Regularized Sample Average Approximation for High-Dimensional Stochastic Optimization Under Low-Rankness

Hongcheng Liu
Charles Hernandez
Hung Yi Lee
Department of Industrial and Systems Engineering
University of Florida
Gainesville, FL 32611, USA

Abstract
This paper concerns a high-dimensional stochastic programming problem minimizing the expected function of a matrix argument. To this problem, one of the most widely applied solution paradigms is the sample average approximation (SAA), which uses the average cost over sampled scenarios as a surrogate to approximate the exact function of expected cost. Traditional SAA theories require the sample size to grow rapidly when the problem dimensionality increases and the resulting demand of samples tends to become prohibitive. Indeed, the required sample size of SAA is quadratic in $p$; more specifically, for a problem of optimizing over a $p$-by-$p$ matrix, the sample complexity of the SAA is given by $\tilde{O}(1) \cdot \frac{p^2}{\epsilon^2} \cdot \text{polylog}(\frac{1}{\epsilon})$ to achieve an $\epsilon$-suboptimality gap, for some almost linear function $\text{polylog}(\cdot)$ and some quantity $\tilde{O}(1)$ independent of dimensionality $p$ and sample size $n$. In contrast, this paper considers a regularized SAA (RSAA) with a low-rankness-inducing penalty. We show that the sample complexity of RSAA can be substantially reduced to $\tilde{O}(1) \cdot \frac{p^3}{\epsilon^3} \cdot \text{polylog}(p, \frac{1}{\epsilon})$, almost linear in $p$. Therefore, RSAA can be more advantageous than SAA especially for larger scale and higher dimensional problems. Due to the close correspondence between stochastic programming and statistical learning, our results also indicate that high-dimensional low-rank matrix recovery is possible generally beyond a linear model, even if the common assumption of restricted strong convexity is completely absent.

Keywords: Stochastic optimization, MCP, folded concave penalty, sample average approximation, high dimensionality, sparsity, low-rankness
1. Introduction

As dimensionality inflates in modern applications of stochastic programming (SP) in order to generate more comprehensive and higher-granular decisions, the sample average approximation (SAA), which is traditionally a common solution paradigm for SP, becomes prohibitively demanding for sample availability. The current SAA theories as per [25], [22], [23], and [24] require that the number of samples should always be greater than the number of decision variables; for optimizing over a $p$-by-$p$ matrix, the sample size $n$ should grow at least quadratically in $p$. Such a sample size requirement tends to be undesirably costly. To reduce the sample complexity, in a precursor of this paper, we have studied a regularized SAA with sparsity-inducing penalty [13]. We have shown that the proposed approach yields a significant reduction of sample size requirement by exploiting sparse structures in the problem. This current paper then seeks to generalize the result therein to the settings where sparsity is replaced by a low-rankness assumption. We will show that a similar level of success can be achieved.

The particular problem of focus is stated as follows: Let $Z \in \mathcal{W}$, for some $\mathcal{W} \subseteq \mathbb{R}^q$ and $q > 0$, be a random vector. Consider a measurable, deterministic function $f : \mathcal{S}_p \times \mathcal{W} \to \mathbb{R}$ where $\mathcal{S}_p$ is the cone of $p$-by-$p$ symmetric and positive semidefinite matrices and $f(X, Z)$ is a cost function with respect to parameter $Z$ and a fixed matrix of decision variables $X$. Assume $p \geq 1$ hereafter. Then the problem of consideration is an SP problem given as

$$X^* \in \arg\min \{ F(X) : X \in \mathcal{S}_p \}.$$  \hspace{1cm} (1)

where $F(X) = \mathbb{E}[f(X, Z)]$ is well-defined and finite-valued for any given $X \in \mathcal{S}_p$. Assume, hereafter, that $\sigma_{\text{max}}(X^*) \leq R$ for some constant $R \geq 1$, where $\sigma_{\text{max}}(\cdot)$ denotes the spectral radius. With some abuse of terminology, we say that the dimensionality of this problem is $p$, since the unknown is a $p$-by-$p$ matrix. We refer to this optimization problem as the “true problem” and $X^*$ as the “true solution”, as they assume the exact knowledge of the underlying distribution and the admissibility of calculating the multi-dimensional integration involved in evaluating the expected value. We would like to remark that (2) subsumes the unconstrained problems since any symmetric matrix can be represented by the difference between two symmetric and positive semidefinite matrices. Furthermore, also subsumed by (2) are problems with non-symmetric and non-square matrices $X$, since they can be transformed into symmetric matrices by the self-adjoint dilation with $X = \begin{bmatrix} 0 & X \\ X^\top & 0 \end{bmatrix}$ with some all-zero matrices $0$’s, with proper dimensions.

Hereafter, let $Z^n = (Z_1, ..., Z_n)$ be a sequence of $n$-many i.i.d. random samples of $Z$. To solve Problem (2), one of the most popular solution schemes, as mentioned above, is to invoke the following SAA formulation as a surrogate:

$$X^{\text{SAA}} \in \arg\min \left\{ \mathcal{F}_n(X, Z^n) := \frac{1}{n} \sum_{i=1}^{n} f(X, Z_i) : X \in \mathcal{S}_p \right\}.$$  \hspace{1cm} (2)
According to the seminal results by [25], $X^{SAA}$ well approximates $X^*$ in the sense that
\[
F(X^{SAA}) - F(X^*) \leq \tilde{O}(1) \cdot \sqrt{\frac{p^2 \cdot \ln n}{n}},
\]
with high probability, where $\tilde{O}(\cdot)$ is some quantity that is independent of $p$ and $n$. Thus, to ensure the same suboptimality gap, it stipulates that the sample size, $n$, must grow quadratically if $p$ increases. For an SP problem where $X^*$ is sparse and $f$ is twice-differentiable almost surely, we have shown in [13] that (7) can be radically sharpened into:
\[
F(X^{RSAA}) - F(X^*) \leq \tilde{O}(1) \cdot \sqrt{\frac{\ln(np)}{n^{1/4}}},
\]
with high probability, where $X^{RSAA}$ is an SAA scheme with sparsity-inducing regularization.

Similar (and potentially more general) results than the above have been reported by [10] and [11] in the context of high-dimensional statistical and machine learning under a sparsity assumption and/or its limited variations.

In contrast, this paper provides a substantial generalization to [13; 10; 11] by weakening the sparsity and twice-differentiability assumptions simultaneously to low-rankness and continuous differentiability. Particularly, our low-rankness assumption is as below:

**Assumption 1** The rank $\text{rk}(\cdot)$ of $X^*$ in Eq. (2) satisfies $s := \text{rk}(X^*) \ll p$ for some $s \geq 1$.

The low-rankness is more general than the sparsity assumption of a vector, since any vector $x$ can be represented by a diagonal matrix, $\text{diag}(x)$, whose diagonal entries equal to $x$. Then, sparsity of $x$ implies that $\text{diag}(x)$ is of low rank. In addition, we assume Lipschitz continuity of the partial derivative of $f$ w.r.t. the eigenvalues of the input matrix, as we will discuss in more detail subsequently.

For this more general problem, our solution paradigm modifies the SAA into the following regularized SAA (RSAA):
\[
X^{RSAA} \in \arg \min_{X \in S_p} \left\{ F_{n,\lambda}(X, Z^n) := F_n(X, Z^n) + \sum_{j=1}^{p} P_\lambda(\sigma_j(X)) \right\},
\]
where $\sigma_j(X)$ stands for the $j$th eigenvalue of $X$ and $P_\lambda$ is a penalty function in the form of the minimax concave penalty (MCP) [26] given as $P_\lambda(x) = \int_0^x \frac{|a \lambda - t|_+}{a} dt$, for some user-specific tuning parameters $a$, $\lambda > 0$. The MCP is a mainstream special form of the folded concave penalty (FCP) first proposed by [7].

Under the above settings, the RSAA formulation is nonconvex and its global solutions are elusive. To ensure computability, this paper considers stationary points that satisfies, what we call, the significant subspace second-order necessary conditions ($S^3$ONC), given as in Definition 2 in the subsequent. The $S^3$ONC herein is an extension to a similar notion presented by [12; 13] and is a special case than the canonical second-order KKT conditions. Hence, any second-order (local optimization) algorithm that computes a second-order KKT solution ensures the $S^3$ONC. The resulting computational effort of an $S^3$ONC solution (a solution that satisfies the $S^3$ONC) is likely tractable.
Let $X^{\ell_1}$ be defined as

$$X^{\ell_1} \in \arg \min_{X \in S_p} F_n(X, Z^n_1) + \lambda \|X\|_s,$$  

(6)

with $\| \cdot \|_s$ denoting the nuclear norm. We show that, under a few standard assumptions in addition to Assumptions 1 for any $S^3$-ONC solution to the RSAA $X^{RSAA}$ which satisfies $F_{n,\lambda}(X^{RSAA}) \leq F_{n,\lambda}(X^{\ell_1})$ a.s., it holds that

$$F(X^{RSAA}) - F(X^*) \leq \tilde{O}(1) \cdot \left( \frac{s \cdot p^{2/3}}{n^{2/3}} + \frac{s \cdot p^{1/3}}{n^{1/3}} \right) \cdot \ln(np),$$  

(7)

with overwhelming probability. The above results is then the promised, almost linear, sample complexity; $n$ should only increase linearly in $p$ to compensate the growth in dimensionality. This indicates that the RSAA would be much more advantageous than the SAA especially for problems with higher dimensions. To compute the desired solution $X^{RSAA}$, one may invoke an $S^3$-ONC-guaranteeing algorithm initialized at $X^{\ell_1}$. Meanwhile, the initial solution, $X^{\ell_1}$, is often polynomial-time computable when $f(\cdot, w)$ is convex for almost every $w \in W$ (although the convexity of $f(\cdot, w)$ is not necessary to prove the almost linear sample complexity).

To our knowledge, our paper presents the first SAA variant that ensures an almost linear complexity under low-rankness. Even though similar results have been achieved by [16], [20] and [6] in the context of high-dimensional low-rank matrix estimation, all those results assume the presence of restricted strong convexity (RSC) or its variations. While the RSC is deemed generally plausible for statistical and/or machine learning, such type of assumptions are often not satisfied by stochastic programming. Furthermore, due to the correspondence between the SAA and matrix estimation problems, our results may also imply that high-dimensional matrix estimation is generally possible under low-rankness; even if the conditions such as the RSC or alike are completely absent, the FCP-based regularization may still ensure a sound generalization error as measured by the excess risk (ER), which coincides in formulation with the suboptimality gap in minimizing the SP. In addition, our results do not assume a linear or generalized linear model in data generation. Even though a few other likely more important error bounds are unavailable herein but are presented by [16], [20] and [6] (most of whom more on linear or generalized linear models under RSC or alike), we believe that the ER is still an important out-of-sample performance measure commonly employed by, e.g., [1], [9], and [5].

