

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

LIU.H@UFL.EDU

CDHERNANDEZ@UFL.EDU

HUNGYILEE@UFL.EDU

Abstract

This paper concerns a high-dimensional stochastic programming problem of minimizing a function of expected cost with 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 expected cost. Traditional SAA theories require the sample size to grow rapidly when the problem dimensionality increases. Indeed, 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 ϵ -suboptimality gap, for some poly-logarithmic 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 demonstrate that the sample complexity of RSAA is $\tilde{O}(1) \cdot \frac{p}{\epsilon^3} \cdot \text{polylog}(p, \frac{1}{\epsilon})$, which is almost linear in p and thus indicates a substantially lower dependence on dimensionality. 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, sometimes tends to be demanding for sample availability. The current SAA theories as per [27], [24], [25] and [26] 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 sample size requirement may be undesirably costly in certain high-dimensional applications. Recently, a regularized SAA with sparsity-inducing penalty has been studied by [13], which shows that significant reduction of sample size requirement may be achieved 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 $f : \mathcal{S}_p \times \mathcal{W} \rightarrow \mathbb{R}$ where \mathcal{S}_p is the cone of p -by- p ($p \geq 1$) symmetric and positive semidefinite matrices and $f(\mathbf{X}, Z)$ is a cost function with respect to parameter Z and a fixed matrix of decision variables \mathbf{X} . Then the problem of consideration is an SP problem given as

$$\mathbf{X}^* \in \arg \min \{ \mathbb{F}(\mathbf{X}) : \mathbf{X} \in \mathcal{S}_p \}. \quad (1)$$

where $\mathbb{F}(\mathbf{X}) = \mathbb{E}[f(\mathbf{X}, Z)]$ is well-defined and finite-valued for any given $\mathbf{X} \in \mathcal{S}_p$. Assume, hereafter, that $\sigma_{\max}(\mathbf{X}^*) \leq R$ for some constant $R \geq 1$, where $\sigma_{\max}(\cdot)$ 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 \mathbf{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 cost. We would like to remark that (1) 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 (1) are problems with non-symmetric and non-square matrices \mathbf{X} , since they can be transformed into symmetric matrices by the self-adjoint dilation with $\bar{\mathbf{X}} = \begin{bmatrix} \mathbf{0} & \mathbf{X} \\ \mathbf{X}^\top & \mathbf{0} \end{bmatrix}$ for some all-zero matrices $\mathbf{0}$ ’s with proper dimensions.

Hereafter, let $\mathbf{Z}_1^n = (Z_1, \dots, Z_n)$ be a sequence of n -many i.i.d. random samples of Z . To solve Problem (1), one of the most popular solution schemes, as mentioned above, is to invoke the following SAA formulation as a surrogate:

$$\mathbf{X}^{SAA} \in \arg \min \left\{ \mathcal{F}_n(\mathbf{X}, \mathbf{Z}_1^n) := \frac{1}{n} \sum_{i=1}^n f(\mathbf{X}, Z_i) : \mathbf{X} \in \mathcal{S}_p \right\}. \quad (2)$$

According to the seminal results by [27], \mathbf{X}^{SAA} well approximates \mathbf{X}^* in the sense that

$$\mathbb{F}(\mathbf{X}^{SAA}) - \mathbb{F}(\mathbf{X}^*) \leq \tilde{O}(1) \cdot \sqrt{\frac{p^2 \cdot \ln n}{n}} \quad (3)$$

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 \mathbf{X}^* is sparse and f is twice-differentiable almost surely, we have shown in [13] that (3) can be sharpened, in terms of its dependence on p , into:

$$\mathbb{F}(\mathbf{X}^{RSAA}) - \mathbb{F}(\mathbf{X}^*) \leq \tilde{O}(1) \cdot \frac{\sqrt{\ln(np)}}{n^{1/4}}, \quad (4)$$

with high probability, where \mathbf{X}^{RSAA} is an SAA scheme with sparsity-inducing regularization. Similar (and potentially stronger) 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 $\mathbf{rk}(\cdot)$ of \mathbf{X}^* in the problem (1) satisfies $s := \mathbf{rk}(\mathbf{X}^*) \ll p$ for some $s \geq 1$.*

The above low-rankness assumption is more general than the sparsity assumption of a vector, since any vector \mathbf{x} can be represented by a diagonal matrix, $\text{diag}(\mathbf{x})$, whose diagonal entries equal to \mathbf{x} . Then, sparsity of \mathbf{x} implies that $\text{diag}(\mathbf{x})$ is of low rank. Furthermore, we generalize the assumption twice-differentiability to Lipschitz continuity of the partial derivatives 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):

$$\mathbf{X}^{RSAA} \in \arg \min_{\mathbf{X} \in \mathcal{S}_p} \left\{ \mathcal{F}_{n,\lambda}(\mathbf{X}, \mathbf{Z}_1^n) := \mathcal{F}_n(\mathbf{X}, \mathbf{Z}_1^n) + \sum_{j=1}^p P_\lambda(\sigma_j(\mathbf{X})) \right\}, \quad (5)$$

where $\sigma_j(\mathbf{X})$ stands for the j th eigenvalue of \mathbf{X} and P_λ is a penalty function in the form of the minimax concave penalty (MCP) [28] given as $P_\lambda(x) = \int_0^x \frac{[a\lambda - t]_+}{a} dt$, for some user-specific tuning parameters $a, \lambda > 0$. Here $[\cdot]_+ = \max\{\cdot, 0\}$. 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 satisfy a set of significant subspace second-order necessary conditions (S^3ONC), given as in Definition 2 in the subsequent. The S^3ONC 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^3ONC . The resulting computational effort of an S^3ONC solution (a solution that satisfies the S^3ONC) is likely tractable.

Let $\mathbf{X}_\lambda^{\ell_1}$ be defined as

$$\mathbf{X}_\lambda^{\ell_1} \in \arg \min_{\mathbf{X} \in \mathcal{S}_p} \mathcal{F}_n(\mathbf{X}, \mathbf{Z}_1^n) + \lambda \|\mathbf{X}\|_*, \quad (6)$$

with $\|\cdot\|_*$ denoting the nuclear norm. We show that, under a few standard assumptions in addition to Assumptions 1, for any S³ONC solution to the RSAA, denoted \mathbf{X}^{RSAA} , which satisfies $\mathcal{F}_{n,\lambda}(\mathbf{X}^{RSAA}) \leq \mathcal{F}_{n,\lambda}(\mathbf{X}_\lambda^{\ell_1})$ a.s., it holds that

$$\mathbb{F}(\mathbf{X}^{RSAA}) - \mathbb{F}(\mathbf{X}^*) \leq \tilde{O}(1) \cdot \left(\frac{s \cdot p^{2/3}}{n^{2/3}} + \frac{s \cdot p^{1/3}}{n^{1/3}} \right) \cdot \ln(np), \quad (7)$$

with overwhelming probability, when our knowledge on the rank of \mathbf{X}^* is completely absent. Furthermore, if we allow the penalty parameter to incorporate knowledge on the rank s of \mathbf{X}^* , as in Assumption 1, then a better choice of λ allows that

$$\mathbb{F}(\mathbf{X}^{RSAA}) - \mathbb{F}(\mathbf{X}^*) \leq \tilde{O}(1) \cdot \left(\frac{s^{1/3} \cdot p^{2/3}}{n^{2/3}} + \frac{s^{2/3} \cdot p^{1/3}}{n^{1/3}} \right) \cdot \ln(np), \quad (8)$$

with overwhelming probability, where the sample size requirement has a lower dependence on s of \mathbf{X}^* compared to (7). The above results are then the promised, almost linear, sample complexity; from both (7) and (8), n should only increase almost 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 \mathbf{X}^{RSAA} , one may invoke an S³ONC-guaranteeing algorithm initialized at $\mathbf{X}_\lambda^{\ell_1}$. Meanwhile, the initial solution, $\mathbf{X}_\lambda^{\ell_1}$, is often polynomial-time computable when $f(\cdot, w)$ is convex for almost every $w \in \mathcal{W}$ (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 a sample complexity that is almost linear in dimensionality under low-rankness. Even though similar results have been achieved previously, e.g., by [17], [21] and [6] in the context of high-dimensional low-rank matrix estimation, most of the existing 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 the low-rankness assumption; even if the conditions such as the RSC or alike are completely absent, the MCP-based regularization may still ensure a sound generalization error as measured by the excess risk, 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 [17], [21] and [6] (most of whom focus more on linear or generalized linear models under RSC or alike), we believe that the excess risk 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 major 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 $\|\cdot\|_*$ the nuclear norm, and by $\|\cdot\|_{\mathbf{p}}$ the \mathbf{p} -norm (with $1 \leq \mathbf{p} \leq \infty$). Let $\sigma_j(\mathbf{X})$ be the j th singular value of matrix \mathbf{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 S³ONC 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*

$$\left| \frac{\partial f(\mathbf{X}, z)}{\partial \sigma_j(\mathbf{X})} \Big|_{\mathbf{X}=\mathbf{X}_1} - \frac{\partial f(\mathbf{X}, z)}{\partial \sigma_j(\mathbf{X})} \Big|_{\mathbf{X}=\mathbf{X}_2} \right| < \mathcal{U}_L \cdot |\sigma_j(\mathbf{X}_1) - \sigma_j(\mathbf{X}_2)| \quad (9)$$

for every $j = 1, \dots, p$, all $\mathbf{X}_1, \mathbf{X}_2 \in \mathcal{S}_p$, and almost every $z \in \mathcal{W}$.

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

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

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

Invoking the well-known Bernstein-type inequality, one has that, for all $\mathbf{X} \in \mathcal{S}_p$, it holds that

$$\mathbb{P} \left(\left| \sum_{i=1}^n a_i \{f(\mathbf{X}, Z_i) - \mathbb{E}[f(\mathbf{X}, Z_i)]\} \right| > K(\|\mathbf{a}\|\sqrt{t} + \|\mathbf{a}\|_{\infty} t) \right) \leq 2 \exp(-ct), \quad \forall t \geq 0, \mathbf{a} = (a_i) \in \mathbb{R}^n, \quad (10)$$

for some absolute constant $c \in (0, \frac{1}{2}]$. [See also 23].

Assumption 4 *For some measurable and deterministic function $\mathcal{C} : \mathcal{W} \rightarrow \mathbb{R}$ with $\mathbb{E}[|\mathcal{C}(Z)|] \leq \mathcal{C}_{\mu}$, for some $\mathcal{C}_{\mu} \geq 1$, the random variable $\mathcal{C}(Z)$ satisfies that*

$\|\mathcal{C}(Z) - \mathbb{E}[\mathcal{C}(Z)]\|_{\psi_1} \leq K_C$ for some $K_C \geq 1$. Furthermore, $|f(\mathbf{X}_1, z) - f(\mathbf{X}_2, z)| \leq \mathcal{C}(z)\|\mathbf{X}_1 - \mathbf{X}_2\|$ for all $\mathbf{X}_1, \mathbf{X}_2 \in \mathcal{S}_p$, 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 [27] 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 $S^3\text{ONC}$ 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 $\mathbf{Z}_1^n \in \mathcal{W}^n$, a vector $\hat{\mathbf{X}} \in \mathcal{S}_p$ is said to satisfy the $S^3\text{ONC}$ (denoted by $S^3\text{ONC}(\mathbf{Z}_1^n)$) of the problem (5) if both of the following sets of conditions are satisfied:

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

$$\nabla \mathcal{F}_{n,\lambda}(\mathbf{X}^{RSAA}, \mathbf{Z}_1^n) = 0, \quad (11)$$

where $\nabla \mathcal{F}_{n,\lambda}(\mathbf{X}^{RSAA}, \mathbf{Z}_1^n)$ is the gradient of $\mathcal{F}_{n,\lambda}(\mathbf{X}^{RSAA}, \mathbf{Z}_1^n)$ at \mathbf{X}^{RSAA} .

