Journal of Global Optimization (2023) 85:257–282
https://doi.org/10.1007/s10898-022-01206-3

Regularized sample average approximation for high-dimensional stochastic optimization under low-rankness

Hung Yi Lee[1] · **Charles Hernandez**[1] · **Hongcheng Liu**[1]

Received: 30 October 2021 / Accepted: 16 June 2022 / Published online: 15 July 2022

Abstract
This paper concerns a high-dimensional stochastic programming (SP) 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 $\widetilde{O}(1) \cdot \frac{p^2}{\epsilon^2} \cdot polylog(\frac{1}{\epsilon})$ to achieve an ϵ-suboptimality gap, for some poly-logarithmic function $polylog(\cdot)$ and some quantity $\widetilde{O}(1)$ 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, when the optimal solution to the SP is of low rank, the sample complexity of RSAA is $\widetilde{O}(1) \cdot \frac{p}{\epsilon^3} \cdot polylog(p, \frac{1}{\epsilon})$, 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 programming · Sample average approximation · Low-rankness

✉ Hung Yi Lee
hungyilee@ufl.edu

Charles Hernandez
cdhernandez@ufl.edu

Hongcheng Liu
liu.h@ufl.edu

[1] Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, USA

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 [19, 20, 22] and [21] 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 [15], 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 substantially 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 $\mathcal{W} \subseteq \Re^q$ and $q > 0$, be a random vector. Consider a measurable, deterministic function $f : \mathcal{S}_p^+ \times \mathcal{W} \to \Re$ 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

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

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.

The problem of interest (1) falls into the general category of the stochastic version of semidefinite programming (SDP), whose many applications have been found in science and engineering (See more discussions in [1, 24, 25]). As compared to the traditional schemes, the proposed RSAA is expected to result in substantial improvement in the efficacy and efficiency of solving the stochastic SDP by reducing the need of simulation replications or data collections. Though, admittedly, many of the SDP applications stipulates constraints that are more sophisticated than our formulation in (1), we argue that this work can serve as the underpinning to the development of more advanced schemes to address those additional sophistication.

We would like to remark that the formulation in (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 $\mathbf{Z}_1^n = (Z_1, ..., Z_n)$ 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:

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

According to the seminal results by [22], $\mathbf{X}^{SAA}$ well approximates $\mathbf{X}^*$ in the sense that

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

with high probability, where $\widetilde{O}(\cdot)$ 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 qradratically if p increases. For an SP problem where $\mathbf{X}^*$ is sparse and f is twice-differentiable almost surely, we have shown in [15] that (3) can be sharpened, in terms of its dependence on p, into:

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

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 [12, 13] 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 [12, 13, 15] 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 $\boldsymbol{x}$ can be represented by a diagonal matrix, $diag(\boldsymbol{x})$, whose diagonal entries equal to $\boldsymbol{x}$. Then, sparsity of $\boldsymbol{x}$ implies that $diag(\boldsymbol{x})$ 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):

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

where $\sigma_j(\mathbf{X})$ stands for the jth eigenvalue of $\mathbf{X}$ and P_λ 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 $a,\ \lambda > 0$. 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 6 in the subsequent. The S^3ONC herein is an extension to a similar notion presented by [14, 15] 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. To add to the existing literature, we will present a new, highly tractable S^3ONC-guaranteeing algorithm.

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

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

with $\|\cdot\|_*$ denoting the nuclear norm. We show that, under a few standard assumptions in addition to Assumption 1, for any S^{3}ONC solution to the RSAA, denoted $\mathbf{X}^{RSAA}$, which satisfies $\mathcal{F}_{n,\lambda}(\mathbf{X}^{RSAA}) \le \mathcal{F}_{n,\lambda}(\mathbf{X}_\lambda^{\ell_1})$ a.s., it holds that

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

with overwhelming probability, when our knowledge on the rank of $\mathbf{X}^*$ is completely absent. Furthermore, as a quick extension to (7), 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 the sample size requirement to entail a lower dependence on s. The above results are then the promised, almost linear, sample complexity; from (7), n should only increase almost linearly in p to compensate the growth in dimensionality. This indicates that the RSAA can be provably much more advantageous than the SAA especially for high-dimensional problems that satisfies $p^2 > \widetilde{O}(1)\cdot n$.

To compute the desired solution $\mathbf{X}^{RSAA}$, one may invoke an S^{3}ONC-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 [6, 17, 18] 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 [6, 17, 18] (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., [2, 5, 11].

The rest of the paper is organized as follows: Sect. 2 presents our assumptions and main results concerning the sample complexity of RSAA. A highly tractable solution scheme to compute a desired S^{3}ONC solution is presented in Sect. 3. Section 4 presents some preliminary numerical experiments to show the consistency between our theory and the corresponding simulation results. Section 5 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 $1 \le \mathbf{p} \le \infty$). Let $\sigma_j(\mathbf{X})$ be

the jth singular value of matrix $\mathbf{X}$. Denote by $\|\cdot\|_F$ the Frobenius norm. $\mathcal{S}_p$ and $\mathcal{S}_p^+$ are the cone of symmetric and symmetric and positive definite matrices, respectively.

2 Sample complexity of RSAA

This section presents our main results in Sect. 2.3 after we introduce our assumptions in Sect. 2.1 as well as the definition of the S^{3}ONC in Sect. 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 The gradient of $f(\mathbf{X},\, z)$ with respect to the singular values, denoted by $\left(\frac{\partial f(\mathbf{X},\, z)}{\partial \sigma_j(\mathbf{X})} :\, j = 1, ..., p\right)$, is well defined and Lipschitz continuous with constant $\mathcal{U}_L \geq 1$ for any $\mathbf{X} \in \mathcal{S}_p^+ :\, \sigma_{\max}(X) \leq R$ and $z \in \mathcal{W}$.

Assumption 3 The family of random variables, $f(\mathbf{X},\, Z_i) - \mathbb{E}[f(\mathbf{X},\, Z_i)]$, $i = 1, ..., n$, 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

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} a_i \left\{f(\mathbf{X},\, Z_i) - \mathbb{E}[f(\mathbf{X},\, Z_i)]\right\}\right| > K\left(\|\mathbf{a}\|\sqrt{t} + \|\mathbf{a}\|_\infty t\right)\right) \leq 2\exp\left(-ct\right), \tag{8}$$

for all $t \geq 0$, $\mathbf{a} = (a_i) \in \Re^n$ and for some absolute constant $c \in (0,\, \frac{1}{2}]$.

Assumption 4 For some measurable and deterministic function $\mathcal{C} :\, \mathcal{W} \to \Re$ 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 5 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. Examples of distributions that satisfy 3 are Gaussian, χ^2, exponential, and uniform distributions as well as any distribution with bounded support. Assumption 3 essentially establishes the probability for the average cost function to be Lipschitz continuous.

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 6 For given $\mathbf{Z}_1^n \in \mathcal{W}^n$, a vector $\widehat{\mathbf{X}} \in \mathcal{S}_p^+$ is said to satisfy the S^{3}ONC (denoted by S^3*ONC*$(\mathbf{Z}_1^n)$) of the problem (5) if the following inequality holds at $\widehat{\mathbf{X}}$ for all $j = 1, ..., p$:

$$\mathcal{U}_L + \left[\frac{\partial^2 P_\lambda(\sigma_j(\mathbf{X}))}{\partial \sigma_j(\mathbf{X})^2}\right]_{\mathbf{X}=\widehat{\mathbf{X}}} \geq 0, \quad \text{if } \sigma_j(\widehat{\mathbf{X}}) \in (0,\, a\lambda), \tag{9}$$

where $\mathcal{U}_L$ is as defined in 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. Thus, many schemes exist to compute such a condition. We will introduce a new, highly tractable S^{3}ONC-guaranteeing algorithm subsequently.

2.3 Main results on sample complexity

Introduce a few short-hand notations: Denote $\widetilde{\Delta} := \ln\left(18R \cdot (K_C + \mathcal{C}_\mu)\right)$ and

$$\lambda(\rho) := \sqrt{\frac{8K(2p+1)^{2/3}s^{-\rho}}{c \cdot a \cdot n^{2/3}}[\ln(n^{1/3}p) + \widetilde{\Delta}]}, \tag{10}$$

for the same c in (8) and a user-specific $\rho \geq 0$. Recall the definition of $\mathbf{X}_\lambda^{\ell_1}$ in (6). We are now ready to present our claimed results.

Theorem 7 *Suppose that Assumption 1 through 4 hold. Specify the penalty parameter* $\lambda := \lambda(\rho)$ *and* $a = (2\mathcal{U}_L)^{-1}$. *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 any* $\Gamma \geq 0$ *and some universal constants* $\widetilde{c},\ C_1 > 0$, *if*

$$n > C_1 \cdot s^{3\rho} \cdot \left[\left(\frac{\Gamma}{K}\right)^3 + 1\right] \cdot p + C_1 \cdot s \cdot p \cdot \left(\ln(n^{1/3}p) + \widetilde{\Delta}\right), \tag{11}$$

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*

$$\begin{aligned}\mathbb{F}(\mathbf{X}^{RSAA}) - \mathbb{F}(\mathbf{X}^*) &\leq \sqrt{\frac{K \cdot s^\rho \cdot p^{1/3} \cdot \Gamma}{n^{1/3}}} + \Gamma \\ &+ C_1 K \cdot \left[\frac{s^{1-\rho} \cdot p^{2/3} \cdot \left(\ln(n^{1/3}p) + \widetilde{\Delta}\right)}{n^{2/3}} + \sqrt{\frac{s \cdot p \cdot \left(\ln(n^{1/3}p) + \widetilde{\Delta}\right)}{n}} + \frac{p^{1/3} \cdot s^\rho}{n^{1/3}}\right],\end{aligned} \tag{12}$$

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

Remark 8 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. $\widetilde{\Delta}$ is some logarithmic term independent of p and n.
4. K and K_C are the upper bounds on the subexponential norm of the underlying distributions. They are alternative measures of the distributions' variances.
5. An equivalent representation of the results in (12) can lead to the following sample size requirement to achieve an out-of-sample suboptimality gap ε with probability at least $1-\alpha$, for any $\varepsilon > \Gamma$ and $\alpha \in (0, 1]$,

$$\widehat{O}(1) \cdot \left(\frac{p \cdot \Gamma^3}{(\varepsilon - \Gamma)^6} + \frac{p}{(\varepsilon - \Gamma)^3} + \ln \frac{p}{\alpha} + \frac{1}{p^2} \left(\ln \frac{1}{\alpha} \right)^3 \right),$$

for some quantity $\widehat{O}(1)$ that is independent of, or at most logarithmic in, p, ϵ, and α.