The rest of the paper is organized as follows: Section 2 presents our assumptions and main results. Section 3 presents the general road map for our proof and main schemes employed. Section 4 then concludes our paper. All technical proofs are presented in the appendix.

### 1.1 Notations

Throughout this paper, we denote by $\| \cdot \|$ the 2-norm of a vector, by $\sigma_{\max}(\cdot)$ the spectral norm, by $\| \cdot \|_s$ the nuclear norm, and by $\| \cdot \|_p$ the $p$-norm (with $1 \leq p \leq \infty$). Let $\sigma_j(X)$ be the $j$th singular value of matrix $X$. Denote by $\| \cdot \|_F$ the Frobenius norm.
2. Sample complexity of the regularized SAA under low-rankness

This section presents our main results in Subsection 2.3 after we introduce our assumptions in Subsection 2.1 as well as the definition of the $S^3$ONC in Subsection 2.2.

2.1 Assumptions.

In addition to the low-rankness structure as in Assumption 1, we will make the following additional assumptions about continuous differentiability (Assumption 2), the tail of the underlying distribution (Assumption 3), and a Lipschitz-like continuity (Assumption 4).

Assumption 2 Let $\mathcal{U}_L \geq 1$. Assume that

$$\left| \frac{\partial f(X, z)}{\partial \sigma_j(X)} \bigg|_{X=X_1} - \frac{\partial f(X, z)}{\partial \sigma_j(X)} \bigg|_{X=X_2} \right| < \mathcal{U}_L \cdot |\sigma_j(X_1) - \sigma_j(X_2)|$$

for every $j = 1, \ldots, p$, all $X_1, X_2 \in S_p$, and almost every $z \in \mathcal{W}$.

The above assumption is standard and easily verifiable. It essentially ensure the continuity of the partial derivatives of the function $f$ w.r.t. its singular values.

Assumption 3 The family of random variables, $f(X, Z_i) - \mathbb{E}[f(X, Z_i)]$, $i = 1, \ldots, n$, are independent and follow sub-exponential distributions; that is

$$\|f(X, Z_i) - \mathbb{E}[f(X, Z_i)]\|_{\psi_1} \leq K,$$

for some $K \geq 1$ for all $X \in S_p : \sigma_{\max}(X) \leq R$, where $\| \cdot \|_{\psi_1}$ is the sub-exponential norm.

Invoking the well-known Bernstein-type inequality, one has that, for all $X \in S_p$, it holds that

$$\mathbb{P}\left( \left| \sum_{i=1}^n a_i \{f(X, Z_i) - \mathbb{E}[f(X, Z_i)]\} \right| > K(\|a\|\sqrt{t} + \|a\|_\infty t) \right) \leq 2 \exp\left(-ct\right),$$

$$\forall t \geq 0, a = (a_i) \in \mathbb{R}^n,$$  \hspace{1cm} (9)

for some absolute constant $c \in (0, \frac{1}{2}]$. [See also 21].

Assumption 4 For some measurable and deterministic function $C : \mathcal{W} \to \mathbb{R}$ with $\mathbb{E}[|C(Z)|] \leq C_\mu$, for some $C_\mu \geq 1$, the random variable $C(Z)$ satisfies that $\|C(Z) - \mathbb{E}[C(Z)]\|_{\psi_1} \leq K_C$ for some $K_C \geq 1$. Furthermore, $|f(X_1, z) - f(X_2, z)| \leq C(z)\|X_1 - X_2\|$ for all $X_1, X_2 \in S_p$, and almost every $z \in \mathcal{W}$.

Remark 1 Assumption 2 is easily verifiable and applies to a flexible set of SP problems. Assumptions 3 and 4 are standard, and, by a close examination, it is essentially equivalent to the assumptions made by [22] in the analysis of the traditional SAA.
2.2 The significant subspace second-order necessary conditions

Our sample complexity results concern critical points that satisfy the \( S^3 \)ONC as per the following definition, where we notice that \( P_\lambda(t) \) is twice differentiable for all \( t \in (0, a\lambda) \).

**Definition 2** For given \( Z^n_1 \in W^n \), a vector \( \hat{X} \in S_p \) is said to satisfy the \( S^3 \)ONC (denoted by \( S^3 \)ONC\((Z^n_1) \)) of the problem (5) if both of the following sets of conditions are satisfied:

a. The first-order KKT condition is satisfied at \( X^{RSAA} \); that is,

\[
\nabla F_{n,\lambda}(X^{RSAA}, Z^n_1) = 0,
\]

where \( \nabla F_{n,\lambda}(X^{RSAA}, Z^n_1) \) is the gradient of \( F_{n,\lambda}(X^{RSAA}, Z^n_1) \) at \( X^{RSAA} \).

b. The following inequality holds at \( X^{RSAA} \) for all \( j = 1, \ldots, p \):

\[
U_L + \left[ \frac{\partial^2 P_\lambda(\sigma_j(X))}{\partial \sigma_j(X)^2} \right]_{X=X^{RSAA}} \geq 0, \quad \text{if} \ \sigma_j(X^{RSAA}) \in (0, a\lambda),
\]

where \( U_L \) is as defined in (8) for Assumption 2.

As mentioned, the above \( S^3 \)ONC is verifiably a weaker condition than the canonical second-order KKT conditions. Therefore, any local optimization algorithm that guarantees the second-order KKT conditions will necessarily ensure the \( S^3 \)ONC.

2.3 Main results

Introduce a few short-hand notations: Denote \( \tilde{\Delta} := \ln (18R \cdot (K_C + C_\mu)) \), let \( X^{\ell} \) be defined as in (6), and specify the parameters, \( \lambda := \sqrt{\frac{8K(2p+1)^{2/3}}{c-a-n^{2/3}}} \left[ \ln(n^{1/3}3p) + \tilde{\Delta} \right] \), for the same \( c \) in (9), and \( a^{-1} = 2U_L \) (and thus \( a < U_L^{-1} \)). We are now ready to present our claimed results.

**Theorem 3** Suppose that Assumptions 1 through 4 hold. Let \( X^{RSAA} \in S_p : \sigma_{\text{max}}(X^{RSAA}) \leq R \) satisfy the \( S^3 \)ONC\((Z^n_1) \) to (5) almost surely. The following statements hold:

(i) For any \( \Gamma \geq 0 \) and some universal constants \( \tilde{c}, C_1 > 0 \), if

\[
n > C_1 \cdot \left[ \left( \frac{\Gamma}{K} \right)^{3/2} + 1 \right] \cdot p + C_1 \cdot s \cdot \left( \ln(n^{1/3}3p) + \tilde{\Delta} \right),
\]

and \( F_{n,\lambda}(X^{RSAA}, Z^n_1) \leq F_{n,\lambda}(X^*, Z^n_1) + \Gamma \) almost surely, then the excess risk is bounded by

\[
F(X^{RSAA}) - F(X^*) \leq \sqrt{\frac{Kp^{1/3}\Gamma}{n^{1/3}}} + \Gamma
\]

\[
+ C_1 K \cdot \left( sp^{2/3} \cdot \left( \ln(n^{1/3}3p) + \tilde{\Delta} \right) \right) + \sqrt{\frac{s \cdot \left( \ln(n^{1/3}3p) + \tilde{\Delta} \right)}{n}} + \frac{p^{1/3}}{n^{1/3}},
\]

6
High-Dimensional Stochastic Optimization Under Low-Rankness

with probability at least 

\[ 1 - 2(p + 1) \exp(-\tilde{c}n) - 6 \exp\left(-2c(2p + 1)^{2/3}n^{1/3}\right). \]

(ii) For some universal constant \( \tilde{c} \), \( C > 0 \), if

\[ n > C \cdot p \cdot U_L \cdot [\ln(n^{1/3}p) + \tilde{\Delta}] \cdot s^{3/2}R^2, \]

and

\[ F_{n,\lambda}(X_{RSAA}^n, Z^n_1) \leq F_{n,\lambda}(X_{\ell_1}^n, Z^n_1) \]

almost surely, where \( X_{\ell_1} \) is as defined in (6), then the excess risk is bounded by

\[
F(X_{RSAA}) - F(X^*) \leq C \cdot s \cdot K \cdot \left[ \frac{p^{2/3} \left( \ln(n^{1/3}p) + \tilde{\Delta} \right)}{n^{1/3}} + \frac{p^{1/3}R \cdot U_L^{1/2} \sqrt{\ln(n^{1/3}p) + \tilde{\Delta}}}{n^{1/3}} \right],
\]

with probability at least 

\[ 1 - 2(p + 1) \exp(-\tilde{c}n) - 6 \exp\left(-2c(2p + 1)^{2/3}n^{1/3}\right). \]

Proof See proof in Section A.1.

Remark 4 We would like to make a few remarks on the above result:

1. Part (i) of the above theorem ensures that all \( S^0 \) ONC solution yields a bounded suboptimality in minimizing the true problem (2). Such suboptimality gap is a function of the solution quality in solving the RSAA formulation.

2. Part (ii) considers the particular sublevel set that has a better objective value (in terms of RSAA formulation) than \( X_{\ell_1} \). In such a case, the suboptimality in minimizing the true problem (2) explicitly vanishes as sample size \( n \) increases.

3. In Part (ii), \( X_{\ell_1} \) is an initial solution often tractably computable if \( f(\cdot, z) \) is convex for almost every \( z \in W \). In such a case, (5) is a convex problem w.p.1. However, our theorem is not contingent on the convexity of \( f(\cdot, z) \).

Remark 5 This theorem is consistent with (7), which ensures the claimed sample complexity almost linear in \( p \). Indeed, for achieving an accuracy of \( \epsilon \), the above bounds indicates a sample complexity \( \tilde{O}(1) \cdot \frac{p}{\epsilon^2} \cdot \text{polylog}(\frac{1}{\epsilon}) \), which is almost linear in \( p \), for some quantities \( \tilde{O}(1) \) that is independent of \( n, \epsilon, \) and \( p \). We note that the dependence of sample size \( n \) on rank \( s \) of the true solution \( X^* \) is cubic, which means that the proposed approach is more powerful when the true solution \( X^* \) is of very low rank. The deterioration may be fast when \( s \) increases. Nonetheless, we believe it possible to significantly reduce the order on \( s \) if any further information below is given: (i) If the \( F_n \) or \( F \) satisfies strong convexity or its certain relaxed forms, dependence on \( s \) is likely reducible, as it has been successful for [14] in sparse stochastic optimization. (ii) If the value of \( s \) can be coarsely predicted in the sense that \( O(1) \cdot s \) for some universal constant \( O(1) \) is given, then one may also properly modify the value of \( \lambda \) to decrease the dependence on \( s \); this has also been achieved in the sparse...
stochastic optimization case by [13]. We will consider those two relatively special cases in future study. Beyond those two scenarios, to our knowledge, no existing result can allow for a sample complexity almost linear in \( p \).

