Math. Program., Ser. A (2019) 178:69–108
https://doi.org/10.1007/s10107-018-1278-0

FULL LENGTH PAPER

Sample average approximation with sparsity-inducing penalty for high-dimensional stochastic programming

Hongcheng Liu[1] · **Xue Wang**[2] · **Tao Yao**[2] · **Runze Li**[3] · **Yinyu Ye**[4]

Received: 3 June 2016 / Accepted: 13 April 2018 / Published online: 3 May 2018

Abstract The theory on the traditional sample average approximation (SAA) scheme for stochastic programming (SP) dictates that the number of samples should be polynomial in the number of problem dimensions in order to ensure proper optimization accuracy. In this paper, we study a modification to the SAA in the scenario where the global minimizer is either sparse or can be approximated by a sparse solution. By making use of a regularization penalty referred to as the folded concave penalty (FCP), we show that, if an FCP-regularized SAA formulation is solved locally, then the required number of samples can be significantly reduced in approximating the global solution of a convex SP: the sample size is only required to be poly-logarithmic in the

✉ Tao Yao
taoyao@psu.edu

Hongcheng Liu
liu.h@ufl.edu

Xue Wang
xzw118@psu.edu

Runze Li
rli@stat.psu.edu

Yinyu Ye
yinyu-ye@stanford.edu

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

[2] Harold and Inge Marcus Department of Industrial and Manufacturing Engineering, The Pennsylvania State University, University Park, PA 16802, USA

[3] Department of Statistics, The Pennsylvania State University, University Park, PA 16802, USA

[4] Department of Management Science and Engineering, Stanford University, Stanford, CA 94305, USA

number of dimensions. The efficacy of the FCP regularizer for nonconvex SPs is also discussed. As an immediate implication of our result, a flexible class of folded concave penalized sparse M-estimators in high-dimensional statistical learning may yield a sound performance even when the problem dimension cannot be upper-bounded by any polynomial function of the sample size.

Keywords Sample average approximation · Folded concave penalty · Second order necessary condition · Stochastic programming

Mathematics Subject Classification 90C15 · 65C05 · 90C26 · 62J07

1 Introduction

We are interested in solving stochastic programming (SP) when the problem dimension is high but the global solution is approximately sparse. Denote by W a random vector with probability distribution $\mathbb{P}$ and support $\mathcal{W} \subseteq \Re^q$ for some $q > 0$. Define by $f(\cdot,\cdot) : \mathcal{X} \times \mathcal{W} \to \Re$ a deterministic mapping, where $\mathcal{X} \subseteq \Re_+^p$ for some integer $p > 0$ is a compact and convex feasible region. Let $\mathbb{E}[f(\mathbf{x}, W)] = \int_{\mathcal{W}} f(\mathbf{x}, w)\mathbb{P}(dw)$. Assume that, for every $\mathbf{x} \in \mathcal{X}$, the function $f(\mathbf{x}, \cdot)$ is measurable and integrable on $\mathcal{W}$. Then, the SP formulation of consideration is given as:

$$\min_{\mathbf{x}\in\mathcal{X}}\{F(\mathbf{x}) := \mathbb{E}[f(\mathbf{x}, W)]\}, \tag{1}$$

Throughout the paper, we assume that $\mathcal{X}$ is defined only by coordinate-wise constraints, that is, $\mathcal{X} := \{\mathbf{x} = (x_i) : x_i \in X_i, i = 1, \ldots, p\}$ for some $X_i \subseteq \Re_+$ for all $i = 1, \ldots, p$. Notice that the non-negativity constraints are not restrictive, in that we may always represent a negative variable by the difference of two non-negative variables.

In addition, we will restrict our discussions to the cases where the solution to the original SP, denoted $\mathbf{x}^{\min} \in \arg\min_{\mathbf{x}\in\mathcal{X}} F(\mathbf{x})$, can be well approximated by a sparse solution. More precisely, we assume that there exists $\hat{\mathbf{x}}^{\min}$ that satisfies

$$F(\hat{\mathbf{x}}^{\min}) - F(\mathbf{x}^{\min}) \le \hat{\varepsilon} \tag{2}$$

for some $\hat{\varepsilon} \ge 0$. We denote that $\mathcal{S} := \{i : \hat{x}_i^{\min} > 0\}$ and $\mathcal{S}^c := \{i : \hat{x}_i^{\min} = 0\}$. Here $\mathcal{S}$ can be understood as the index set for the most contributing dimensions with $|\mathcal{S}|$ assumed small and satisfying $|\mathcal{S}| << p$ and $|\mathcal{S}| < n$. In the special case when $\hat{\varepsilon} = 0$, we know that $\hat{\mathbf{x}}^{\min}$ is an exact solution to (1).

Under the above setting, one of the most commonly used techniques to solve the SP, the sample average approximation (SAA), is undesirably restrictive on the sample size in some scenarios. The SAA approximates the objective function of (1) by

$$F_n(\mathbf{x}) := \frac{1}{n}\sum_{j=1}^{n} f(\mathbf{x}, W^j) \tag{3}$$

where $\{W^j : 1, \ldots, n\}$ is a sequence of independently and identically distributed (i.i.d.) random samples of W. Denote that $\mathbf{x}^{SAA} \in \arg\min_{\mathbf{x}\in\mathcal{X}} F_n(\mathbf{x})$. Much literature has discussed the efficacy of $\mathbf{x}^{SAA}$ in approximating $\mathbf{x}^{\min}$ (see [9,21]). It has been shown in the celebrated work by Shapiro et al. [14,20,21] that to ensure the optimization accuracy, the required number of samples should be larger than the number of dimensions and should grow polynomially with the increase of dimensionality. In specific, to ensure

$$\mathbf{P}\left[F(\mathbf{x}^{SAA}) - F(\mathbf{x}^{\min}) \le \epsilon\right] \ge 1 - \alpha, \tag{4}$$

for any $\epsilon \in (0, 1]$ and $\alpha \in (0, 1]$, the sample size n should satisfy

$$n \gtrsim \frac{p}{\epsilon^2} \ln \frac{1}{\epsilon} + \frac{1}{\epsilon^2} \ln \frac{1}{\alpha}, \tag{5}$$

where $x \gtrsim y$ for any $x, y \in \Re$ means $x \ge \tilde{c}y$, for some constant $\tilde{c} > 0$ that are independent of α, ϵ, p, and $|\mathcal{S}|$, but may depend polynomially on some other problem quantities. Consider (5) in a problem with perhaps hundreds of thousands of dimensions, which is not rare in actual applications of SP. The SAA then likely requires more than millions or even tens of millions of samples for the SAA to perform properly. The overhead in generating these samples, before conducting any optimization-related computation, may have already become prohibitive. Especially considering the case where the most contributing dimensions are in tens or hundreds, such a sample size requirement seems unreasonably demanding.[1]

Seeking to address the above issue, this paper studies a modification to (3) by adding a regularization term to encourage sparsity. This term is in the form of a folded concave penalty (FCP) as first introduced by [10,27] to some statistical learning problems. We refer to this modification the regularized SAA (RSAA), which is formulated as

$$\min_{\mathbf{x}\in\mathcal{X}} \left\{ F_{n,\lambda}(\mathbf{x}) := F_n(\mathbf{x}) + \sum_{i=1}^{p} P_\lambda(x_i) \right\} \tag{6}$$

where P_λ with parameters $a > 0$ and $\lambda > 0$ is a special form of FCP called the minimax concave penalty (MCP) [27]:

$$P_\lambda(\tau) := \int_0^\tau \frac{(a\lambda - t)_+}{a}\, dt = \begin{cases} \lambda\tau - \frac{\tau^2}{2a} & \text{if } 0 \le \tau \le a\lambda; \\ \frac{1}{2}a\lambda^2 & \text{if } \tau > a\lambda. \end{cases} \tag{7}$$

We show in this paper that the RSAA allows the dimension to be (much) more than the sample size. In specific, when $\hat{\varepsilon} = 0$, to achieve the same optimization quality in (4), the sample size requirement for the global minimizer to RSAA is

[1] This is because, if only we would know which dimensions are nonzero, we may equivalently reduce the problem to one that has only tens or hundreds of dimensions. Then, according to (5), the required sample size would likely be only in thousands.

$$n \gtrsim \frac{|\mathcal{S}|}{\epsilon^3}\left(\ln \frac{p}{\epsilon}\right)^{1.5} + \frac{1}{\epsilon^2}\ln\frac{1}{\alpha}, \tag{8}$$

under no assumption of convexity. Compared to (5), the required sample size of RSAA only depends polynomially on $|\mathcal{S}|$ and $\ln p$, instead of p. Although, as a tradeoff, the dependency on ϵ becomes worse after regularization, we believe that such a tradeoff can be well compensated by the efficiency in handling high dimensionality at least for some applications.