- b. The following inequality holds at \mathbf{X}^{RSAA} for all $j = 1, \dots, p$:

$$\mathcal{U}_L + \left[\frac{\partial^2 P_\lambda(\sigma_j(\mathbf{X}))}{[\partial \sigma_j(\mathbf{X})]^2} \right]_{\mathbf{X}=\mathbf{X}^{RSAA}} \geq 0, \quad \text{if } \sigma_j(\mathbf{X}^{RSAA}) \in (0, a\lambda), \quad (12)$$

where \mathcal{U}_L is as defined in (9) for Assumption 2.

As mentioned, the above $S^3\text{ONC}$ 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 $S^3\text{ONC}$.

2.3 Main results

Introduce a few short-hand notations: Denote $\tilde{\Delta} := \ln(18R \cdot (K_C + \mathcal{C}_\mu))$ and $\lambda(\rho) := \sqrt{\frac{8K(2p+1)^{2/3}s^{-\rho}}{c \cdot a \cdot n^{2/3}}} [\ln(n^{1/3}p) + \tilde{\Delta}]$, for the same c in (10) and a user-specific $\rho \geq 0$. Recall the definition of $\mathbf{X}_\lambda^{\ell_1}$ in (6) and specify $a^{-1} = 2\mathcal{U}_L$ (and thus $a < \mathcal{U}_L^{-1}$). We are now ready to present our claimed results.

Theorem 3 Suppose that Assumptions 1 through 4 hold. Specify the penalty parameter $\lambda := \lambda(\rho)$. Let $\mathbf{X}^{RSAA} \in \mathcal{S}_p : \sigma_{\max}(\mathbf{X}^{RSAA}) \leq R$ satisfy the $S^3\text{ONC}(\mathbf{Z}_1^n)$ to (5) almost surely. For any $\Gamma \geq 0$ and some universal constants $\tilde{c}, C_1 > 0$, if

$$n > C_1 \cdot s^{3\rho} \cdot \left[\left(\frac{\Gamma}{K} \right)^3 + 1 \right] \cdot p + C_1 \cdot s \cdot p \cdot \left(\ln(n^{1/3}p) + \tilde{\Delta} \right), \quad (13)$$

and $\mathcal{F}_{n,\lambda}(\mathbf{X}^{RSAA}, \mathbf{Z}_1^n) \leq \mathcal{F}_{n,\lambda}(\mathbf{X}^*, \mathbf{Z}_1^n) + \Gamma$ almost surely, then the excess risk is bounded by

$$\begin{aligned} \mathbb{F}(\mathbf{X}^{RSAA}) - \mathbb{F}(\mathbf{X}^*) &\leq \sqrt{\frac{K \cdot s^\rho \cdot p^{1/3} \cdot \Gamma}{n^{1/3}}} + \Gamma \\ &+ C_1 K \cdot \left[\frac{s^{1-\rho} \cdot p^{2/3} \cdot (\ln(n^{1/3}p) + \tilde{\Delta})}{n^{2/3}} + \sqrt{\frac{s \cdot p \cdot (\ln(n^{1/3}p) + \tilde{\Delta})}{n}} + \frac{p^{1/3} \cdot s^\rho}{n^{1/3}} \right], \end{aligned} \quad (14)$$

with probability at least $1 - 2(p+1)\exp(-\tilde{c}n) - 6\exp(-2c(2p+1)^{2/3}n^{1/3})$.

Proof See proof in Section A.1. ■

Remark 4 Some explanations on the notations are below:

1. Γ measures the solution quality in solving the (in-sample) RSAA formulation; that is, Γ is the suboptimality gap of minimizing the RSAA, which is the surrogate model for the true SP problem in (1). We refer to Γ as “in-sample suboptimality gap” hereafter.
2. More important to us is a second type of suboptimality gap, which we refer to as the “out-of-sample suboptimality gap”, calculated as $\mathbb{F}(\mathbf{X}) - \mathbb{F}(\mathbf{X}^*)$ for a feasible solution \mathbf{X} . The out-of-sample suboptimality gap measures how well the solution \mathbf{X} optimizes the true SP problem in (1).
3. $\tilde{\Delta}$ is some logarithmic terms independent of p and n .
4. K and K_C are subexponential norms of the underlying distributions. They are alternative measures of the distributions’ variances.

Remark 5 Some intuitions on the above theorem are as follows:

1. Theorem 3 ensures that all S^3 ONC solutions to the RSAA formulation yield a bounded out-of-sample suboptimality gap in minimizing the true problem (1).
2. Furthermore, the out-of-sample suboptimality gap is consistent with the in-sample suboptimality gap in the sense that the former deteriorates as Γ increases. When Γ is relatively large, the deterioration is dominated by a linear rate.

We may well control the in-sample suboptimality gap Γ by properly initializing the search for an S^3 ONC solution. Indeed, as is shown in the corollary below, using $\mathbf{X}_\lambda^{\ell_1}$ defined in (6) to warm-start any S^3 ONC-guaranteeing local optimization algorithm ensures the promised sample complexity.

Corollary 6 Suppose that Assumptions 1 through 4 hold. Specify the penalty parameter $\lambda = \lambda(0)$ (that is, $\rho = 0$) in both formulations (6) and (5). Let $\mathbf{X}^{RSAA} \in \mathcal{S}_p : \sigma_{\max}(\mathbf{X}^{RSAA}) \leq R$ satisfy the S^3 ONC(\mathbf{Z}_1^n) to (5) almost surely. For some universal constant \tilde{c} , $C_2 > 0$, if

$$n > C_2 \cdot p \cdot \mathcal{U}_L \cdot [\ln(n^{1/3}p) + \tilde{\Delta}] \cdot s^{\frac{3}{2}} R^{\frac{3}{2}}, \quad (15)$$

and

$$\mathcal{F}_{n,\lambda}(\mathbf{X}^{RSAA}, \mathbf{Z}_1^n) \leq \mathcal{F}_{n,\lambda}(\mathbf{X}_\lambda^{\ell_1}, \mathbf{Z}_1^n) \quad (16)$$

almost surely, where $\mathbf{X}_\lambda^{\ell_1}$ is as defined in (6), then the excess risk is bounded by

$$\begin{aligned} & \mathbb{F}(\mathbf{X}^{RSAA}) - \mathbb{F}(\mathbf{X}^*) \\ & \leq C_2 \cdot s \cdot K \cdot \left[\frac{p^{2/3} \left(\ln(n^{\frac{1}{3}}p) + \tilde{\Delta} \right)}{n^{\frac{2}{3}}} + \frac{p^{1/3} R \cdot \mathcal{U}_L^{1/2} \sqrt{\ln(n^{\frac{1}{3}}p) + \tilde{\Delta}}}{n^{\frac{1}{3}}} \right], \quad (17) \end{aligned}$$

with probability at least $1 - 2(p+1)\exp(-\tilde{c}n) - 6\exp(-2c(2p+1)^{2/3}n^{1/3})$.

Proof See proof in Section A.2. ■

Remark 7 We would like to make a few remarks on the above result:

1. Corollary 6 above establishes our claimed result of almost linear complexity at an S^3 ONC solution generated with a proper initialization.
2. The same corollary considers the particular sublevel set that has a better objective value (in terms of RSAA formulation) than $\mathbf{X}_\lambda^{\ell_1}$. In such a case, the suboptimality in minimizing the true problem (1) explicitly vanishes as sample size n increases.
3. $\mathbf{X}_\lambda^{\ell_1}$ is an initial solution often tractably computable under the common assumption that $f(\cdot, z)$ is convex for almost every $z \in \mathcal{W}$. However, our results in Theorem 3 is not contingent on the convexity of $f(\cdot, z)$, although generating $\mathbf{X}_\lambda^{\ell_1}$ may be intractable when convexity of $f(\cdot, z)$ is not in presence.
4. Corollary 6 above is consistent with the claimed sample complexity in (7), which is almost linear in p . Indeed, for achieving an accuracy of ϵ , the above bounds indicate a sample complexity $\tilde{O}(1) \cdot \frac{p}{\epsilon^3} \cdot \text{polylog}(p, \frac{1}{\epsilon})$, which is almost linear in p , for some quantities $\tilde{O}(1)$ that is independent of n , ϵ , and p .

We note that the dependence of sample size n on rank s of the true solution \mathbf{X}^* is cubic, which means that the proposed approach is more powerful when the true solution \mathbf{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 \mathcal{F}_n or \mathbb{F} satisfies strong convexity or its certain relaxed forms, dependence on s is likely reducible, as it has been successful for [13] in stochastic optimization under sparsity. (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 λ to decrease the dependence on s . We will consider the relatively special case in (i) in future study. Nonetheless, our claim in (ii) above is provided in Corollary 8 below.

Corollary 8 *Suppose that Assumptions 1 through 4 hold. Specify the penalty parameter $\lambda = \lambda(\frac{2}{3})$ (that is, $\rho = \frac{2}{3}$) in both formulations (6) and (5). Let $\mathbf{X}^{RSAA} \in \mathcal{S}_p : \sigma_{\max}(\mathbf{X}^{RSAA}) \leq R$ satisfy the $S^3\text{ONC}(\mathbf{Z}_1^n)$ to (5) almost surely. For some universal constant \tilde{c} , $C_3 > 0$, if*

$$n > C_3 \cdot p \cdot \mathcal{U}_L \cdot [\ln(n^{\frac{1}{3}}p) + \tilde{\Delta}] \cdot s^2 \cdot R^{\frac{3}{2}}, \quad (18)$$

and (16) holds almost surely, where $\mathbf{X}_\lambda^{\ell_1}$ is as defined in (6), then the excess risk is bounded by

$$\begin{aligned} & \mathbb{F}(\mathbf{X}^{RSAA}) - \mathbb{F}(\mathbf{X}^*) \\ & \leq C_3 \cdot K \cdot \left[\frac{s^{1/3} p^{2/3} (\ln(n^{\frac{1}{3}}p) + \tilde{\Delta})}{n^{\frac{2}{3}}} + \frac{s^{2/3} p^{1/3} \cdot R \cdot \mathcal{U}_L^{1/2} \cdot \sqrt{\ln(n^{\frac{1}{3}}p) + \tilde{\Delta}}}{n^{\frac{1}{3}}} \right], \end{aligned} \quad (19)$$

with probability at least $1 - 2(p+1)\exp(-\tilde{c}n) - 6\exp(-2c(2p+1)^{2/3}n^{1/3})$.

Proof See proof in Section A.3. ■

Remark 9 *The Corollary 8, similar to Corollary 6, establishes our claimed result of almost linear complexity at a computable $S^3\text{ONC}$ solution generated with a proper initialization, $\mathbf{X}_\lambda^{\ell_1}$, which can be tractable when $f(\cdot, z)$ is convex for almost every z .*

Remark 10 *In contrast to Corollary 6, Corollary 8 yields a sample complexity with much reduced dependence on s ; quadratic instead of cubic in s . We suppose that this dependence is no longer improvable. This is because, even if we are given the exact knowledge to correctly reduce the “redundant” dimensions of the problem, the traditional SAA to the reduced problem will still require a sample size quadratically dependent on s .*

Remark 11 *There is strong correspondence between the SP and statistical learning as formerly noted by [13; 16]. More specifically, the SAA formulation (5) can be considered as an M -estimation problem and the suboptimality gap $\mathbb{F}(\mathbf{X}^{RSAA}) - \mathbb{F}(\mathbf{X}^*)$ has the same formulation as the excess risk discussed by [1], [9], and [5]. We therefore argue that the results in Theorem (3) and Corollaries 6 and 8 indicate 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 12 *We would like to remark again that, to obtain the desired results, the incurred computational ramification can be reasonably small. This is because \mathbf{X}^{RSAA} is only a stationary point that satisfies (16). First, the stationarity can be ensure by invoking local optimization algorithms. Second, the stipulated inequality in (16) can be ensured by initializing the local algorithm with $\mathbf{X}_\lambda^{\ell_1}$. Such an initializer often can be generated within polynomial time under the common assumption that $f(\cdot, w)$ 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 of our proof is straightforward and focuses on addressing the question: *how to show that an S^3 ONC solution has low rank*. If this question is answered, then the desired results can be almost evident by analyzing the ϵ -net for all the low-rank subspaces. Such an analysis is available in Lemma 3.1 of [3] and is restated (with minor modifications) in Lemma 20 herein.

