2\\&\Leftrightarrow\ (4).\end{aligned}$$

Finally, by the triangle inequality, we have

$$|\|\boldsymbol{x}^*+\boldsymbol{\delta}\|-\|\boldsymbol{x}^*\||\le\|\boldsymbol{\delta}\|\le\|-\boldsymbol{x}^*\|+\|\boldsymbol{x}^*+\boldsymbol{\delta}\|,$$

which in turn verifies $(1-u)^2\le v\le(1+u)^2$. This completes our proof of the lemma. □

Next, to extend the result in Lemma 1 to any $\boldsymbol{x}\in\mathbb{R}\setminus\{0\}$, we substitute $\boldsymbol{x}^*=\boldsymbol{x}/\|\boldsymbol{x}\|$ and $\boldsymbol{\delta}=\boldsymbol{\delta}/\|\boldsymbol{x}\|$ into (2) and obtain

$$\sup_{\boldsymbol{y}\in\Pi_{\mathcal{C}}(\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}+\frac{\boldsymbol{\delta}}{\|\boldsymbol{x}\|})}\left\|\boldsymbol{y}-\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}-\Big(\boldsymbol{I}_n-\frac{\boldsymbol{x}\boldsymbol{x}^\top}{\|\boldsymbol{x}\|^2}\Big)\frac{\boldsymbol{\delta}}{\|\boldsymbol{x}\|}\right\|\le 2\frac{\|\boldsymbol{\delta}\|^2}{\|\boldsymbol{x}\|^2}. \tag{5}$$

Since the projection onto the unit sphere is scale-invariant,

$$\Pi_{\mathcal{C}}\Big(\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}+\frac{\boldsymbol{\delta}}{\|\boldsymbol{x}\|}\Big)=\Pi_{\mathcal{C}}(\boldsymbol{x}+\boldsymbol{\delta}). \tag{6}$$

Substituting (6) into (5) yields (6). Thus, by Definition 2, for any $\boldsymbol{x}\neq\boldsymbol{0}$ we obtain

$$\begin{aligned}\nabla\mathcal{P}_{\mathcal{C}}(\boldsymbol{x})&=\frac{1}{\|\boldsymbol{x}\|}\Big(\boldsymbol{I}_n-\frac{\boldsymbol{x}\boldsymbol{x}^\top}{\|\boldsymbol{x}\|^2}\Big),\\c_1(\boldsymbol{x})=\infty,\quad c_2(\boldsymbol{x})&=\frac{2}{\|\boldsymbol{x}\|^2}.\end{aligned}$$

III. Details of Application B - Iterative Hard Thresholding for Sparse Recovery

A. Proof of (42)

In this subsection, we first show that any $\boldsymbol{x}^*\in\Phi_{=s}$ and $\boldsymbol{x}\in\mathcal{B}(\boldsymbol{x}^*,|x^*_{[s]}|/\sqrt{2})$ share the same index set of s-largest elements (in magnitude), i.e., $\Omega_s(\boldsymbol{x}^*)$. Then, we construct a counter-example to demonstrate that $|x^*_{[s]}|/\sqrt{2}$ is the largest possible radius so that (42) holds.

First, we show that for any $i\in\Omega_s(\boldsymbol{x}^*)$ and $j\in\{1,\dots,n\}\setminus\Omega_s(\boldsymbol{x}^*)$, $|x_j|<|x_i|$ as follows. In particular, we have

$$\begin{aligned}|x_j-x_j^*|+|x_i-x_i^*|&\le\sqrt{2((x_j-x^*_{[j]})^2+(x_i-x_i^*)^2)}\\&\le\sqrt{2\|\boldsymbol{x}-\boldsymbol{x}^*\|^2}<|x^*_{[s]}|,\end{aligned}$$

where the last inequality stems from the fact that $\|\boldsymbol{x}-\boldsymbol{x}^*\| < |x^*_{[s]}|/\sqrt{2}$. Now, since $x^*_j = 0$ for all $j \in \{1,\dots,n\}\setminus\Omega_s(\boldsymbol{x}^*)$, we have

$$\begin{aligned}|x_j| &= |x_j - x^*_j| \\ &< |x^*_{[s]}| - |x_i - x^*_i| \\ &\le |x^*_i| - |x_i - x^*_i| \\ &\le |x^*_i + (x_i - x^*_i)| = |x_i|, \end{aligned} \tag{7}$$