Perhaps more importantly, we further consider stationary points that satisfy the significant subspace second-order necessary condition (S^{3}ONC) [16], which is weaker than the second-order KKT condition. When $\hat{\varepsilon} = 0$, we show that, if an S^{3}ONC solution is achieved by a(n arbitrary) descent local algorithm starting at an all-zero vector, then the sample size is required to be

$$n \gtrsim \frac{|\mathcal{S}|^{2.5}}{\epsilon^4}\left(\ln \frac{p}{\epsilon}\right)^{2} + \frac{1}{\epsilon^2}\ln\frac{1}{\alpha}, \tag{9}$$

if $f(\cdot, W)$ is convex for almost every $W \in \mathcal{W}$. Furthermore, assume in addition that F is differentiable and strongly convex. Then a smaller sample size is allowed, that is,

$$n \gtrsim \frac{|\mathcal{S}|^{1.5}}{\epsilon^3}\left(\ln \frac{p}{\epsilon}\right)^{1.5} + \frac{1}{\epsilon^2}\ln\frac{1}{\alpha}. \tag{10}$$

Both bounds are worse than (8) in terms of $|\mathcal{S}|$ and/or ϵ, but present similar levels of efficacy in addressing high dimensionality as in (8). Meanwhile, the computational overhead in solving for an S^{3}ONC solution is largely reduced compared to that in solving for a global solution.

Furthermore, it is worthwhile to mention a special case to demonstrate RSAA's efficacy. Assume again that $f(\cdot, W)$ is convex for almost every $W \in \mathcal{W}$, function F is differentiable and strongly convex, and $\hat{\varepsilon} = 0$. If all of the most contributing dimensions have a reasonably large magnitude that differentiates them from zero, that is, the value of $\min_{i\in\mathcal{S}} |x_i^{\min}|$ is above a certain threshold dependent only on $|\mathcal{S}|$ and the modulus of strong convexity, then the required sample size becomes as small as

$$n \gtrsim \frac{|\mathcal{S}|}{\epsilon^2}\ln\frac{p}{\epsilon} + \frac{1}{\epsilon^2}\ln\frac{1}{\alpha}, \tag{11}$$

for an S^{3}ONC solution. In contrast, under the same set of assumptions, the best known bound on the performance of traditional SAA is still (5), this means that, at least for some scenarios, the proposed RSAA may achieve a non-trivial improvement to SAA in handling high dimensionality without any compromise in terms of dependencies on $|\mathcal{S}|$, ϵ, and α. A summary of comparisons between RSAA and SAA is provided in Table 1 given $\hat{\varepsilon} = 0$.

When the exact global solution to the SP is not sparse but can be approximated by a sparse solution, i.e., $\hat{\varepsilon} > 0$, it turns out that the sample size should grow polynomially in $\hat{\varepsilon}$ and that there can also be a residual suboptimality gap linear in $\hat{\varepsilon}$. However,

Table 1 A summary of sample size requirement to guarantee optimization quality of (4) when $\hat{\varepsilon} = 0$ as defined in (2)

	$n \gtrsim$	Global	$f(\cdot, W)$ convex	F strongly convex and differentiable	$\min_{i \in \mathcal{S}} \hat{x}_i^{\min} \geq$ threshold
SAA	$\frac{p}{\epsilon^2} \ln \frac{1}{\epsilon} + \frac{1}{\epsilon^2} \ln \frac{1}{\alpha}$	✓	×	×	×
RSAA	$\frac{\lvert\mathcal{S}\rvert}{\epsilon^3} \left(\ln \frac{p}{\epsilon}\right)^{1.5} + \frac{1}{\epsilon^2} \ln \frac{1}{\alpha}$	✓	×	×	×
	$\frac{\lvert\mathcal{S}\rvert^{2.5}}{\epsilon^4} \left(\ln \frac{p}{\epsilon}\right)^{2} + \frac{1}{\epsilon^2} \ln \frac{1}{\alpha}$	×	✓	×	×
	$\frac{\lvert\mathcal{S}\rvert^{1.5}}{\epsilon^3} \left(\ln \frac{p}{\epsilon}\right)^{1.5} + \frac{1}{\epsilon^2} \ln \frac{1}{\alpha}$	×	✓	✓	×
	$\frac{\lvert\mathcal{S}\rvert}{\epsilon^2} \ln \frac{p}{\epsilon} + \frac{1}{\epsilon^2} \ln \frac{1}{\alpha}$	×	✓	✓	✓

The "Global" column indicates whether the approximation formulation being solved globally (✓) or locally (×) is one of the conditions for the bounds on "n" of the same row; the "$f(\cdot, W)$ convex" and the "$\min_{i \in \mathcal{S}} \hat{x}_i^{\min} \geq$ threshold" columns indicate whether (✓) or not (×) Function $f(\cdot, W)$ being convex for a.e. $W \in \mathcal{W}$ and $\min_{i \in \mathcal{S}} \hat{x}_i^{\min}$ being above a certain threshold are conditions for the corresponding bounds on "n", respectively

the poly-logarithmic dependency of sample size requirement on the dimensionality is maintained.

Since second-order KKT condition implies S^{3}ONC, all numerical algorithms that ensure the second-order KKT condition (e.g., [4,7,19,25,26]) guarantee S^{3}ONC. Some of these algorithms such as the interior point methods in [4] are fully polynomial-time approximation schemes (FPTAS). Meanwhile, as we will illustrate later, computing the global minimizer may also be possible via a mixed integer programming reformulation.

Regularizing the SP solution schemes with a sparsity-inducing penalty for an important class of SP formulations has been discussed by some literature, such as [1], which focuses on the computational complexity when a stochastic optimization algorithm incorporates an ℓ_1-norm penalty. To our knowledge, no theoretical analysis has been established to qualify the performance of the sparsity-inducing penalties in terms of approximating the true SP problem by the sample average approximation.

Our results may also have implications to the understanding of a flexible class of high-dimensional sparse learning problems for M-estimation with the FCP. In fact, the SAA (3) can be considered as a formulation of an M-estimator with f representing a statistical loss function, and the SP problem (1) is the corresponding population version of the learning problem with F measuring the generalization error. Such a correspondence is also noted by [2]. Following this correspondence, the RSAA (6) is then the formulation of the sparse learning problem that incorporates the FCP as a regularizer. Our findings imply that high-dimensional M-estimation is possible through the regularization of the FCP, even if the problem dimension cannot be bounded by any polynomial function of the sample size. While most existing literature on high-dimensional learning such as [5,6,10,16–18,23,24,27–29] either focuses on linear regression models or relies on additional conditions such as the (restricted) strong convexity, our analyses do not rely on those assumptions and may apply to a more general M-estimation problem. We would also like to comment that much literature has been devoted to studying an alternative regularizer, the ℓ_1-norm regularizer, or a.k.a., the Lasso. For many reported simulated experiments, numerical comparisons between Lasso and FCP have been reported by [10,12,15,16,23,24] in supportive of relative outperformance of the latter. Some theoretical explanations of such outperformance are also provided by [10,12,16] in some special cases of high-dimensional learning.

The rest of this paper is organized as following: Sect. 2 presents our assumptions and the necessary optimality conditions. Section 3 summarizes our major results. Proofs for those results are presented in Sect. 4. Section 5 discusses different approaches in solving for a desired local/global solution. Section 6 presents some preliminary numerical results. Finally, Sect. 7 concludes the paper. Throughout the paper we will denote by $\|\cdot\|$, $|\cdot|$, and $\|\cdot\|_{\mathbf{p}}$ $(1 \leq \mathbf{p} \leq \infty)$ for a vector the ℓ_2, ℓ_1, and $\ell_{\mathbf{p}}$ norm, while $|\cdot|$ for a finite set denotes the cardinality of the set. For any scalars x and y, we denote by $x \bigvee y$ (and by $x \bigwedge y$) the larger (smaller, resp.) number between the two. We will also use "a.s." as an abbreviation for "almost surely", and "a.e." for "almost every".