To that end, we utilize a unique property of the MCP function, which ensure that the stationary points that satisfy the S^3 ONC solutions \mathbf{X}^{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 \mathbf{X}^{RSAA} , an additional penalty of value $\frac{a\lambda^2}{2}$ is added to the objective function of the RSAA, and, therefore, the total penalty incurred by the low-rankness-inducing regularization is $\sum_{j=1}^p P_\lambda(\sigma_j(\mathbf{X}^{RSAA})) = \mathbf{rk}(\mathbf{X}^{RSAA}) \cdot \frac{a\lambda^2}{2}$. Now, consider those stationary points whose suboptimality gaps (in minimizing the RSAA) are smaller than a user-specific quantity Γ , and therefore, $\mathcal{F}_{n,\lambda}(\mathbf{X}^{RSAA}, \mathbf{Z}_1^n) = \mathcal{F}_n(\mathbf{X}^{RSAA}, \mathbf{Z}_1^n) + \mathbf{rk}(\mathbf{X}^{RSAA}) \cdot \frac{a\lambda^2}{2} \leq \mathcal{F}_{n,\lambda}(\mathbf{X}^*, \mathbf{Z}_1^n) + \Gamma$. The rank of such \mathbf{X}^{RSAA} rank must be bounded from above by a function of Γ . Such a function can be explicated via a peeling technique discussed by [22]. Some relative details are provided below.

3.2 Proof Roadmap

The proof of Theorem 3 is motivated by [11] but involves substantial generalization from an SP problem under sparsity in [11] to an SP problem under low-rankness herein. To understand the non-trivial step involved in this generalization, one may observe the fundamental differences between those two problems: While low-rankness can be represented by sparsity via a linear transformation, the linear operator involved in this transformation is completely unknown. More specifically, by singular value decomposition, one may write $\mathbf{X}^* := U D^* V^\top$ for some proper unitary matrices U and V . Apparently, as per Assumption 1, the diagonal matrix D^* must be sparse and U^\top and V are linear operators that project \mathbf{X}^* to a sparse domain; indeed, $D^* = U^\top \mathbf{X}^* V$. Nonetheless, the knowledge on U^\top and V are completely absent, which leads to significant ramifications in analysis.

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 $\mathcal{U}_L < a^{-1}$, in Proposition 13, we show that, for an S^3 ONC solution to the RSAA formulation, denoted \mathbf{X}^{RSAA} , a thresholding rule of $\sigma_j(\mathbf{X}^{RSAA})$, for all j , is that $\sigma_j(\mathbf{X}^{RSAA}) \neq 0 \implies \sigma_j(\mathbf{X}^{RSAA}) \geq a\lambda$, where a and λ are the tuning parameters of the MCP function, P_λ . This can be demonstrated by observing that the definition of the S^3 ONC, which is $\mathcal{U}_L - P'_\lambda(\sigma_j(\mathbf{X}^{RSAA})) = \mathcal{U}_L - \frac{1}{a} \geq 0$ if $\sigma_j(\mathbf{X}^{RSAA}) \in (0, a\lambda)$, contradicts with the assumption that $\mathcal{U}_L < a^{-1}$. Therefore, it holds that $\sigma_j(\mathbf{X}^{RSAA}) \geq a\lambda$, unless $\sigma_j(\mathbf{X}^{RSAA}) = 0$.

Step 2. *ϵ -net argument for low-rank subspaces.* We apply the well-known ϵ -net argument to show a point-wise error bound for $|\mathcal{F}_{n,\lambda}(\mathbf{X}, \mathbf{Z}_1^n) - \mathbb{F}(\mathbf{X})| \leq \epsilon$ for all $\mathbf{X} \in \mathcal{S}_p : \sigma_{\max}(\mathbf{X}) \leq R$ 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 ϵ -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 $\binom{p}{\tilde{p}}$ -many rank- \tilde{p} subspaces, the total covering number for all possible rank- \tilde{p} subspaces is

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

Combining this covering number, the Bernstein-like inequality, and Lipschitz-like inequality in (10), we have that, for any $t \geq 0$,

$$|\mathcal{F}_n(\mathbf{X}, \mathbf{Z}_1^n) - \mathbb{F}(\mathbf{X})| > \tilde{O}(1) \cdot \frac{t}{n} + \tilde{O}(1) \cdot \sqrt{\frac{t}{n}} + \epsilon, \quad \forall \mathbf{X} \in \mathcal{S}_p : \sigma_{\max}(\mathbf{X}) \leq R : \mathbf{rk}(\mathbf{X}) \leq \tilde{p}, \quad (20)$$

with probability at most $\left(\tilde{O}(1) \cdot \frac{p}{\epsilon} \right)^{(2p+1)\tilde{p}} \exp(-ct) + \exp(-\tilde{O}(1) \cdot n)$ for some universal constant $c \in (0, 1/2]$. We may choose to let $t = 2\tilde{p}(2p+1) \ln \left(\frac{\tilde{O}(1) \cdot p}{\epsilon} \right)$, as well as $\epsilon = n^{-\frac{1}{3}}$, 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 \mathbf{X}^{RSAA} be an $\mathbf{S}^3\text{ONC}$ solution to the RSAA formulation in (5). Assume that \mathbf{X}^{RSAA} is within the Γ -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., $\sigma_j(\mathbf{X}) \geq a\lambda$, unless $\sigma_j(\mathbf{X}) = 0$),

$$\mathcal{F}_n(\mathbf{X}^{RSAA}, \mathbf{Z}_1^n) + \mathbf{rk}(\mathbf{X}^{RSAA}) \cdot \frac{a\lambda^2}{2} \leq \mathcal{F}_n(\mathbf{X}^*, \mathbf{Z}_1^n) + \frac{a\lambda^2 \cdot s}{2} + \Gamma. \quad (21)$$

If $\mathbf{rk}(\mathbf{X}^{RSAA}) \leq \tilde{p}$, the result from Step 2 can be invoked to bound the differences, $\mathbb{F}(\mathbf{X}^{RSAA}) - \mathcal{F}_n(\mathbf{X}^{RSAA}, \mathbf{Z}_1^n)$ and $\mathcal{F}_n(\mathbf{X}^*, \mathbf{Z}_1^n) - \mathbb{F}(\mathbf{X}^*)$, to be smaller than a desired level. In particular, as we choose to let $t = 2\tilde{p}(2p+1) \ln \left(\frac{\tilde{O}(1) \cdot p}{\epsilon} \right)$, as well as $\epsilon = n^{-\frac{1}{3}}$, in (20) and $\lambda = \tilde{O}(1) \cdot \frac{p^{1/3} \sqrt{\ln(np)}}{n^{1/3}}$ in (21). After some algebraic simplification, we obtain that

$$\begin{aligned} & \mathbb{F}(\mathbf{X}^{RSAA}) - \mathbb{F}(\mathbf{X}^*) \\ & \leq -\frac{a\lambda^2}{2} \mathbf{rk}(\mathbf{X}^{RSAA}) + \tilde{O}(1) \cdot \frac{sp^{2/3} \ln(pn)}{n^{2/3}} + \tilde{O}(1) \cdot \sqrt{\frac{\tilde{p}}{n} \ln(pn)} + \frac{p^{1/3}}{n^{1/3}} + \Gamma \end{aligned} \quad (22)$$

$$\leq \tilde{O}(1) \cdot \frac{\tilde{p} \cdot p \ln(pn)}{n} + \tilde{O}(1) \cdot \sqrt{\frac{\tilde{p} \cdot p}{n} \ln(pn)} + \frac{1}{n^{1/3}} + \Gamma \quad (23)$$

with probability at least $1 - \exp(-\tilde{O}(1) \cdot \tilde{p} \cdot p \cdot \ln(np)) - \exp(-\tilde{O}(1) \cdot n)$. Recalling that \tilde{p} is an upper bound on the rank of \mathbf{X}^{RSAA} , the above result in (23) is now close to the desired “almost linear” sample complexity results if \tilde{p} much smaller than p . As it turns out,

it is indeed the case. As is demonstrated in Theorem 3, we can show that $\mathbf{rk}(\mathbf{X}^{RSAA}) \leq \tilde{p} := \tilde{O}(1) \cdot \left(s + \frac{n^{1/3}}{p^{1/3}} + \frac{n^{1/3}}{p^{1/3}} \cdot \Gamma\right)$, which is to be explained subsequently.

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

$$\mathbf{rk}(\mathbf{X}^{RSAA}) \leq \tilde{O}(1) \cdot \left(s + \frac{n^{1/3}}{p^{1/3}} + \frac{n^{1/3}}{p^{1/3}} \cdot \Gamma\right). \quad (24)$$

To that end, we may invoke a scheme motivated by the peeling technique discussed by [22]. We will show in Proposition 15 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 (22) cannot be satisfied given $\{\mathbf{rk}(\mathbf{X}^{RSAA}) \geq \tilde{p}\}$; 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 (22) a negative quantity, which implies $\mathbb{F}(\mathbf{X}^{RSAA}) - \mathbb{F}(\mathbf{X}^*) < 0$ and contradicts with the fact that \mathbf{X}^* minimizes \mathbb{F} by definition. Since $\{(22) \text{ holds}\} \cap \{\mathbf{rk}(\mathbf{X}^{RSAA}) \geq \tilde{p}\} \supseteq \{\mathbf{rk}(\mathbf{X}) = \tilde{p}\} \cap \{\text{The complement to (20) holds with given } \tilde{p}\}$, it then implies that, for all $\tilde{p} \geq \tilde{p}_u$,

$$0 = \mathbb{P}[\{\mathbf{rk}(\mathbf{X}^{RSAA}) = \tilde{p}\} \cap \{\text{The complement to (20) holds with given } \tilde{p}\}].$$

As an immediate result, $\mathbb{P}[\mathbf{rk}(\mathbf{X}) = \tilde{p}] \leq \mathbb{P}[\{(20) \text{ holds with given } \tilde{p}\}]$ for all $\tilde{p} : \tilde{p} \geq \tilde{p}_u$. Therefore, invoking union bound and De Morgan's law, $\mathbb{P}[\mathbf{rk}(\mathbf{X}) \leq \tilde{p}_u - 1] \geq 1 - \sum_{\tilde{p}=\tilde{p}_u}^p \mathbb{P}[\mathbf{rk}(\mathbf{X}) = \tilde{p}] \geq 1 - \sum_{\tilde{p}=\tilde{p}_u}^p \mathbb{P}[\{(20) \text{ holds with given } \tilde{p}\}]$. By our choice of parameters for t and ϵ as in Step 2, the **Observation** (\star) (which is defined in Step 2) leads to a simplification of the probability bound by noting $\sum_{\tilde{p}=\tilde{p}_u}^p \mathbb{P}[\{(20) \text{ holds with given } \tilde{p}\}]$ 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 Theorem 3 after some algebraic simplification.

Step 5. *To show Corollaries 6 and 8.* Both corollaries can be shown by noticing that $\mathbf{X}_\lambda^{\ell_1}$ yields a suboptimality gap of no more than $\tilde{O}(1) \cdot \lambda \cdot s \cdot R$ when we choose $\lambda = \tilde{O}(1) \cdot \frac{p^{1/3} \sqrt{\ln(np)}}{n^{1/3} s^{-\rho/2}}$ in (6) (which share the same λ value as in (5)). Specifically, Corollary 6 is shown with $\rho = 0$ and Corollary 8 is shown with $\rho = 2/3$.

