THE NONCONVEX GEOMETRY OF LOW-RANK MATRIX OPTIMIZATIONS WITH GENERAL OBJECTIVE FUNCTIONS

Qiuwei Li and Gongguo Tang

Department of Electrical Engineering, Colorado School of Mines, Golden, CO USA

ABSTRACT

This work considers the minimization of a general convex function $f(X)$ over the cone of positive semi-definite matrices whose optimal solution $X^\star$ is of low-rank. Standard first-order convex solvers require performing an eigenvalue decomposition in each iteration, severely limiting their scalability. A natural nonconvex reformulation of the problem factors the variable X into the product of a rectangular matrix with fewer columns and its transpose. For a special class of matrix sensing and completion problems with quadratic objective functions, local search algorithms applied to the factored problem have been shown to be much more efficient and, in spite of being nonconvex, to converge to the global optimum. The purpose of this work is to extend this line of study to general convex objective functions $f(X)$ and investigate the geometry of the resulting factored formulations. Specifically, we prove that when $f(X)$ satisfies the restricted well-conditioned assumption, each critical point of the factored problem either corresponds to the optimal solution $X^\star$ or a strict saddle where the Hessian matrix has a strictly negative eigenvalue. Such a geometric structure of the factored formulation ensures that many local search algorithms can converge to the global optimum with random initializations.

Index Terms— Burer-Monteiro factorization, low-rank matrix optimization, nonconvex optimization, strict saddle property

1. INTRODUCTION

Consider a general semi-definite program (SDP) where a convex objective function $f(X)$ is minimized over the cone of positive semi-definite (PSD) matrices:

$$\underset{X\in\mathbb{R}^{n\times n}}{\text{minimize}}\, f(X) \text{ subject to } X \succeq 0. \tag{1}$$

For this problem, even fast first-order methods, such as the projected gradient descent algorithm [2], require performing an expensive eigenvalue decomposition in each iteration. These expensive operations form the major computational bottleneck of the algorithms and prevent them from scaling to scenarios with millions of variables, a typical situation in a diverse of applications, including quantum state tomography [3], user preferences prediction [4], and pairwise distances estimation in sensor localization [5].

When the SDP (1) admits a low-rank solution $X^\star$, in their pioneer work [6], Burer and Monteiro proposed to factorize the variable $X = UU^T$, where $U \in \mathbb{R}^{n\times r}$ with $r \ll n$, and solved a factored nonconvex problem

$$\underset{U\in\mathbb{R}^{n\times r}}{\text{minimize}}\, g(U), \text{ where } g(U) := f(UU^T). \tag{2}$$

There, they dealt with standard SDPs with a linear objective function and several linear constraints, and argued that when the factorization $X = UU^T$ is overparameterized, *i.e.*, $r > r^\star := \text{rank}(X^\star)$, any local minimum of (2) corresponds to the solution $X^\star$, provided some regularity conditions are satisfied. Unfortunately, these regularity conditions are generally hard to verify for specific SDPs arising in applications. Our work differs in that the convex objective function $f(X)$ is generally not linear and there are no additional linear constraints.

The past few years have seen renewed interest in the Burer-Monteiro factorization for solving low-rank matrix recovery inverse problems. With technical innovations in analyzing the nonconvex landscape of the factored objective function, several recent works have shown that with exact parameterization (*i.e.*, $r = r^\star$) the factored objective function $g(U)$ in has no spurious local minima or degenerate saddle points [7–12]. An important implication is that local search algorithms, such as gradient descent and its variants, are able to converge to the global optimum with even random initialization [13].

We generalize this line of work by assuming a general objective function $f(X)$ in the optimization (1). Viewing the factored problem (2) as a way to solve the convex optimization (1) to the global optimum, frees us from rederiving the statistical performances of the factored optimization (2). Instead, its performance inherits from that of the convex optimization (1), whose performance can be developed using a suite of powerful convex analysis techniques accumulated from several decades of research. As a specific example, the optimal sampling complexity [14] and minimax denoising rate [15] need not to be rederived once one knows the equivalence between the convex and the factored formulations.

Full version appears as [1]. This work was supported by NSF grant CCF-1464205. Email: {qiuli, gtang}@mines.edu.

2. MAIN THEOREM

Before presenting our main result, we provide several necessary definitions. We call a vector $\boldsymbol{x}$ a *critical point* of some differentiable function $f(\cdot)$ if the gradient $\nabla f(\boldsymbol{x}) = \mathbf{0}$. When $f(\cdot)$ is twice continuously differentiable, a critical point $\boldsymbol{x}$ is called a *strict saddle* or *riddable saddle* [16] if the Hessian has a strictly negative eigenvalue, i.e., $\lambda_{\min}(\nabla^2 f(\boldsymbol{x})) < 0$. A twice continuously differentiable function satisfies the *strict saddle property* if every critical point is either a local minimum or is a strict saddle [7].

Heuristically, the strict saddle property describes a geometric structure of the landscape: if a critical point is not a local minimum, then it is a strict saddle, which implies the Hessian matrix at this point has a strictly negative eigenvalue. Hence, we can continue to decrease the function value at this point along the negative-curvature direction.