Remark 9 Some intuitions on the above theorem are as follows:

1. Theorem 7 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). The out-of-sample performance of these S^{3}ONC solutions differentiates as their in-sample optimality gaps vary; Eq. (12) indicates that the out-of-sample optimality gap is strictly increasing in the in-sample optimality gap Γ.
2. 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 10 *Suppose that Assumption 1 through 4 hold. Specify the penalty parameter $\lambda = \lambda(0)$ (that is, $\rho = 0$) and $a = (2\mathcal{U}_L)^{-1}$ in both formulations (6) and (5). Let $\mathbf{X}^{RSAA} \in \mathcal{S}_p^+ : \sigma_{\max}(\mathbf{X}^{RSAA}) \le R$ satisfy the S^{3}ONC$(\mathbf{Z}_1^n)$ to (5) almost surely. For some universal constant $\widetilde{c}$, $C_2 > 0$, if*

$$n > C_2 \cdot p \cdot \mathcal{U}_L \cdot [\ln(n^{\frac{1}{3}} p) + \widetilde{\Delta}] \cdot s^{\frac{3}{2}} R^{\frac{3}{2}}, \tag{13}$$

and

$$\mathcal{F}_{n,\lambda}(\mathbf{X}^{RSAA}, \mathbf{Z}_1^n) \le \mathcal{F}_{n,\lambda}(\mathbf{X}_\lambda^{\ell_1}, \mathbf{Z}_1^n) \tag{14}$$

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

$$\begin{aligned} &\mathbb{F}(\mathbf{X}^{RSAA}) - \mathbb{F}(\mathbf{X}^*) \\ &\le C_2 \cdot s \cdot K \cdot \left[\frac{p^{2/3} \left(\ln(n^{\frac{1}{3}} p) + \widetilde{\Delta} \right)}{n^{\frac{2}{3}}} + \frac{p^{1/3} R \cdot \mathcal{U}_L^{1/2} \sqrt{\ln(n^{\frac{1}{3}} p) + \widetilde{\Delta}}}{n^{\frac{1}{3}}} \right], \end{aligned} \tag{15}$$

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

Remark 11 We would like to make a few remarks on the above result:

1. Corollary 10 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 7 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 10 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 $\widetilde{O}(1) \cdot \frac{p}{\epsilon^3} \cdot polylog(p, \frac{1}{\epsilon})$, which is almost linear in p, for some quantities $\widetilde{O}(1)$ that is independent of n, ϵ, and p.
5. One may equivalently represent (15) into the following sample size requirement to achieve an out-of-sample suboptimality gap ε with probability at least $1-\alpha$, for any $\varepsilon > \Gamma$ and $\alpha \in (0, 1]$,

$$\widehat{O}(1) \cdot \left(\frac{p}{\varepsilon^3} + \ln \frac{p}{\alpha} + \frac{1}{p^2} \left(\ln \frac{1}{\alpha} \right)^3 \right), \tag{16}$$

for some quantity $\widehat{O}(1)$ that is independent of, or at most logarithmic in, p, ϵ, and α.

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 for certain applications. 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 [15] 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 12 below.

Corollary 12 *Suppose that Assumption 1 through 4 hold. Specify the penalty parameter* $\lambda = \lambda(\frac{2}{3})$ *(that is,* $\rho = \frac{2}{3}$*) and* $a = (2\mathcal{U}_L)^{-1}$ *in both formulations (6) and (5). Let* $\mathbf{X}^{RSAA} \in \mathcal{S}_p^+ : \sigma_{\max}(\mathbf{X}^{RSAA}) \le R$ *satisfy the S*^{3}*ONC*$(\mathbf{Z}_1^n)$ *to (5) almost surely. For some universal constant* $\widetilde{c}$, $C_3 > 0$, *if*

$$n > C_3 \cdot p \cdot \mathcal{U}_L \cdot [\ln(n^{\frac{1}{3}} p) + \widetilde{\Delta}] \cdot s^2 \cdot R^{\frac{3}{2}}, \tag{17}$$

and (14) holds almost surely, where $\mathbf{X}_\lambda^{\ell_1}$ *is as defined in (6), then the excess risk is bounded by*

$$\begin{aligned} &\mathbb{F}(\mathbf{X}^{RSAA}) - \mathbb{F}(\mathbf{X}^*) \\ &\le C_3 \cdot K \cdot \left[\frac{s^{1/3} p^{2/3} \left(\ln(n^{\frac{1}{3}} p) + \widetilde{\Delta} \right)}{n^{\frac{2}{3}}} + \frac{s^{2/3} p^{1/3} \cdot R \cdot \mathcal{U}_L^{1/2} \cdot \sqrt{\ln(n^{\frac{1}{3}} p) + \widetilde{\Delta}}}{n^{\frac{1}{3}}} \right], \end{aligned} \tag{18}$$

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

Remark 13 The Corollary 12, similar to Corollary 10, establishes our claimed result of almost linear complexity at a computable S^{3}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 14 In contrast to Corollaries 10 and 12 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. As a direct implication of (18), the sample requirement follows the same rate as in (16) in terms of the dependence on p, out-of-sample suboptimality gap ε, and confidence level α.

Remark 15 There is strong correspondence between the SP and statistical learning as formerly noted by [15, 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 [2, 5, 11]. We therefore argue that the results in Theorem (7) and Corollaries 10 and 12 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 16 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 (14). First, the stationarity can be ensured by invoking local optimization algorithms. Second, the stipulated inequality in (14) 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.

Remark 17 In practice, exact optimal solutions to the convex program (6) may be challenging to compute. It is nonetheless easy to see that a good approximation to the optimal solution is sufficient to ensure a comparable out-of-sample performance; in particular, one may easily verify the following by the same argument as in proving Corollary 12: If $\mathbf{X}_\lambda^{\ell_1}$ incurs a suboptimality gap of $\delta > 0$ in solving (6), then the right-hand-side of error bound in (18) will need to involve an additional error term of $\sqrt{\frac{K \cdot s^\rho \cdot p^{1/3} \cdot \delta}{n^{1/3}}} + \delta$ in the worse case. Thus, if $\mathbf{X}_\lambda^{\ell_1}$ is a $\widetilde{O}(1)\frac{p^{1/3}\ln p}{n^{1/3}}$-suboptimal solution to the convex problem in (6), then same error rate as in (18) can be maintained.

3 A highly tractable S^{3}ONC-guaranteeing algorithm

Below we introduce a computing procedure to generate an S^{3}ONC solution. For convenience, we consider a more abstract form than (5) given as:

$$\min_{\mathbf{X} \in \mathcal{S}_p^+} \mathcal{G}(\mathbf{X}) := f(\mathbf{X}) + \sum_{j=1}^{p} P_\lambda(\sigma_j(X)). \tag{19}$$

The pseudo-code of the proposed algorithm is provided below.

The (nonconvex) subproblem in (20) admits a closed form solution and provably generates an S^{3}ONC solution as shown in the theorem below.

Theorem 18 *Let L satisfy that $\mathcal{U}_L \le L < \frac{1}{a}$. A closed form solution to (20) is that $\widehat{\mathbf{X}} = Q\,\mathrm{diag}\left(\{y_j : j = 1, ..., p\}\right) Q^{-1}$ where*

$$y_j = \begin{cases} \sigma_j(\mathbf{X}^0 - \frac{1}{L} \cdot \nabla f(\mathbf{X}^0)) & \text{if } \sigma_j(\mathbf{X}^0 - \frac{1}{L} \cdot \nabla f(\mathbf{X}^0)) \ge a\lambda; \\ 0 & \text{otherwise.} \end{cases}$$

Algorithm 1 An S^{3}ONC-guaranteeing algorithm

1: Initialize a feasible $\mathbf{X}^0$ (e.g., we may let $\mathbf{X}^0 := \mathbf{X}_\lambda^{\ell_1}$ as in (6)).
2: Solve the following nonconvex (but tractable) optimization problem, and output its optimal solution $\widehat{\mathbf{X}}$

$$\widehat{\mathbf{X}} \in \arg\min_{\mathbf{X} \in \mathcal{S}_p^+} \left\langle \nabla f(\mathbf{X}^0),\, \mathbf{X} - \mathbf{X}^0 \right\rangle + \frac{L}{2} \left\| \mathbf{X} - \mathbf{X}^0 \right\|_F^2 + \sum_{j=1}^{p} P_\lambda(\sigma_j(\mathbf{X})). \tag{20}$$

and $Q \Lambda Q^{-1}$*, with* $\Lambda = \mathrm{diag}(\{\lambda_j : j = 1, ..., p\})$*, is the eigenvalue decomposition of the matrix* $\mathbf{Y} := \mathbf{X}^0 - \frac{1}{L}\nabla f(\mathbf{X}^0)$*. Furthermore,* $\widehat{\mathbf{X}}$ *is an S*^{3}*ONC solution to* (19) *and satisfies that* $\mathcal{G}(\widehat{\mathbf{X}}) \le \mathcal{G}(\mathbf{X}^0)$.

Remark 19 Based on the closed form solution provided in (18), the proposed scheme is equivalent to one iteration of an iterative hard thresholding algorithm, as discussed by, e.g., [3, 10, 17, 23], following the computation of a nuclear norm-regularized problem. Both components have been studied previously. Nonetheless, existing theories on out-of-sample performance (e.g., by [17, 23]) concerning each scheme assume conditions such as RSC or its variations, which are not stipulated by our results. More on the RSC have been provided in the introduction and in Remark 15.

Remark 20 If one choses a different initial point than X_0, Theorem 7 still applies to the output of the algorithm; the in-sample suboptimality gap of the solution generated by Algorithm 1 is then the value of Γ in Eq. (12)

4 Preliminary numerical experiments

Our numerical experiment was focused on a special matrix recovery problem using simulated data formulated as:

$$\min_{\mathbf{X}} \left\{ \mathbb{E}[(X_{ij} - T_{ij})^2] : \ \mathbf{X} \in \mathcal{S}_p^+ \right\}, \tag{21}$$

where T_{ij} is a randomly selected entry from a matrix $T := \mathbf{T}^{true} + W$ for some target matrix $\mathbf{T}^{true} \in \mathcal{S}_p^+$ and white noise $W \in \Re^{p \times p}$. The expectation is over both the random indices $i, j \in \{1, ..., p\}$ and the white noise W. Assume that $\mathbf{T}^{true}$ is of rank two. We can invoke both SAA and RSAA to solve (21) and compare their performance. All the experiments were implemented in Matlab and run on a PC with 2.2 GHz Intel Core i7 and 16 GB Memory. The SAA in (2) and the convex problem (6) in Step 1 of Algorithm 1 are both solved by invoking CVX 2.2 [8, 9] via Matlab. Then, Step 2 of Algorithm 1 invokes a closed form given in Theorem 18.