4. Conclusions

This paper proposes a regularized SAA (RSAA), which is incorporates a low-rankness-exploiting regularization into the traditional SAA framework, 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 at least quadratic in p . The reduced 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 $z \in \mathcal{W}$. Our RSAA theory also implies that, under the low-rankness assump-

tion, 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

The proof follows the argument of Proposition 1 in [11] and makes important generalizations from handling sparsity to low-rankness. Furthermore, much more flexible choices of penalty parameters λ is enabled. We follow the same set of notations in Proposition 16 in defining \tilde{p}_u , ϵ , and $\Delta_1(\epsilon) := \ln\left(\frac{18pR \cdot (K_C + C_\mu)}{\epsilon}\right)$. Furthermore, we will let $\epsilon := \frac{1}{n^{1/3}}$ and $\tilde{\Delta} := \ln(18 \cdot R \cdot (K_C + C_\mu))$. Then $\Delta_1(\epsilon) = \ln\left(\frac{18 \cdot (K_C + C_\mu) \cdot p \cdot R}{\epsilon}\right) = \ln(n^{1/3}p) + \tilde{\Delta} > 0$ and $\lambda = \sqrt{\frac{8 \cdot s^{-\rho} \cdot K \cdot (2p+1)^{2/3} \cdot \Delta_1(\epsilon)}{c \cdot a \cdot n^{2/3}}} = \sqrt{\frac{8 \cdot s^{-\rho} \cdot K \cdot (2p+1)^{2/3}}{c \cdot a \cdot n^{2/3}}} [\ln(n^{1/3}p) + \tilde{\Delta}]$. We will denote by $O(1)$'s universal constants, which may be different in each of their occurrence.

To show the desired results, it suffices to simplify the results in Proposition 16. We will first derive an explicit form for \tilde{p}_u . To that end, we let $P_X := \tilde{p}_u$ and $T_1 := 2P_\lambda(a\lambda) - \frac{8K \cdot (2p+1)}{cn} \Delta_1(\epsilon)$. We then solve the following inequality, which is equivalent to (44) of Proposition 16, for a feasible P_X ,

$$\frac{T_1}{2} \cdot P_X - \frac{2K}{\sqrt{n}} \sqrt{\frac{2P_X \cdot (2p+1) \Delta_1(\epsilon)}{c}} > \Gamma + 2\epsilon + sP_\lambda(a\lambda), \quad (25)$$

for the same $c \in (0, 0.5]$ in (10). Solving the above inequality in terms of P_X , we have $\sqrt{P_X} > \frac{2K}{T_1 \sqrt{n}} \sqrt{\frac{2(2p+1) \cdot \Delta_1(\epsilon)}{c}} + \frac{\sqrt{\frac{2(2K)^2 \cdot (2p+1) \cdot \Delta_1(\epsilon)}{cn} + 2T_1[\Gamma + 2\epsilon + sP_\lambda(a\lambda)]}}{T_1}$. To find a feasible P_X , we may as well let $P_X > \frac{32K^2 \cdot (2p+1) \cdot \Delta_1(\epsilon)}{cT_1^2 \cdot n} + 8T_1^{-1}[\Gamma + 2\epsilon + sP_\lambda(a\lambda)]$. For $\lambda = \sqrt{\frac{8K \cdot s^{-\rho} \cdot \Delta_1(\epsilon) \cdot (2p+1)^{2/3}}{c \cdot a \cdot n^{2/3}}} = \sqrt{\frac{8K \cdot s^{-\rho} \cdot (2p+1)^{2/3}}{c \cdot a \cdot n^{2/3}}} [\ln(n^{1/3}p) + \tilde{\Delta}]$ with $\tilde{\Delta} := \ln(18 \cdot R \cdot (K_C + C_\mu))$, we have $P_\lambda(a\lambda) = \frac{a\lambda^2}{2} = \frac{4K \cdot s^{-\rho} \cdot (2p+1)^{2/3}}{c \cdot n^{2/3}} \cdot \Delta_1(\epsilon)$. Furthermore, $2P_\lambda(a\lambda) = \frac{8K \cdot s^{-\rho} \cdot (2p+1)^{2/3} \cdot \Delta_1(\epsilon)}{c \cdot n^{2/3}} > \frac{4 \cdot s^{-\rho} K \cdot \Delta_1(\epsilon) \cdot (2p+1)^{2/3}}{c \cdot n^{2/3}} + \frac{8K \cdot (2p+1)}{nc} \Delta_1(\epsilon)$ as per our assumption (i.e., (13) implies that $n^{1/3} > 2s^\rho$). Therefore, $T_1 = 2P_\lambda(a\lambda) - \frac{8K \cdot (2p+1)}{nc} \Delta_1(\epsilon) > \frac{4K \cdot s^{-\rho} \cdot \Delta_1(\epsilon) \cdot (2p+1)^{2/3}}{c \cdot n^{2/3}}$. Hence, if we recall $\epsilon = n^{-1/3}$, to satisfy (25), it suffices to let P_X be any integer that satisfies $P_X \geq \frac{2cn^{1/3}s^{2\rho}}{\Delta_1(n^{-\frac{1}{3}}) \cdot (2p+1)^{2/3}} + \frac{2cn^{2/3}s^\rho}{K \Delta_1(n^{-\frac{1}{3}}) \cdot (2p+1)^{2/3}} \cdot \left[\Gamma + \frac{2}{n^{1/3}} + sP_\lambda(a\lambda)\right]$, which is satisfied by letting $P_X \geq \tilde{p}_u$ with

$$\tilde{p}_u := \left\lceil \frac{2cn^{1/3}s^{2\rho}}{\Delta_1(n^{-\frac{1}{3}}) \cdot (2p+1)^{1/3}} + \frac{2cn^{2/3}s^\rho}{K \cdot \Delta_1(n^{-\frac{1}{3}}) \cdot (2p+1)^{2/3}} \cdot \left(\Gamma + \frac{2}{n^{1/3}}\right) + 8s \right\rceil. \quad (26)$$

In the meantime, verifiably, $\tilde{p}_u > s$. Since the above is a sufficient to ensure (25), we know that (44) in Proposition 16 holds for any \tilde{p} : $\tilde{p}_u \leq \tilde{p} \leq p$. Due to Proposition 16,

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

$$\begin{aligned} \mathbb{F}(\mathbf{X}^{RSA}) - \mathbb{F}(\mathbf{X}^*) &\leq s \cdot P_\lambda(a\lambda) + \frac{2K}{\sqrt{n}} \sqrt{\frac{2\tilde{p}_u(2p+1)}{c} \Delta_1(n^{-\frac{1}{3}})} \\ &\quad + \frac{4K}{n} \frac{\tilde{p}_u(2p+1)}{c} \Delta_1(n^{-\frac{1}{3}}) + 2\epsilon + \Gamma, \end{aligned} \quad (27)$$

in which \tilde{p}_u is as per (26).

The following simplifies the formula while seeking to preserve the rates in n and p . Firstly, we have

$$\begin{aligned} &\sqrt{\frac{2\tilde{p}_u \cdot (2p+1)}{cn} \Delta_1(n^{-\frac{1}{3}})} \quad (28) \\ &\leq \sqrt{\frac{4 \cdot (2p+1) s^{2\rho}}{cn \cdot (2p+1)^{1/3}} \Delta_1(n^{-\frac{1}{3}}) \cdot \frac{cn^{1/3}}{\Delta_1(n^{-\frac{1}{3}})} + \frac{4cn^{2/3}(2p+1)s^\rho}{K(2p+1)^{2/3}\Delta_1(n^{-\frac{1}{3}})} \left(\Gamma + \frac{2}{n^{1/3}}\right) \cdot \frac{\Delta_1(n^{-\frac{1}{3}})}{cn}} \\ &\quad + \sqrt{\frac{2}{cn} \Delta_1(n^{-\frac{1}{3}}) \cdot (8s+1) \cdot (2p+1)} \\ &\leq \sqrt{\frac{4(2p+1)^{2/3}s^{2\rho}}{n^{2/3}} + \frac{4s^\rho \cdot (\Gamma + \frac{2}{n^{1/3}}) \cdot (2p+1)^{1/3}}{Kn^{1/3}}} + \sqrt{\frac{2}{nc} \Delta_1(n^{-\frac{1}{3}}) \cdot (8s+1) \cdot (2p+1)}, \end{aligned} \quad (29)$$

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

Similar to the above, we obtain

$$\begin{aligned} &\frac{3\tilde{p}_u \cdot (2p+1)}{cn} \Delta_1(n^{-\frac{1}{3}}) \\ &\leq \frac{4 \cdot (2p+1)^{2/3} s^{2\rho}}{n^{2/3}} + \frac{2}{nc} \Delta_1(n^{-\frac{1}{3}}) (8s+1) \cdot (2p+1) + \frac{4 \cdot s^\rho \cdot (\Gamma + \frac{2}{n^{1/3}})}{K \cdot n^{1/3}} \cdot (2p+1)^{1/3}. \end{aligned} \quad (30)$$

Since (13) and $\Delta_1(n^{-\frac{1}{3}}) = \ln(np) + \tilde{\Delta}$, we have $\frac{4(2p+1)^{2/3}s^{2\rho}}{n^{2/3}} + \frac{4(\Gamma + \frac{2}{n^{1/3}}) \cdot (2p+1)^{1/3}s^\rho}{Kn^{1/3}} \leq O(1)$ and $\frac{2}{nc} \Delta_1(n^{-\frac{1}{3}}) [8s+1] \cdot (2p+1) \leq O(1)$. Therefore, it holds that $\frac{2\tilde{p}_u}{cn} \Delta_1(n^{-\frac{1}{3}}) (2p+1) \leq O(1) \cdot \sqrt{\frac{(2p+1)^{2/3}s^{2\rho}}{n^{2/3}} + \frac{(\Gamma + \frac{2}{n^{1/3}}) \cdot (2p+1)^{1/3}s^\rho}{Kn^{1/3}}} + O(1) \cdot \sqrt{\frac{\Delta_1(n^{-\frac{1}{3}})}{nc} \cdot (8s+1) \cdot (2p+1)}$. Combining the above with (29) and (30), the inequality in (27) can be simplified into $\mathbb{F}(\mathbf{X}^{RSA}) - \mathbb{F}(\mathbf{X}^*) \leq O(1)s^{1-\rho} \cdot \frac{K \cdot \Delta_1(n^{-\frac{1}{3}}) \cdot p^{2/3}}{c \cdot n^{2/3}} + O(1) \cdot K \cdot \sqrt{\frac{p^{2/3}s^{2\rho}}{n^{2/3}} + \frac{(\Gamma + \frac{2}{n^{1/3}}) \cdot p^{1/3}s^\rho}{Kn^{1/3}}} + O(1) \cdot K \sqrt{\frac{sp}{nc} \Delta_1(n^{-\frac{1}{3}})} +$

$\frac{2}{n^{1/3}} + \Gamma$. Together with $\Delta_1(n^{-\frac{1}{3}}) \geq \ln 2$, $K \geq 1$, and $0 < c \leq 0.5$, the above becomes

$$\begin{aligned} \mathbb{F}(\mathbf{X}^{RSAA}) - \mathbb{F}(\mathbf{X}^*) &\leq O(1) \cdot \left(\frac{s^{1-\rho} \cdot \Delta_1(n^{-1/3}) \cdot p^{2/3}}{n^{2/3}} + \frac{p^{1/3} \cdot s^\rho}{n^{1/3}} + \sqrt{\frac{s \cdot p \cdot \Delta_1(n^{-1/3})}{n}} \right) \cdot K \\ &\quad + O(1) \cdot \sqrt{\frac{K \cdot s^\rho \cdot p^{1/3} \cdot \Gamma}{n^{1/3}}} + \Gamma, \quad (31) \end{aligned}$$

which then shows Theorem 3 since $\Delta_1(n^{-\frac{1}{3}}) := \ln(18n^{1/3}(K_C + \mathcal{C}_\mu) \cdot p \cdot R)$. \blacksquare