Theorem 1 (Local convergence [13, 17, 18]). *The strict saddle property allows many local search algorithms to escape all the saddle points and converge to a local minimum.*

Our governing assumption on the objective function $f(X)$ is the $(2r, 4r)$-restricted well-conditioned assumption:

$$m \leq [\nabla^2 f(X)](D, D)/\|D\|_F^2 \leq M \text{ with } \frac{M}{n} \leq 1.5 \qquad (3)$$

for any D of $\operatorname{rank}(D) \leq 4r$ and any PSD matrix X with $\operatorname{rank}(X) \leq 2r$. Here, $[\nabla^2 f(X)](D, D)$ is the directional curvature along D, defined as $\sum_{i,j,l,k} \frac{\partial^2 f(X)}{\partial X_{ij} \partial X_{lk}} D_{ij} D_{lk}$. This restricted well-conditioned assumption (3) is standard in matrix inverse problem [19, 20]. We show that if the original objective function $f(X)$ is $(2r, 4r)$-restricted well conditioned, then each critical point of the factored objective function $g(U)$ either corresponds to the low-rank global solution of the original convex program or is a strict saddle where the Hessian $\nabla^2 g(U)$ has a strictly negative eigenvalue. This implies the factored objective function $g(U)$ satisfies the strict saddle property.

Theorem 2 (Global landscape). *Suppose the function $f(X)$ in (1) is twice continuously differentiable and restricted well-conditioned (3). Assume $X^\star$ is an optimal solution of the minimization (1) with $\operatorname{rank}(X^\star) = r^\star$. Set $r \geq r^\star$ in (2). Let U be any critical point of $g(U)$ satisfying $\nabla g(U) = \mathbf{0}$. Then U either corresponds to a square-root factor of $X^\star$,* i.e.,

$$X^\star = UU^T; \qquad (4)$$

or is a strict saddle of the factored problem (2):

$$\lambda_{\min}(\nabla^2 g(U)) \leq \begin{cases} -0.24 m\tau & \text{when } r > r^\star \\ -0.19 m\rho(X^\star) & \text{when } r = r^\star \\ -0.24 m\rho(X^\star) & \text{when } U = \mathbf{0} \end{cases} \qquad (5)$$

with $\tau := \min\{\rho(U)^2, \rho(X^\star)\}$ and $\rho(W)$ denoting the smallest nonzero singular value.

Remarks. First, the matrix D is the direction from the saddle point U to its closest globally optimal factor $U^\star R$ of the same size as U. Second, our result covers both over-parameterization where $r > r^\star$ and exact parameterization where $r = r^\star$. Third, this strict saddle property ensures that many iterative algorithms, for example, stochastic gradient descent [17], trust-region method [18], and gradient descent with sufficiently small stepsize [13], all converge to a square-root factor of $X^\star$, even with random initialization.

3. APPLICATIONS

Our main result only relies on the restricted well-conditioned property. Therefore, in addition to the traditional low-rank matrix recovery problems with a quadratic loss function, it is also applicable to a lot of other low-rank matrix optimization problems with possibly non-quadratic loss functions. We compiled the following list of applications that are covered by our theory.

Weighted PCA Problem. Formally, in the weighted-PCA problem, given a pointwisely-weighted observation of a PSD matrix X, *i.e.*, $Y = W \odot X$ where $\odot$ is the Hadamard product or its perturbed version with W being the sensing matrix, one aims to recover the principle component U by minimizing the nonconvex objective function $g(U) = \|Y - W \odot (UU^T)\|_F^2$. The weighted-PCA problem has no known analytic solution and it is shown to be NP-hard [21]. Fortunately, by defining $f(X) = \|Y - W \odot X\|_F^2$, we can compute its directional curvature as $[\nabla^2 f(X)](D, D) = \|W \odot D\|_F^2$. Hence, as long as the weights have a smaller dynamic range: $\frac{\max W_{ij}^2}{\min W_{ij}^2} \leq 1.5$, it is guaranteed to recover U through local search algorithms.

Symmetric Robust PCA. In the symmetric variant of robust PCA, the observed matrix $Y = X + S$ with S being sparse and X being PSD. Traditionally, we recover X by minimizing $||Y - X||_1 = \sum_{ij} |Y_{ij} - S_{ij}|$ subject to a PSD constraint. However, this formulation doesn't fit into our framework naively due to the non-smoothness of the ℓ_1 norm. An interesting bypass would be solving X by minimizing $\sum_{ij} h_a(Y_{ij} - S_{ij})$ where $h_a(.)$ is chosen to be a convex smooth approximation to the absolute value function. A possible choice is $h_a(x) = a \log((\exp(x/a) + \exp(-x/a))/2)$, which is shown to be strictly convex and smooth in [22, Lemma A.1].

1-Bit Matrix Recovery. Given quantized measurements: $y_j = \operatorname{bit}(A_j \bullet X^\star)$ where $\bullet$ denotes the inner product and $\operatorname{bit}(x)$ outputs 0 or 1 in a probabilistic manner, we attempt to recover $X^\star \in \mathbb{R}^{n \times m}$ by minimizing $f(X) = -\sum_j (y_j \log(\sigma(A_j \bullet X)) + (1 - y_j) \log(1 - \sigma(A_j \bullet X)))$, where $\sigma(x) = \frac{e^x}{1+e^x}$ is the logistic regression function [23]. Moreover, the Hessian quadratic form of $f(X)$ is $[\nabla^2 f(X)](D, D)$