To generate a low-rank target matrix $\mathbf{T}^{true}$, we first simulated a random matrix $\mathbf{A} \in \Re^{r \times p}$ with i.i.d. entries following uniform distribution on [0, 1]. We then calculated a matrix of $\mathbf{A}\mathbf{A}^\top$ which admits an eigenvalue decomposition of $Q_A V Q_A^{-1}$. We set all the nonzero diagonals of V to be 2 and obtained a new diagonal matrix $\widetilde{V}$. Then, the target matrix was generated by $\mathbf{T}^{true} = Q_A \widetilde{V} Q_A^{-1}$.

To generate the samples for both SAA and RSAA, we first simulated W as a random matrix with i.i.d. normal gaussian entries. Then, we downsampled $T := (T_{ij}) = \mathbf{T}^{true} + W$ by selecting entries of T following a discrete uniform distribution repetitively. This process

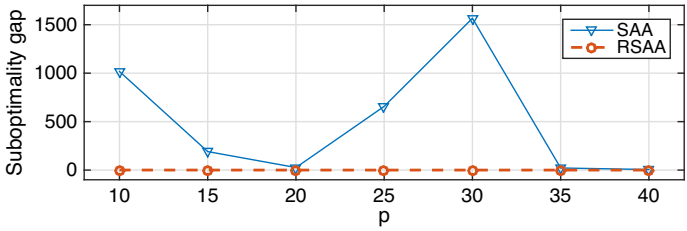

Fig. 1 Comparison between SAA and RSAA in terms of suboptimality gaps, when dimensionality p varies and sampling rate is set as $0.4/p^2 \cdot 100\%$

Table 1 Average computational time (s) for solving SAA and RSAA formulations for different values of p

p	10	20	30	40	50	60
SAA	1.50	3.89	10.49	32.03	68.40	160.51
RSAA	1.45	4.19	11.22	33.55	68.64	158.81

was replicated for a pre-specified n-many rounds and one entry was selected in each round. Thus, we obtained n-many samples for both SAA and RSAA. We refer to $\frac{n}{p^2} \times 100\%$ as the sampling rate. In all our experiments, we set $a = 1$ and $\lambda = 0.01 \cdot \frac{1}{n^{1/3}}$ for RSAA.

We report two sets of test results. The first involves a sampling rate of $40/p^2 \times 100\%$ and various values of dimensionality p. Notice that, this way, the number of sampled entries is fixed at $n = 40$ for any p. To see this, we note that $40/p^2 \times 100\% = \frac{n}{p^2} \times 100\% \iff n = 40$. For each combination of sampling rate and dimensionality, five problem instances were generated and solved by both SAA and RSAA. Figure 1 reports the average suboptimality gaps (bad) achieved by both schemes. In the figure, each data point represents the average suboptimality gap over the five replications of problem instances for the same n-p combination. We see that as the dimensionality increases from $p = 10$ (100 unknowns) to $p = 60$ (3600 unknowns), SAA became unstable and may perform significantly worse than the RSAA. Yet, the latter outperformed the former for all the combinations of n and p involved in the experiment. Table 1 reports the average computational time on solving both SAA and RSAA out of the five replications for different p. From this table, one may observe that the computational effort incurred by the proposed RSAA is comparable to that of RSAA.

The second test set involves different sampling rates {2%, 4%, 6%, 8%, 10%, 12%, 14%, 16%, 20%, 50%, 80%} and a fixed dimensionality with $p = 60$. Thus, there are 3600 unknowns. Again, five random replications were done for each combination of settings.

Figure 2 reports the average of (logarithm of) suboptimality gaps. We can see that, when the sampling rate is as low as 2%, the RSAA yields a significantly better performance. Although the difference in performance reduces when the sampling rate increases, the RSAA is consistently better than the SAA. Table 2 reports the average computational time spent on solving SAA and RSAA when the sampling rates change from 2% to 80%. Again, we observe that the RSAA yielded comparable computational cost with the SAA for all the different sampling rates.

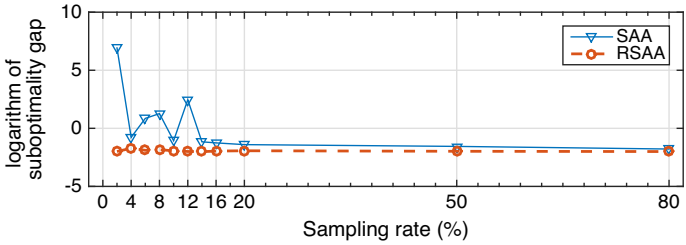

Fig. 2 Comparison between SAA and RSAA in terms of (the logarithm of) suboptimality gaps, when sampling rate varies and dimensionality is set as $p = 60$

Table 2 Average computational time (s) for solving SAA and RSAA formulations for different sampling rates

Sampling rate (%)	SAA	RSAA
2	138.5	130.1
4	142.1	139.6
6	155.7	161.5
8	147.6	155.8
10	171.4	162.7
12	149.7	159.3
14	153.4	157.2
16	135.7	149.6
20	170.3	183.8
50	188.5	179.4
80	173.1	168.9

5 Conclusions

This paper proposes a regularized SAA (RSAA), which 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 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.

Acknowledgements The authors would like to thank the anonymous reviewers and editors for their constructive comments that have helped improve this paper. This research is partially supported by NSF grant CMMI-2016571.

A Technical proofs

A.1 Proof of results on sample complexity

A.1.1 General ideas

The general idea of our proof is focused on addressing the question: *how to show that an* S^3ONC *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 [4] and is restated (with minor modifications) in Lemma 28 herein. Proposition 22 then establishes a point-wise bound between the average cost and the expected cost for all the low-rank subspaces.

To bound the rank of an S^3ONC solution, we utilize a unique property of the MCP function, which ensures that the S^3ONC solutions $\mathbf{X}^{RSAA}$ must obey a thresholding rule: for all the singular values, they must be either 0 or greater than $a\lambda$, where a and λ are hyper-parameters of the penalty. Proposition 21 formalizes this thresholding rule.

By this rule and the definition of the MCP, for each nonzero singular value in the S^3ONC solution $\mathbf{X}^{RSAA}$, the total value of penalty incurred by the MCP-based low-rankness-inducing regularization becomes $\sum_{j=1}^{p} P_\lambda(\sigma_j(\mathbf{X}^{RSAA})) = \mathbf{rk}(\mathbf{X}^{RSAA}) \cdot \frac{a\lambda^2}{2}$. Now, consider those S^3ONC points whose suboptimality gaps in terms of minimizing the RSAA are smaller than a user-specific quantity Γ. These solutions should satisfy

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

By this inequality, we may observe that the rank of $\mathbf{X}^{RSAA}$ must be bounded from above. It is also easy to see that this upper bound should be a function of Γ. This function is explicated by Proposition 23. Then, the desired results in Theorem 7 immediately follows the combination of Propositions 22 and 23. Finally, the value of Γ can be well contained and explicated by proper (and tractable) initializations, as shown in and Corollaries 10 and 12. The proof of these corollaries are based on Lemma 27, which shows that $\mathbf{X}_\lambda^{\ell_1}$ yields a small value of Γ.

A.1.2 Proof of Theorem 7

Proof This proof substantially generalizes the argument of Proposition 1 in [13] from handling sparsity to handling low-rankness. Meanwhile, much more flexible choices of penalty parameters λ is enabled. We follow the same set of notations in Proposition 24 in defining $\widetilde{p}_u$, 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 $\widetilde{\Delta} := \ln\big(18 \cdot R \cdot (K_C + C_\mu)\big)$. Then $\Delta_1(\epsilon) = \ln\left(\frac{18\cdot(K_C+C_\mu)\cdot p\cdot R}{\epsilon}\right) = \ln(n^{1/3}p) + \widetilde{\Delta} > 0$ and $\lambda = \sqrt{\frac{8\cdot s^{-\rho}\cdot K\cdot(2p+1)^{2/3}\cdot\Delta_1(\epsilon)}{c\cdot a\cdot n^{2/3}}} = \sqrt{\frac{8\cdot s^{-\rho}\cdot K\cdot(2p+1)^{2/3}}{c\cdot a\cdot n^{2/3}}[\ln(n^{1/3}p) + \widetilde{\Delta}]}$. We will denote by $O(1)$'s the universal constants, which may be different in each occurence.

To show the desired results, it suffices to simplify the results in Proposition 24. We will first derive an explicit form for $\widetilde{p}_u$. To that end, we let $P_X := \widetilde{p}_u$ 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 (41) of Proposition 24, for a feasible P_X,

$$\frac{T_1}{2} \cdot P_X - \frac{2K}{\sqrt{n}}\sqrt{\frac{2P_X\cdot(2p+1)\Delta_1(\epsilon)}{c}} > \Gamma + 2\epsilon + sP_\lambda(a\lambda), \tag{22}$$

for the same $c \in (0, 0.5]$ in (8). 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 $\lambda = \sqrt{\frac{8K\cdot s^{-\rho}\cdot\Delta_1(\epsilon)\cdot(2p+1)^{2/3}}{c\cdot a\cdot n^{2/3}}} = \sqrt{\frac{8K\cdot s^{-\rho}\cdot(2p+1)^{2/3}}{c\cdot a\cdot n^{2/3}}[\ln(n^{1/3}p)+\widetilde{\Delta}]}$ with $\widetilde{\Delta} := \ln\big(18\cdot R\cdot(K_C+\mathcal{C}_\mu)\big)$, 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., (11) 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 (22), 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 $P_X \geq \widetilde{p}_u$ with

$$\widetilde{p}_u := \left\lceil \frac{2cn^{1/3}s^{2\rho}}{\Delta_1(n^{-\frac{1}{3}})\cdot(2p+1)^{1/3}} + \frac{2cn^{2/3}s^\rho}{K\cdot\Delta_1(n^{-\frac{1}{3}})\cdot(2p+1)^{2/3}}\cdot\left(\Gamma+\frac{2}{n^{1/3}}\right) + 8s \right\rceil. \tag{23}$$

In the meantime, verifiably, $\widetilde{p}_u > s$. Since the above is a sufficient to ensure (22), we know that (41) in Proposition 24 holds for any $\widetilde{p} : \widetilde{p}_u \leq \widetilde{p} \leq p$. Due to Proposition 24, with probability at least $P^* := 1 - 6\exp\left(-\widetilde{p}_u\cdot(2p+1)\cdot\Delta_1(n^{-\frac{1}{3}})\right) - 2(p+1)\exp(-\widetilde{c}n) \geq 1 - 6\exp(-2c\cdot(2p+1)^{2/3}\cdot n^{1/3}) - 2(p+1)\exp(-\widetilde{c}n)$, it holds that

$$\begin{aligned}\mathbb{F}(\mathbf{X}^{RSAA}) - \mathbb{F}(\mathbf{X}^*) \leq\ & s\cdot P_\lambda(a\lambda) + \frac{2K}{\sqrt{n}}\sqrt{\frac{2\widetilde{p}_u(2p+1)}{c}\Delta_1(n^{-\frac{1}{3}})} \\ & + \frac{4K}{n}\frac{\widetilde{p}_u(2p+1)}{c}\Delta_1(n^{-\frac{1}{3}}) + 2\epsilon + \Gamma,\end{aligned} \tag{24}$$