A.2 Proof of Corollary 6

Lemma 19 implies that $\mathcal{F}_{n,\lambda}(\mathbf{X}^{RSAA}, \mathbf{Z}_1^n) \leq \mathcal{F}_{n,\lambda}(\mathbf{X}^*, \mathbf{Z}_1^n) + \lambda \|\mathbf{X}^*\|_*$ almost surely. Below we invoke the results from Theorem 3 with $\Gamma = \lambda \|\mathbf{X}^*\|_*$ and assumption that $\rho = 0$ and $\lambda = \lambda(0)$. Note that it is assumed that

$$n > C_2 \cdot p \cdot \mathcal{U}_L \cdot [\ln(np) + \tilde{\Delta}] \cdot s^{3/2} R^{3/2} > O(1) \cdot p \cdot a^{-1} \cdot [\ln(np) + \tilde{\Delta}] \cdot s^{3/2} R^{3/2}, \quad (32)$$

and $\frac{\Gamma}{K} \leq \frac{\lambda \|\mathbf{X}^*\|_*}{K} \leq \frac{\|\mathbf{X}^*\|_* \cdot \sqrt{\frac{8K \cdot (2p+1)^{2/3}}{c \cdot a \cdot n^{2/3}} [\ln(n^{1/3}p) + \tilde{\Delta}]}}{K}$ (as well as $K \geq 1$). In view of (32), it then holds under Assumption 1 that $\frac{\Gamma}{K} \leq Rs \cdot \sqrt{\frac{8(2p+1)^{2/3}}{cK \cdot a \cdot n^{2/3}} [\ln(n^{1/3}p) + \tilde{\Delta}]} \leq O(1) \cdot \sqrt{\frac{Rs}{a^{1/3}} [\ln(n^{1/3}p) + \tilde{\Delta}]^{1/3}}$. Therefore, $(\frac{\Gamma}{K})^3 \leq \left(O(1) \cdot \sqrt{\frac{Rs}{a^{1/3}} [\ln(n^{1/3}p) + \tilde{\Delta}]^{1/3}} \right)^3 \leq O(1) \cdot R^{3/2} s^{3/2} \sqrt{a^{-1} \cdot [\ln(n^{1/3}p) + \tilde{\Delta}]}$, for some universal constants $O(1)$. Furthermore, since $a < \mathcal{U}_L^{-1} \leq 1$, it holds that, if n satisfies (15) for some universal constant C_2 , then $n > O(1) \cdot p \cdot a^{-1} \cdot [\ln(n^{1/3}p) + \tilde{\Delta}] \cdot s^{3/2} R^{3/2} \geq O(1) \cdot p \cdot R^{3/2} s^{3/2} \sqrt{a^{-1} \cdot [\ln(n^{1/3}p) + \tilde{\Delta}]} + O(1) \cdot p + C_1 \cdot s \cdot p \cdot (\ln(n^{1/3}p) + \tilde{\Delta}) \geq C_1 \cdot \left[\left(\frac{\Gamma}{K} \right)^3 p + p + s \cdot p \cdot (\ln(n^{1/3}p) + \tilde{\Delta}) \right]$. Therefore, Theorem 3 is met and thus (14) in Theorem 3 implies that

$$\begin{aligned} \mathbb{F}(\mathbf{X}^{RSAA}) - \mathbb{F}(\mathbf{X}^*) &\leq O(1) \cdot K \cdot \left(\frac{sp^{2/3} \Delta_1(n^{-1/3})}{n^{2/3}} + \sqrt{\frac{sp \Delta_1(n^{-\frac{1}{3}})}{n}} + \frac{p^{1/3}}{n^{1/3}} \right) \\ &\quad + O(1) \cdot \sqrt{\frac{K p^{1/3} (\lambda \|\mathbf{X}^*\|_*)}{n^{1/3}}} + \lambda \|\mathbf{X}^*\|_*, \end{aligned}$$

with probability at least $1 - 2(2p+1) \exp(-\tilde{c}n) - 6 \exp(-2cn^{1/3} \cdot (2p+1)^{2/3})$. Note that $a < 1$, $K \geq 1$, $p \geq 1$, $[\ln(n^{1/3}p) + \tilde{\Delta}] \geq 1$ and $\sqrt{\frac{sp \Delta_1(n^{-\frac{1}{3}})}{n}} \leq \frac{s(2p+1)^{1/3} \cdot \sqrt{\Delta_1(n^{-\frac{1}{3}})}}{n^{1/3}}$ (due to (15) again). Hence, $\mathbb{F}(\mathbf{X}^{RSAA}) - \mathbb{F}(\mathbf{X}^*) \leq O(1) \cdot K \cdot \left[\frac{sp^{2/3} \cdot (\ln(np) + \tilde{\Delta})}{n^{2/3}} + \frac{p^{1/3}}{n^{1/3}} \right] + O(1) \cdot \frac{sRK \cdot (2p+1)^{1/3}}{\min\{a^{1/2}n^{1/3}, a^{1/4}n^{1/3}\}} [\ln(n^{1/3}p) + \tilde{\Delta}]^{1/2}$, which shows Part (ii) by further noticing that $a = \frac{1}{2\mathcal{U}_L}$ and $\mathcal{U}_L \geq 1$. \blacksquare

A.3 Proof of Corollary 8

The proof follows almost the same argument as in Section A.2 for proving Corollary 6, except that the choice of user-specific parameters are different. Again, Lemma 19 implies that $\mathcal{F}_{n,\lambda}(\mathbf{X}^{RSAA}, \mathbf{Z}_1^n) \leq \mathcal{F}_{n,\lambda}(\mathbf{X}^*, \mathbf{Z}_1^n) + \lambda \|\mathbf{X}^*\|_*$ almost surely. As the same in Part (ii), below we invoke the results from Theorem 3 with $\Gamma = \lambda \|\mathbf{X}^*\|_*$ and assumption that $\rho = 2/3$ and $\lambda = \lambda(\frac{2}{3})$. Note that it is assumed that

$$n > C_3 \cdot p \cdot \mathcal{U}_L \cdot [\ln(np) + \tilde{\Delta}] \cdot s^2 R^{3/2} > O(1) \cdot p \cdot a^{-1} \cdot [\ln(np) + \tilde{\Delta}] \cdot s^2 R^{3/2}, \quad (33)$$

and $\frac{\Gamma}{K} \leq \frac{\lambda \|\mathbf{X}^*\|_*}{K} \leq \frac{\|\mathbf{X}^*\|_* \cdot \sqrt{\frac{8K \cdot (2p+1)^{2/3} \cdot s^{-2/3}}{c \cdot a \cdot n^{2/3}} [\ln(n^{1/3}p) + \tilde{\Delta}]}}{K}$ (as well as $K \geq 1$). In view of (33), it then holds under Assumption 1 that $\frac{\Gamma}{K} \leq Rs \cdot \sqrt{\frac{8(2p+1)^{2/3} s^{-2/3}}{cK \cdot a \cdot n^{2/3}} [\ln(n^{1/3}p) + \tilde{\Delta}]} \leq O(1) \cdot \sqrt{\frac{R}{a^{1/3}} [\ln(n^{1/3}p) + \tilde{\Delta}]^{1/3}}$. Therefore, $(\frac{\Gamma}{K})^3 \leq \left(O(1) \cdot \sqrt{\frac{R}{a^{1/3}} [\ln(n^{1/3}p) + \tilde{\Delta}]^{1/3}}\right)^3 \leq O(1) \cdot R^{3/2} \sqrt{a^{-1} \cdot [\ln(n^{1/3}p) + \tilde{\Delta}]}$, for some universal constants $O(1)$. Furthermore, since $a < \mathcal{U}_L^{-1} \leq 1$, it holds that, if n satisfies (18), then $n > O(1) \cdot p \cdot a^{-1} \cdot [\ln(n^{1/3}p) + \tilde{\Delta}] \cdot s^2 R^{3/2} \geq O(1) \cdot p \cdot R^{3/2} s^2 \sqrt{a^{-1} \cdot [\ln(n^{1/3}p) + \tilde{\Delta}]} + O(1) \cdot s^2 \cdot p + C_1 s \cdot p \cdot (\ln(n^{1/3}p) + \tilde{\Delta}) \geq C_1 \cdot \left[s^2 \left(\frac{\Gamma}{K}\right)^3 p + s^2 \cdot p + s \cdot p \cdot (\ln(n^{1/3}p) + \tilde{\Delta})\right]$. Therefore, (13) in Theorem 3 is met and thus (14) in Theorem 3 implies that

$$\begin{aligned} \mathbb{F}(\mathbf{X}^{RSAA}) - \mathbb{F}(\mathbf{X}^*) &\leq O(1) \cdot K \cdot \left(\frac{s^{1/3} p^{2/3} \Delta_1(n^{-1/3})}{n^{2/3}} + \sqrt{\frac{sp \Delta_1(n^{-1/3})}{n}} + \frac{p^{1/3} \cdot s^{2/3}}{n^{1/3}} \right) \\ &\quad + O(1) \cdot \sqrt{\frac{K p^{1/3} \cdot s^{2/3} \cdot (\lambda \|\mathbf{X}^*\|_*)}{n^{1/3}}} + \lambda \|\mathbf{X}^*\|_*, \end{aligned}$$

with probability at least $1 - 2(2p+1) \exp(-\tilde{c}n) - 6 \exp(-2cn^{1/3} \cdot (2p+1)^{2/3})$. Note that $a < 1$, $K \geq 1$, $p \geq s \geq 1$, $[\ln(n^{1/3}p) + \tilde{\Delta}] \geq 1$ and $\sqrt{\frac{sp \Delta_1(n^{-1/3})}{n}} \leq \frac{(2p+1)^{1/3} \cdot \sqrt{\Delta_1(n^{-1/3})}}{n^{1/3}}$ (in view of (18) again). Hence, $\mathbb{F}(\mathbf{X}^{RSAA}) - \mathbb{F}(\mathbf{X}^*) \leq O(1) \cdot K \cdot \left[\frac{s^{1/3} p^{2/3} \cdot (\ln(np) + \tilde{\Delta})}{n^{2/3}} + \frac{s^{2/3} \cdot p^{1/3}}{n^{1/3}} \right] + O(1) \cdot \frac{s^{2/3} R K \cdot (2p+1)^{1/3}}{\min\{a^{1/2} n^{1/3}, a^{1/4} n^{1/3}\}} [\ln(n^{1/3}p) + \tilde{\Delta}]^{1/2}$, which shows Part (iii) by further noticing that $a = \frac{1}{2\mathcal{U}_L}$. \blacksquare

A.4 Auxiliary results

Proposition 13 *Suppose that $a < \mathcal{U}_L^{-1}$. Assume that the $S^3 ONC(\mathbf{Z}_1^n)$ is satisfied almost surely at $\mathbf{X}^{RSAA} \in \mathcal{S}_p$. Then,*

$$\mathbb{P}[\{|\sigma_j(\mathbf{X}^{RSAA})| \notin (0, a\lambda) \text{ for all } j\}] = 1.$$

Proof Since \mathbf{X}^{RSAA} satisfies the $S^3\text{ONC}(\mathbf{Z}_1^n)$ almost surely, Eq. (12) implies that for any $j \in \{1, \dots, p\}$, if $\sigma_j(\mathbf{X}^{RSAA}) \in (0, a\lambda)$, then

$$0 \leq U_L + \left[\frac{\partial^2 P_\lambda(|\sigma_j(\mathbf{X})|)}{[\partial \sigma_j(\mathbf{X})]^2} \right]_{\mathbf{X}=\mathbf{X}^{RSAA}} = U_L - \frac{1}{a}. \quad (34)$$

Further observe that $\frac{\partial^2 P_\lambda(t)}{\partial t^2} = -a^{-1}$ for $t \in (0, a\lambda)$. Therefore, (34) contradicts with the assumption that $U_L < \frac{1}{a}$. This contradiction implies that

$$\begin{aligned} & \mathbb{P}[\{\mathbf{X}^{RSAA} \text{ satisfies the } S^3\text{ONC}(\mathbf{Z}_1^n)\} \cap \{|\sigma_j(\mathbf{X}^{RSAA})| \in (0, a\lambda)\}] = 0 \\ \implies & 0 \geq 1 - \mathbb{P}[\{\mathbf{X}^{RSAA} \text{ does not satisfy the } S^3\text{ONC}(\mathbf{Z}_1^n)\}] - \mathbb{P}[|\sigma_j(\mathbf{X}^{RSAA})| \notin (0, a\lambda)]. \end{aligned}$$