2 Settings and necessary conditions

2.1 Assumptions

Our analysis relies on the following assumptions.

Assumption A

A.1 For any $\mathbf{x} \in \mathcal{X}$, the following inequality holds

$$\mathbb{E}[\exp\left(t\left[f(\mathbf{x},\ W) - F(\mathbf{x})\right]\right)] \leq \exp\left(\frac{\sigma^2 t^2}{2}\right), \quad \forall\, t \in \Re,$$

for some $\sigma > 0$.

A.2 There exists a measurable and deterministic function $L : \mathcal{W} \to \Re$ such that

$$\mathbb{E}[\exp\left(t\left[L(W) - L_\mu\right]\right)] \leq \exp\left(\frac{\sigma_L^2 t^2}{2}\right), \quad \forall\, t \in \Re,$$

for some $\sigma_L > 0$ and $L_\mu := \mathbb{E}[L(W)] \geq 1$ and that

$$\sup_{\mathbf{x}_1,\,\mathbf{x}_2 \in \mathcal{X}} \{|f(\mathbf{x}_1,\ W) - f(\mathbf{x}_2,\ W)| - L(W)\|\mathbf{x}_1 - \mathbf{x}_2\|\} \leq 0, \quad a.e.\ W \in \mathcal{W}.$$

A.3 For almost every $W \in \mathcal{W}$, function $f(\mathbf{x},\ W)$ is twice differentiable in $\mathbf{x}$ and satisfies

$$\frac{\partial^2 f(\mathbf{x},\ W)}{(\partial x_i)^2} \leq L_{\mathcal{H}}, \quad \forall\, i \in \{1, \ldots, p\},\ \mathbf{x} = (x_i) \in \mathcal{X}$$

for some $L_{\mathcal{H}} > 0$.

A.4 Assume that $\mathcal{X}$ is defined by coordinate-wise constraints with $\mathcal{X} := \{\mathbf{x} = (x_i) : x_i \in X_i, i = 1, \ldots, p\}$ for some $X_i \subseteq \Re_+$ for all $i = 1, \ldots, p$, and that there exist two hypercubes $\mathbb{H}(0, R) := \{\mathbf{x} \in \Re_+^p : \mathbf{x} \leq R\}$, for some $R \geq 1$, and $\mathbb{H}(0, 1) := \{\mathbf{x} \in \Re_+^p : \mathbf{x} \leq 1\}$ such that $\mathbb{H}(0, 1) \subseteq \mathcal{X} \subseteq \mathbb{H}(0, R)$.

A.5 Function $f(\cdot, W)$ is convex for almost every $W \in \mathcal{W}$.

We will also make stipulations on the choices of the penalty parameters a and λ.

Condition B. Let the penalty parameters (a, λ) of the MCP as in (7) satisfy that $a < L_{\mathcal{H}}^{-1}$, $a \leq 1$ and $\lambda > 0$.

Assumption A.1 and A.2 are essentially subgaussian. The same set of assumptions are standard for sample complexity analyses of the conventional SAA as in [21]. Meanwhile, A.3 and A.5 are verifiable regularities of the objective function. More specifically, Assumption A.3 essentially assumes that the largest eigenvalue of the Hessian matrix of the SAA formulation is bounded from above almost surely and Assumption A.5 requires that the SAA formulation is almost surely convex. Assumption A.4 requires that the constraints are component-wise rectangle constraints. In addition, it is also required that the feasible region contain an inner hypercube and is compact. For some of our theoretical results (as in Theorem 1), Assumption A.5 is not required. Condition B is non-restrictive, since the parameters a and λ are user-specified.

Under Assumption A.2, there exists another measurable and deterministic function, denoted by $L_{|\mathcal{S}|} : \mathcal{W} \to \Re$, and a constant, denoted by $L_{\mu,s} : 1 \leq L_{\mu,s} \leq L_\mu$, such that

$$\mathbb{E}\left[\exp\left(t\left[L_{|\mathcal{S}|}(W) - L_{\mu,s}\right]\right)\right] \leq \exp\left(\frac{\sigma_L^2 t^2}{2}\right), \tag{12}$$

for all $t \in \Re$, and that $\sup_{\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X} \cap \{\mathbf{x}: x_i = 0,\, j \in \mathcal{S}^c\}} \{|f(\mathbf{x}_1, W) - f(\mathbf{x}_2, W)| - L_{|\mathcal{S}|}(W)\|\mathbf{x}_1 - \mathbf{x}_2\|\} \leq 0$ for almost every $W \in \mathcal{W}$. In some cases, such as when F_n is quadratic, $L_{\mu,s}$ may be nontrivially smaller than L_μ especially if p is large.

2.2 Necessary conditions for local minimality

We focus on local solutions to (6) that satisfy some necessary conditions for local minimality. Telling from (7), $P_\lambda(t)$ is twice differentiable in t for all $t \in [0, a\lambda)$. In the meantime, $F_n(\mathbf{x})$ is almost surely twice differentiable under Assumption A.3 for any $\mathbf{x} \in \mathcal{X}$. We consider the following necessary conditions:

First-order necessary condition (FONC): The solution $\mathbf{x}^* \in \mathcal{X}$ satisfies that

$$\langle \nabla F_n(\mathbf{x}^*) + (P_\lambda'(x_i^*) : 1 \leq i \leq p), \mathbf{x} - \mathbf{x}^* \rangle \geq 0, \quad \forall\, \mathbf{x} \in \mathcal{X}. \tag{13}$$

Significant subspace second-order necessary condition (S^3*ONC):* The solution $\mathbf{x}^* := (x_i^* : 1 \leq i \leq p) \in \mathcal{X}$ satisfies FONC. Furthermore, for all $i \in \{i : x_i^* \in (0, \min\{1, a\lambda\})\}$, it holds that $\left.\frac{\partial^2[F_n(\mathbf{x}) + \sum_{i=1}^p P_\lambda(x_i)]}{(\partial x_i)^2}\right|_{\mathbf{x}=\mathbf{x}^*} \geq 0$.

The S^{3}ONC is derived from the observation that a local minimal solution to the original problem must be a local minimizer in the subspace that considers only a single nonzero variable (see also [8,16]). One may easily check that any second-order KKT point satisfies the S^{3}ONC.

3 Major results

Our major results concern two propositions and four theorems. Propositions 1 and 2 provide sample size estimates for all S^{3}ONC solutions within the set $\{\mathbf{x} : F(\mathbf{x}) - F(\hat{\mathbf{x}}^{\min}) \le \Gamma\}$ for some prescribed $\Gamma \ge 0$. Those bounds vary with different regularities on f or F. Then Theorems 1, 2, and 3 discuss some special S^{3}ONC solutions: the global solutions or the local solutions generated with some naive initialization. Finally, Theorem 4 presents the special case where the RSAA improves over the conventional SAA nearly without any compromise.

3.1 Sample size estimates for all S^{3}ONC solutions

We will use the following short-hand notation:

$$N^*(c_1) := \frac{\sigma^2}{\epsilon^2} \ln \frac{c_1}{\alpha} + \frac{\sigma^2 |\mathcal{S}|}{\epsilon^2} \ln \frac{c_1 R L_\mu p}{\epsilon} + \sigma_L^2 \cdot \ln \frac{c_1 p}{\alpha}, \tag{14}$$

where $c_1 > 0$.