Therefore, every $\boldsymbol{x} \in \mathcal{B}(\boldsymbol{x}^*, |x^*_{[s]}|/\sqrt{2})$ shares the same index set of s-largest (in magnitude) elements with $\boldsymbol{x}^*$, i.e., $\Omega_s(\boldsymbol{x}) = \Omega_s(\boldsymbol{x}^*)$, which implies (42).

We now construct the counter-example as a point $\boldsymbol{x}$ such that $\Omega_s(\boldsymbol{x}) \neq \Omega_s(\boldsymbol{x}^*)$ and $\boldsymbol{x}$ is not in $\mathcal{B}(\boldsymbol{x}^*, |x^*_{[s]}|/\sqrt{2})$ but arbitrarily close to its boundary. Without loss of generality, assume that $|x^*_1| \ge \ldots \ge |x^*_s| > |x^*_{s+1}| = \ldots = |x^*_n| = 0$. For arbitrarily small $\epsilon > 0$, define $\boldsymbol{x}$ as

$$x_i = \begin{cases} x^*_s/2 & \text{if } i = s, \\ x^*_s/2 + \epsilon & \text{if } i = s+1, \\ x_i & \text{otherwise.} \end{cases}$$

Then, since $x_{s+1} < x_s$, $\boldsymbol{x}$ does not shares the same index set of s-largest (in magnitude) elements with $\boldsymbol{x}^*$. On the other hand, as $\epsilon \to 0$, we have

$$\begin{aligned}\|\boldsymbol{x}-\boldsymbol{x}^*\| &= \sqrt{\sum_{i=1}^{n}(x_i - x^*_i)^2} \\ &= \sqrt{\left(-\frac{x^*_s}{2}\right)^2 + \left(\frac{x^*_s}{2}+\epsilon\right)^2} \to \frac{1}{\sqrt{2}}|x^*_{[s]}|. \end{aligned}$$

This means $\boldsymbol{x} \notin \mathcal{B}(\boldsymbol{x}^*, |x^*_{[s]}|/\sqrt{2})$ but it can approach the boundary of the ball as ϵ decreases to 0.

B. Proof of Remark 6

In the following, we show any stationary point $\boldsymbol{x}^*$ of (40) is also a local minimum by proving that the objective function does not decrease if we add any perturbation to $\boldsymbol{x}^*$ on $\mathcal{C}$. Let us consider any perturbation $\boldsymbol{\delta}$ such that $\boldsymbol{\delta} \in \mathcal{B}(\boldsymbol{0}, c_1(\boldsymbol{x}^*))$ and $\boldsymbol{x} = \boldsymbol{x}^* + \boldsymbol{\delta} \in \mathcal{C}$. Since $\boldsymbol{x} \in \mathcal{B}(\boldsymbol{x}^*, c_1(\boldsymbol{x}^*))$, using (7), we have $|x_{[1]}| \ge \ldots |x_{[s]}| > 0$. On the other hand, since $\boldsymbol{x}$ has no more than s non-zero entries, it must hold that $|x_{[s+1]}| = \ldots = |x_{[n]}| = 0$. Therefore, $\boldsymbol{x} = \boldsymbol{S}_{\boldsymbol{x}^*}\boldsymbol{S}^\top_{\boldsymbol{x}^*}\boldsymbol{x}$, which implies $\boldsymbol{\delta} = \boldsymbol{S}_{\boldsymbol{x}^*}\boldsymbol{S}^\top_{\boldsymbol{x}^*}\boldsymbol{\delta}$. Now we represent the change in the objective function as