Since $\mathbb{P}[\{\mathbf{X}^{RSAA} \text{ satisfies the } S^3\text{ONC}(\mathbf{Z}_1^n)\}] = 1$, it holds that $\mathbb{P}[|\sigma_j(\mathbf{X}^{RSAA})| \notin (0, a\lambda)] = 1$ for all $j = 1, \dots, n$, which immediately leads to the desired result. \blacksquare

Proposition 14 Suppose that Assumptions 3 and 4 hold. Let $\epsilon \in (0, 1]$, $\tilde{p} : \tilde{p} > s$, $\Delta_1(\epsilon) := \ln\left(\frac{18 \cdot (K_C + C_\mu) \cdot p \cdot R}{\epsilon}\right)$, and $\mathcal{B}_{\tilde{p}, R} := \{\mathbf{X} \in \mathcal{S}_p : \sigma_{\max}(\mathbf{X}) \leq R, \mathbf{rk}(\mathbf{X}) \leq \tilde{p}\}$. Then, for the same $c \in (0, 0.5]$ as in (10) and for some $\tilde{c} > 0$,

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

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

Proof We will follow the “ ϵ -net” argument similar to Shapiro et al. [27] to construct a net of discretization grids $\mathcal{G}(\epsilon) := \{\tilde{\mathbf{X}}^k\} \subseteq \mathcal{B}_{\tilde{p}, R}$ such that for any $\mathbf{X} \in \mathcal{B}_{\tilde{p}, R}$, there is $\tilde{\mathbf{X}}^k \in \mathcal{G}(\epsilon)$ that satisfies $\|\mathbf{X}^k - \mathbf{X}\| \leq \frac{\epsilon}{2K_C + 2C_\mu}$ for any fixed $\epsilon \in (0, 1]$.

Invoking Lemma 20, for an arbitrary $\mathbf{X} \in \mathcal{B}_{\tilde{p}, R}$, to ensure that there always exists $\tilde{\mathbf{X}}^k \in \mathcal{G}(\epsilon)$ that ensures $\|\mathbf{X} - \tilde{\mathbf{X}}^k\| \leq \frac{\epsilon}{(2K_C + 2C_\mu)}$, it is sufficient to have the number of grids to be no more than $\left(\frac{18R\sqrt{\tilde{p}} \cdot (K_C + C_\mu)}{\epsilon}\right)^{(2p+1)\tilde{p}}$. Now, we may observe

$$\begin{aligned} & \mathbb{P} \left[\max_{\mathbf{X}^k \in \mathcal{G}(\epsilon)} \left| \frac{1}{n} \sum_{i=1}^n f(\mathbf{X}^k, Z_i) - \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n f(\mathbf{X}^k, Z_i) \right] \right| \leq K \sqrt{\frac{t}{n}} + \frac{Kt}{n} \right] \\ = & \mathbb{P} \left[\bigcap_{\mathbf{X}^k \in \mathcal{G}(\epsilon)} \left\{ \left| \frac{1}{n} \sum_{i=1}^n f(\mathbf{X}^k, Z_i) - \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n f(\mathbf{X}^k, Z_i) \right] \right| \leq K \sqrt{\frac{t}{n}} + \frac{Kt}{n} \right\} \right] \\ \geq & 1 - \sum_{\mathbf{X}^k \in \mathcal{G}(\epsilon)} \mathbb{P} \left[\left| \frac{1}{n} \sum_{i=1}^n f(\mathbf{X}^k, Z_i) - \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n f(\mathbf{X}^k, Z_i) \right] \right| > K \sqrt{\frac{t}{n}} + \frac{Kt}{n} \right]. \quad (35) \end{aligned}$$

Further invoking Eq. (10), for the same c as in (10), it holds that

$$\begin{aligned} & \mathbb{P} \left[\max_{\mathbf{X}^k \in \mathcal{G}(\epsilon)} \left| \frac{1}{n} \sum_{i=1}^n f(\mathbf{X}^k, Z_i) - \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n f(\mathbf{X}^k, Z_i) \right] \right| \leq K \sqrt{\frac{t}{n}} + \frac{Kt}{n} \right] \\ & \geq 1 - |\mathcal{G}(\epsilon)| \cdot 2 \exp(-ct) \geq 1 - 2 \left(\frac{18R\sqrt{\tilde{p}} \cdot (K_C + \mathcal{C}_\mu)}{\epsilon} \right)^{(2p+1)\tilde{p}} \cdot \exp(-ct). \end{aligned}$$

Combined with Lemma 17 and Lemma 18,

$$\begin{aligned} & \max_{\mathbf{X} \in \mathcal{B}_{\tilde{p}, R}, \mathbf{X}^k \in \mathcal{G}(\epsilon)} \left\{ \left| \frac{1}{n} \sum_{i=1}^n f(\mathbf{X}, Z_i) - \frac{1}{n} \sum_{i=1}^n f(\mathbf{X}^k, Z_i) \right| \right. \\ & \quad \left. + \left| \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n f(\mathbf{X}, Z_i) \right] - \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n f(\mathbf{X}^k, Z_i) \right] \right| \right\} \\ & \leq 2(K_C + \mathcal{C}_\mu) \cdot \frac{\epsilon}{2K_C + 2\mathcal{C}_\mu} = \epsilon, \quad (36) \end{aligned}$$

with probability at least $1 - 2 \exp(-\tilde{c} \cdot n)$ for some problem independent $\tilde{c} > 0$ and any fixed $\tau > 0$. Observe that for any $\mathbf{X} \in \mathcal{B}_{\tilde{p}, R}$ and $\mathbf{X}^k \in \mathcal{G}(\epsilon)$, it holds that $|\mathcal{F}_n(\mathbf{X}, \mathbf{Z}_1^n) - \mathbb{E}[\mathcal{F}_n(\mathbf{X}, \mathbf{Z}_1^n)]| \leq |\mathcal{F}_n(\mathbf{X}^k, \mathbf{Z}_1^n) - \mathbb{E}[\mathcal{F}_n(\mathbf{X}^k, \mathbf{Z}_1^n)]| + |\mathcal{F}_n(\mathbf{X}, \mathbf{Z}_1^n) - \mathcal{F}_n(\mathbf{X}^k, \mathbf{Z}_1^n)| + |\mathbb{E}[\mathcal{F}_n(\mathbf{X}, \mathbf{Z}_1^n)] - \mathbb{E}[\mathcal{F}_n(\mathbf{X}^k, \mathbf{Z}_1^n)]|$. Therefore, with probability at least $1 - 2 \exp(-\tilde{c} \cdot n)$ for some positive constant $\tilde{c} > 0$,

$$\max_{\mathbf{X} \in \mathcal{B}_{\tilde{p}, R}, \mathbf{X}^k \in \mathcal{G}(\epsilon)} \left\{ |\mathcal{F}_n(\mathbf{X}, \mathbf{Z}_1^n) - \mathbb{E}[\mathcal{F}_n(\mathbf{X}, \mathbf{Z}_1^n)]| - |\mathcal{F}_n(\mathbf{X}^k, \mathbf{Z}_1^n) - \mathbb{E}[\mathcal{F}_n(\mathbf{X}^k, \mathbf{Z}_1^n)]| \right\} \leq \epsilon. \quad (37)$$

Further invoking (35), we now obtain that

$$\max_{\mathbf{X} \in \mathcal{B}_{\tilde{p}, R}, \mathbf{X}^k \in \mathcal{G}(\epsilon)} |\mathcal{F}_n(\mathbf{X}, \mathbf{Z}_1^n) - \mathbb{F}(\mathbf{X})| \leq \epsilon + K \sqrt{\frac{t}{n}} + \frac{Kt}{n},$$

with probability at least $1 - 2 \left(\frac{18R\sqrt{\tilde{p}} \cdot (K_C + \mathcal{C}_\mu)}{\epsilon} \right)^{(2p+1)\tilde{p}} \cdot \exp(-ct) - 2 \exp(-\tilde{c} \cdot n)$. Finally, we may let $t := \frac{2\tilde{p}}{c} \cdot (2p+1) \cdot \Delta_1(\epsilon)$, where $\Delta_1(\epsilon) := \ln \left(\frac{18 \cdot (K_C + \mathcal{C}_\mu) \cdot p \cdot R}{\epsilon} \right)$, and obtain the desired result. \blacksquare

Proposition 15 *Suppose that Assumptions 1 through 3 hold, the solution $\mathbf{X}^{RSAA} \in \mathcal{S}_p : \sigma_{\max}(\mathbf{X}^{RSAA}) \leq R$ satisfies $S^3 \text{ONC}(\mathbf{Z}_1^n)$ almost surely,*

$$\mathcal{F}_{n,\lambda}(\mathbf{X}^{RSAA}, \mathbf{Z}_1^n) \leq \mathcal{F}_{n,\lambda}(\mathbf{X}^*, \mathbf{Z}_1^n) + \Gamma, \quad w.p.1. \quad (38)$$

where $\Gamma \geq 0$, $\epsilon \in (0, 1]$, $\Delta_1(\epsilon) := \ln \left(\frac{18 \cdot (K_C + C_\mu) \cdot p \cdot R}{\epsilon} \right)$. For a positive integer $\tilde{p}_u : \tilde{p}_u > s$, if

$$(\hat{p} - s) \cdot P_\lambda(a\lambda) > \frac{4K}{cn} \Delta_1(\epsilon) \cdot \hat{p} \cdot (2p + 1) + \frac{2K}{\sqrt{n}} \sqrt{\frac{2\hat{p} \cdot (2p + 1)}{c}} \Delta_1(\epsilon) + \Gamma + 2\epsilon, \quad (39)$$

for all $\hat{p} : \tilde{p}_u \leq \hat{p} \leq p$, then $\mathbb{P}[\mathbf{rk}(\mathbf{X}^{RSAA}) \leq \tilde{p}_u - 1] \geq 1 - 2p \exp(-\tilde{c}n) - 4 \exp(-\tilde{p}_u(2p + 1)\Delta_1(\epsilon))$ for the same c in (10) and some $\tilde{c} > 0$.