Proposition 1 *Suppose that Assumptions* A.1–A.3*, and Condition B hold. Let* $|\mathcal{S}| \ge 1, 4p^2 \ge n, \lambda = \frac{\sigma^{2\delta}}{n^\delta |\mathcal{S}|^\rho}$ *for arbitrary* $\delta : 0 < \delta < 1/2$ *and* $\rho : 0 \le \rho \le 1/2$*. Consider an* S^3ONC *solution* $\mathbf{x}^*$ *to* (6) *that satisfies* $F_{n,\lambda}(\mathbf{x}^*) \le F_{n,\lambda}(\hat{\mathbf{x}}^{\min}) + \Gamma$ *almost surely for some* $\Gamma \ge 0$*. For any* $\alpha : 0 < \alpha \le 1$*,* $\epsilon : 0 < \epsilon \le 1$ *and* $\hat{\varepsilon} \ge 0$*:*

1. *if it holds that, for some problem-independent constant* $c_2 > 0$*,*

$$n \ge N_1 \bigvee c_2 \cdot N^*(c_2) \tag{15}$$

where $N_1 := \sigma^2 \left(\frac{1}{\epsilon}\right)^{\frac{1}{2\delta}} |\mathcal{S}|^{\frac{1-2\rho}{2\delta}} \bigvee \sigma^2 |\mathcal{S}|^{\frac{2\rho}{1-2\delta}} \left(c_2 \frac{1+\Gamma+\hat{\varepsilon}}{a^2\epsilon^2} \ln \frac{c_2 R L_\mu p}{\min\{\epsilon, \sigma^{2\delta}\}}\right)^{\frac{1}{1-2\delta}}$*, then* $F(\mathbf{x}^*) - F(\mathbf{x}^{\min}) \le 2\epsilon + \hat{\varepsilon} + \Gamma$ *with probability lower bounded by* $1 - \alpha$*;*

2. *if Assumption* A.5 *is satisfied and it holds that, for some problem-independent constant* $c_2 > 0$*,*

$$n \ge N_2 \bigvee c_2 \cdot N^*(c_2), \tag{16}$$

where $N_2 := \sigma^2 \cdot |\mathcal{S}|^{\frac{1-\rho}{\delta}} \left(\frac{R}{\epsilon}\right)^{\frac{1}{\delta}} \bigvee \sigma^2 |\mathcal{S}|^{\frac{2\rho}{1-2\delta}} \cdot \left(c_2 \frac{1+\Gamma+\hat{\varepsilon}}{a^2\epsilon^2} \ln \frac{c_2 R L_\mu p}{\min\{\epsilon, \sigma^{2\delta}\}}\right)^{\frac{1}{1-2\delta}}$*, then* $F(\mathbf{x}^*) - F(\mathbf{x}^{\min}) \le 2\epsilon + \hat{\varepsilon}$ *with probability lower bounded by* $1 - \alpha$*.*

Proof The proof is postponed till Sect. 4.2.3. □

We assume in the following proposition that F is differentiable and strongly convex with constant $\mathcal{U}_{\mathcal{H}}$ such that, for any $\mathbf{x}_1,\ \mathbf{x}_2 \in \mathcal{X}$,

$$F(\mathbf{x}_1) - F(\mathbf{x}_2) \geq \langle \nabla F(\mathbf{x}_2),\ \mathbf{x}_1 - \mathbf{x}_2 \rangle + \frac{\mathcal{U}_{\mathcal{H}}}{2} \|\mathbf{x}_1 - \mathbf{x}_2\|^2, \tag{17}$$

for some $U_{\mathcal{H}} > 0$, where $\nabla F(\mathbf{x}_2)$ is a gradient of F at $\mathbf{x}_2$. Due to the increased regularity, we may have a different sample size requirement.

Proposition 2 *Consider an S^3ONC solution $\mathbf{x}^*$ to (6) that satisfies $F_{n,\lambda}(\mathbf{x}^*) \leq F_{n,\lambda}(\hat{\mathbf{x}}^{\min}) + \Gamma$ almost surely for some $\Gamma \geq 0$. Suppose that Assumption A and Condition B hold. Let $4p^2 \geq n$, $|\mathcal{S}| \geq 1$ and $\lambda = \frac{\sigma^{2\delta}}{n^\delta |\mathcal{S}|^\rho}$ for arbitrary $\delta : 0 < \delta < 1/2$ and $\rho : 0 \leq \rho \leq 1/2$. Assume, in addition, that F is differentiable and strongly convex to satisfy (17). For any $\alpha : 0 < \alpha \leq 1$, $\epsilon : 0 < \epsilon \leq 1$, and $\hat{\varepsilon} \geq 0$, if it holds that, for some problem-independent constant $c_3 > 0$,*

$$n \geq c_3 \cdot N^*(c_3) \bigvee N_3 \tag{18}$$

where $N_3 := \frac{\sigma^2 |\mathcal{S}|^{\frac{1-2\rho}{2\delta}}}{\mathcal{U}_{\mathcal{H}}^{\frac{1}{2\delta}}} \left[\left(\frac{c_3}{\epsilon}\right)^{\frac{1}{2\delta}} + \left(\frac{c_3 \hat{\varepsilon}}{\epsilon^2}\right)^{\frac{1}{2\delta}} \right] \bigvee \frac{\sigma^2}{|\mathcal{S}|^{\frac{2\rho}{2\delta-1}}} \left(c_3 \frac{1+\Gamma+\hat{\varepsilon}}{a^2\epsilon^2} \ln \frac{c_3 R L_\mu p}{\min\{\epsilon, \sigma^{2\delta}\}} \right)^{\frac{1}{1-2\delta}}$,
then $F(\mathbf{x}^*) - F(\mathbf{x}^{\min}) \leq 3(\epsilon + \hat{\varepsilon})$ with probability lower bounded by $1 - \alpha$.

Proof The proof is postponed till Sect. 4.2.4. □

Remark 1 The assumption of $4p^2 \geq n$ can be easily relaxed but is imposed for notational simplification in our derivations. Meanwhile, it is possible that (17) is satisfied but $F_n(\,\cdot\,) = \frac{1}{n} \sum_{i=1}^{j} f(\,\cdot\,, W^j)$ is not strongly convex. For an example, we may consider the case of linear regression, which is often solved with the SAA in the form of the least squares problem. When $n < p$, the least squares problem may not be strongly convex, but the population version of the linear regression problem (which is the corresponding SP problem) usually have a strongly convex objective.

Remark 2 Consider the global minimizer, denoted $\mathbf{x}^{SAA}$, to the conventional SAA formulation in (3) within the feasible region $\mathcal{X}$. In [21], it is shown (after some immediate conversion of notations from Theorem 5.18 therein) that to achieve an optimization accuracy of $F(\mathbf{x}^{SAA}) - F(\mathbf{x}^{\min}) \leq \epsilon$ with lower-bounded probability $1 - \alpha$, the stipulated sample size follows

$$n \geq \frac{c_a \sigma^2}{\epsilon^2} \left[p \ln \frac{c_a L_\mu R}{\epsilon} + \ln \frac{c_a}{\alpha} \right] \bigvee \sigma_L^2 \cdot \ln \frac{c_a}{\alpha} =: N_{SAA}. \tag{19}$$

for some constants $c_a > 0$. In contrast, Propositions 1 and 2 indicate that, in nonconvex, convex, and strongly convex cases, RSAA requires the sample sizes to be at least $N_1 \bigvee c_2 N^*(c_2)$ in (15), $N_2 \bigvee c_2 N^*(c_2)$ in (16), or $N_3 \bigvee c_3 N^*(c_3)$ in (18), respectively. For all the three cases, it is easily verifiable that N^* is always dominantly better than N_{SAA} in terms of dependency, while as a tradeoff, N_1, N_2, and N_3 may become more sensitive to the reduction in ϵ than the conventional SAA. A detailed comparison will be made in the next subsection.

3.2 Sample size estimates for some special S^{3}ONC solutions

We consider, in Theorem 1, the performance of a global minimal solution $\mathbf{x}^*$, in the sense that $F_{n,\lambda}(\mathbf{x}^*) = \inf_{\mathbf{x}\in\mathcal{X}} F_{n,\lambda}(\mathbf{x})$ almost surely. Then in Theorems 2, 3, and 4, we study the S^{3}ONC solutions with a better objective value than an all-zero vector, denoted by $\mathbf{0}$. In particular, Theorem 4 identifies the best performing case for RSAA.

Recalling the definition of N^* in (14), we have the following results on the global solution.