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

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

$$\begin{aligned}&\sqrt{\frac{2\widetilde{p}_u\cdot(2p+1)}{cn}\Delta_1(n^{-\frac{1}{3}})} \qquad (25)\\ &\leq \sqrt{\frac{4\cdot(2p+1)s^{2\rho}}{cn\cdot(2p+1)^{1/3}}\Delta_1(n^{-\frac{1}{3}})\cdot\frac{cn^{1/3}}{\Delta_1(n^{-\frac{1}{3}})} + \frac{4cn^{2/3}(2p+1)s^\rho}{K(2p+1)^{2/3}\Delta_1(n^{-\frac{1}{3}})}\left(\Gamma+\frac{2}{n^{1/3}}\right)\cdot\frac{\Delta_1(n^{-\frac{1}{3}})}{cn}} \\ &\quad + \sqrt{\frac{2}{cn}\Delta_1(n^{-\frac{1}{3}})\cdot(8s+1)\cdot(2p+1)} \\ &\leq \sqrt{\frac{4(2p+1)^{2/3}s^{2\rho}}{n^{2/3}} + \frac{4s^\rho\cdot(\Gamma+\frac{2}{n^{1/3}})\cdot(2p+1)^{1/3}}{Kn^{1/3}}} + \sqrt{\frac{2}{nc}\Delta_1(n^{-\frac{1}{3}})\cdot(8s+1)\cdot(2p+1)}, \qquad (26)\end{aligned}$$

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

$$\begin{aligned}&\frac{3\widetilde{p}_u\cdot(2p+1)}{cn}\Delta_1(n^{-\frac{1}{3}})\\&\le\frac{4\cdot(2p+1)^{2/3}s^{2\rho}}{n^{2/3}}+\frac{2}{nc}\Delta_1(n^{-\frac{1}{3}})\,(8s+1)\cdot(2p+1)\\&\quad+\frac{4\cdot s^{\rho}\cdot(\Gamma+\frac{2}{n^{1/3}})}{K\cdot n^{1/3}}\cdot(2p+1)^{1/3}.\end{aligned}\tag{27}$$

Since (11) and $\Delta_1(n^{-\frac{1}{3}})=\ln(np)+\widetilde{\Delta}$, we have $\frac{4(2p+1)^{2/3}s^{2\rho}}{n^{2/3}}+\frac{4(\Gamma+\frac{2}{n^{1/3}})\cdot(2p+1)^{1/3}s^{\rho}}{Kn^{1/3}}\le O(1)$ and $\frac{2}{nc}\Delta_1(n^{-\frac{1}{3}})\,[8s+1]\cdot(2p+1)\le O(1)$. Therefore, it holds that $\frac{2\widetilde{p}_u}{cn}\Delta_1(n^{-\frac{1}{3}})(2p+1)\le O(1)\cdot\sqrt{\frac{(2p+1)^{2/3}s^{2\rho}}{n^{2/3}}+\frac{(\Gamma+\frac{2}{n^{1/3}})\cdot(2p+1)^{1/3}\cdot s^{\rho}}{Kn^{1/3}}}+O(1)\cdot\sqrt{\frac{\Delta_1(n^{-\frac{1}{3}})}{nc}\cdot(8s+1)\cdot(2p+1)}$. Combining the above with (26) and (27), the inequality in (24) can be simplified into $\mathbb{F}(\mathbf{X}^{RSAA})-\mathbb{F}(\mathbf{X}^*)\le O(1)s^{1-\rho}\cdot\frac{K\cdot\Delta_1(n^{-\frac{1}{3}})\cdot p^{2/3}}{c\cdot n^{2/3}}+O(1)\cdot K\cdot\sqrt{\frac{p^{2/3}s^{2\rho}}{n^{2/3}}+\frac{(\Gamma+\frac{2}{n^{1/3}})\cdot p^{1/3}s^{\rho}}{Kn^{1/3}}}+O(1)\cdot K\sqrt{\frac{sp}{nc}\Delta_1(n^{-\frac{1}{3}})}+\frac{2}{n^{1/3}}+\Gamma$. Together with $\Delta_1(n^{-\frac{1}{3}})\ge\ln 2$, $K\ge 1$, and $0<c\le 0.5$, the above becomes

$$\begin{aligned}&\mathbb{F}(\mathbf{X}^{RSAA})-\mathbb{F}(\mathbf{X}^*)\\&\le O(1)\cdot\left(\frac{s^{1-\rho}\cdot\Delta_1(n^{-1/3})\cdot p^{2/3}}{n^{2/3}}+\frac{p^{1/3}\cdot s^{\rho}}{n^{1/3}}+\sqrt{\frac{s\cdot p\cdot\Delta_1(n^{-1/3})}{n}}\right)\cdot K\\&\quad+O(1)\cdot\sqrt{\frac{K\cdot s^{\rho}\cdot p^{1/3}\cdot\Gamma}{n^{1/3}}}+\Gamma,\end{aligned}\tag{28}$$

which then shows Theorem 7 since $\Delta_1(n^{-\frac{1}{3}}):=\ln\left(18n^{1/3}(K_C+\mathcal{C}_\mu)\cdot p\cdot R\right)$. □

A.1.3 Proof of Corollary 10

Proof Lemma 27 implies that $\mathcal{F}_{n,\lambda}(\mathbf{X}^{RSAA},\,\mathbf{Z}_1^n)\le\mathcal{F}_{n,\lambda}(\mathbf{X}^*,\,\mathbf{Z}_1^n)+\lambda\|\mathbf{X}^*\|_*$ almost surely. Below we invoke the results from Theorem 7 with $\Gamma=\lambda\|\mathbf{X}^*\|_*$ and assumption that $\rho=0$ and $\lambda=\lambda(0)$. Note that it is assumed that

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

and $\frac{\Gamma}{K}\le\frac{\lambda\|\mathbf{X}^*\|_*}{K}\le\frac{\|\mathbf{X}^*\|_*\cdot\sqrt{\frac{8K\cdot(2p+1)^{2/3}}{c\cdot a\cdot n^{2/3}}[\ln(n^{1/3}p)+\widetilde{\Delta}]}}{K}$ (as well as $K\ge 1$). In view of (29), it then holds under Assumption 1 that $\frac{\Gamma}{K}\le Rs\cdot\sqrt{\frac{8(2p+1)^{2/3}}{cK\cdot a\cdot n^{2/3}}[\ln(n^{1/3}p)+\widetilde{\Delta}]}\le O(1)\cdot\sqrt{\frac{Rs}{a^{1/3}}[\ln(n^{1/3}p)+\widetilde{\Delta}]^{1/3}}$. Therefore, $\left(\frac{\Gamma}{K}\right)^3\le\left(O(1)\cdot\sqrt{\frac{Rs}{a^{1/3}}[\ln(n^{1/3}p)+\widetilde{\Delta}]^{1/3}}\right)^3\le O(1)\cdot R^{3/2}s^{3/2}\sqrt{a^{-1}\cdot[\ln(n^{1/3}p)+\widetilde{\Delta}]}$, for some universal constants $O(1)$. Furthermore, since $a<\mathcal{U}_L^{-1}\le 1$, it holds that, if n satisfies (13) for some universal constant C_2, then $n>O(1)\cdot p\cdot a^{-1}\cdot[\ln(n^{1/3}p)+\widetilde{\Delta}]\cdot s^{3/2}R^{3/2}\ge O(1)\cdot p\cdot R^{3/2}s^{3/2}\sqrt{a^{-1}\cdot[\ln(n^{1/3}p)+\widetilde{\Delta}]}+O(1)\cdot p+C_1\cdot s\cdot p\cdot\left(\ln(n^{1/3}p)+\widetilde{\Delta}\right)\ge C_1\cdot\left[\left(\frac{\Gamma}{K}\right)^3p+p+s\cdot p\cdot\left(\ln(n^{1/3}p)+\widetilde{\Delta}\right)\right]$.

Therefore, Theorem 7 is met and thus (12) in Theorem 7 implies that

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

with probability at least $1-2(2p+1)\exp(-\widetilde{c}n) - 6\exp\left(-2cn^{1/3}\cdot(2p+1)^{2/3}\right)$. Note that $a<1$, $K\ge 1$, $p\ge 1$, $\left[\ln(n^{1/3}p)+\widetilde{\Delta}\right]\ge 1$ and $\sqrt{\frac{sp\Delta_1(n^{-\frac{1}{3}})}{n}} \le \frac{s(2p+1)^{1/3}\cdot\sqrt{\Delta_1(n^{-\frac{1}{3}})}}{n^{1/3}}$ (due to (13) again). Hence, $\mathbb{F}(\mathbf{X}^{RSAA}) - \mathbb{F}(\mathbf{X}^*) \le O(1)\cdot K\cdot\left[\frac{sp^{2/3}\cdot(\ln(np)+\widetilde{\Delta})}{n^{2/3}} + \frac{p^{1/3}}{n^{1/3}}\right] + O(1)\cdot\frac{sRK\cdot(2p+1)^{1/3}}{\min\{a^{1/2}n^{1/3},\, a^{1/4}n^{1/3}\}}\left[\ln(n^{1/3}p)+\widetilde{\Delta}\right]^{1/2}$, which shows Part (ii) by further noticing that $a=\frac{1}{2\mathcal{U}_L}$ and $\mathcal{U}_L\ge 1$. □