Proof This proof generalizes Proposition EC.3 from [11] bounding the sparsity of an S³ONC solution to bounding the rank of an S³ONC solution. Though the argument is similar, details are quite different and thus the result is different. Define $\mathcal{B}_R := \{\mathbf{X} \in \mathcal{S}_p : \sigma_{\max}(\mathbf{X}) \leq R\}$. Define a few events:

$$\begin{aligned} \mathcal{E}_1 &:= \left\{ (\tilde{\mathbf{X}}, \tilde{\mathbf{Z}}_1^n) \in \mathcal{B}_R \times \mathcal{W}^n : \mathcal{F}_{n,\lambda}(\tilde{\mathbf{X}}, \tilde{\mathbf{Z}}_1^n) \leq \mathcal{F}_{n,\lambda}(\mathbf{X}^*, \tilde{\mathbf{Z}}_1^n) + \Gamma \right\}, \\ \mathcal{E}_2 &:= \left\{ \tilde{\mathbf{X}} \in \mathcal{B}_R : |\sigma_j(\tilde{\mathbf{X}})| \notin (0, a\lambda) \text{ for all } j \right\}, \\ \mathcal{E}_{3,\hat{p}} &:= \left\{ \tilde{\mathbf{X}} \in \mathcal{B}_R : \mathbf{rk}(\tilde{\mathbf{X}}) = \hat{p} \right\}, \end{aligned}$$

where c in $\mathcal{E}_{5,\hat{p}}$ is a universal constant defined to be the same as in (10), $\hat{p} : \tilde{p}_u \leq \hat{p} \leq p$ and (thus $\hat{p} > s$ by the assumption that $\tilde{p}_u > s$). For any $(\tilde{\mathbf{X}}, \tilde{\mathbf{Z}}_1^n) \in \{(\tilde{\mathbf{X}}, \tilde{\mathbf{Z}}_1^n) \in \mathcal{E}_1\} \cap \{\tilde{\mathbf{X}} \in \mathcal{E}_2 \cap \mathcal{E}_{3,\hat{p}}\}$, where $\tilde{\mathbf{Z}}_1^n = (\tilde{Z}_1, \dots, \tilde{Z}_n)$, since $\tilde{\mathbf{X}} \in \mathcal{E}_{3,\hat{p}} \cap \mathcal{E}_2$, which means that $\tilde{\mathbf{X}}$ has \hat{p} -many non-zero singular values and each must not be within the interval $(0, a\lambda)$, it holds that

$$\mathcal{F}_n(\tilde{\mathbf{X}}, \tilde{\mathbf{Z}}_1^n) + \hat{p}P_\lambda(a\lambda) \leq \frac{1}{n}\mathcal{F}_n(\mathbf{X}^*, \tilde{\mathbf{Z}}_1^n) + sP_\lambda(a\lambda) + \Gamma, \quad (40)$$

Notice that $\mathbf{X}^* \in \mathcal{B}_R : \mathbf{rk}(\mathbf{X}^*) = s < \hat{p}$ by Assumption 1. We may obtain that, for all $\tilde{\mathbf{X}} \in \mathcal{E}_{3,\hat{p}}$,

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n f(\mathbf{X}^*, \tilde{Z}_i) - \frac{1}{n} \sum_{i=1}^n f(\tilde{\mathbf{X}}, \tilde{Z}_i) \\ &= \left[\frac{1}{n} \sum_{i=1}^n f(\mathbf{X}^*, \tilde{Z}_i) - \mathbb{F}(\mathbf{X}^*) \right] + \left[\mathbb{F}(\tilde{\mathbf{X}}) - \frac{1}{n} \sum_{i=1}^n f(\tilde{\mathbf{X}}, \tilde{Z}_i) \right] + \left[\mathbb{F}(\mathbf{X}^*) - \mathbb{F}(\tilde{\mathbf{X}}) \right] \\ &\leq 2 \max_{\mathbf{X} \in \mathcal{E}_{3,p}} \left| \frac{1}{n} \sum_{i=1}^n f(\mathbf{X}, \tilde{Z}_i) - \mathbb{F}(\mathbf{X}) \right| + \mathbb{F}(\mathbf{X}^*) - \mathbb{F}(\tilde{\mathbf{X}}) \\ &\leq 2 \max_{\mathbf{X} \in \mathcal{E}_{3,p}} \left| \frac{1}{n} \sum_{i=1}^n f(\mathbf{X}, \tilde{Z}_i) - \mathbb{F}(\mathbf{X}) \right|, \end{aligned} \quad (41)$$

where the last inequality is due to $\mathbb{F}(\mathbf{X}^*) \leq \mathbb{F}(\mathbf{X})$ for all $\mathbf{X} \in \mathcal{S}_p$ by the definition of \mathbf{X}^* . Define that

$$\begin{aligned}\mathcal{E}_4 &:= \left\{ (\tilde{\mathbf{X}}, \tilde{\mathbf{Z}}_1^n) \in \mathcal{B}_R \times \mathcal{W}^n : \tilde{\mathbf{X}} \text{ satisfies } S^3\text{ONC}(\tilde{\mathbf{Z}}_1^n) \right\} \\ \mathcal{E}_{5,\hat{p}} &:= \left\{ \tilde{\mathbf{Z}}_1^n \in \mathcal{W}^n : \max_{\mathbf{X} \in \mathcal{B}_R : \mathbf{rk}(\mathbf{X}) \leq \hat{p}} \left| \mathcal{F}_n(\mathbf{X}, \tilde{\mathbf{Z}}_1^n) - \mathbb{F}(\mathbf{X}) \right| \leq \frac{K}{\sqrt{n}} \sqrt{\frac{2\hat{p}(2p+1)}{c}} \Delta_1(\epsilon) \right. \\ &\quad \left. + \frac{K}{n} \cdot \frac{2\hat{p}(2p+1)}{c} \Delta_1(\epsilon) + \epsilon \right\},\end{aligned}$$

Now let us examine the following set:

$$\Lambda = \{(\tilde{\mathbf{X}}, \tilde{\mathbf{Z}}_1^n) : (\tilde{\mathbf{X}}, \tilde{\mathbf{Z}}_1^n) \in \mathcal{E}_1 \cap \mathcal{E}_4\} \cap \{(\tilde{\mathbf{X}}, \tilde{\mathbf{Z}}_1^n) : \tilde{\mathbf{X}} \in \mathcal{E}_{3,p} \cap \mathcal{E}_2\} \cap \{(\tilde{\mathbf{X}}, \tilde{\mathbf{Z}}_1^n) : \tilde{\mathbf{Z}}_1^n \in \mathcal{E}_{5,\hat{p}}\}.$$

Combined with (40) and (41), $\Lambda \neq \emptyset \implies (\hat{p} - s) \cdot P_\lambda(a\lambda) \leq \frac{2K}{\sqrt{n}} \sqrt{\frac{2\hat{p}(2p+1)}{c}} \Delta_1(\epsilon) + \frac{2K}{n} \cdot \frac{2\hat{p}(2p+1)}{c} \Delta_1(\epsilon) + 2\epsilon + \Gamma$, which contradicts with (39) for all $\hat{p} : \tilde{p}_u \leq \hat{p} \leq p$. Now we recall the definition of $\mathbf{X}^{RSAA} \in \mathcal{B}_R$, which is a solution that satisfies the $S^3\text{ONC}(\mathbf{Z}_1^n)$, w.p.1., and $\mathcal{F}_{n,\lambda}(\mathbf{X}^{RSAA}, \tilde{\mathbf{Z}}_1^n) \leq \mathcal{F}_{n,\lambda}(\mathbf{X}^*, \tilde{\mathbf{Z}}_1^n) + \Gamma$, w.p.1. Invoking Proposition 13, we have $\mathbb{P}[(\mathbf{X}^{RSAA}, \mathbf{Z}_1^n) \in \mathcal{E}_1 \cap \mathcal{E}_4, \mathbf{X}^{RSAA} \in \mathcal{E}_2] = 1$. Hence,

$$\begin{aligned}0 &= \mathbb{P}[\Lambda] \\ &\geq 1 - \mathbb{P}[\mathbf{X}^{RSAA} \notin \mathcal{E}_{3,p}] - \mathbb{P}[\mathbf{Z}_1^n \notin \mathcal{E}_{5,\hat{p}}] - \{1 - \mathbb{P}[(\mathbf{X}^{RSAA}, \mathbf{Z}_1^n) \in \mathcal{E}_1 \cap \mathcal{E}_4, \mathbf{X}^{RSAA} \in \mathcal{E}_2]\},\end{aligned}$$

for all $\hat{p} : \tilde{p}_u \leq \hat{p} \leq p$. The above then implies that $\mathbb{P}[\mathbf{Z}_1^n \notin \mathcal{E}_{5,\hat{p}}] \geq \mathbb{P}[\mathbf{X}^{RSAA} \in \mathcal{E}_{3,p}]$ for all $\hat{p} : \tilde{p}_u \leq \hat{p} \leq p$. Therefore, $\mathbb{P}[\mathbf{rk}(\mathbf{X}^{RSAA}) = \hat{p}] \leq 1 - \mathbb{P}[\mathbf{Z}_1^n \in \mathcal{E}_{5,\hat{p}}]$ for all $\hat{p} : \tilde{p}_u \leq \hat{p} \leq p$. Together with Proposition 14, we have that

$$\begin{aligned}\mathbb{P}[\mathbf{rk}(\mathbf{X}^{RSAA}) \leq \tilde{p}_u - 1] &= \mathbb{P}[\mathbf{rk}(\mathbf{X}^{RSAA}) \notin \{\tilde{p}_u, \tilde{p}_u + 1, \dots, p\}] \\ &= 1 - \mathbb{P}\left[\bigcup_{\hat{p}=\tilde{p}_u}^p \{\mathbf{rk}(\mathbf{X}^{RSAA}) = \hat{p}\}\right] \geq 1 - \sum_{\hat{p}=\tilde{p}_u}^p \mathbb{P}[\mathbf{rk}(\mathbf{X}^{RSAA}) = \hat{p}] \geq 1 - \sum_{\hat{p}=\tilde{p}_u}^p (1 - \mathbb{P}[\mathbf{Z}_1^n \in \mathcal{E}_{5,\hat{p}}]) \\ &\geq 1 - 2(p - \tilde{p}_u + 1) \exp(-\tilde{c}n) - \sum_{\hat{p}=\tilde{p}_u}^p 2 \exp(-\hat{p}(2p+1) \cdot \Delta_1(\epsilon)).\end{aligned}\tag{42}$$

where $\tilde{c} > 0$ is some universal constant. Observing that $\Delta_1(\epsilon) = \ln\left(\frac{18 \cdot (K_C + C_\mu) \cdot p \cdot R}{\epsilon}\right) > 1$ by observing that the above (42) involves a geometric sequence, we have

$$\mathbb{P}[\mathbf{rk}(\mathbf{X}^{RSAA}) \leq \tilde{p}_u - 1] \geq 1 - \frac{2 \exp(-\tilde{p}_u(2p+1)\Delta_1(\epsilon))}{1 - \exp(-(2p+1)\Delta_1(\epsilon))} - 2p \exp(-\tilde{c}n).\tag{43}$$

Further noting that $\frac{2 \exp(-\tilde{p}_u(2p+1)\Delta_1(\epsilon))}{1 - \exp(-(2p+1)\Delta_1(\epsilon))} \leq 4 \exp(-\tilde{p}_u(2p+1)\Delta_1(\epsilon))$, we then have the desired result. \blacksquare

Proposition 16 *Let*

$$\Delta_1(\epsilon) := \ln \left(\frac{18 \cdot (K_C + C_\mu) \cdot p \cdot R}{\epsilon} \right).$$

Assume that (i) the solution \mathbf{X}^{RSAA} satisfies $S^3\text{ONC}(\mathbf{Z}_1^n)$ almost surely; (ii) $\mathcal{F}_{n,\lambda}(\mathbf{X}^{RSAA}, \mathbf{Z}_1^n) \leq \mathcal{F}_{n,\lambda}(\mathbf{X}^, \mathbf{Z}_1^n) + \Gamma$ with probability one; and (iii) for some integer $\tilde{p}_u : \tilde{p}_u > s$, it holds that*

$$\hat{p} > s + \frac{4K \cdot \hat{p} \cdot (2p+1)}{cn \cdot P_\lambda(a\lambda)} \Delta_1(\epsilon) + \frac{2K}{\sqrt{n} \cdot P_\lambda(a\lambda)} \sqrt{\frac{2\hat{p} \cdot (2p+1)}{c} \Delta_1(\epsilon)} + \frac{\Gamma + 2\epsilon}{P_\lambda(a\lambda)}, \quad (44)$$

for all $\tilde{p} : \tilde{p}_u \leq \tilde{p} \leq p$, any $\Gamma \geq 0$, and any $\epsilon \in (0, 1]$. It then holds that

$$\begin{aligned} \mathbb{F}(\mathbf{X}^{RSAA}) - \mathbb{F}(\mathbf{X}^*) &\leq \frac{4K \cdot \hat{p} \cdot (p+1)}{cn} \Delta_1(\epsilon) \\ &\quad + \frac{2K}{\sqrt{n}} \sqrt{\frac{2\hat{p} \cdot (2p+1)}{c} \Delta_1(\epsilon)} + \Gamma + 2\epsilon + sP_\lambda(a\lambda), \end{aligned} \quad (45)$$