Theorem 1 *Suppose that Assumptions* A.1–A.3*, and Condition B hold. Let* $4p^2 \ge n$*,* $|\mathcal{S}| \ge 1$*, and* $\lambda = \frac{\sigma^{1/3}}{n^{1/6}|\mathcal{S}|^{1/4}}$*. Consider a global solution* $\mathbf{x}^*$ *to* (6)*. For any* $\alpha : 0 < \alpha \le 1$*,* $\epsilon : 0 < \epsilon \le 1$*, and* $\hat{\varepsilon} \ge 0$*, if*

$$n \ge \frac{c_4\sigma^2|\mathcal{S}|}{\epsilon^3}\left[1 + \frac{(1+\hat{\varepsilon})^{\frac{3}{2}}}{a^3}\left(\ln\frac{c_4 R L_\mu p}{\min\{\epsilon,\, \sigma^{1/3}\}}\right)^{\frac{3}{2}}\right] \bigvee c_4 \cdot N^*(c_4), \tag{20}$$

is satisfied for some problem-independent constant $c_4 > 0$*, then* $F(\mathbf{x}^*) - F(\mathbf{x}^{\min}) \le 2\epsilon + \hat{\varepsilon}$ *with probability lower bounded by* $1 - \alpha$*.*

Proof Since the global solution is also a local minimal solution, $\mathbf{x}^*$ also satisfies the S^{3}ONC almost surely. In addition, since $F(\mathbf{x}^*) \le F(\mathbf{x}^{\min}) \le F(\hat{\mathbf{x}}^{\min})$, we may invoke Part 1 of Proposition 1 with $\Gamma = 0$, $\delta = \frac{1}{6}$, and $\rho = \frac{1}{3}$ to obtain the desired results. □

Remark 3 Theorem 1 stipulates the minimal assumptions on F_n, but, as a tradeoff, it requires the global optimization of (6). Computing (6) globally is challenging, because the MCP is nonconvex. [13] showed that (6) in some special cases is strongly NP-hard. This motivates us to further consider a class of solutions that only satisfy certain necessary conditions for local minimality.

Theorem 2 *Suppose that Assumption A and Condition B hold. Let* $4p^2 \ge n$*,* $|\mathcal{S}| \ge 1$*, and* $\lambda = \frac{\sigma^{1/2}}{n^{1/4}|\mathcal{S}|^{3/8}}$*. Consider an S*^{3}*ONC solution* $\mathbf{x}^*$ *to* (6) *that satisfies* $F_{n,\lambda}(\mathbf{x}^*) \le F_{n,\lambda}(\mathbf{0})$ *almost surely. For any* $\alpha : 0 < \alpha \le \frac{1}{2}$*,* $\epsilon : 0 < \epsilon \le 1$ *and* $\hat{\varepsilon} \ge 0$*, if*

$$n \ge \frac{c_5\sigma^2|\mathcal{S}|^{\frac{5}{2}}}{\epsilon^4}\left[R^4 + \frac{(1 + L_{\mu,s}R + \hat{\varepsilon})^2}{a^4}\left(\ln\frac{c_5 R L_\mu p}{\min\{\epsilon,\, \sigma^{1/2}\}}\right)^2\right] \bigvee c_5 N^*(c_5) \tag{21}$$

is satisfied for some problem-independent constants $c_5 > 0$*, then* $F(\mathbf{x}^*) - F(\mathbf{x}^{\min}) \le 2\epsilon + \hat{\varepsilon}$ *with probability lower bounded by* $1 - 2\alpha$*.*

Proof The proof is postponed till Sect. 4.2.5. □

Theorem 3 *Suppose that Assumption A and Condition B hold. Let* $4p^2 \ge n$*,* $|\mathcal{S}| \ge 1$*, and* $\lambda = \frac{\sigma^{1/3}}{n^{1/6}|\mathcal{S}|^{1/4}}$*. Also assume that* F *is differentiable and strongly convex as in* (17)*.*

Consider an S^3*ONC solution* $\mathbf{x}^*$ *to* (6) *that satisfies* $F_{n,\lambda}(\mathbf{x}^*) \leq F_{n,\lambda}(\mathbf{0})$ *almost surely. For any* $\alpha : 0 < \alpha \leq \frac{1}{2}$, $\epsilon : 0 < \epsilon \leq 1$, *and* $\hat{\varepsilon} \geq 0$, *if*

$$n \geq \frac{c_6\sigma^2|\mathcal{S}|^{\frac{3}{2}}}{\epsilon^3}\left[\frac{1}{\mathcal{U}_{\mathcal{H}}^3} + \frac{\hat{\varepsilon}^3}{\mathcal{U}_{\mathcal{H}}^3\epsilon^3} + \frac{(1 + L_{\mu,s}R + \hat{\varepsilon})^{\frac{3}{2}}}{a^3}\left(\ln\frac{c_6RL_\mu p}{\min\{\epsilon,\, \sigma^{1/3}\}}\right)^{\frac{3}{2}}\right] \bigvee c_6N^*(c_6), \tag{22}$$

is satisfied for some problem-independent constant $c_6 > 0$, *then* $F(\mathbf{x}^*) - F(\mathbf{x}^{\min}) \leq 3(\epsilon + \hat{\varepsilon})$ *with probability lower bounded by* $1 - 2\alpha$.

Proof The proof is postponed till Sect. 4.2.6. □

Theorem 4 *Consider an* S^3*ONC solution* $\mathbf{x}^*$ *to* (6). *Suppose that the same set of assumptions hold as in Theorem* 3. *Let* $\lambda = \frac{1}{|\mathcal{S}|^{1/4}}$. *Assume additionally* $\hat{\varepsilon} = 0$ *and* $\min_{i\in\mathcal{S}} |\hat{x}_i^{\min}| > \frac{|\mathcal{S}|^{1/4}+\sqrt{|\mathcal{S}|^{1/2}+2\mathcal{U}_{\mathcal{H}}}}{\mathcal{U}_{\mathcal{H}}}$, *where* $\hat{x}_i^{\min}$ *is the* i*-th dimension of* $\hat{\mathbf{x}}^{\min}$. *For any* $\alpha : 0 < \alpha \leq \frac{1}{2}$ *and* $\epsilon : 0 < \epsilon \leq 1$, *if*

$$n \geq \frac{c_7\sigma^2|\mathcal{S}|}{\epsilon^2}\left(\frac{1 + L_{\mu,s}R}{a^2}\ln\frac{c_6RL_\mu p}{\epsilon}\right) \bigvee c_7N^*(c_7), \tag{23}$$

for some problem-independent constant $c_7 > 0$, *then* $F(\mathbf{x}^*) - F(\mathbf{x}^{\min}) \leq \epsilon$ *with probability lower bounded by* $1 - 2\alpha$.

Proof The proof is postponed till Sect. 4.2.7. □

Remark 4 We notice that the choices of λ are different among the above theorems. At the minimum, the above theorems ensure the existence of proper λ's that ensure the sound performance of the RSAA in all the scenarios discussed above. In practice, λ can also be determined by a simple cross-validation procedure, which is a commonly adopted scheme in penalized statistical learning to tune the parameter of the sparsity-inducing penalties.

Remark 5 We would like to compare the sample size requirement of the RSAA as presented in the results above with that of the conventional SAA.

- We see that N_{SAA} as in (19) depends polynomially in the problem dimension p. In contrast, Theorems 1, 2, 3, and 4 reveal that the global solutions and some computable local solutions to RSAA require the sample size to be polynomial in $\ln p$ and $|\mathcal{S}|$. We regard it as a demonstration of the RSAA's capability in handling high dimensionality, as now exponentially increased p can be compensated by polynomially increasing n.
- As a tradeoff to the potential advantage mentioned above, the RSAA's performance has a worse dependency on ϵ than the conventional SAA in general. More specifically, N_{SAA} increases at a rate of $O(\frac{1}{\epsilon^2}\ln\frac{1}{\epsilon})$. In contrast, RSAA follows a rate of $O(\frac{1}{\epsilon^3}\cdot(\ln\frac{1}{\epsilon})^{3/2})$ if minimized globally (under Assumptions A.1–A.3),

or $O(\frac{1}{\epsilon^4}\cdot(\ln\frac{1}{\epsilon})^2)$ if solved locally with a naive initialization (additionally under Assumption A.5). Furthermore, under some assumption of differentiability and strong convexity, if $\hat{\varepsilon}\leq O(1)\cdot\epsilon$ for some problem-independent constant $O(1)$, then a local solution with a naive initialization retains the rate of $O(\frac{1}{\epsilon^3}\cdot(\ln\frac{1}{\epsilon})^{3/2})$, which is the same as the global minimizer. We think that compromising the dependency on ϵ to achieve a non-trivial reduction in the dependency on p can be worthwhile in many high dimensional SP applications, where p can be redundantly very large but the suboptimality gap ϵ is not required to be very small.