**Remark 6** There are strong correspondence between the SP and statistical learning as formerly noted by. Indeed, without changing any assumptions, the SAA formulation (2) is essentially equivalent to high-dimensional low-rank matrix M-estimation problem and the suboptimality gap \( F(X_{\text{RSAA}}) - F(X^*) \) is the same as the excess risk as discussed by [1], [9], and [5]. The result in Theorem 3 then indicates that M-estimation with high dimensions is generally possible under a low-rankness assumption. In particular, since our analysis does not assume any form of RSC, we believe that our results then provides perhaps the first out-of-sample performance guarantee for high-dimensional low-rank estimation beyond RSC.

**Remark 7** We would like to remark again that, to obtain the desired results, the incurred computational ramification can be reasonably small. This is because \( X^{\text{RSAA}} \) is only a stationary point that satisfies (15). First, the stationarity can be ensure by invoking local optimization algorithms. Second, the stipulated inequality in (15) can be ensured by initializing the local algorithm with \( X^{\ell_1} \). Such an initializer often can be generated within polynomial time when \( f(\cdot, w) \) is convex for almost every \( w \in \mathcal{W} \), although the convexity of \( f(\cdot, w) \) is not necessary for proving the claimed almost linear sample complexity.

### 3. Proof Overview and Techniques

#### 3.1 General ideas

The general idea is straightforward and focus on attacking one primary question: how to demonstrate the low-rankness of a stationary point to the RSAA formulation, without the assumption of RSC or alike. If this question is answered, then the desired results can be almost evident by some \( \epsilon \)-net-based analysis utilizing Lemma 15 (which is almost the same as Lemma 3.1 of [3]).

To that end, we utilize a unique property of the MCP function, which ensure that the stationary points that satisfy the \( S^3 \)ONC solution \( X^{\text{RSAA}} \) must obey a thresholding rule: for all the singular values, they must be either 0 or greater than \( a\lambda \). This means that for each nonzero singular value in the \( S^3 \)ONC solution \( X^{\text{RSAA}} \), an additional penalty of value \( a^2 \) is added to the objective function of the RSAA, and, therefore, the total penalty incurred by the low-rankness-inducing regularization is \( \sum_{j=1}^{p} P_\lambda(\sigma_j(X^{\text{RSAA}})) = \text{rk}(X^{\text{RSAA}}) \cdot \frac{a^2}{2} \). Now, consider those stationary points whose suboptimality gaps (in minimizing the RSAA) are smaller than a user-specific quantity \( \Gamma \), and therefore, \( F_{n,\lambda}(X^{\text{RSAA}}, Z^n_1) = F_n(X^{\text{RSAA}}, Z^n_1) + \text{rk}(X^{\text{RSAA}}) \cdot \frac{a^2}{2} \leq F_{n,\lambda}(X^*, Z^n_1) + \Gamma \). The rank of such \( X^{\text{RSAA}} \) rank must be bounded from above by a function of \( \Gamma \). Such a function can be explicated via a peeling technique discussed by [19]. Some relative details are provided below.

#### 3.2 Proof Roadmap

The proof of Theorem 3 is motivated by [11] but contain substantial generalizations from a sparse SP problem into a low-rank SP problem. The following are general explanations on
the key steps, where $\tilde{O}(1)$’s denote (potentially different) quantities that are independent of $p$ and $n$:

**Step 1.** The thresholding rule of the MCP. Under the assumption that $U_L < a^{-1}$, in Proposition 8, we show that, for any $S^3ONC$ solution, a thresholding rule of $\sigma_j(X)$, for all $j$, is that $\sigma_j(X) \neq 0 \implies \sigma_j(X) \geq a\lambda$, where $a$ and $\lambda$ are the tuning parameters of the MCP function, $P_\lambda$. This can be demonstrated by observing that the definition of the $S^3ONC$, which is $U_L - P'_\lambda(\sigma_j(X)) = U_L - \frac{1}{\lambda} \geq 0$ if $\sigma_j(X) \in (0, a\lambda)$, contradicts with the assumption that $U_L < a^{-1}$. Therefore, it holds that $\sigma_j(X) \geq a\lambda$, unless $\sigma_j(X) = 0$.

**Step 2.** $\epsilon$-net argument for low-rank subspaces. We apply the well-known $\epsilon$-net argument to show a point-wise error bound for $|F_n,\lambda(X, Z^n) - F(X)| \leq \epsilon$ for all $X \in S_p : \sigma_{\max}(X) \leq R$ in all rank-$\tilde{p}$ subspaces, whose elements have rank no greater than a given $\tilde{p}$. To that end, first observe that, for any rank-$\tilde{p}$ subspace, the standard $\epsilon$-net argument results in a covering number of $\tilde{O}(1)\left(\sqrt{\tilde{p}} \cdot \frac{\tilde{O}(1)}{\epsilon}\right)^{(2p+1)\tilde{p}}$. Second, since there can be $(\frac{p}{\tilde{p}})$-many rank-$\tilde{p}$ subspaces, the total covering number for all possible rank-$\tilde{p}$ subspaces is

$$\left(\frac{p}{\tilde{p}}\right) \cdot \left(\sqrt{\tilde{p}} \cdot \frac{\tilde{O}(1)}{\epsilon}\right)^{\tilde{p}} \leq \left(\tilde{O}(1) \cdot \frac{p}{\tilde{p}}\right)^{(2p+1)\tilde{p}}.$$  

Combining this covering number, the Bernstein-like inequality, and Lipschitz-like inequality in [1], we have that, for any $t \geq 0$,

$$|F_n(X, Z^n) - F(X)| > \tilde{O}(1) \cdot \frac{t}{n} + \tilde{O}(1) \cdot \sqrt{\frac{t}{n}} + \epsilon, \quad \forall X \in S_p : \sigma_{\max}(X) \leq R : \operatorname{rk}(X) \leq \tilde{p}, \quad (17)$$

with probability at most $\left(\frac{\tilde{O}(1) \cdot \tilde{p}}{\epsilon}\right)^{(2p+1)\tilde{p}} \exp(-ct) + \exp(-\tilde{O}(1) \cdot n)$ for some universal constant $c \in (0, 1/2]$. We may choose to let $t = 2\tilde{p}(2p + 1) \ln \left(\frac{\tilde{O}(1) \cdot \tilde{p}}{\epsilon}\right)$, as well as $\epsilon = n^{-\frac{1}{4}}$, then, observe that the probability the fact (we will call it Observation $(\star)$, to be useful later in Step 4) that the first term in the probability is vanishing exponentially fast to zero as $\tilde{p}$ increases and the second term is independent of $\tilde{p}$.

**Step 3.** An implication of Step 2. Let $X^{RSAA}$ be an $S^3$ONC solution to the RSAA formulation in [3]. Assume that $X^{RSAA}$ is within the $\Gamma$-sublevel set for some $\Gamma \geq 0$. Then, (cf. Assumption [1]) it is straightforward to obtain from the fact that $0 \leq P_\lambda(\cdot) \leq \frac{a\lambda^2}{2}$ and the results of Step 1 (i.e., $\sigma_j(X) \geq a\lambda$, unless $\sigma_j(X) = 0$),

$$F_n(X^{RSAA}, Z^n) + \operatorname{rk}(X^{RSAA}) \cdot \frac{a\lambda^2}{2} \leq F_n(X^*, Z^n) + \frac{a\lambda^2 \cdot s}{2} + \Gamma.$$  

(18)

If $\operatorname{rk}(X^{RSAA}) \leq \tilde{p}$, the result from Step 2 can be invoked to bound the differences, $F(X^{RSAA}) - F_n(X^{RSAA}, Z^n)$ and $F_n(X^*, Z^n) - F(X^*)$, to be smaller than a desired level. In particular, as we choose to let $t = 2\tilde{p}(2p + 1) \ln \left(\frac{\tilde{O}(1) \cdot \tilde{p}}{\epsilon}\right)$, as well as $\epsilon = n^{-\frac{1}{4}}$, in (17) and
\[ \lambda = \tilde{O}(1) \cdot \frac{p^{1/3} \sqrt{\ln(np)}}{n^{1/3}} \] in (18). After some algebraic simplification, we obtain that

\[
\begin{align*}
P(X_{RSAA}^\ast) - P(X^\ast) &\leq - \frac{a \lambda^2}{2} \text{rk}(X_{RSAA}) + \tilde{O}(1) \cdot \frac{s p^{2/3} \ln(pn)}{n^{2/3}} + \tilde{O}(1) \cdot \sqrt{\frac{p}{n} \ln(pn)} + \frac{p^{1/3}}{n^{1/3}} + \Gamma \\
&\leq \tilde{O}(1) \cdot \frac{\tilde{p} \cdot p \ln(pn)}{n} + \tilde{O}(1) \cdot \sqrt{\frac{p \cdot p}{n} \ln(pn)} + \frac{1}{n^{1/3}} + \Gamma
\end{align*}
\]

with probability at least \( 1 - \exp \left( -\tilde{O}(1) \cdot \tilde{p} \cdot p \cdot \ln(n p) \right) \). Recalling that \( \tilde{p} \) is an upper bound on the rank of \( X_{RSAA} \), the above result in (20) is now close to the desired “almost linear” sample complexity results if \( \tilde{p} \) much smaller than \( p \) is small. As it turns out, it is indeed the case. As is demonstrated in Theorem 3 we can show that

\[ \text{rk}(X_{RSAA}) \leq \tilde{O}(1) \cdot \left( s + \frac{n^{1/3}}{p^{1/3}} + \frac{n^{1/3}}{p^{1/3}} \cdot \Gamma \right) \]

which is to be explained subsequently.