A.1.4 Proof of Corollary 12

Proof The proof follows almost the same argument as in Sect. A.1.3 for proving Corollary 10, except that the choice of user-specific parameters are different. Again, Lemma 27 implies that $\mathcal{F}_{n,\lambda}(\mathbf{X}^{RSAA},\, \mathbf{Z}_1^n) \le \mathcal{F}_{n,\lambda}(\mathbf{X}^*,\, \mathbf{Z}_1^n) + \lambda\|\mathbf{X}^*\|_*$ almost surely. As the same in Part (ii), below we invoke the results from Theorem 7 with $\Gamma = \lambda\|\mathbf{X}^*\|_*$ and assumption that $\rho = 2/3$ and $\lambda = \lambda(\frac{2}{3})$. Note that it is assumed that

$$n > C_3\cdot p\cdot\mathcal{U}_L\cdot[\ln(np)+\widetilde{\Delta}]\cdot s^2R^{3/2} > O(1)\cdot p\cdot a^{-1}\cdot[\ln(np)+\widetilde{\Delta}]\cdot s^2R^{3/2}, \quad (30)$$

and $\frac{\Gamma}{K} \le \frac{\lambda\|\mathbf{X}^*\|_*}{K} \le \frac{\|\mathbf{X}^*\|_*\cdot\sqrt{\frac{8K\cdot(2p+1)^{2/3}\cdot s^{-2/3}}{c\cdot a\cdot n^{2/3}}[\ln(n^{1/3}p)+\widetilde{\Delta}]}}{K}$ (as well as $K\ge 1$). In view of (30), it then holds under Assumption 1 that $\frac{\Gamma}{K} \le Rs\cdot\sqrt{\frac{8(2p+1)^{2/3}s^{-2/3}}{cK\cdot a\cdot n^{2/3}}[\ln(n^{1/3}p)+\widetilde{\Delta}]} \le O(1)\cdot\sqrt{\frac{R}{a^{1/3}}[\ln(n^{1/3}p)+\widetilde{\Delta}]^{1/3}}$. Therefore, $\left(\frac{\Gamma}{K}\right)^3 \le \left(O(1)\cdot\sqrt{\frac{R}{a^{1/3}}[\ln(n^{1/3}p)+\widetilde{\Delta}]^{1/3}}\right)^3 \le O(1)\cdot R^{3/2}\sqrt{a^{-1}\cdot[\ln(n^{1/3}p)+\widetilde{\Delta}]}$, for some universal constants $O(1)$. Furthermore, since $a<\mathcal{U}_L^{-1}\le 1$, it holds that, if n satisfies (17), then $n > O(1)\cdot p\cdot a^{-1}\cdot[\ln(n^{1/3}p)+\widetilde{\Delta}]\cdot s^2R^{3/2} \ge O(1)\cdot p\cdot R^{3/2}s^2\sqrt{a^{-1}\cdot[\ln(n^{1/3}p)+\widetilde{\Delta}]} + O(1)\cdot s^2\cdot p + C_1 s\cdot p\cdot\left(\ln(n^{1/3}p)+\widetilde{\Delta}\right) \ge C_1\cdot\left[s^2\left(\frac{\Gamma}{K}\right)^3 p + s^2\cdot p + s\cdot p\cdot\left(\ln(n^{1/3}p)+\widetilde{\Delta}\right)\right]$. Therefore, (11) in Theorem 7 is met and thus (12) in Theorem 7 implies that

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

with probability at least $1-2(2p+1)\exp(-\widetilde{c}n) - 6\exp\left(-2cn^{1/3}\cdot(2p+1)^{2/3}\right)$. Note that $a<1$, $K\ge 1$, $p\ge s\ge 1$, $\left[\ln(n^{1/3}p)+\widetilde{\Delta}\right]\ge 1$ and $\sqrt{\frac{sp\Delta_1(n^{-\frac{1}{3}})}{n}} \le \frac{(2p+1)^{1/3}\cdot\sqrt{\Delta_1(n^{-\frac{1}{3}})}}{n^{1/3}}$ (in

view of (17) again). Hence, $\mathbb{F}(\mathbf{X}^{RSAA})-\mathbb{F}(\mathbf{X}^*) \le O(1)\cdot K\cdot\left[\frac{s^{1/3}p^{2/3}\cdot(\ln(np)+\widetilde{\Delta})}{n^{2/3}}+\frac{s^{2/3}\cdot p^{1/3}}{n^{1/3}}\right]+ O(1)\cdot\frac{s^{2/3}RK\cdot(2p+1)^{1/3}}{\min\{a^{1/2}n^{1/3},\, a^{1/4}n^{1/3}\}}\left[\ln(n^{1/3}p)+\widetilde{\Delta}\right]^{1/2}$, which shows Part (iii) by further noticing that $a=\frac{1}{2\mathcal{U}_L}$. □

A.1.5 Pillar results for sample complexity

Proposition 21 *Suppose that $a < {\mathcal{U}_L}^{-1}$. Assume that the $S^3ONC(\mathbf{Z}_1^n)$ 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 $\mathrm{S}^3ONC(\mathbf{Z}_1^n)$ almost surely, Eq. (9) implies that for any $j \in \{1, ..., p\}$, if $\sigma_j(\mathbf{X}^{RSAA}) \in (0,\, a\lambda)$, then

$$0 \le \mathcal{U}_L + \left[\frac{\partial^2 P_\lambda(|\sigma_j(\mathbf{X})|)}{[\partial\sigma_j(\mathbf{X})]^2}\right]_{\mathbf{X}=\mathbf{X}^{RSAA}} = \mathcal{U}_L - \frac{1}{a}. \tag{31}$$

Further observe that $\frac{\partial^2 P_\lambda(t)}{\partial t^2} = -a^{-1}$ for $t \in (0,\, a\lambda)$. Therefore, (31) contradicts with the assumption that $\mathcal{U}_L < \frac{1}{a}$. This contradiction implies that

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

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

Proposition 22 *Suppose that Assumptions 3 and 4 hold. Let $\epsilon \in (0,\, 1]$, $\widetilde{p} : \widetilde{p} > s$, $\Delta_1(\epsilon) := \ln\left(\frac{18\cdot(K_C+\mathcal{C}_\mu)\cdot p\cdot R}{\epsilon}\right)$, and $\mathcal{B}_{\widetilde{p},R} := \left\{\mathbf{X} \in \mathcal{S}_p : \sigma_{\max}(\mathbf{X}) \le R,\ \mathbf{rk}(\mathbf{X}) \le \widetilde{p}\right\}$. Then, for the same $c \in (0,\, 0.5]$ as in (8) and for some $\widetilde{c} > 0$,*

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

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

Proof We will follow the "ϵ-net" argument similar to [22] to construct a net of discretization grids $\mathcal{G}(\epsilon) := \{\widetilde{\mathbf{X}}^k\} \subseteq \mathcal{B}_{\widetilde{p},R}$ such that for any $\mathbf{X} \in \mathcal{B}_{\widetilde{p},R}$, there is $\mathbf{X}^k \in \mathcal{G}(\epsilon)$ that satisfies $\|\mathbf{X}^k - \mathbf{X}\| \le \frac{\epsilon}{2K_C+2\mathcal{C}_\mu}$ for any fixed $\epsilon \in (0,\, 1]$.

Invoking Lemma 28, for an arbitrary $\mathbf{X} \in \mathcal{B}_{\widetilde{p},R}$, to ensure that there always exists $\widetilde{\mathbf{X}}^k \in \mathcal{G}(\epsilon)$ that ensures $\left\|\mathbf{X} - \widetilde{\mathbf{X}}^k\right\| \le \frac{\epsilon}{(2K_C+2\mathcal{C}_\mu)}$, it is sufficient to have the number of grids to be no more than $\left(\frac{18R\sqrt{\widetilde{p}}\cdot(K_C+\mathcal{C}_\mu)}{\epsilon}\right)^{(2p+1)\widetilde{p}}$. Now, we may observe

$$\mathbb{P}\left[\max_{\mathbf{X}^k\in\mathcal{G}(\epsilon)}\left|\frac{1}{n}\sum_{i=1}^n f(\mathbf{X}^k, Z_i) - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n f(\mathbf{X}^k, Z_i)\right]\right| \le K\sqrt{\frac{t}{n}} + \frac{Kt}{n}\right]$$
$$= \mathbb{P}\left[\bigcap_{\mathbf{X}^k\in\mathcal{G}(\epsilon)}\left\{\left|\frac{1}{n}\sum_{i=1}^n f(\mathbf{X}^k, Z_i) - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n f(\mathbf{X}^k, Z_i)\right]\right| \le K\sqrt{\frac{t}{n}} + \frac{Kt}{n}\right\}\right]$$

$$\geq 1 - \sum_{\mathbf{X}^k \in \mathcal{G}(\epsilon)} \mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{X}^k, Z_i) - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{X}^k, Z_i)\right]\right| > K\sqrt{\frac{t}{n}} + \frac{Kt}{n}\right]. \quad (32)$$

Further invoking Eq. (8), for the same c as in (8), it holds that

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

Combined with Lemmas 25 and 26,

$$\max_{\mathbf{X} \in \mathcal{B}_{\widetilde{p},R},\, \mathbf{X}^k \in \mathcal{G}(\epsilon)} \left\{\left|\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{X}, Z_i) - \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{X}^k, Z_i)\right| + \left|\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{X}, Z_i)\right] - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{X}^k, Z_i)\right]\right|\right\}$$
$$\leq 2(K_C + \mathcal{C}_\mu) \cdot \frac{\epsilon}{2K_C + 2\mathcal{C}_\mu} = \epsilon, \quad (33)$$

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

$$\max_{\mathbf{X} \in \mathcal{B}_{\widetilde{p},R},\, \mathbf{X}^k \in \mathcal{G}(\epsilon)} \left\{\left|\mathcal{F}_n(\mathbf{X}, \mathbf{Z}_1^n) - \mathbb{E}\left[\mathcal{F}_n(\mathbf{X}, \mathbf{Z}_1^n)\right]\right| - \left|\mathcal{F}_n(\mathbf{X}^k, \mathbf{Z}_1^n) - \mathbb{E}\left[\mathcal{F}_n(\mathbf{X}^k, \mathbf{Z}_1^n)\right]\right|\right\} \leq \epsilon. \quad (34)$$

Further invoking (32), we now obtain that

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

with probability at least $1 - 2\left(\frac{18R\sqrt{\widetilde{p}} \cdot (K_C + \mathcal{C}_\mu)}{\epsilon}\right)^{(2p+1)\widetilde{p}} \cdot \exp(-ct) - 2\exp(-\widetilde{c} \cdot n)$. Finally, we may let $t := \frac{2\widetilde{p}}{c} \cdot (2p+1) \cdot \Delta_1(\epsilon)$, 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. □

Proposition 23 *Suppose that Assumptions 1 through 3 hold, the solution* $\mathbf{X}^{RSAA} \in \mathcal{S}_p : \sigma_{\max}(\mathbf{X}^{RSAA}) \leq R$ *satisfies* $S^3ONC(\mathbf{Z}_1^n)$ *almost surely,*

$$\mathcal{F}_{n,\lambda}(\mathbf{X}^{RSAA}, \mathbf{Z}_1^n) \leq \mathcal{F}_{n,\lambda}(\mathbf{X}^*, \mathbf{Z}_1^n) + \Gamma, \quad w.p.1. \quad (35)$$

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

$$(\widehat{p} - s) \cdot P_\lambda(a\lambda) > \frac{4K}{cn}\Delta_1(\epsilon) \cdot \widehat{p} \cdot (2p+1) + \frac{2K}{\sqrt{n}}\sqrt{\frac{2\widehat{p} \cdot (2p+1)}{c}\Delta_1(\epsilon)} + \Gamma + 2\epsilon, \quad (36)$$

for all $\widehat{p}: \widetilde{p}_u \le \widehat{p} \le p$, *then* $\mathbb{P}[\mathbf{rk}(\mathbf{X}^{RSAA}) \le \widetilde{p}_u - 1] \ge 1 - 2p\exp(-\widetilde{c}n) - 4\exp(-\widetilde{p}_u(2p+1)\Delta_1(\epsilon))$ *for the same* c *in* (8) *and some* $\widetilde{c} > 0$.