- Theorem 4 identifies a case where RSAA non-trivially reduces the dependency on p while the growth of the required sample size maintains at the same rate as the conventional SAA in terms of ϵ.
- The RSAA's dependencies on σ and σ_L are almost the same as those of the SAA. Meanwhile, RSAA becomes dependent on some other quantities that originally do not influence the SAA's performance: a, $|\mathcal{S}|$, and $\mathcal{U}_{\mathcal{H}}$. Moreover, in some cases, the RSAA may be more sensitive to the increase in the Lipschitz-like constant $L_{\mu,s}$ as defined in (12) and the radius of the feasible region, R. Nonetheless, those dependencies all maintain to be polynomial.

Remark 6 By allowing $\hat{\varepsilon}\geq 0$, our results apply to the cases where the exact solution to the SP is dense, but can be approximated by a sparse solution. We can see that, when $\hat{\varepsilon}>0$, RSAA will require more samples and may incur a residual suboptimality gap no greater than $O(1)\cdot\hat{\varepsilon}$.

Remark 7 Our results may also have potentially important implications to high-dimensional M-estimation. One may consider the following correspondence between our setting and the setting for a high-dimensional learning problem: (i) Eq. (3) can be thought of as an in-sample statistical loss function; (ii) the (global/local) solution to RSAA formulation (6) can be considered as a folded concave penalized sparse estimator; (iii) the SP formulation (1) can be considered as the population version of the (unpenalized) learning problem (a.k.a., expected risk or generalization error); and (iv) The suboptimality gap $F(\mathbf{x}^*)-F(\mathbf{x}^{\min})$ is then a performance measure[2] of the estimator $\mathbf{x}^*$. The above conversion is also noted by [2]. Under this conversion, we can easily tell from Theorems 1, 2, and 3 that a global solution or an S^{3}ONC solution initialized at an all-zero vector can achieve a reasonable upper bound on the $F(\mathbf{x}^*)-F(\mathbf{x}^{\min})$ even in the undesirable scenarios where the dimension p cannot be upper bounded by any polynomial of n. The same setting has been discussed by [11] for the linear regression model, by [12] for several M-estimation models, and by [17,24] under restricted strong convexity (RSC, which is some variation of strong convexity in certain subset of the feasible region). In contrast, our results may be applicable to a wider class of M-estimators without the RSC assumption. In particular, if we consider the estimator that globally minimizes the RSAA, nonconvexity in the statistical loss function is also allowed.

Remark 8 We would also like to remark that the sparsity of an S^{3}ONC solution is dependent on λ and Γ. The correlations between those quantities and the sparsity level

[2] $F(\mathbf{x}^*)-F(\mathbf{x}^{\min})$ is also referred to as the "excess risk" in a learning problem. See for example [3].

are in fact characterized by Lemma 4 in the subsequent section. Although the formula seem nontrivial, we think that the general trend is clear; that is, larger λ, and smaller Γ may result in fewer nonzeros in the S^{3}ONC solution. Our numerical experiments in Sect. 6 also show that the number of nonzero dimensions can be well constrained at an S^{3}ONC solution.

4 Technical proofs

We will first present a set of preliminary results in Sect. 4.1 and then provide the proofs for the claimed results in Sect. 4.2. A sketch of proof is provided in Sect. 4.2.1.

4.1 Some preliminary results

In this subsection, we present a couple of observations that are useful to our proofs. Firstly, we observe that MCP as in (7) has the following properties:

(i) $P_\lambda(t)$ is non-decreasing and concave in $t \in \Re_+$ with $P_\lambda(0) = 0$ and $P_\lambda(t) > 0$ if $t > 0$;
(ii) $P_\lambda(t)$ is differentiable for all $t \in \Re_+$ and twice differentiable for any $t \in [0,\ a\lambda) \cup (a\lambda,\ \infty)$;
(iii) The first derivative $P'_\lambda(t) = 0$ for any $t \geq a\lambda$;
(iv) $0 \leq P'_\lambda(t) \leq \lambda$ and $0 \leq P_\lambda(t) \leq P_\lambda(a\lambda) = \frac{a\lambda^2}{2}$ for any $t \geq 0$;
(v) The second derivative $P''_\lambda(t) = -\frac{1}{a}$ for any $t \in [0,\ a\lambda)$ and $P''_\lambda(t) = 0$ for any $t > a\lambda$.

Secondly, consider an S^{3}ONC solution $\mathbf{x}^* \in \mathcal{X}$ under Assumption A.5. Recall that S^{3}ONC implies FONC. Then, from the definition of FONC in Eq. (13) and Assumption A.5, we know that, if $\mathbf{x}^*$ satisfies the FONC, then it holds that

$$F_n(\mathbf{x}^*) + \sum_{i=1}^{p} P'_\lambda(x_i^*)x_i^* \leq F_n(\mathbf{x}) + \sum_{i=1}^{p} P'_\lambda(x_i^*)x_i, \ \forall\, \mathbf{x} = (x_i) \in \mathcal{X}, \ \ a.s., \quad (24)$$

which immediately yields that

$$F_n(\mathbf{x}^*) + \sum_{i=1}^{p} P'_\lambda(x_i^*)x_i^* \leq F_n(\hat{\mathbf{x}}^{\min}) + \sum_{i=1}^{p} P'_\lambda(x_i^*)\hat{x}_i^{\min}, \quad a.s.$$

Together with (a) $\hat{x}_i^{\min} = 0$ for all $i \in \mathcal{S}^c$, (b) $\mathbf{x}^* \geq 0$, and (c) Property (iv) of P_λ, it is then straightforward to obtain:

$$\begin{aligned} F_n(\mathbf{x}^*) - F_n(\hat{\mathbf{x}}^{\min}) &\leq \sum_{i=1}^{p} P'_\lambda(x_i^*)(\hat{x}_i^{\min} - x_i^*) \\ &\leq \sum_{i\in\mathcal{S}} P'_\lambda(x_i^*)|\hat{x}_i^{\min} - x_i^*| + \sum_{i\in\mathcal{S}^c} P'_\lambda(x_i^*)(\hat{x}_i^{\min} - x_i^*) \end{aligned}$$

$$\begin{aligned}&\overset{(a)}{=}\sum_{i\in\mathcal{S}}P_\lambda'(x_i^*)|\hat{x}_i^{\min}-x_i^*|+\sum_{i\in\mathcal{S}^c}P_\lambda'(x_i^*)\cdot(-x_i^*)\\&\overset{(b),(c)}{\leq}\lambda\sum_{i\in\mathcal{S}}|\hat{x}_i^{\min}-x_i^*|,\quad a.s.\end{aligned}\tag{25}$$

Similarly, with (a) $\hat{x}_i^{\min}=0$ for all $i\in\mathcal{S}^c$, (b) $\mathbf{x}^*\geq 0$, and (c) Property (iv) of P_λ, again,

$$\begin{aligned}F_n(\mathbf{x}^*)-F_n(\hat{\mathbf{x}}^{\min})&\leq\sum_{i=1}^{p}P_\lambda'(x_i^*)(\hat{x}_i^{\min}-x_i^*)\\&\leq\sum_{i\in\mathcal{S}}P_\lambda'(x_i^*)(\hat{x}_i^{\min}-x_i^*)+\sum_{i\in\mathcal{S}^c}P_\lambda'(x_i^*)(\hat{x}_i^{\min}-x_i^*)\\&\overset{(a)}{=}\sum_{i\in\mathcal{S}}P_\lambda'(x_i^*)(\hat{x}_i^{\min}-x_i^*)+\sum_{i\in\mathcal{S}^c}P_\lambda'(x_i^*)\cdot(-x_i^*)\\&\overset{(b)}{\leq}\sum_{i\in\mathcal{S}}P_\lambda'(x_i^*)(\hat{x}_i^{\min})\overset{(c)}{\leq}\lambda\sum_{i\in\mathcal{S}}|\hat{x}_i^{\min}|,\quad a.s.\end{aligned}\tag{26}$$