**Step 4. Upper bound on \( \text{rk}(X_{RSAA}) \).** From Step 3, we observe that the desired result in Theorem 3 can be shown by proving that

\[ \text{rk}(X_{RSAA}) \leq \tilde{O}(1) \cdot \left( s + \frac{n^{1/3}}{p^{1/3}} + \frac{n^{1/3}}{p^{1/3}} \cdot \Gamma \right). \]  

To that end, we may invoke a scheme motivated by the peeling technique discussed by [19]. We will show in Proposition 10 that, for some integer \( \tilde{p}_u := \tilde{O}(1) \cdot \left( s + \frac{n^{1/3}}{p^{1/3}} + \frac{n^{1/3}}{p^{1/3}} \cdot \Gamma \right) \), it holds that, for all \( \tilde{p} \geq \tilde{p}_u \), the inequality in (19) can never be satisfied; this is because the first (negative) term therein would have too large a magnitude and render the whole composite on the right-hand-side of (19) a negative quantity, which implies \( P(X_{RSAA}) - P(X^\ast) < 0 \) and contradicts with the fact that \( X^\ast \) minimizes \( P \) by definition. Since \( \{19\} \) holds \( \supseteq \{\text{rk}(X) = \tilde{p}\} \cap \{\text{The complement to (17) holds with given } \tilde{p}\} \), it then implies that, for all \( \tilde{p} \geq \tilde{p}_u \),

\[ 0 = P \left[ \{\text{rk}(X_{RSAA}) = \tilde{p}\} \cap \{\text{The complement to (17) holds with given } \tilde{p}\} \right]. \]

As an immediate result, we may invoke union bound and De Morgan’s law to obtain that \( \text{P}[\text{rk}(X) = \tilde{p}] \leq \text{P}[\{17\} \text{ holds with given } \tilde{p}] \) for all \( \tilde{p} \geq \tilde{p}_u \). Therefore, invoking union bound and De Morgan’s law again, \( \text{P}[\text{rk}(X) \leq \tilde{p}_u - 1] \geq 1 - \sum_{\tilde{p} = \tilde{p}_u}^p \text{P}[\text{rk}(X) = \tilde{p}] \geq 1 - \sum_{\tilde{p} = \tilde{p}_u}^p \text{P}[\{17\} \text{ holds with given } \tilde{p}] \). By our choice of parameters for \( t \) and \( \epsilon \) as in Step 2, the Observation (*) (which is defined in Step 2) leads to a simplification of the probability bound by noting \( \sum_{\tilde{p} = \tilde{p}_u}^p \text{P}[\{17\} \text{ holds with given } \tilde{p}] \) involves the sum of a geometric sequence plus a term vanishing exponentially in \( n \). Combining the results from Step 4 with Step 3, we can then show Part (i) of Theorem 3 after some algebraic simplification.
Step 5. To show Part (ii) of Theorem 5. The second part of Theorem 5 can be shown by noticing that \( X^{\ell_1} \) yields a suboptimality gap of \( \tilde{O}(1) \cdot \frac{sp^{1/3}\sqrt{\ln(np)}}{n^{1/3}} \). When we choose \( \lambda = \tilde{O}(1) \cdot \frac{p^{1/3}\sqrt{\ln(np)}}{n^{1/3}} \) in (3) (which share the same \( \lambda \) value as in (5)).

4. Conclusions

This paper proposes a low-rankness-exploiting regularization SAA variant, referred to as the RSAA, to solve high-dimensional SP problems of minimizing an expected function over a \( p \)-by-\( p \) matrix argument. We prove that certain stationary points ensure an almost linear sample complexity: the RSAA only requires a sample size almost linear in \( p \) to achieve sound optimization quality, while, in contrast, the required sample size for the traditional SAA is quadratic in \( p \). Such sample complexity can be obtained at certain stationary points without incurring a significant computational effort, especially when the cost function \( f(\cdot, z) \) is convex for almost every \( z \in W \). Our RSAA theory also implies that, under the low-rankness assumption, high-dimensional matrix estimation is generally possible beyond linear and generalized linear models even if \( p \), the size of the matrix to be estimated, is large and the RSC is absent. Future research will focus on generalizing our paradigm to problems with general linear and nonlinear constraints. Furthermore, we will investigate the (non-)tightness of our bound on sample complexity.

Appendix A. Technical proofs

A.1 Proof of Theorem 3

We follow the same set of notations in Proposition 11 in defining \( \hat{p}_u \), \( \epsilon \), and \( \Delta_1(\epsilon) := \ln \left( \frac{18pR(KC+C\mu)}{\epsilon} \right) \). Furthermore, we will let \( \epsilon := \frac{1}{n^{1/3}} \) and \( \tilde{\Delta} := \ln (18 \cdot R \cdot (KC + C\mu)) \).

Then \( \Delta_1(\epsilon) = \ln \left( \frac{18(KC+C\mu)pR}{\epsilon} \right) = \ln(n^{1/3}/p) + \tilde{\Delta} > 0 \) and \( \lambda = \sqrt{\frac{8K(2p+1)^2/3 \cdot \Delta_1(\epsilon)}{c \cdot a \cdot n^{2/3}}} \) for the same \( c \) in (3).

To show the desired results, it suffices to simplify the results in Proposition 11. We will first derive an explicit form for \( \hat{p}_u \). To that end, we let \( P_X := \hat{p}_u \) and \( T_1 := 2P_X(a\lambda) - \frac{8K(2p+1)^2}{cn} \cdot \Delta_1(\epsilon) \). Then we solve the following inequality, which is equivalent to (11) of Proposition 11 for a feasible \( P_X \),

\[
\frac{T_1}{2} \cdot P_X - \frac{2K}{\sqrt{n}} \sqrt{\frac{2P_X \cdot (2p+1)\Delta_1(\epsilon)}{c}} > \Gamma + 2\epsilon + sP_X(a\lambda),
\]

for the same \( c \in (0, 0.5) \) in (3). Solving the above inequality in terms of \( P_X \), we have \( \sqrt{P_X} > \frac{2K}{\Gamma_1 \sqrt{n}} \sqrt{\frac{2(2p+1)\Delta_1(\epsilon)}{c}} + \frac{2(2K)^2(2p+1)\Delta_1(\epsilon)}{cT_1} + 2T_1[\Gamma + 2\epsilon + sP_X(a\lambda)] \). To find a feasible \( P_X \), we may as well let \( P_X > \frac{32K^2(2p+1)\Delta_1(\epsilon)}{cT_1} + 8T_1^{-1}[\Gamma + 2\epsilon + sP_X(a\lambda)] \). For \( \lambda = \sqrt{\frac{8K(2p+1)^2/3}{c \cdot a \cdot n^{2/3}}} \cdot \Delta_1(\epsilon) + \tilde{\Delta} \) with \( \tilde{\Delta} := \ln (18 \cdot R \cdot (KC + C\mu)) \), we have \( P_X(a\lambda) = \frac{a^2}{2} = \frac{4K(2p+1)^2/3}{c \cdot a \cdot n^{2/3}} \cdot \Delta_1(\epsilon) \). Furthermore, \( 2P_X(a\lambda) = \frac{8K(2p+1)^2/3 \cdot \Delta_1(\epsilon)}{c \cdot a \cdot n^{2/3}} > \)
\[
\frac{4K\Delta_1(\epsilon)(2p+1)^{2/3}}{cn^{2/3}} + \frac{8K^2(2p+1)}{nc} \Delta_1(\epsilon) \quad \text{as per our assumption (i.e., (12) implies that } n^{1/3} > 2) .
\]

Therefore, \( T_1 = 2P_\lambda(a\lambda) - \frac{8K^2(2p+1)}{nc} \Delta_1(\epsilon) > \frac{4K\Delta_1(\epsilon)(2p+1)^{2/3}}{cn^{2/3}} \). Hence, to satisfy (22), it suffices to let \( P_X \) be any integer that satisfies
\[
P_X \geq \frac{2cn^{1/3}}{\Delta_1(n^{-\frac{2}{3}})} \cdot \frac{K}{\Delta_1(n^{-\frac{2}{3}})} \cdot (2p+1)^{2/3} \cdot \left( \Gamma + \frac{2}{n^{1/3}} \right) + \frac{2cn^{2/3}}{\Delta_1(n^{-\frac{2}{3}})} \cdot (2p+1)^{2/3} .
\]

Due to Proposition 11, in Proposition 11 holds for any \( \tilde{p} : \tilde{p}_u \leq \tilde{p} \leq p \). Due to Proposition 11 with probability at least \( P^* := 1 - 6\exp\left(-\tilde{p}_u \cdot (2p+1) \cdot \Delta_1(n^{-\frac{1}{3}}) \right) - 2(p+1)\exp(-\tilde{c}n) \geq 1 - 6\exp(-2c \cdot (2p+1)^{2/3} \cdot n^{1/3}) - 2(p+1)\exp(-\tilde{c}n) \), it holds that
\[
\mathbb{P}(X^{RSAA}) - \mathbb{P}(X^\ast) \leq s \cdot P_\lambda(a\lambda) + \frac{2K}{\sqrt{n}} \sqrt{\frac{2\tilde{p}_u(2p+1)}{c} \Delta_1(n^{-\frac{1}{3}})} + \frac{4K \tilde{p}_u(2p+1)}{c} \Delta_1(n^{-\frac{2}{3}}) + 2\epsilon + \Gamma ,
\]

in which \( \tilde{p}_u \) is as per (23).

The following simplifies the formula while seeking to preserve the rates in \( n \) and \( p \).

Firstly, we have
\[
\sqrt{\frac{2\tilde{p}_u(2p+1)}{cn} \Delta_1(n^{-\frac{1}{3}})} \leq \frac{4 \cdot (2p+1)}{cn \cdot (2p+1)^{1/3}} \Delta_1(n^{-\frac{1}{3}}) \cdot \frac{cn^{1/3}}{\Delta_1(n^{-\frac{1}{3}})} + \frac{4cn^{2/3}(2p+1)}{K(2p+1)^{2/3}\Delta_1(n^{-\frac{2}{3}})} \left( \Gamma + \frac{2}{n^{1/3}} \right) \cdot \frac{\Delta_1(n^{-\frac{2}{3}})}{cn} + \sqrt{\frac{2}{cn} \Delta_1(n^{-\frac{1}{3}})} \cdot [s + 1],
\]

which is due to \( \sqrt{x+y} \leq \sqrt{x} + \sqrt{y} \) for any \( x, y \geq 0 \) and the relations that \( 0 < a < u_{\ell}^{-1} \leq 1, 0 < c \leq 0.5, K \geq 1, \) and \( \Delta_1(n^{-\frac{2}{3}}) \geq \ln 36 \).