$$\begin{aligned}&\frac{1}{2}\|\boldsymbol{A}(\boldsymbol{x}^*+\boldsymbol{\delta}) - \boldsymbol{b}\|^2 - \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}\|^2 \\ &= \frac{1}{2}\boldsymbol{\delta}^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{\delta} + \boldsymbol{\delta}^\top\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}) \\ &= \frac{1}{2}\boldsymbol{\delta}^\top\boldsymbol{S}_{\boldsymbol{x}^*}\boldsymbol{S}^\top_{\boldsymbol{x}^*}\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{S}_{\boldsymbol{x}^*}\boldsymbol{S}^\top_{\boldsymbol{x}^*}\boldsymbol{\delta} + \boldsymbol{\delta}^\top\boldsymbol{S}_{\boldsymbol{x}^*}\boldsymbol{S}^\top_{\boldsymbol{x}^*}\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}) \\ &= \frac{1}{2}\boldsymbol{\delta}^\top\boldsymbol{S}_{\boldsymbol{x}^*}(\boldsymbol{S}^\top_{\boldsymbol{x}^*}\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{S}_{\boldsymbol{x}^*})\boldsymbol{S}^\top_{\boldsymbol{x}^*}\boldsymbol{\delta} \ge 0, \end{aligned} \tag{8}$$

where the last equality uses the stationarity condition in (43). From (8), we conclude $\boldsymbol{x}^*$ is a local minimum of (40).

REFERENCES

[1] D. G. Luenberger, Y. Ye *et al.*, *Linear and nonlinear programming*. Springer, 1984, vol. 2.

[2] B. T. Polyak, *Introduction to optimization. Optimization software*. Inc., Publications Division, New York, 1987, vol. 1.

[3] D. P. Bertsekas, "Nonlinear programming," *J. Oper. Res. Soc.*, vol. 48, no. 3, pp. 334–334, 1997.

[4] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2003, vol. 87.

[5] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[6] A. Beck, *First-order methods in optimization*. SIAM, 2017.

[7] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Appl. Comput. Harmon. Anal.*, vol. 27, no. 3, pp. 265–274, 2009.

[8] C. Ma, K. Wang, Y. Chi, and Y. Chen, "Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution," in *Found. Comput. Math.*, 2020, pp. 451—-632.

[9] R. Khanna and A. Kyrillidis, "IHT dies hard: Provable accelerated iterative hard thresholding," in *Proc. Int. Conf. Artif. Intell. Stat.*, 2018, pp. 188–198.

[10] A. Beck and Y. Vaisbourd, "Globally solving the trust region subproblem using simple first-order methods," *SIAM J. Optim.*, vol. 28, no. 3, pp. 1951–1967, 2018.

[11] B. T. Polyak, "Gradient methods for minimizing functionals," *Zhurnal vychislitel'noi matematiki i matematicheskoi fiziki*, vol. 3, no. 4, pp. 643–653, 1963.

[12] ——, "Some methods of speeding up the convergence of iteration methods," *USSR Comput. Math. Math. Phys.*, vol. 4, no. 5, pp. 1–17, 1964.

[13] J. W. Daniel, "The conjugate gradient method for linear and nonlinear operator equations," *SIAM J. Numer. Anal.*, vol. 4, no. 1, pp. 10–26, 1967.

[14] L. V. Kantorovich, "Functional analysis and applied mathematics," *Uspekhi Matematicheskikh Nauk*, vol. 3, no. 6, pp. 89–185, 1948.

[15] D. G. Luenberger, "The gradient projection method along geodesics," *Manag. Sci.*, vol. 18, no. 11, pp. 620–631, 1972.

[16] A. Lichnewsky, "Minimisation des fonctionnelles définies sur une variété par la methode du gradiënt conjugué," Ph.D. dissertation, These de Doctorat d'Etat. Paris: Université de Paris-Sud, 1979.

[17] D. Gabay, "Minimizing a differentiable function over a differential manifold," *J. Optim. Theory Appl.*, vol. 37, no. 2, pp. 177–219, 1982.