Thirdly, consider an S^{3}ONC solution $\mathbf{x}^*\in\mathcal{X}$ again. One has that

$$x_i^*\notin(0,\ \min\{a\lambda,\ 1\})\text{ for any }i=\{1,\ldots,p\},\text{ almost surely.}\tag{27}$$

To see this, suppose that for an arbitrary dimension $i\in\{1,\ldots,p\}$, it holds that $x_i^*\in(0,\ \min\{a\lambda,\ 1\})$. Since $\frac{\partial^2F_n(\mathbf{x})}{(\partial x_i)^2}\leq L_{\mathcal{H}}$ for all $\mathbf{x}\in\mathcal{X}$ almost surely as an immediate result of Assumption A.3, combined with $a<L_{\mathcal{H}}^{-1}$ under Condition B and Property (v) of P_λ, we have that $\left.\frac{\partial^2F_{n,\lambda}(\mathbf{x})}{(\partial x_i)^2}\right|_{\mathbf{x}=\mathbf{x}^*}=\left[\frac{\partial^2F_n(\mathbf{x})}{(\partial x_i)^2}-\frac{1}{a}\right]_{\mathbf{x}=\mathbf{x}^*}<0$, almost surely. The satisfaction of this inequality contradicts with the S^{3}ONC, that is, for all $i=1,\ldots,p$,

$$\mathbb{P}\left[\left\{\frac{\partial^2F_n(\mathbf{x}^*)}{(\partial x_i)^2}\leq L_{\mathcal{H}}\right\}\cap\{\mathbf{x}^*\text{ satisifies S}^3\text{ONC}\}\cap\{x_i^*\in(0,\ \min\{a\lambda,\ 1\})\}\right]=0.$$

Notice that

$$\begin{aligned}&\mathbb{P}\left[\left(\left\{\frac{\partial^2F_n(\mathbf{x}^*)}{(\partial x_i)^2}\leq L_{\mathcal{H}}\right\}\cap\{\mathbf{x}^*\text{ satisfies S}^3\text{ONC}\}\right)\cup\{x_i^*\in(0,\ \min\{a\lambda,\ 1\})\}\right]\\&\quad=\mathbb{P}\left[\{x_i^*\in(0,\ \min\{a\lambda,\ 1\})\}\right]+\mathbb{P}\left[\left\{\frac{\partial^2F_n(\mathbf{x}^*)}{(\partial x_i)^2}\leq L_{\mathcal{H}}\right\}\cap\{\mathbf{x}^*\text{ satisfies S}^3\text{ONC}\}\right]\\&\quad\quad-\mathbb{P}\left[\left\{\frac{\partial^2F_n(\mathbf{x}^*)}{(\partial x_i)^2}\leq L_{\mathcal{H}}\right\}\cap\{\mathbf{x}^*\text{ satisfies S}^3\text{ONC}\}\cap\{x_i^*\in(0,\ \min\{a\lambda,\ 1\})\}\right]\end{aligned}$$

which means that

$$\begin{aligned}1 &= \mathbb{P}\left[\{x_i^* \in (0,\ \min\{a\lambda,\ 1\})\}\right] + 1 - 0, \quad \forall\, i = 1, \ldots, p\\ &\Longrightarrow \mathbb{P}\left[\{x_i^* \in (0,\ \min\{a\lambda,\ 1\})\}\right] = 0, \quad \forall\, i = 1, \ldots, p\\ &\Longrightarrow \mathbb{P}\left[\{x_i^* \notin (0,\ \min\{a\lambda,\ 1\}),\ \forall\, i = 1, \ldots, p\}\right] = 1\end{aligned}$$

Combined with Properties (i) and (iii) of P_λ, it further implies that

$$\begin{aligned}P_\lambda(a\lambda)\|\mathbf{x}^*\|_0 &\geq \sum_{i=1}^{p} P_\lambda(x_i^*) \geq\ P_\lambda(\min\{a\lambda,\ 1\})\|\mathbf{x}^*\|_0\\ &= \left(\lambda \min\{a\lambda,\ 1\} - \frac{\min\{a^2\lambda^2,\ 1\}}{2a}\right)\|\mathbf{x}^*\|_0, \quad a.s. \qquad (28)\end{aligned}$$

Fourthly, the following two useful lemmas are some quick results from Assumption A.2 and are taken from [21] after some slight changes.

Lemma 1 (a) *Under Assumption A.2, for any $t > 0$,*

$$\sup_{\mathbf{x}_1,\,\mathbf{x}_2 \in \mathcal{X}} \left\{\left|\sum_{j=1}^{n} f(\mathbf{x}_1, W^j)/n - \sum_{j=1}^{n} f(\mathbf{x}_2, W^j)/n\right| - (L_\mu + t)\|\mathbf{x}_1 - \mathbf{x}_2\|\right\} \leq 0,$$

with probability at least $1 - 2\exp\left(-\frac{nt^2}{2\sigma_L^2}\right)$.

(b) *Under Assumption A.2, for any $t > 0$,*

$$\sup_{\substack{\mathbf{x}_1,\,\mathbf{x}_2 \in \mathcal{X}\cap\\ \{\mathbf{x}:\, x_i=0,\, i\in\mathcal{S}^c\}}} \left\{\left|\sum_{j=1}^{n} f(\mathbf{x}_1, W^j)/n - \sum_{j=1}^{n} f(\mathbf{x}_2, W^j)/n\right| - (L_{\mu,s} + t)\|\mathbf{x}_1 - \mathbf{x}_2\|\right\} \leq 0,$$

with probability at least $1 - 2\exp\left(-\frac{nt^2}{2\sigma_L^2}\right)$.

Proof To show (a): Firstly, by Assumption A.2, one has $\sup_{\mathbf{x}_1,\,\mathbf{x}_2\in\mathcal{X}}\{|f(\mathbf{x}_1, W^j) - f(\mathbf{x}_2, W^j)| - L(W^j)\|\mathbf{x}_1 - \mathbf{x}_2\|\} \leq 0$ for all $j = 1, \ldots, n$ almost surely. Combining the inequalities for all $j = 1, \ldots, n$, we obtain

$$\sup_{\mathbf{x}_1,\,\mathbf{x}_2 \in \mathcal{X}} \left\{\sum_{j=1}^{n} |f(\mathbf{x}_1, W^j) - f(\mathbf{x}_2, W^j)| - \sum_{j=1}^{n} L(W^j)\|\mathbf{x}_1 - \mathbf{x}_2\|\right\} \leq\ 0, \quad a.s.$$

By triangular inequality and dividing both sides by n, we have

$$\sup_{\mathbf{x}_1,\,\mathbf{x}_2\in\mathcal{X}}\left\{\left|\sum_{j=1}^{n} f(\mathbf{x}_1,\,W^j)/n-\sum_{j=1}^{n} f(\mathbf{x}_2,\,W^j)/n\right|\right.$$
$$\left.-\sum_{j=1}^{n} n^{-1}L(W^j)\|\mathbf{x}_1-\mathbf{x}_2\|\right\}\leq 0\quad a.s.$$

By the second part of Assumption A.2, we can invoke the well-known large deviation theorem on subgaussian i.i.d. random variables and obtain

$$\mathbb{P}\left[\left|n^{-1}\sum_{j=1}^{n} L(W^j)-L_\mu\right|\geq t\right]\leq 2\exp\left(-\frac{nt^2}{2\sigma_L^2}\right) \tag{29}$$

for any $t>0$. Combining the above,

$$\sup_{\mathbf{x}_1,\,\mathbf{x}_2\in\mathcal{X}}\left\{\left|\sum_{j=1}^{n} f(\mathbf{x}_1,\,W^j)/n-\sum_{j=1}^{n} f(\mathbf{x}_2,\,W^j)/n\right|-(L_\mu+t)\|\mathbf{x}_1-\mathbf{x}_2\|\right\}\leq 0,$$

with probability at least $1-2\exp\left(-\frac{nt^2}{2\sigma_L^2}\right)$, as claimed.