Similar to the above, we obtain
\[
3\tilde{p}_u \cdot (2p+1) \Delta_1(n^{-\frac{1}{3}}) \leq \frac{4 \cdot (2p+1)^{2/3}}{n^{2/3}} + \frac{2}{nc} \Delta_1(n^{-\frac{1}{3}}) \cdot [s + 1] + \frac{4 \cdot (\Gamma + \frac{2}{n^{1/3}})}{Kn^{1/3}} \cdot (2p+1)^{1/3} .
\]
Since (12) and \( \Delta_1(n^{-\frac{1}{4}}) = \ln(np) + \hat{\Delta} \), we have \( \frac{4(2p+1)^{2/3}}{n^{2/3}} + \frac{4(\Gamma + \frac{2p}{Kn^{1/3}})^{(2p+1)^{1/3}}}{n^{2/3}} \leq O(1) \) and \( \frac{2}{nc} \Delta_1(n^{-\frac{1}{4}})[8s + 1] \leq O(1) \) for some universal constants \( O(1) \). Therefore, for some universal constant \( O(1) \), it holds that \( \frac{2\hat{\Delta}_n}{nc} \Delta_1(n^{-\frac{1}{4}})(2p + 1) \leq O(1) \cdot \sqrt{\frac{(2p+1)^{2/3}}{n^{2/3}}} + \frac{(\Gamma + \frac{2p}{Kn^{1/3}})^{(2p+1)^{1/3}}}{n^{2/3}} \) and \( O(1) \cdot \sqrt{\Delta_1(n^{-\frac{1}{4}})}(8s + 1) \). Combining the above with (26) and (27), (24) can be simplified into

\[
\mathbb{F}(X^{RSAA}) - \mathbb{F}(X^*) \leq O(1) \cdot \left( \frac{s \cdot \Delta_1(n^{-1/3}) \cdot p^{2/3}}{n^{2/3}} + \frac{p^{1/3}}{n^{1/3}} + \sqrt{\frac{s \cdot \Delta_1(n^{-1/3})}{n}} \right) \cdot K \\
+ O(1) \cdot \sqrt{\frac{Kp^{1/3}}{n^{1/3}}} + \Gamma, \tag{28}
\]

which then shows Part (i) since \( \Delta_1(n^{-\frac{1}{4}}) = \ln(18n^{1/3}(K_C + C_p) \cdot p \cdot R) \).

For Part (ii), Lemma 14 implies that \( \mathcal{F}_{n, \lambda}(X^{RSAA}, Z_1^*) \leq \mathcal{F}_{n, \lambda}(X^*, Z_1^*) + \lambda \|X^*\|_* \) almost surely. Below we invoke the results from Part (i) with \( \Gamma = \|X^*\|_* \). Note that it is assumed that

\[
n > C_2 \cdot p \cdot U_L \cdot [\ln(np) + \hat{\Delta}] \cdot s^{3/2} R^{3/2} > O(1) \cdot p \cdot a^{-1} \cdot [\ln(np) + \hat{\Delta}] \cdot s^{3/2} R^{3/2}, \tag{29}
\]

and \( \frac{\Gamma}{K} \leq \frac{\|X^*\|_*}{K} \leq \frac{\|X^*\|_* \cdot \sqrt{s K (2p+1)^{2/3}}}{c \cdot a^{n^{2/3}}} \cdot [\ln(n^{1/3}p) + \hat{\Delta}] \) (as well as \( K \geq 1 \)). It then holds under Assumption 1 that \( \frac{\Gamma}{K} \leq \frac{Rs}{a^{1/3}} \cdot [\ln(n^{1/3}p) + \hat{\Delta}]^{1/3} \). Therefore, \( \left( \frac{\Gamma}{K} \right)^3 \leq \left( O(1) \cdot \sqrt{\frac{Rs}{a^{1/3}}} \cdot [\ln(n^{1/3}p) + \hat{\Delta}] \right)^3 \leq O(1) \cdot R^{3/2} s^{3/2} \sqrt{a^{-1} \cdot [\ln(n^{1/3}p) + \hat{\Delta}]} \), for some universal constants \( O(1) \). Furthermore, since \( a < U_L \leq 1 \), it holds that, if \( a \) satisfies (14) for some universal constant \( C_2 \), then

\[
n > O(1) \cdot p \cdot a^{-1} \cdot [\ln(n^{1/3}p) + \hat{\Delta}] \cdot s^{3/2} R^{3/2} \geq O(1) \cdot p \cdot R^{3/2} s^{3/2} \sqrt{a^{-1} \cdot [\ln(n^{1/3}p) + \hat{\Delta}]} + O(1) \cdot p + C_1 s \cdot [\ln(n^{1/3}p) + \hat{\Delta}] \geq C_1 \cdot \left( \left( \frac{\Gamma}{K} \right)^3 p + p + s \cdot [\ln(n^{1/3}p) + \hat{\Delta}] \right). \tag{12}
\]

in Part (i) is met and thus (28) in Part (i) implies that

\[
\mathbb{F}(X^{RSAA}) - \mathbb{F}(X^*) \leq O(1) \cdot K \cdot \left( \frac{sp^{2/3} \Delta_1(n^{-1/3})}{n^{2/3}} + \frac{sp \Delta_1(n^{-1/3})}{n} + \frac{p^{1/3}}{n^{1/3}} \right) \\
+ O(1) \cdot \sqrt{\frac{Kp^{1/3}(\lambda \|X^*\|_*)}{n^{1/3}}} + \lambda \|X^*\|_*,
\]
with probability at least $1 - 2(2p + 1)\exp(-\tilde{c}n) - 6\exp\left(-2cn^{1/3} \cdot (2p + 1)^{2/3}\right)$. Note that $a < 1$, $K \geq 1$, $p \geq 1$, $\left[\ln(n^{1/3}/p) + \tilde{\Delta}\right] \geq 1$ and $\sqrt{2p\Delta_1(n^{-1/3})} \leq s(2p+1)^{1/3}\sqrt{\Delta_1(n^{-1/3})/n^{1/3}}$. Hence, $\mathbb{F}(X_{RSAA}) - \mathbb{F}(X^*) \leq O(1) \cdot K \cdot \left[\frac{s(2p+1)^{1/3}\left(\ln(n^{1/3}) + \tilde{\Delta}\right)}{n^{2/3}} + \frac{p^{1/3}}{n^{1/3}}\right] + O(1) \cdot \min\left\{\frac{K}{a}, \frac{K}{a^{1/2}}\right\}^{1/2} \left[\ln(n^{1/3}/p) + \tilde{\Delta}\right]^{1/2}$, which shows Part (ii) by further noticing that $a = \frac{\tilde{\Delta}c}{\sqrt{n}}$.

\section*{A.2 Auxiliary results}

\textbf{Proposition 8} Suppose that $a < U_L^{-1}$. Assume that the $S^3$ ONC($Z^n_i$) is satisfied almost surely at $X_{RSAA} \in \mathcal{S}_p$. Then,

$$\mathbb{P}\{\{\sigma_j(X_{RSAA}) \notin (0, a\lambda) \text{ for all } j\} \} = 1.$$  

\textbf{Proof} Since $X_{RSAA}$ satisfies the $S^3$ ONC($Z^n_i$) almost surely, Eq. (11) implies that for any $j \in \{1, ..., p\}$, if $\sigma_j(X_{RSAA}) \in (0, a\lambda)$, then

$$0 \leq U_L + \left[\frac{\partial^2 P_\lambda(\sigma_j(X))}{\partial \sigma_j(X)^2}\right]_{X=X_{RSAA}} = U_L - \frac{1}{a}. \quad (30)$$

Further observe that $\frac{\partial^2 P_\lambda(t)}{\partial^2 t} = -a^{-1}$ for $t \in (0, a\lambda)$. Therefore, (30) contradicts with the assumption that $U_L < \frac{1}{a}$. This contradiction implies that

$$\mathbb{P}\{X_{RSAA} \text{ satisfies the } S^3 \text{ ONC} (Z^n_i) \} \cap \{\{\sigma_j(X_{RSAA}) \in (0, a\lambda)\} \} = 0 \implies 0 \geq 1 - \mathbb{P}\{X_{RSAA} \text{ does not satisfy the } S^3 \text{ ONC} (Z^n_i) \} - \mathbb{P}\{\{\sigma_j(X_{RSAA}) \notin (0, a\lambda)\} \} = 1 \text{ for all } j = 1, ..., n$$

which immediately leads to the desired result.

\textbf{Proposition 9} Suppose that Assumptions 3 and 4 hold. Let $\epsilon \in (0, 1]$, $\bar{p} : \bar{p} > s$, $\Delta_1(\epsilon) := \ln\left(\frac{18(Kc+C_\mu)p-R}{\epsilon}\right)$, and $\mathcal{B}_{\bar{p},R} := \{X \in \mathcal{S}_p : |\sigma_{\max}(X)| \leq R, \text{rk}(X) \leq \bar{p}\}$. Then, for the same $c \in (0, 0.5]$ as in (9) and for some $\tilde{c} > 0$,

$$\max_{X \in \mathcal{B}_{\bar{p},R}} \left|\frac{1}{n} \sum_{i=1}^{n} f(X, Z_i) - \mathbb{E}(F)\right| \leq \frac{K}{\sqrt{n}} \sqrt{\frac{2\bar{p}(2p+1)}{c} \Delta_1(\epsilon)} + \frac{K}{n} \frac{2\bar{p}(2p+1)}{c} \Delta_1(\epsilon) + \epsilon$$

with probability at least $1 - 2\exp(-\tilde{c}(2p+1)\Delta_1(\epsilon)) - 2\exp(-\tilde{c}n)$.

\textbf{Proof} We will follow the “$\epsilon$-net” argument similar to Shapiro et al. [25] to construct a net of discretization grids $\mathcal{G}(\epsilon) := \{\bar{X}^k\} \subseteq \mathcal{B}_{\bar{p},R}$ such that for any $X \in \mathcal{B}_{\bar{p},R}$, there is $X^k \in \mathcal{G}(\epsilon)$ that satisfies $\|X^k - X\| \leq \frac{2K}{\sqrt{2c} + 2\mu}$ for any fixed $\epsilon \in (0, 1]$. 