Proof This proof generalizes Proposition EC.3 from [13] bounding the sparsity of an S^{3}ONC solution to bounding the rank of an S^{3}ONC solution. Though the argument is similar, details are quite different and thus the result is different. Define $\mathcal{B}_R := \{\mathbf{X} \in \mathcal{S}_p : \sigma_{\max}(\mathbf{X}) \le R\}$. Define a few events:

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

where c in $\mathcal{E}_{5,\widehat{p}}$ is a universal constant defined to be the same as in (8), $\widehat{p}: \widetilde{p}_u \le \widehat{p} \le p$ and (thus $\widehat{p} > s$ by the assumption that $\widetilde{p}_u > s$). For any $(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Z}}_1^n) \in \{(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Z}}_1^n) \in \mathcal{E}_1\} \cap \{\widetilde{\mathbf{X}} \in \mathcal{E}_2 \cap \mathcal{E}_{3,p}\}$, where $\widetilde{\mathbf{Z}}_1^n = (\widetilde{Z}_1, ..., \widetilde{Z}_n)$, since $\widetilde{\mathbf{X}} \in \mathcal{E}_{3,p} \cap \mathcal{E}_2$, which means that $\widetilde{\mathbf{X}}$ has $\widehat{p}$-many non-zero singular values and each must not be within the interval $(0,\ a\lambda)$, it holds that

$$\mathcal{F}_n(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Z}}_1^n) + \widehat{p}P_\lambda(a\lambda) \le \frac{1}{n}\mathcal{F}_n(\mathbf{X}^*, \widetilde{\mathbf{Z}}_1^n) + sP_\lambda(a\lambda) + \Gamma, \tag{37}$$

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

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n f(\mathbf{X}^*, \widetilde{Z}_i) - \frac{1}{n}\sum_{i=1}^n f(\widetilde{\mathbf{X}}, \widetilde{Z}_i)\\
&= \left[\frac{1}{n}\sum_{i=1}^n f(\mathbf{X}^*, \widetilde{Z}_i) - \mathbb{F}(\mathbf{X}^*)\right] + \left[\mathbb{F}(\widetilde{\mathbf{X}}) - \frac{1}{n}\sum_{i=1}^n f(\widetilde{\mathbf{X}}, \widetilde{Z}_i)\right] + \left[\mathbb{F}(\mathbf{X}^*) - \mathbb{F}(\widetilde{\mathbf{X}})\right]\\
&\le 2\max_{\mathbf{X}\in\mathcal{E}_{3,p}}\left|\frac{1}{n}\sum_{i=1}^n f(\mathbf{X}, \widetilde{Z}_i) - \mathbb{F}(\mathbf{X})\right| + \mathbb{F}(\mathbf{X}^*) - \mathbb{F}(\widetilde{\mathbf{X}})\\
&\le 2\max_{\mathbf{X}\in\mathcal{E}_{3,p}}\left|\frac{1}{n}\sum_{i=1}^n f(\mathbf{X}, \widetilde{Z}_i) - \mathbb{F}(\mathbf{X})\right|,
\end{aligned} \tag{38}$$

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

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

Now let us examine the following set:

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

Combined with (37) and (38), $\Lambda \ne \emptyset \Longrightarrow (\widehat{p} - s)\cdot P_\lambda(a\lambda) \le \frac{2K}{\sqrt{n}}\sqrt{\frac{2\widehat{p}(2p+1)}{c}\Delta_1(\epsilon)} + \frac{2K}{n}\cdot\frac{2\widehat{p}(2p+1)}{c}\Delta_1(\epsilon) + 2\epsilon + \Gamma$, which contradicts with (36) for all $\widehat{p}: \widetilde{p}_u \le \widehat{p} \le p$. Now we recall the definition of $\mathbf{X}^{RSAA} \in \mathcal{B}_R$, which is a solution that satisfies the S^3*ONC*$(\mathbf{Z}_1^n)$,

w.p.1., and $\mathcal{F}_{n,\lambda}(\mathbf{X}^{RSAA},\ \widetilde{\mathbf{Z}}_1^n) \le \mathcal{F}_{n,\lambda}(\mathbf{X}^*,\ \widetilde{\mathbf{Z}}_1^n)+\Gamma$, w.p.1. Invoking Proposition 21, we have $\mathbb{P}\left[(\mathbf{X}^{RSAA}, \mathbf{Z}_1^n) \in \mathcal{E}_1 \cap \mathcal{E}_4,\ \mathbf{X}^{RSAA} \in \mathcal{E}_2\right] = 1$. Hence,

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

for all $\widehat{p}:\ \widetilde{p}_u \le \widehat{p} \le p$. The above then implies that $\mathbb{P}\left[\mathbf{Z}_1^n \notin \mathcal{E}_{5,\widehat{p}}\right] \ge \mathbb{P}\left[\mathbf{X}^{RSAA} \in \mathcal{E}_{3,p}\right]$ for all $\widehat{p}:\ \widetilde{p}_u \le \widehat{p} \le p$. Therefore, $\mathbb{P}[\mathbf{rk}(\mathbf{X}^{RSAA}) = \widehat{p}] \le 1 - \mathbb{P}\left[\mathbf{Z}_1^n \in \mathcal{E}_{5,\widehat{p}}\right]$ for all $\widehat{p}:\ \widetilde{p}_u \le \widehat{p} \le p$. Together with Proposition 22, we have that

$$\begin{aligned} \mathbb{P}[\mathbf{rk}(\mathbf{X}^{RSAA}) \le \widetilde{p}_u - 1] &= \mathbb{P}[\mathbf{rk}(\mathbf{X}^{RSAA}) \notin \{\widetilde{p}_u,\ \widetilde{p}_u + 1, ..., p\}] \\ &= 1 - \mathbb{P}\left[\bigcup_{\widehat{p}=\widetilde{p}_u}^{p} \{\mathbf{rk}(\mathbf{X}^{RSAA}) = \widehat{p}\}\right] \\ &\ge 1 - \sum_{\widehat{p}=\widetilde{p}_u}^{p} \mathbb{P}[\mathbf{rk}(\mathbf{X}^{RSAA}) = \widehat{p}] \\ &\ge 1 - \sum_{\widehat{p}=\widetilde{p}_u}^{p} \left(1 - \mathbb{P}\left[\mathbf{Z}_1^n \in \mathcal{E}_{5,\widehat{p}}\right]\right) \\ &\ge 1 - 2(p - \widetilde{p}_u + 1)\exp(-\widetilde{c}n) \\ &\quad - \sum_{\widehat{p}=\widetilde{p}_u}^{p} 2\exp\left(-\widehat{p}(2p+1)\cdot\Delta_1(\epsilon)\right). \end{aligned} \tag{39}$$

where $\widetilde{c} > 0$ is some universal constant. Observing that $\Delta_1(\epsilon) = \ln\left(\frac{18\cdot(K_C+\mathcal{C}_\mu)\cdot p\cdot R}{\epsilon}\right) > 1$ by observing that the above (39) involves a geometric sequence, we have

$$\mathbb{P}[\mathbf{rk}(\mathbf{X}^{RSAA}) \le \widetilde{p}_u - 1] \ge 1 - \frac{2\exp\left(-\widetilde{p}_u(2p+1)\Delta_1(\epsilon)\right)}{1-\exp\left(-(2p+1)\Delta_1(\epsilon)\right)} - 2p\exp(-\widetilde{c}n). \tag{40}$$

Further noting that $\frac{2\exp(-\widetilde{p}_u(2p+1)\Delta_1(\epsilon))}{1-\exp(-(2p+1)\Delta_1(\epsilon))} \le 4\exp\left(-\widetilde{p}_u(2p+1)\Delta_1(\epsilon)\right)$, we then have the desired result. □

Proposition 24 *Let* $\Delta_1(\epsilon) := \ln\left(\frac{18\cdot(K_C+\mathcal{C}_\mu)\cdot p\cdot R}{\epsilon}\right)$. *Assume that (i) the solution* $\mathbf{X}^{RSAA}$ *satisfies* $S^3ONC(\mathbf{Z}_1^n)$ *almost surely; (ii)* $\mathcal{F}_{n,\lambda}(\mathbf{X}^{RSAA},\ \mathbf{Z}_1^n) \le \mathcal{F}_{n,\lambda}(\mathbf{X}^*,\ \mathbf{Z}_1^n)+\Gamma$ *with probability one; and (iii) for some integer* $\widetilde{p}_u:\ \widetilde{p}_u > s$, *it holds that*

$$\widehat{p} > s + \frac{4K\cdot\widehat{p}\cdot(2p+1)}{cn\cdot P_\lambda(a\lambda)}\Delta_1(\epsilon) + \frac{2K}{\sqrt{n}\cdot P_\lambda(a\lambda)}\sqrt{\frac{2\widehat{p}\cdot(2p+1)}{c}\Delta_1(\epsilon)} + \frac{\Gamma+2\epsilon}{P_\lambda(a\lambda)}, \tag{41}$$

for all $\widetilde{p}:\ \widetilde{p}_u \le \widetilde{p} \le p$, *any* $\Gamma \ge 0$, *and any* $\epsilon \in (0,\ 1]$. *It then holds that*

$$\begin{aligned} &\mathbb{F}(\mathbf{X}^{RSAA}) - \mathbb{F}(\mathbf{X}^*) \\ &\le \frac{4K\cdot\widehat{p}\cdot(p+1)}{cn}\Delta_1(\epsilon) + \frac{2K}{\sqrt{n}}\sqrt{\frac{2\widehat{p}\cdot(2p+1)}{c}\Delta_1(\epsilon)} + \Gamma + 2\epsilon + sP_\lambda(a\lambda), \end{aligned} \tag{42}$$

with probability at least $P^* := 1 - 2(p+1)\exp(-\widetilde{c}n) - 6\exp\left(-\widetilde{p}_u(2p+1)\Delta_1(\epsilon)\right)$ *for some universal constant* $\widetilde{c} > 0$.