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 $0 < \epsilon \leq 1$. By assumption,

$$\mathcal{F}_{n,\lambda}(\mathbf{X}^{RSAA}, \mathbf{Z}_1^n) \leq \mathcal{F}_{n,\lambda}(\mathbf{X}^*, \mathbf{Z}_1^n) + \Gamma,$$

w.p.1., $P_\lambda(t) \geq 0$ for all $t \geq 0$, and $\text{rk}(\mathbf{X}^*) = s$, yields that $\frac{1}{n} \sum_{i=1}^n f(\mathbf{X}^{RSAA}, Z_i) \leq \frac{1}{n} \sum_{i=1}^n f(\mathbf{X}^*, Z_i) + sP_\lambda(a\lambda) + \Gamma$, a.s. Furthermore, conditioning on the events that (a) $\text{rk}(\mathbf{X}^{RSAA}) \leq \tilde{p}_u$, (b) $\max_{\mathbf{X} \in \mathcal{B}_{\tilde{p}_u, R}} \left| \frac{1}{n} \sum_{i=1}^n f(\mathbf{X}, Z_i) - \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n f(\mathbf{X}, Z_i) \right] \right| \leq \frac{K}{\sqrt{n}} \sqrt{\frac{\tilde{p}_u \cdot (2p+1)}{c} \Delta_1(\epsilon)} + \frac{K}{n} \frac{\tilde{p}_u \cdot (2p+1)}{c} \Delta_1(\epsilon) + \epsilon$, we obtain that $\mathbb{F}(\mathbf{X}^{RSAA}) - \mathbb{F}(\mathbf{X}^*) \leq s \cdot P_\lambda(a\lambda) + \frac{2K}{\sqrt{n}} \sqrt{\frac{2\tilde{p}_u \cdot (2p+1)}{c} \Delta_1(\epsilon)} + \frac{4K}{n} \frac{\tilde{p}_u \cdot (2p+1)}{c} \Delta_1(\epsilon) + 2\epsilon + \Gamma$, a.s. Further invoking Propositions 14 and 15, we have that both events hold simultaneously with probability at least as in P^* , which verifiably implies the claimed results. \blacksquare

A.5 Useful Lemmata

Lemma 17 *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*

$$\begin{aligned} \max_{\mathbf{X}_1, \mathbf{X}_2 \in \mathcal{S}_p} \{ |\mathcal{F}_n(\mathbf{X}_1, \mathbf{Z}_1^n) - \mathcal{F}_n(\mathbf{X}_2, \mathbf{Z}_1^n)| \} &\leq (2K_C + C_\mu) \cdot \tau. \\ \cap \{ \mathbf{X} : \sigma_{\max}(\mathbf{X}) \leq R, \\ &\quad \|\mathbf{X}_1 - \mathbf{X}_2\| \leq \tau \} \end{aligned}$$

for any given $\tau \geq 0$.

Proof This proof follows a closely similar lemma by [27]. Similar proof has also been provided by [11], but some subtle differences in the problem context present and thus we redo the the proof herein. By Assumption 4, for some $c > 0$,

$$\mathbb{P} \left(\left| \sum_{i=1}^n \frac{1}{n} \{ \mathcal{C}(Z_i) - \mathbb{E}[\mathcal{C}(Z_i)] \} \right| > K_C \left(\frac{t}{n} + \sqrt{\frac{t}{n}} \right) \right) \leq 2 \exp(-ct), \quad \forall t \geq 0.$$

If we let $t := n$ and observe that $\mathbb{E}[\mathcal{C}(Z_i)] \leq \mathcal{C}_\mu$, we immediately have that

$$\mathbb{P} \left(\sum_{i=1}^n \frac{\mathcal{C}(Z_i)}{n} \leq 2K_C + \mathcal{C}_\mu \right) \leq 1 - 2 \exp(-cn). \quad (46)$$

If we invoke Assumption 4 again given the event that $\left\{ \sum_{i=1}^n \frac{\mathcal{C}(Z_i)}{n} \leq 2K_C + \mathcal{C}_\mu \right\}$, we have that for any $\mathbf{X}_1, \mathbf{X}_2 \in \mathcal{S}_p$,

$$\begin{aligned} & \max_{\substack{\mathbf{X}_1, \mathbf{X}_2 \in \mathcal{S}_p \\ \cap \{ \mathbf{X}: \sigma_{\max}(\mathbf{X}) \leq R, \\ \|\mathbf{X}_1 - \mathbf{X}_2\| \leq \tau \}}} \left| \frac{1}{n} \sum_{i=1}^n f(\mathbf{X}_1, Z_i) - \frac{1}{n} \sum_{i=1}^n f(\mathbf{X}_2, Z_i) \right| \\ & \leq \max_{\substack{\mathbf{X}_1, \mathbf{X}_2 \in \mathcal{S}_p \\ \cap \{ \mathbf{X}: \sigma_{\max}(\mathbf{X}) \leq R, \\ \|\mathbf{X}_1 - \mathbf{X}_2\| \leq \tau \}}} \frac{1}{n} \sum_{i=1}^n \|f(\mathbf{X}_1, Z_i) - f(\mathbf{X}_2, Z_i)\| \\ & \leq \max_{\substack{\mathbf{X}_1, \mathbf{X}_2 \in \mathcal{S}_p \\ \cap \{ \mathbf{X}: \sigma_{\max}(\mathbf{X}) \leq R, \\ \|\mathbf{X}_1 - \mathbf{X}_2\| \leq \tau \}}} \frac{1}{n} \sum_{i=1}^n \mathcal{C}(Z_i) \|\mathbf{X}_1 - \mathbf{X}_2\| \leq (2K_C + \mathcal{C}_\mu) \cdot \tau \end{aligned}$$

We have the desired result by combining the above with (46). ■

Lemma 18 *Under Assumption 4, for all*

$$\mathbf{X}_1, \mathbf{X}_2 \in \mathcal{S}_p : \max\{\sigma_{\max}(\mathbf{X}_1), \sigma_{\max}(\mathbf{X}_2)\} \leq R,$$

it holds that

$$|\mathbb{E}[\mathcal{F}_n(\mathbf{X}_1, \mathbf{Z}_1^n)] - \mathbb{E}[\mathcal{F}_n(\mathbf{X}_2, \mathbf{Z}_1^n)]| \leq \mathcal{C}_\mu \cdot \|\mathbf{X}_1 - \mathbf{X}_2\|. \quad (47)$$

Proof This proof follows a closely similar lemma by [27]. Again, a similar proof has also been provided by [11], but some subtle differences make it necessary to conduct the repetition herein. As per Assumption 4, it holds that

$$\mathbb{E} [|\mathcal{F}_n(\mathbf{X}_1, \mathbf{Z}_1^n) - \mathcal{F}_n(\mathbf{X}_2, \mathbf{Z}_1^n)|] \leq \mathbb{E} \left[\sum_{i=1}^n \frac{\mathcal{C}(Z_i)}{n} \|\mathbf{X}_1 - \mathbf{X}_2\| \right].$$

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

$$\begin{aligned} |\mathbb{E}[\mathcal{F}_n(\mathbf{X}_1, \mathbf{Z}_1^n)] - \mathbb{E}[\mathcal{F}_n(\mathbf{X}_2, \mathbf{Z}_1^n)]| &\leq \mathbb{E} \left[\sum_{i=1}^n \frac{\mathcal{C}(Z_i)}{n} \|\mathbf{X}_1 - \mathbf{X}_2\| \right] \\ &= \mathbb{E} \left[\sum_{i=1}^n \frac{\mathcal{C}(Z_i)}{n} \right] \cdot \|\mathbf{X}_1 - \mathbf{X}_2\|. \end{aligned}$$

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

Lemma 19 *Denote that $\mathbf{X}_\lambda^{\ell_1} \in \arg \min_{\mathbf{X} \in \mathbb{S}^p} \mathcal{F}_n(\mathbf{X}, \mathbf{Z}_1^n) + \lambda \|\mathbf{X}\|_*$, it holds that $\mathcal{F}_{n,\lambda}(\mathbf{X}_\lambda^{\ell_1}, \mathbf{Z}_1^n) \leq \mathcal{F}_{n,\lambda}(\mathbf{X}^*, \mathbf{Z}_1^n) + \lambda \|\mathbf{X}^*\|_*$.*

Proof This proof generalizes a similar one in [11] from sparsity-inducing penalty to low-rankness-inducing penalty; that is, from ℓ_1 regularization to nuclear norm-based regularization. As per Assumption 4, it holds that We first invoke the definition of P_λ to obtain

$$0 \leq P_\lambda(t) = \int_0^t \frac{[a\lambda - \theta]_+}{a} d\theta \leq \int_0^t \frac{a\lambda}{a} d\theta = \lambda \cdot t. \quad (48)$$

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

$$\mathcal{F}_n(\mathbf{X}_\lambda^{\ell_1}, \mathbf{Z}_1^n) + \lambda \|\mathbf{X}_\lambda^{\ell_1}\|_* \leq \mathcal{F}_n(\mathbf{X}^*, \mathbf{Z}_1^n) + \lambda \|\mathbf{X}^*\|_*. \quad (49)$$

Combining (48) and (49), it holds that

$$\begin{aligned} \mathcal{F}_n(\mathbf{X}_\lambda^{\ell_1}, \mathbf{Z}_1^n) + \sum_{j=1}^p P_\lambda(|\sigma_j(\mathbf{X}_\lambda^{\ell_1})|) &\leq \mathcal{F}_n(\mathbf{X}_\lambda^{\ell_1}, \mathbf{Z}_1^n) + \sum_{j=1}^p \lambda \cdot |\sigma_j(\mathbf{X}_\lambda^{\ell_1})| \\ &\leq \mathcal{F}_n(\mathbf{X}^*, \mathbf{Z}_1^n) + \sum_{j=1}^p P_\lambda(|\sigma_j(\mathbf{X}^*)|) + \lambda \|\mathbf{X}^*\|_*, \end{aligned}$$

as desired. \blacksquare

Lemma 20 *Let $S_{r,R} := \{X \in \mathbb{R}^{p \times p} : \mathbf{rk}(X) \leq r, \sigma_{\max}(X) \leq R\}$. Then, in terms of the Frobenius norm, there exists an ϵ -net \bar{S}_r obeying $|\bar{S}_r| \leq \left(\frac{9\sqrt{r}R}{\epsilon}\right)^{(2p+1)r}$.*

Proof The proof follows a closely similar result by [3, Lemma 3.1]. Denote by $X := U\Sigma V^\top$ 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 \bar{D} be an $\frac{\epsilon}{3}$ -net (in terms of Frobenius norm) for D with $|\bar{D}| \leq \left(\frac{9\sqrt{r}R}{\epsilon}\right)^r$.

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

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

References

- [1] P.L. Bartlett PL, M.I. Jordan 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. Candes 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. Cl  men  on, 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] E. Moulines, E. and F.R. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. in *Advances in Neural Information Processing Systems*. 2011.
- [17] 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
- [18] 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.
- [19] 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.
- [20] S. F. Nielsen. Empirical processes in m -estimation. *J. of Appl. Stat.*, 27(8):1067–1068, 11 2000.
- [21] Rohde, A. and Tsybakov, A.B. Estimation of high-dimensional low-rank matrices. *Ann. of Stat.*, 39(2), 887-930, 2011
- [22] G. Raskutti, W.J. Wainwright, and B. Yu, Minimax Rates of Estimation for high-dimensional linear regression over ℓ_q -balls. *IEEE Transactions on Information Theory*, 57(10) 6976 - 6994, 2011.

- [23] 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.
- [24] A. Ruszczyński 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](https://doi.org/10.1016/S0927-0507(03)10001-1).
- [25] A. Ruszczyński 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](https://doi.org/10.1016/S0927-0507(03)10002-3).
- [26] 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>.
- [27] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory, Second Edition*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2014. ISBN 1611973422, 9781611973426.
- [28] C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.*, 38(2):894–942, 04 2010. doi: 10.1214/09-AOS729.
- [29] 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.