14
Invoking Lemma 15 for an arbitrary $X \in B_{\bar{p},R}$, to ensure that there always exists $\tilde{X}^k \in G(\epsilon)$ such that
\[
\|X - \tilde{X}^k\| \leq \frac{\epsilon}{(2KC + 2C_{\mu})},
\] (31)
it is sufficient to have the number of grids to be no more than $\left(\frac{18R\sqrt{p}}{C_{\mu}}\right)^{(2p+1)\bar{p}}$. Now, we may observe
\[
\mathbb{P}\left[ \max_{X^k \in G(\epsilon)} \left| \frac{1}{n} \sum_{i=1}^{n} f(X^k, Z_i) - \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} f(X^k, Z_i) \right] \right| \leq K \sqrt{\frac{t}{n}} + \frac{Kt}{n} \right]
\leq \mathbb{P}\left[ \bigcap_{X^k \in G(\epsilon)} \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} f(X^k, Z_i) - \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} f(X^k, Z_i) \right] \right| \leq K \sqrt{\frac{t}{n}} + \frac{Kt}{n} \right\} \right]
\geq 1 - \sum_{X^k \in G(\epsilon)} \mathbb{P}\left[ \left| \frac{1}{n} \sum_{i=1}^{n} f(X^k, Z_i) - \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} f(X^k, Z_i) \right] \right| > K \sqrt{\frac{t}{n}} + \frac{Kt}{n} \right],
\]
(32)
Further invoking Eq. (31), for the same $c$ as in (31), it holds that
\[
\mathbb{P}\left[ \max_{X^k \in G(\epsilon)} \left| \frac{1}{n} \sum_{i=1}^{n} f(X^k, Z_i) - \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} f(X^k, Z_i) \right] \right| \leq K \sqrt{\frac{t}{n}} + \frac{Kt}{n} \right]
\geq 1 - |G(\epsilon)| \cdot 2\exp(-ct) \geq 1 - 2 \left( \frac{18R\sqrt{p}}{C_{\mu}} \right)^{(2p+1)\bar{p}} \cdot \exp(-ct).
\]
Combined with Lemma 12
\[
\left| \frac{1}{n} \sum_{i=1}^{n} f(X, Z_i) - \frac{1}{n} \sum_{i=1}^{n} f(X^k, Z_i) \right| + \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} f(X, Z_i) \right] - \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} f(X^k, Z_i) \right]
\leq 2(KC + C_{\mu})\|X - X^k\|, \quad (33)
\]
with probability at least $1 - 2\exp(-\tilde{c} \cdot n)$ for some positive constant $\tilde{c} > 0$. Therefore, for any $X \in B_{\bar{p},R}$ and $X^k \in G(\epsilon)$,
\[
|F_n(X, Z^n_1) - \mathbb{E}[F_n(X, Z^n_1)]| \leq |F_n(X^k, Z^n_1) - \mathbb{E}[F_n(X^k, Z^n_1)]| + |F_n(X, Z^n_1) - F_n(X^k, Z^n_1)| + |\mathbb{E}[F_n(X, Z^n_1)] - \mathbb{E}[F_n(X^k, Z^n_1)]|.
\]
(34)
Therefore, with probability at least $1 - 2\exp(-\tilde{c} \cdot n)$ for some positive constant $\tilde{c} > 0$,
\[
|F_n(X, Z^n_1) - \mathbb{E}[F_n(X, Z^n_1)]| \leq 2(KC + C_{\mu})\|X - X^k\| + |F_n(X^k, Z^n_1) - \mathbb{E}[F_n(X^k, Z^n_1)]|.
\]
(35)
Further invoking (32), it yields that
\[ |F_n(X, Z^n) - F(X)| \leq 2(K_C + C_\mu)\|X - X^k\| + K \sqrt{\frac{t}{n} + \frac{Kt}{n}}, \]
with probability at least 1 - 2 \left( \frac{18R\sqrt{p}(K_C + C_\mu)}{\epsilon} \right)^{(2p+1)\tilde{p}} \cdot \exp(-ct) - 2\exp(-\tilde{c} \cdot n). We may always choose the closest \( X^k \) to \( X \). Therefore, invoking (31), for any \( \epsilon : 0 < \epsilon \leq 1 \):
\[
P\left[ \max_{X \in B_{\tilde{p}, R}} |F_n(X, Z^n) - F(X)| \leq K \sqrt{\frac{t}{n} + \frac{Kt}{n} + \epsilon} \right] \\
\geq 1 - 2 \cdot \exp \left( (2p + 1) \cdot \tilde{p} \cdot \ln \frac{18(K_C + C_\mu)Rp}{\epsilon} - ct \right) - 2\exp(-\tilde{c}n).
\]
Finally, we may let \( t := \frac{2\tilde{p}}{c}(2p+1)\cdot\Delta_1(\epsilon) \), where we recall that \( \Delta_1(\epsilon) := \ln \left( \frac{18(K_C + C_\mu)\tilde{p} \cdot R}{\epsilon} \right) \), we then obtain the desired result. \( \blacksquare \)

**Proposition 10** Let \( \Gamma \geq 0 \), \( \epsilon \in (0, 1) \), \( \Delta_1(\epsilon) := \ln \left( \frac{18(K_C + C_\mu)\tilde{p} \cdot R}{\epsilon} \right) \). Suppose that Assumptions \( \mathfrak{A} \) through \( \mathfrak{B} \) hold, the solution \( X^{RSAA} \in S_p : \sigma_{\max}(X^{RSAA}) \leq R \) satisfies \( S^3\text{ONC}(Z^n) \) almost surely, and
\[
F_{n, \lambda}(X^{RSAA}, Z^n) \leq F_{n, \lambda}(X^*, Z^n) + \Gamma, \ \text{w.p.1.} \tag{36}
\]
For a positive integer \( \bar{p}_u : \bar{p}_u > s \), if
\[
(\bar{p} - s) \cdot P_X(a\lambda) > \frac{4K}{cn} \Delta_1(\epsilon) \cdot \tilde{p} \cdot (2p + 1) + \frac{2K}{\sqrt{n}} \sqrt{\frac{2\tilde{p} \cdot (2p + 1)}{c}} \Delta_1(\epsilon) + \Gamma + 2\epsilon, \tag{37}
\]
for all \( \bar{p} : \tilde{p}_u \leq \bar{p} \leq p \), then \( \mathbb{P}[\text{rk}(X^{RSAA}) \leq \bar{p}_u - 1] \geq 1 - 2p \exp(-\tilde{c}n) - 4 \exp(-\tilde{p}_u(2p + 1)\Delta_1(\epsilon)) \) for the same \( c \) in (9) and some \( \tilde{c} > 0 \).

**Proof** Let \( B_R := \{ X \in S_p : \sigma_{\max}(X) \leq R \} \). Consider an arbitrary \( \bar{p} : \bar{p}_u \leq \bar{p} \leq p \). Since \( \bar{p} > s \) by the assumption that \( \bar{p}_u > s \), we may consider the following events:
\[
\mathcal{E}_1 := \left\{ (\bar{X}, \bar{Z}^n) \in B_R \times W^n : F_{n, \lambda}(\bar{X}, \bar{Z}^n) \leq F_{n, \lambda}(X^*, \bar{Z}^n) + \Gamma \right\},
\]
\[
\mathcal{E}_2 := \left\{ \bar{X} \in B_R : |\sigma_j(\bar{X})| \notin (0, a\lambda) \text{ for all } j \right\},
\]
\[
\mathcal{E}_3 := \left\{ (\bar{X}, \bar{Z}^n) \in B_R \times W^n : \bar{X} \text{ satisfies } S^3\text{ONC}(\bar{Z}^n) \right\}.
\]
\[
\mathcal{E}_{4, \bar{p}} := \left\{ \tilde{X} \in B_R : \text{rk}(\tilde{X}) = \bar{p} \right\},
\]
\[
\mathcal{E}_{5, \bar{p}} := \left\{ \bar{Z}^n \in W^n : \max_{X \in B_R : \text{rk}(X) \leq \bar{p}} \left| F_n(X, \bar{Z}^n) - F(X) \right| \leq \frac{K}{\sqrt{n}} \sqrt{\frac{2\tilde{p} \cdot (2p + 1)}{c}} \Delta_1(\epsilon) + \Gamma + \frac{K}{n} \cdot \frac{2\tilde{p} \cdot (2p + 1)}{c} \Delta_1(\epsilon) + \epsilon \right\},
\]
where \( c \) in \( \mathcal{E}_{5, \hat{p}} \) is a universal constant defined to be the same as in (38). For any \((\hat{X}, \hat{Z}_n) \in \{(\hat{X}, \hat{Z}_n) \in \mathcal{E}_1 \} \cap \{(\hat{X} \in \mathcal{E}_2 \cap \mathcal{E}_{4, \hat{p}})\} \), where \( \hat{Z}_n = (\hat{Z}_1, ..., \hat{Z}_n) \), since \( \hat{X} \in \mathcal{E}_{4, \hat{p}} \cap \mathcal{E}_2 \), which means that \( \hat{X} \) has \( \hat{p} \)-many non-zero singular values and each must not be within the interval \((0, a\lambda)\), it holds that

\[
\mathcal{F}_n(\hat{X}, \hat{Z}_n) + \hat{p}P_{\lambda}(a\lambda) \leq \frac{1}{n} \mathcal{F}_n(X^*, \hat{Z}_n) + sP_{\lambda}(a\lambda) + \Gamma, \tag{38}
\]

Notice that \( X^* \in \mathcal{B}_R : \text{rk}(X^*) = s < \hat{p} \) by Assumption [1]. We may obtain that, for all \( \hat{X} \in \mathcal{E}_{4, \hat{p}} \),

\[
\frac{1}{n} \sum_{i=1}^{n} f(X^*, \hat{Z}_i) - \frac{1}{n} \sum_{i=1}^{n} f(\hat{X}, \hat{Z}_i) = \left[ \frac{1}{n} \sum_{i=1}^{n} f(X^*, \hat{Z}_i) - \mathcal{F}(X^*) \right] + \left[ \mathcal{F}(\hat{X}) - \frac{1}{n} \sum_{i=1}^{n} f(\hat{X}, \hat{Z}_i) \right] + \left[ \mathcal{F}(X^*) - \mathcal{F}(\hat{X}) \right] \leq 2 \max_{\hat{X} \in \mathcal{E}_{4, \hat{p}}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X, \hat{Z}_i) - \mathcal{F}(X) \right| + \mathcal{F}(X^*) - \mathcal{F}(\hat{X}) \tag{39}
\]

where the last inequality is due to \( \mathcal{F}(X^*) \leq \mathcal{F}(X) \) for all \( X \in \mathcal{S}_p \) by the definition of \( X^* \). Combined with (38) and (39), if it holds that

\[
\emptyset \neq \{(\hat{X}, \hat{Z}_n) : (\hat{X}, \hat{Z}_n) \in \mathcal{E}_1 \cap \mathcal{E}_3 \} \cap \{(\hat{X}, \hat{Z}_n) : \hat{X} \in \mathcal{E}_{4, \hat{p}} \cap \mathcal{E}_2 \} \cap \{(\hat{X}, \hat{Z}_n) : \hat{Z}_n \in \mathcal{E}_{5, \hat{p}} \},
\]

then \( (\hat{p} - s) \cdot P_{\lambda}(a\lambda) \leq \frac{2K}{\sqrt{n}} \sqrt{\frac{2\hat{p}(2p+1)}{\epsilon}} \Delta_1(\epsilon) + \frac{2K}{n} \cdot \frac{2\hat{p}(2p+1)}{\epsilon} \Delta_1(\epsilon) + 2\epsilon + \Gamma \), which contradicts with the assumed inequality (37) for all \( \hat{p} : \hat{p}_u \leq \hat{p} \leq p \). Now we recall the definition of \( X^{RSAA} \in \mathcal{B}_R \), which is a solution that satisfies the \( S^8_{\text{ONC}}(Z_1^n) \), w.p.1., and \( \mathcal{F}_{n, \lambda}(X^{RSAA}, Z^n) \leq \mathcal{F}_{n, \lambda}(X^*, Z^n) + \Gamma \), w.p.1. Invoking Proposition \( \mathbb{S} \) we have \( \mathbb{P} \left[ (X^{RSAA}, Z^n) \in \mathcal{E}_1 \cap \mathcal{E}_3, X^{RSAA} \in \mathcal{E}_2 \right] = 1 \). Hence,