Proof We first observe that $\Delta_1(\epsilon) := \ln\left(\frac{18\cdot(K_C+\mathcal{C}_\mu)\cdot p\cdot R}{\epsilon}\right) \geq \ln 36$ because $p \geq 1$, $K_C,\ \mathcal{C}_\mu,\ R \geq 1$ 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 $\mathbf{rk}(\mathbf{X}^*) = s$, 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) $\mathbf{rk}(\mathbf{X}^{RSAA}) \leq \widetilde{p}_u$, (b) $\max_{\mathbf{X}\in\mathcal{B}_{\widetilde{p}_u,R}} \left|\frac{1}{n}\sum_{i=1}^n f(\mathbf{X}, Z_i) - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n f(\mathbf{X}, Z_i)\right]\right| \leq \frac{K}{\sqrt{n}}\sqrt{\frac{\widetilde{p}_u\cdot(2p+1)}{c}\Delta_1(\epsilon)} + \frac{K}{n}\frac{\widetilde{p}_u\cdot(2p+1)}{c}\Delta_1(\epsilon) + \epsilon$, we obtain that $\mathbb{F}(\mathbf{X}^{RSAA}) - \mathbb{F}(\mathbf{X}^*) \leq s\cdot P_\lambda(a\lambda) + \frac{2K}{\sqrt{n}}\sqrt{\frac{2\widetilde{p}_u\cdot(2p+1)}{c}\Delta_1(\epsilon)} + \frac{4K}{n}\frac{\widetilde{p}_u\cdot(2p+1)}{c}\Delta_1(\epsilon) + 2\epsilon + \Gamma$, a.s. Further invoking Propositions 22 and 23, we have that both events hold simultaneously with probability at least as in P^*, which verifiably implies the claimed results. □

A.1.6 Useful Lemmata

Lemma 25 *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*

$$\max_{\substack{\mathbf{X}_1,\,\mathbf{X}_2\in\mathcal{S}_p \\ \cap\{\mathbf{X}:\,\sigma_{\max}(\mathbf{X})\leq R,\\ \|\mathbf{X}_1-\mathbf{X}_2\|\leq\tau\}}} \{|\mathcal{F}_n(\mathbf{X}_1, \mathbf{Z}_1^n) - \mathcal{F}_n(\mathbf{X}_2, \mathbf{Z}_1^n)|\} \leq \left(2K_C + \mathcal{C}_\mu\right)\cdot\tau.$$

for any given $\tau \geq 0$.

Proof This proof follows a closely similar lemma by [22]. Similar proof has also been provided by [13], but some subtle differences in the problem context present and thus we redo the the proof herein. By Assumption 4, for some $c > 0$,

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

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

$$\mathbb{P}\left(\sum_{i=1}^n \frac{\mathcal{C}(Z_i)}{n} \leq 2K_C + \mathcal{C}_\mu\right) \leq 1 - 2\exp\left(-cn\right). \tag{43}$$

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 (43). □

Lemma 26 *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

$$\left|\mathbb{E}[\mathcal{F}_n(\mathbf{X}_1, \mathbf{Z}_1^n)] - \mathbb{E}[\mathcal{F}_n(\mathbf{X}_2, \mathbf{Z}_1^n)]\right| \leq \mathcal{C}_\mu \cdot \|\mathbf{X}_1 - \mathbf{X}_2\|. \tag{44}$$

Proof This proof follows a closely similar lemma by [22]. Again, a similar proof has also been provided by [13], but some subtle differences make it necessary to conduct the repetition herein. As per Assumption 4, it holds that

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

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

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

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 $i = 1, ..., n$, which immediately leads to the desired result. □

Lemma 27 *Denote that* $\mathbf{X}_\lambda^{\ell_1} \in \underset{\mathbf{X}\in\mathcal{S}^p}{\arg\min}\, \mathcal{F}_n(\mathbf{X},\ \mathbf{Z}_1^n) + \lambda\,\|\mathbf{X}\|_*$, *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 [13] 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

$$0 \leq P_\lambda(t) = \int_0^t \frac{[a\lambda - \theta]_+}{a} d\theta \leq \int_0^t \frac{a\lambda}{a} d\theta = \lambda \cdot t. \tag{45}$$

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

$$\mathcal{F}_n(\mathbf{X}_\lambda^{\ell_1},\ \mathbf{Z}_1^n) + \lambda\|\mathbf{X}_\lambda^{\ell_1}\|_* \leq \mathcal{F}_n(\mathbf{X}^*,\ \mathbf{Z}_1^n) + \lambda\|\mathbf{X}^*\|_*. \tag{46}$$

Combining (45) and (46), it holds that

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

as desired. □

Lemma 28 *Let* $S_{r,R} := \{X \in \Re^{p\times p} : \mathbf{rk}(X) \le r,\ \sigma_{\max}(X) \le R\}$. *Then, in terms of the Frobenius norm, there exists an* ϵ*-net* $\bar{S}_r$ *obeying* $|\bar{S}_r| \le \left(\frac{9\sqrt{r}R}{\epsilon}\right)^{(2p+1)r}$.

Proof The proof is closely similar to Lemma 3.1 of [4] with minor differences. We still present the proof here for completeness and for ensuring the minor difference would not result in a gap in our theory. 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 terns of Frobenius norm) for D with $|\bar{D}| \le \left(\frac{9\sqrt{r}R}{\epsilon}\right)^r$.

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

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 $\|U-\bar{U}\|_{1,2} \le \epsilon/(3\sqrt{r}R)$, $\|V-\bar{V}\|_{1,2} \le \epsilon/(3\sqrt{r}R)$, and $\|\Sigma-\bar{\Sigma}\|_F \le \epsilon/3$. This gives $\|X-\bar{X}\|_F = \|U\Sigma V^\top - \bar{U}\bar{\Sigma}\bar{V}^\top\|_F = \|U\Sigma V^\top - \bar{U}\Sigma V^\top + \bar{U}\Sigma V^\top - \bar{U}\bar{\Sigma}V^\top + \bar{U}\bar{\Sigma}V^\top - \bar{U}\bar{\Sigma}\bar{V}^\top\|_F \le \|(U-\bar{U})\Sigma V^\top\|_F + \|\bar{U}(\Sigma-\bar{\Sigma})V^\top\|_F + \|\bar{U}\bar{\Sigma}(V-\bar{V})\|_F$. Since V is orthonormal matrix, $\|(U-\bar{U})\Sigma V^\top\|_F = \|(U-\bar{U})\Sigma\|_F = \sqrt{\sum_{1\le j\le r}[\sigma_j(X)]^2\cdot\|\bar{U}_j - U_j\|_2^2} \le \sqrt{\|\Sigma\|_F^2\cdot\|U-\bar{U}\|_{1,2}^2} \le \epsilon/3$, where U_j is the jth column of U. By a symmetric argument, we may also obtain that $\|\bar{U}\bar{\Sigma}(V-\bar{V})^\top\|_F \le \epsilon/3$. To bound the second term, we also notice that $\|\bar{U}(\Sigma-\bar{\Sigma})V^\top\|_F = \|\Sigma-\bar{\Sigma}\|_F \le \epsilon/3$. Combining the above provides the desired result. □

A.2 Proof of results concerning the computation of an S^{3}ONC solution

Proof of Theorem 18 **To show the closed-form solution.** Without loss of generality, we may reduce (20) into the following problem: For a fixed $\mathbf{Y} \in \mathcal{S}_p$, solve the minimization problem of

$$\min_{\mathbf{X}\in\mathcal{S}_p^+} G(\mathbf{X}) := \frac{L}{2}\|\mathbf{X}-\mathbf{Y}\|_F^2 + \sum_{j=1}^p P_\lambda(\sigma_j(\mathbf{X})). \tag{47}$$

In fact, when $\mathbf{Y} := \mathbf{X}^0 - \frac{1}{L}\nabla f(\mathbf{X}^0)$, we recover the exact form of (20).

We will divide the rest of the proof of the closed-form solution into three steps.

Step 1. This step shows a useful inequality $\widehat{\sigma}_j \notin (0,\ a\lambda)$ for all $j = 1, ..., p$, where $\widehat{\sigma}_j$ is the optimal solution to (53).

Observe that the optimal solution must satisfy the second-order necessary conditions, when the second-order derivative exists. In particular, when $\widehat{\sigma}_j \in (0, a\lambda)$, the second-order necessary condition is written as $\left[\frac{\partial^2\left[\frac{L}{2}(\sigma_j(\mathbf{Y})-\sigma_j)^2+P_\lambda(\sigma_j)\right]}{\partial\sigma_j^2}\right]_{\sigma_j=\widehat{\sigma}_j} = L - \frac{1}{a} \ge 0$. However, the last inequality here contradicts with the assumption that $\frac{1}{a} > L$. This contradiction indicates that

$$\widehat{\sigma}_j \notin (0,\ a\cdot\lambda), \tag{48}$$

as desired in this step.

Step 2. We will further derive a sequence of equivalent reformulations to (20). These reformulations will eventually lead to the conclusion that (20) is equivalent to solving a sequence of one-dimensional problems.

Because $\mathbf{Y}$ is symmetric, it admits eigndecomposition $\mathbf{Y} = Q\Sigma_{\mathbf{Y}}Q^{-1}$. By the orthogonal invariance of the Frobenius norm, (47) is reformulated into

$$\min_{\mathbf{X}\in\mathcal{S}_p^+} \frac{L}{2}\|\Sigma_{\mathbf{Y}} - Q^{-1}\mathbf{X}Q\|_F^2 + \sum_{j=1}^{p} P_\lambda(\sigma_j(\mathbf{X})). \tag{49}$$

In view of the results from Step 1 and the fact that $P_\lambda(t) = \frac{a\lambda^2}{2}$ for all $t \geq a\lambda$, the reformulation in (49) is equivalent to

$$\min_{\mathbf{X}\in\mathcal{S}_p^+} \frac{L}{2}\|\Sigma_{\mathbf{Y}} - Q^{-1}\mathbf{X}Q\|_F^2 + \sum_{j=1}^{p} \frac{a\lambda^2}{2}\cdot\mathbb{I}(\sigma_j(\mathbf{X}) \neq 0). \tag{50}$$

where $\mathbb{I}(\sigma_j(\mathbf{X}) \neq 0)$ is the index function that outputs the value 1 if $\sigma_j(\mathbf{X}) \neq 0$ and outputs 0, otherwise.

The optimal solution $\widehat{\mathbf{X}}$ is an element of $\mathcal{S}_p^+$ (that is, $\widehat{\mathbf{X}}$ must be symmetric and positive semidefinite) and thus admits an eigendecomposition. We may verify that $\widehat{\mathbf{X}} = Q\Lambda_{\widehat{\mathbf{X}}}Q^{-1}$, where the same Q is shared by both decompositions. This is because, for any feasible solution $\mathbf{X}$, we may write $Q^{-1}\mathbf{X}Q := \Lambda_{\mathbf{X}} + H$, where $\Lambda_{\mathbf{X}}$ is a diagonal matrix and H is a hollow matrix, and thus $\|\Sigma_{\mathbf{Y}} - \Lambda_{\mathbf{X}}\|_F^2 = \|\Sigma_{\mathbf{Y}} - \Lambda_{\mathbf{X}}\|_F^2 + \|H\|_F^2$. If $\|H\|_F \neq 0$, then one may always construct a solution that has a smaller objective function value for (50). Therefore, (47) is equivalently rewritten as

$$\min_{\Sigma_{\mathbf{X}}} \left\{ \frac{L}{2}\|\Sigma_{\mathbf{Y}} - \Sigma_{\mathbf{X}}\|_F^2 + \sum_{j=1}^{p} P_\lambda(\sigma_j(\mathbf{X})) : \right.$$
$$\left. \Sigma_{\mathbf{X}} \text{ is a positive semidefinite and diagonal matrix} \right\}. \tag{51}$$

If the optimal solution is $\Sigma_{\widehat{\mathbf{X}}}$, then the optimal solution to (47) can be recovered as $Q\Sigma_{\widehat{\mathbf{X}}}Q^{-1}$. By noticing that both $\Sigma_{\mathbf{Y}}$ and $\Sigma_{\widehat{\mathbf{X}}}$ are diagonal matrices, we may further reformulate the above problem into the below:

$$\min_{(\sigma_j)} \left\{ \frac{L}{2}\sum_{j=1}^{p}\big(\sigma_j(\mathbf{Y}) - \sigma_j\big)^2 + \sum_{j=1}^{p} P_\lambda(\sigma_j) : \sigma_j \geq 0,\ \forall j = 1, ..., p \right\}. \tag{52}$$

Let $\widehat{\sigma}_j$, $j = 1, ..., p$, be the optimal solution. Then, the optimal solution to (47) can be recovered as $Q\,\mathrm{diag}(\{\widehat{\sigma}_j : j = 1, ..., p\})Q^{-1}$. Furthermore, (52) can be solved by solving a sequence of one-dimensional optimization problems: For all $j = 1, ..., p$,

$$\min \left\{ g(\sigma_j) := \frac{L}{2}\big(\sigma_j(\mathbf{Y}) - \sigma_j\big)^2 + P_\lambda(\sigma_j) : \sigma_j \geq 0 \right\}. \tag{53}$$

This completes Step 1 of our proof. This one-dimensional optimization formulation will be essential to the rest of the proof.

Step 3. We now start proving the correctness of the claimed closed form solution. We will consider three different cases.

- **Case 3.1.** If $\sigma_j(\mathbf{Y}) < a\lambda$, we will show below that it must hold that $\widehat{\sigma}_j = 0$. Suppose the otherwise; that is, $\widehat{\sigma}_j > 0$. Then (48) implies that $\widehat{\sigma}_j \geq a\lambda$. Then (recalling g defined in (53)) $g(\widehat{\sigma}_j) - g(0) = \frac{L}{2}(\sigma_j(\mathbf{Y}) - \widehat{\sigma}_j)^2 + \frac{a\lambda^2}{2} - \frac{L}{2}[\sigma_j(\mathbf{Y})]^2 = \frac{L}{2}\widehat{\sigma}_j^2 - L\sigma_j(\mathbf{Y}) \cdot \widehat{\sigma}_j + \frac{a\lambda^2}{2} > \frac{L}{2}\widehat{\sigma}_j^2 - La\lambda \cdot \widehat{\sigma}_j + \frac{a^2\lambda^2 \cdot L}{2} = \frac{L}{2}(\widehat{\sigma}_j - a\lambda)^2$, where the second from last relationship is due to $\sigma_j(\mathbf{Y}) < a\lambda$ and $aL < 1$ by assumption. To summarize, we have shown that $g(\widehat{\sigma}_j) > g(0)$, which means that we have identified a strictly better solution 0 than a solution $\widehat{\sigma}_j > 0$. Further invoking (48) again, we know that the only choice of the optimal solution is then $\widehat{\sigma}_j = 0$.
- **Case 3.2.** If $\sigma_j(\mathbf{Y}) \geq a\lambda$, we will show below that $\widehat{\sigma}_j = \sigma_j(\mathbf{Y})$. To that end, we will first show that $\sigma_j(\mathbf{Y})$ is a better solution than 0. This can be seen by observing that $g(\sigma_j(\mathbf{Y})) - g(0) = \frac{a\lambda^2}{2} - \frac{L}{2} \cdot [\sigma_j(\mathbf{Y})]^2$. Because $\sigma_j(\mathbf{Y}) \geq a\lambda$, we know that $g(\sigma_j(\mathbf{Y})) - g(0) = \frac{a\lambda^2}{2} - \frac{L}{2}a^2 \cdot \lambda^2 > 0$, where the last inequality is due to the assumption that $aL < 1$.

  Now that $\sigma_j(\mathbf{Y})$ is a better solution than 0, we can further verify the closed form that $\widehat{\sigma}_j = \sigma_j(\mathbf{Y})$. Because $\widehat{\sigma}_j \neq 0$, by (48), it must hold that $\widehat{\sigma}_j \geq a\lambda$. In such a case, Problem (52), is equivalent to the optimal solution to $\min\left\{\frac{L}{2}\left(\sigma_j(\mathbf{Y}) - \sigma_j\right)^2 + \frac{a\lambda^2}{2} : \sigma_j \geq 0\right\}$, which can be further rewritten into $\min\left\{\frac{L}{2}\left(\sigma_j(\mathbf{Y}) - \sigma_j\right)^2 : \sigma_j \geq 0\right\}$ by removing a constant term. Therefore, the optimal solution must be $\widehat{\sigma}_j = \sigma(\mathbf{Y})$.

This then completes the proof of the closed-form solution.

To show the satisfaction of S^3ONC. For the second part of the theorem, we observe that the closed form solution obeys that $\sigma_j(\widehat{\mathbf{X}}) \notin (0,\ a\lambda)$ for all j. By definition, it is an S^3ONC solution. □

References

1. Ariyawansa, K.A., Zhu, Y.: Stochastic semidefinite programming: a new paradigm for stochastic optimization. 4OR **4**(3), 239–253 (2006)
2. Bartlett, P.L., Jordan, M.I., McAuliffe, J.D.: Convexity, classification, and risk bounds. J. Am. Stat. Assoc. **101**(473), 138–156 (2006)
3. Cai, J.-F., Candés, E.J., Shen, Z.: A singular value thresholding algorithm for matrix completion. SIAM J. optimization **20**(4), 1956–1982 (2010)
4. Candes, E.J., Plan, Y.: 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). https://doi.org/10.1109/TIT.2011.2111771
5. Clémençon, S., Lugosi, G., Vayatis, N.: Ranking and empirical minimization of U-statistics. Ann. Stat. **36**(2), 844–874 (2008)
6. Elsener, A., van de Geer, S.: Robust low-rank matrix estimation. Ann. of Stat. **46**(6B), 3481–3509 (2018)
7. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Stat. Assoc. **96**(456), 1348–1360 (2001). https://doi.org/10.1198/016214501753382273
8. Grant, M., Boyd, S.: CVX: Matlab software for disciplined convex programming, version 2.0 beta. Online at: http://cvxr.com/cvx, (2013)
9. Grant, M., Boyd, S.: Graph implementations for nonsmooth convex programs, Recent Advances in Learning and Control (a tribute to M. Vidyasagar), V. Blondel, S. Boyd, and H. Kimura, editors, pages 95-110, Lecture Notes in Control and Information Sciences, Springer, (2008). http://stanford.edu/~boyd/graph_dcp.html
10. Jain, P., Tewari, A., Kar, P.: On iterative hard thresholding methods for high-dimensional m-estimation. In *Advances in Neural Information Processing Systems*, pp. 685-693, (2014)
11. Koltchinskii, V.: Rademacher complexities and bounding the excess risk in active learning. J. Mach. Learn. Res. **11**, 2457–2485 (2010)
12. Liu, H., Lee, H. Y., Huo, Z.: Linearly constrainted high-dimensional learning. working paper. (2019)
13. Liu, H., Ye, Y.: High-dimensional learning under approximate sparsity: Towards a unified framework for nonsmooth learning and regularized neural networks. *arXiv e-prints*, art. arXiv:1903.00616, (2019)

14. Liu, H., Yao, T., Li, R., Ye, Y.: Folded concave penalized sparse linear regression: sparsity, statistical performance, and algorithmic theory for local solutions. Math. Program. **166**(1), 207–240 (2017). https://doi.org/10.1007/s10107-017-1114-y
15. Liu, H., Wang, X., Yao, T., Li, R., Ye, Y.: Sample average approximation with sparsity-inducing penalty for high-dimensional stochastic programming. Math. Program. (2018). https://doi.org/10.1007/s10107-018-1278-0
16. Moulines, E., Bach, F.: Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pp. 451-459. (2011)
17. Negahban, S., Wainwright, M.J.: Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. J. Mach. Learn. Res. **13**(May), 1665–1697 (2012)
18. Rohde, A., Tsybakov, A.B.: Estimation of high-dimensional low-rank matrices. Ann. of Stat. **39**(2), 887–930 (2011)
19. Ruszczyński, A., Shapiro, A.: Stochastic programming models. In *Stochastic Programming*, volume 10 of *Handbooks in Operations Research and Management Science*, 1 – 64. Elsevier, (2003a). https://doi.org/10.1016/S0927-0507(03)10001-1
20. Ruszczyński, A., Shapiro, A.: Optimality and duality in stochastic programming. In *Stochastic Programming*, volume 10 of *Handbooks in Operations Research and Management Science*, 65 – 139. Elsevier, (2003b). https://doi.org/10.1016/S0927-0507(03)10002-3
21. Shapiro, A., Xu, H.: Uniform laws of large numbers for set-valued mappings and subdifferentials of random functions. J. Math. Anal. Appl. **325**(2), 1390–1399 (2007). https://doi.org/10.1016/j.jmaa.2006.02.078
22. Shapiro, A., Dentcheva, D., Ruszczynski, A.: Lectures on Stochastic Programming: Modeling and Theory, 2nd edn. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA (2014)1611973422, 9781611973426
23. Tanner, J., Wei, K.: Normalized iterative hard thresholding for matrix completion. SIAM J. Sci. Comput. **35**(5), S104–S125 (2013)
24. Vandenberghe, L., Boyd, S.: Applications of semidefinite programming. Appl. Numer. Math. **29**(3), 283–299 (1999)
25. Vandenberghe, L., Boyd, S.: Semidefinite programming. SIAM Rev. **38**(1), 49–95 (1996)
26. Zhang, C.-H.: Nearly unbiased variable selection under minimax concave penalty. Ann. Stat. **38**(2), 894–942 (2010). https://doi.org/10.1214/09-AOS729