\[
0 = \mathbb{P} \left[ \{(X^{RSAA}, Z^n) \in \mathcal{E}_1 \cap \mathcal{E}_3 \} \cap \{X^{RSAA} \in \mathcal{E}_{4, \hat{p}} \cap \mathcal{E}_2 \} \cap \{Z^n \in \mathcal{E}_{5, \hat{p}} \} \right] 
\geq 1 - \mathbb{P} \left[ X^{RSAA} \notin \mathcal{E}_{4, \hat{p}} \right] - \mathbb{P} \left[ Z^n \notin \mathcal{E}_{5, \hat{p}} \right] - \left\{ 1 - \mathbb{P} \left[ (X^{RSAA}, Z^n) \in \mathcal{E}_1 \cap \mathcal{E}_3, X^{RSAA} \in \mathcal{E}_2 \right] \right\},
\]

for all \( \hat{p} : \hat{p}_u \leq \hat{p} \leq p \). The above then implies that \( \mathbb{P} [Z^n \notin \mathcal{E}_{5, \hat{p}}] \geq \mathbb{P} [X^{RSAA} \in \mathcal{E}_{4, \hat{p}}] \) for all \( \hat{p} : \hat{p}_u \leq \hat{p} \leq p \). Therefore, \( \mathbb{P} [\text{rk}(X^{RSAA}) = \hat{p}] \leq 1 - \mathbb{P} [Z^n \in \mathcal{E}_{5, \hat{p}}] \) for all \( \hat{p} : \hat{p}_u \leq \hat{p} \leq p \).
This, combined with Proposition 9 yields that
\[
P[\text{rk}(X^{RSAA}) \leq \tilde{p}_u - 1] = P[\text{rk}(X^{RSAA}) \notin \{\tilde{p}_u, \tilde{p}_u + 1, \ldots, p\}]
\]
\[
= 1 - P \left[ \bigcup_{\tilde{p} = \tilde{p}_u}^p \{\text{rk}(X^{RSAA}) = \tilde{p}\} \right] \geq 1 - \sum_{\tilde{p} = \tilde{p}_u}^p P[\text{rk}(X^{RSAA}) = \tilde{p}] \geq 1 - \sum_{\tilde{p} = \tilde{p}_u}^p (1 - P[Z^n_j \in \mathcal{E}_{5, \tilde{p}}])
\]
\[
\geq 1 - 2(p - \tilde{p}_u + 1) \exp(-\tilde{c}n) - \sum_{\tilde{p} = \tilde{p}_u}^p 2 \exp(-\tilde{p}(2p + 1) \cdot \Delta_1(\epsilon)).
\]

where \(\tilde{c} > 0\) is some universal constant. Observing that \(\Delta_1(\epsilon) = \ln \left(\frac{18 \cdot (K_C + C_\mu) \cdot p \cdot R}{\epsilon}\right) > 1\) (since \(p > 2, R, K_C, C_\mu \geq 1, \) and \(\epsilon \leq 1\)) and \(\sum_{\tilde{p} = \tilde{p}_u}^p 2 \exp(-\tilde{p}(2p + 1) \cdot \Delta_1(\epsilon))\) is the sum of a geometric sequence, we have
\[
P[\text{rk}(X^{RSAA}) \leq \tilde{p}_u - 1] \geq 1 - \frac{2 \exp(-\tilde{p}_u (2p + 1) \Delta_1(\epsilon))}{1 - \exp((-2p + 1) \Delta_1(\epsilon))} - 2p \exp(-\tilde{c}n). \quad (40)
\]
The above can be simplified into \(P[\text{rk}(X^{RSAA}) \leq \tilde{p}_u - 1] \geq 1 - 4 \exp(-\tilde{p}_u (2p + 1) \Delta_1(\epsilon)) - 2p \exp(-\tilde{c}n).\)

**Proposition 11** Let \(\Gamma \geq 0\) and \(\epsilon \in (0, 1]\) be arbitrary scalars and let
\[
\Delta_1(\epsilon) := \ln \left(\frac{18 \cdot (K_C + C_\mu) \cdot p \cdot R}{\epsilon}\right).
\]
Assume that (i) the solution \(X^{RSAA}\) satisfies \(S^{\text{ONC}}(Z^n_i)\) almost surely; (ii) \(\mathcal{F}_{n, \lambda}(X^{RSAA}, Z^n_i) \leq \mathcal{F}_{n, \lambda}(X^*, Z^n_i) + \Gamma\) with probability one; and (iii) for some integer \(\tilde{p}_u: \tilde{p}_u > s\), it holds that
\[
(\tilde{p} - s) \cdot P_\lambda(a\lambda) > \frac{4K}{cn} \Delta_1(\epsilon) \cdot \tilde{p} \cdot (2p + 1) + \frac{2K}{\sqrt{n}} \sqrt{\frac{2\tilde{p} \cdot (2p + 1)}{c}} \Delta_1(\epsilon) + \Gamma + 2\epsilon, \quad (41)
\]
for all \(\tilde{p} : \tilde{p}_u \leq \tilde{p} \leq p\). It then holds that
\[
\mathbb{P}(X^{RSAA}) - \mathbb{P}(X^*) \leq sP_\lambda(a\lambda) + \frac{4K}{cn} \Delta_1(\epsilon) \cdot \tilde{p} \cdot (p + 1)
\]
\[
+ \frac{2K}{\sqrt{n}} \sqrt{\frac{2\tilde{p} \cdot (2p + 1)}{c}} \Delta_1(\epsilon) + \Gamma + 2\epsilon, \quad (42)
\]
with probability at least \(P^* := 1 - 2(p + 1) \exp(-\tilde{c}n) - 6 \exp(-\tilde{p}_u (2p + 1) \Delta_1(\epsilon))\) for some universal constant \(\tilde{c} > 0\).

**Proof** We first observe that \(\Delta_1(\epsilon) := \ln \left(\frac{18 \cdot (K_C + C_\mu) \cdot p \cdot R}{\epsilon}\right) \geq \ln 36\) because \(p \geq 1, K_C, C_\mu, R \geq 1\) and \(0 < \epsilon \leq 1\). Observe that, by definition, \(\sum_{j \in Q} P_\lambda(t) \geq 0\) for all
$t \geq 0$. This, combined with the assumption that
\[
\mathcal{F}_{n,\lambda}(X^{RSAA}, Z^n) \leq \mathcal{F}_{n,\lambda}(X^*, Z^n) + \Gamma,
\]
w.p.1., and the assumption that $\text{rk}(X^*) = s$, yields that $\frac{1}{n} \sum_{i=1}^{n} f(X^{RSAA}, Z_i) \leq \frac{1}{n} \sum_{i=1}^{n} f(X^*, Z_i) + s P_{\lambda}(a\lambda) + \Gamma$, a.s. Furthermore, conditioning on the events that
\[
E_{\tilde{p}_u}^a := \left\{ \max_{X \in B_{\tilde{p}_u, R}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X, Z_i) - \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} f(X, Z_i) \right] \right| \leq K \sqrt{\frac{2\tilde{p}_u \cdot (2p + 1)}{e} \Delta_1(\epsilon)} + K \frac{\tilde{p}_u \cdot (2p + 1)}{c} \Delta_1(\epsilon) + 2\epsilon + \Gamma, \quad \forall \tilde{p}_u \right\}
\]
and $E_{\tilde{p}_u}^a := \{ \text{rk}(X^{RSAA}) \leq \tilde{p}_u \}$ with $\tilde{p}_u > s$, where $B_{\tilde{p}_u, R} := \{ X \in S_p : \text{rk}(X) \leq \tilde{p}_u, \sigma_{\max}(X) \leq R \}$ and $X^*, X^{RSAA} \in B_{\tilde{p}_u, R}$, we obtain that $\mathbb{P}(\mathcal{F}(X^{RSAA}) - \mathcal{F}(X^*) \leq s \cdot P_{\lambda}(a\lambda) + 2K C + C \mu \|X_1 - X_2\|) \leq 0$.

A.3 Useful Lemmata

**Lemma 12** Under Assumption 4, it holds that, for some universal constant $c > 0$, with probability at least $1 - 2\exp(-c \cdot n)$, it holds that
\[
|\mathcal{F}_n(X_1, Z^n) - \mathcal{F}_n(X_2, Z^n)| - (2K C + C \mu) \|X_1 - X_2\| \leq 0.
\]
for all $X_1, X_2 \in S_p \cap \{ X : \sigma_{\max}(X) \leq R \}$.

**Proof** This proof follows a closely similar lemma by [25]. Due to Assumption 4, for some $c > 0$,
\[
\mathbb{P} \left( \left| \sum_{i=1}^{n} \frac{1}{n} \{ C(Z_i) - \mathbb{E}[C(Z_i)] \} \right| > K C \left( \frac{t}{n} + \sqrt{\frac{t}{n}} \right) \right) \leq 2 \exp(-ct), \quad \forall t \geq 0.
\]
If we let $t := n$ and observe that $\mathbb{E}[C(Z_i)] \leq C \mu$, we immediately have that
\[
\mathbb{P} \left( \sum_{i=1}^{n} \frac{C(Z_i)}{n} \leq 2K C + C \mu \right) \leq 1 - 2 \exp(-cn).
\]
\[
19
\]
If we invoke Assumption 4 again given the event that \( \left\{ \sum_{i=1}^{n} \frac{C(Z_i)}{n} \leq 2K_C + C_\mu \right\} \), we have that for any \( X_1, X_2 \in S_p \),

\[
\left\| \frac{1}{n} \sum_{i=1}^{n} f(X_1, Z_i) - \frac{1}{n} \sum_{i=1}^{n} f(X_2, Z_i) \right\| \\
\leq \frac{1}{n} \sum_{i=1}^{n} \| f(X_1, Z_i) - f(X_2, Z_i) \| \\
\leq \frac{1}{n} \sum_{i=1}^{n} C(Z_i) \| X_1 - X_2 \| \leq (2K_C + C_\mu) \| X_1 - X_2 \|
\]

We have the desired result by combining the above with (44).

Lemma 13 \[ \text{Under Assumption 4 for all} \]

\[ X_1, X_2 \in S_p : \text{max}\{\sigma_{\max}(X_1), \sigma_{\max}(X_2)\} \leq R, \]

it holds that

\[
|E[F_n(X_1, Z_1^\alpha)] - E[F_n(X_2, Z_1^\alpha)]| \leq C_\mu \cdot \| X_1 - X_2 \|. \tag{45}
\]

Proof This proof follows a closely similar lemma by [25]. As per Assumption 4 it holds that

\[
E \left[ |F_n(X_1, Z_i^\alpha) - F_n(X_2, Z_i^\alpha)| \right] \leq E \left[ \sum_{i=1}^{n} C(Z_i) / n \| X_1 - X_2 \| \right].
\]

Due to the convexity of the function \( |\cdot| \), it therefore holds that

\[
|E[F_n(X_1, Z_i^\alpha)] - E[F_n(X_2, Z_i^\alpha)]| \leq E \left[ \sum_{i=1}^{n} C(Z_i) / n \| X_1 - X_2 \| \right]
\]

\[
= E \left[ \sum_{i=1}^{n} C(Z_i) / n \right] \cdot \| X_1 - X_2 \|.
\]

Invoking Assumption 4 again, it holds that \( E \left[ \sum_{i=1}^{n} C(Z_i) / n \right] = \frac{\sum_{i=1}^{n} E[C(Z_i)]}{n} \leq C_\mu \) for all \( i = 1, \ldots, n \), which immediately leads to the desired result.

Lemma 14 \[ \text{Denote that} \]

\[ X_1^\ell = \arg \min_{X \in S_p} F_n(X, Z_1^\alpha) + \lambda \| X \|_*, \text{it holds that} \]

\[ F_{n, \lambda}(X_1^\ell, Z_1^\alpha) \leq F_{n, \lambda}(X^*, Z_1^\alpha) + \lambda \| X^* \|_* . \]

20
Proof We first invoke the definition of $P_\lambda$ to obtain

$$0 \leq P_\lambda(t) = \int_0^t \frac{a\lambda - \theta}{a} d\theta \leq \int_0^t \frac{a\lambda}{a} d\theta = \lambda \cdot t.$$  \hfill (46)

for all $t \geq 0$. Secondly, by the definition of $X^{\ell_1}$,

$$\mathcal{F}_n(X^{\ell_1}, Z^n_1) + \lambda \|X^{\ell_1}\|_* \leq \mathcal{F}_n(X^*, Z^n_1) + \lambda \|X^*\|_*.$$  \hfill (47)

Combining (46) and (47), it holds that

$$\mathcal{F}_n(X^{\ell_1}, Z^n_1) + \sum_{j=1}^P P_\lambda(|\sigma_j(X^{\ell_1})|) \leq \mathcal{F}_n(X^{\ell_1}, Z^n_1) + \sum_{j=1}^P \lambda \cdot |\sigma_j(X^{\ell_1})|$$

$$\leq \mathcal{F}_n(X^*, Z^n_1) + \sum_{j=1}^P P_\lambda(|\sigma_j(X^*)|) + \lambda \|X^*\|_*,$$

as desired.

Lemma 15 Let $S_{r,R} := \{X \in \mathbb{R}^{p \times p} : \text{rk}(X) \leq r, \sigma_{\max}(X) \leq R\}$. Then, in terms of the Frobenius norm, there exists an $\epsilon$-net $\tilde{S}_r$ obeying $|\tilde{S}_r| \leq \left(\frac{9\sqrt{r}R}{\epsilon} \right)^{(2p+1)r}$.

Proof The proof follows a closely similar result by [3, Lemma 3.1]. Denote by $X := U\Sigma V^T$ the singular value decomposition (SVD) of a matrix in $S_{r,R}$. Let $D$ be the set of rank-$r$ diagonal matrices with nonnegative diagonal entries and nuclear norm smaller than $R$, and thus any matrix within set $D$ has the Frobenius norm smaller than $\sqrt{r} \cdot R$. We take $\tilde{D}$ be an $\frac{\epsilon}{3\sqrt{r}}$-net (in terms of Frobenius norm) for $D$ with $|\tilde{D}| \leq \left(\frac{9\sqrt{r}R}{\epsilon} \right)^r$.

Let $O_{p,r} := \{U \in \mathbb{R}^{p \times r} : U^T U = I\}$. For the convenience of analysis on $O_{p,r}$, we may as well consider $\tilde{O}_{p,r} := \{X \in \mathbb{R}^{p \times r} : \|X\|_{1,2} \leq 1\}$ and $\|X\|_{1,2} = \max_j \|X_j\|$, where $X_j$ denotes the $j$th column of $X$. Verifiably, $O_{p,r} \subset \tilde{O}_{p,r}$. We may create an $\frac{\epsilon}{3\sqrt{r}}$-net for $\tilde{O}_{p,r}$, denoted by $\tilde{O}_{p,r}$, which satisfies that $|\tilde{O}_{p,r}| \leq (9\sqrt{r}R/\epsilon)^{pr}$.

For any $X \in S_{r,R}$, one may decompose $X$ and obtain $X = U\Sigma V^T$. There exists $\hat{X} = \hat{U}\hat{\Sigma}\hat{V}^T \in S_{r,R}$ with $\hat{U}, \hat{V} \in \tilde{O}_{p,r}$, and $\hat{\Sigma} \in \tilde{D}$ such that $\|U - \hat{U}\|_{1,2} \leq \epsilon/(3\sqrt{r})$, $\|V - \hat{V}\|_{1,2} \leq \epsilon/(3\sqrt{r})$, and $\|\hat{\Sigma} - \Sigma\|_F \leq \epsilon/3$. This gives $\|X - \hat{X}\|_F = \|U\Sigma V^T - \hat{U}\hat{\Sigma}\hat{V}^T\|_F \leq \|U\Sigma V^T - \hat{U}\hat{\Sigma}\hat{V}^T\|_F + \|\hat{U}(\Sigma - \hat{\Sigma})V^T\|_F \leq \|U - \hat{U}\|_{1,2} + \sqrt{\sum_{1 \leq j \leq r} |\sigma_j(X)|^2} \cdot \|U_j - \hat{U}_j\|_2^2 \leq \epsilon/3$, where $U_j$ is the $j$th column of $U$. By a symmetric argument, we may also obtain that $\|\hat{U}(\Sigma - \hat{\Sigma})V^T\|_F \leq \epsilon/3$. To bound the second term, we also notice that $\|\hat{U}(\Sigma - \hat{\Sigma})V^T\|_F = \|\Sigma - \hat{\Sigma}\|_F \leq \epsilon/3$. Combining the above provides the desired result.
References

[1] P.L. Bartlett PL, M.I. Jordan, and J.D. McAuliffe Convexity, classification, and risk bounds. *J. Am. Stat. Assoc.* 101(473):138–156, 2006.

[2] W. Bian, X. Chen, and Y. Ye. Complexity analysis of interior point algorithms for non-lipschitz and nonconvex minimization. *Math. Program.*, 149:301–327, 2015.

[3] E. J. Candès and Y. Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Trans. Inf. Theor.*, 57(4): 2342–2359, 2011. doi: 10.1109/TIT.2011.2111771.

[4] C. Cartis, N. I. M. Gould, and P. L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. part i: motivation, convergence and numerical results. *Math. Program.*, 127(2):245–295, 2011. doi: 10.1007/s10107-009-0286-5.

[5] S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and empirical minimization of $U$-statistics. *Ann. Stat.* 36(2):844–874, 2008.

[6] A. Elsener and S. van de Geer. Robust low-rank matrix estimation. *Ann. of Stat.*, 46(6B), 3481-3509, 2018

[7] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001. doi: 10.1198/016214501753382273.

[8] G. Haeser, H. Liu, and Y. Ye. Optimality condition and complexity analysis for linearly-constrained optimization without differentiability on the boundary. *Math. Program.*, 2018. doi: 10.1007/s10107-018-1290-4.

[9] V. Koltchinskii. (2010) Rademacher complexities and bounding the excess risk in active learning. *J. Mach. Learn. Res.* 11 2457–2485.

[10] H. Liu, H. Y. Lee, and Z. Huo. Linearly constrained high-dimensional learning. 2019. working paper.

[11] H. Liu and Y. Ye. High-Dimensional Learning under Approximate Sparsity: A Unifying Framework for Nonsmooth Learning and Regularized Neural Networks. *arXiv e-prints*, art. arXiv:1903.00616, 2019.

[12] H. Liu, T. Yao, R. Li, and Y. Ye. Folded concave penalized sparse linear regression: sparsity, statistical performance, and algorithmic theory for local solutions. *Math. Program.*, 166(1):207–240, 2017. doi: 10.1007/s10107-017-1114-y.

[13] H. Liu, X. Wang, T. Yao, R. Li, and Y. Ye. Sample average approximation with sparsity-inducing penalty for high-dimensional stochastic programming. *Math. Program.*, 2018. doi: 10.1007/s10107-018-1278-0.

[14] P.-L. Loh. Statistical consistency and asymptotic normality for high-dimensional robust $m$-estimators. *Ann. Stat.*, 45(2):866–896, 04 2017. doi: 10.1214/16-AOS1471.
[15] P.-L. Loh and M. J. Wainwright. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16:559–616, 2015.

[16] S. Negahban and M. J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13(May), 1665-1697, 2012

[17] S. Negahban, B. Yu, M. J. Wainwright, and P. K. Ravikumar. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, 1348–1356, 2009.

[18] Y. Nesterov and B. Polyak. Cubic regularization of newton method and its global performance. *Math. Program.*, 108(1):177–205, Aug 2006. doi: 10.1007/s10107-006-0706-8.

[19] S. F. Nielsen. Empirical processes in m-estimation. *J. of Appl. Stat.*, 27(8):1067–1068, 11 2000.

[20] Rohde, A. and Tsybakov, A.B. Estimation of high-dimensional low-rank matrices. *Ann. of Stat.*, 39(2), 887-930, 2011

[21] M. Rudelson and R. Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electron. Commun. Probab.*, 18:9 pp., 2013. doi: 10.1214/ECP.v18-2865.

[22] A. Ruszczyski and A. Shapiro. Stochastic programming models. In *Stochastic Programming*, volume 10 of *Handbooks in Operations Research and Management Science*, 1 – 64. Elsevier, 2003. doi: https://doi.org/10.1016/S0927-0507(03)10001-1.

[23] A. Ruszczyski and A. Shapiro. Optimality and duality in stochastic programming. In *Stochastic Programming*, volume 10 of *Handbooks in Operations Research and Management Science*, 65 – 139. Elsevier, 2003. doi: https://doi.org/10.1016/S0927-0507(03)10002-3.

[24] A. Shapiro and H. Xu. Uniform laws of large numbers for set-valued mappings and sub-differentials of random functions. *Journal of Mathematical Analysis and Applications*, 325(2):1390 – 1399, 2007. doi: https://doi.org/10.1016/j.jmaa.2006.02.078.

[25] A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory, Second Edition*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2014. ISBN 1611973422, 9781611973426.

[26] C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.*, 38(2):894–942, 04 2010. doi: 10.1214/09-AOS729.

[27] Y. Zhang, M. J. Wainwright, and M. I. Jordan. Optimal prediction for sparse linear models? lower bounds for coordinate-separable m-estimators. *Electron. J. Statist.*, 11 (1):752–799, 2017. doi: 10.1214/17-EJS1233.