To show (b): Under Assumption A.2, it obtains that (12) holds. Then, the same argument to prove Part (a) immediately leads to the desired result in Part (b). □

Lemma 2 (a) *Under Assumption A.2, for any fixed* $\mathbf{x}_1,\,\mathbf{x}_2\in\mathcal{X}$*, it holds that* $|F(\mathbf{x}_1)-F(\mathbf{x}_2)|\leq L_\mu\|\mathbf{x}_1-\mathbf{x}_2\|$.

(b) *Under Assumption* A.2*, for any fixed* $\mathbf{x}_1,\,\mathbf{x}_2\in\mathcal{X}\cap\{\mathbf{x}:\,x_i=0,\,i\in\mathcal{S}^c\}$*, it holds that* $|F(\mathbf{x}_1)-F(\mathbf{x}_2)|\leq L_{\mu,s}\|\mathbf{x}_1-\mathbf{x}_2\|$.

Proof To show (a): By Assumption A.2, we have,

$$\begin{aligned}L_\mu\|\mathbf{x}_1-\mathbf{x}_2\|&=\mathbb{E}[L(W)\cdot\|\mathbf{x}_1-\mathbf{x}_2\|]\geq\ \mathbb{E}[|f(\mathbf{x}_1,\,W)-f(\mathbf{x}_2,\,W)|]\\&\geq|\mathbb{E}[f(\mathbf{x}_1,\,W)]-\mathbb{E}[f(\mathbf{x}_2,\,W)]|=\ |F(\mathbf{x}_1)-F(\mathbf{x}_2)|,\end{aligned}$$

which is immediately the claimed result.

To show (b): Under Assumption A.2, Inequality (12) holds. Then, with the same argument to prove Part (a), we immediately obtain the desired result in Part (b). □

4.2 Proof of major results

This section presents the proofs for our claimed theoretical results. We first present a sketch of the proof in Sect. 4.2.1. Then, two useful lemmas that serve as the pillar of our analysis are presented in Sect. 4.2.2. The proofs for the aforementioned propositions and theorems as our major results are provided subsequently in Subsections from 4.2.3 to 4.2.7.

4.2.1 Sketch of proof

Our proof is organized as following:

Step 1 In Lemma 3, we show how well the objective function of the SP problem F can be approximated by the objective function of the SAA problem F_n at a feasible solution that satisfies the sparsity assumption in addition to the standard assumptions for the SAA (Assumptions A.1 and A.2). More specifically, we derive a bound on the probability for the point-wise difference between $F(\mathbf{x})$ and $F_n(\mathbf{x})$ to be contained within a prescribed level $\epsilon > 0$ when $\|\mathbf{x}\|_0 \leq \tilde{p}$ for any $\tilde{p} : 1 \leq \tilde{p} \leq p$. It turns out that, if sparsity holds (i.e., if $\tilde{p}$ is small), the approximation quality is less sensitive to the problem dimension p compared to the conventional SAA by [20–22].

Step 2 To exploit the results from Step 1, Lemma 4 then shows that, once Assumption A.3 holds (i.e., the diagonals of the Hessian matrix of the SAA formulation is bounded from the above), we can guarantee that any S^{3}ONC solution is sparse. Furthermore, the number of nonzeros can be controlled by tuning the penalty parameters a and λ. As a result, through properly choosing the values for a and λ, we ensure that $\tilde{p}$ can indeed be a small number at the S^{3}ONC solution. Lemma 4 also explicates the number of nonzeros at an S^{3}ONC solution as a function in parameterization of a, λ, and the global suboptimality of that S^{3}ONC solution.

Step 3 Combining results from Steps 1 and 2, we may obtain the claimed results for Propositions 1 and 2 in Sect. 4.2.3 by choosing the proper pair of parameters (a, λ). The bounds derived in both propositions are in parameterization of the suboptimality gap Γ in solving the RSAA. Note that Proposition 2 makes use of additional inequalities from strong convexity and thus provides a sharper bound than Proposition 1.

Step 4 Employing bounds on the approximation quality from Propositions 1 and 2, which are in parameterization of Γ, we then consider the S^{3}ONC solutions where Γ can be explicated. In particular, we focus on two cases. (i) We first consider the global solutions where $\Gamma = 0$. By employing the propositions shown in Step 3, we can immediately derive Theorem 1 by properly choosing a and λ. (ii) Under Assumption A.5 (i.e., the unpenalized SAA formulation is convex) we then look at those solutions that have a better objective value than an all-zero solution. This immediately leads to all our results in Theorems 2–4.

4.2.2 Two pillar lemmas

This section provide two pillar lemmas that lay the foundation of our analyses and constitutes Step 1 of our proof sketch in Sect. 4.2.1.

Lemma 3 *Suppose that Assumptions* A.1 *and* A.2 *hold. For any scalar* $t > 0$ *and any integer* $\tilde{p} : p \geq \tilde{p} > 0$*, the following inequality holds:*

$$\mathbf{P}\left[\sup_{\mathbf{x}\in\mathcal{X}:\,\|\mathbf{x}\|_0\le\tilde{p}}\left|\frac{1}{n}\sum_{j=1}^{n}f(\mathbf{x},\ W^j)-F(\mathbf{x})\right|\le t\right]$$
$$\ge\ 1-2\left\lceil\left(\frac{12\sqrt{\tilde{p}}RL_\mu}{t}\right)^{\tilde{p}}\binom{p}{\tilde{p}}\right\rceil\cdot\exp\left(-\frac{nt^2}{8\sigma^2}\right)-2\exp\left(-\frac{nL_\mu^2}{2\sigma_L^2}\right).$$

Proof We can divide the feasible region $\mathcal{X}$ by a net of finitely many grids $V(t):=\{\mathbf{x}^k,\ k=1,2,\ldots\}\subseteq\mathcal{X}$, such that for any $\mathbf{x}\in\mathcal{X}\cap\{\mathbf{x}:\ \|\mathbf{x}\|_0\le\tilde{p}\}$, there always exists an $\mathbf{x}^k\in V(t)$ that satisfies $\|\mathbf{x}^k-\mathbf{x}\|\le\frac{t}{6L_\mu}$. Since $\mathcal{X}\subseteq\mathbb{H}(0,R)$, it is easily verifiable that one can always find such a net of grids if $|V(t)|=\left\lceil(\frac{12\sqrt{\tilde{p}}RL_\mu}{t})^{\tilde{p}}\binom{p}{\tilde{p}}\right\rceil$. Corresponding to every grid $\mathbf{x}^k$, there is a subset of the feasible region $\mathcal{X}_k:=\left\{\mathbf{x}\in\mathcal{X}:\ \|\mathbf{x}-\mathbf{x}^k\|\le\frac{t}{6L_\mu}\right\}$. As per our construction, we know that $\mathcal{X}\cap\{\mathbf{x}:\ \|\mathbf{x}\|_0\le\tilde{p}\}=\left(\cup_{\mathbf{x}^k\in V(t)}\mathcal{X}_k\right)\cap\{\mathbf{x}:\ \|\mathbf{x}\|_0\le\tilde{p}\}$. Therefore, it holds surely that

$$\sup_{\mathbf{x}\in\mathcal{X}\cap\{\mathbf{x}:\,\|\mathbf{x}\|_0\le\tilde{p}\}}\left|\frac{1}{n}\sum_{j=1}^{n}f(\mathbf{x},\ W^j)-F(\mathbf{x})\right|$$
$$\le\max_{k=1,\ldots,|V(t)|}\sup_{\mathbf{x}\in\mathcal{X}_k}\left|\frac{1}{n}\sum_{j=1}^{n}f(\mathbf{x},\ W^j)-F(\mathbf{x})\right| \tag{30}$$

Now, consider the following events:

$$\mathcal{E}_1(t):=\left\{\max_{\mathbf{y}\in V(t)}\left|\frac{1}{n}\sum_{j=1}^{n}f(\mathbf{y},\ W^j)-F(\mathbf{y})\right|\le t/2\right\}$$
$$\mathcal{E}_2:=\left\{\sup_{\mathbf{x}_1,\,\mathbf{x}_2\in\mathcal{X}}\left|\sum_{j=1}^{n}f(\mathbf{x}_1,W^j)/n-\sum_{j=1}^{n}f(\mathbf{x}_2,W^j)/n\right|-2L_\mu\|\mathbf{x}_1-\mathbf{x}_2\|\le 0\right\}$$
$$\mathcal{E}_3(k):=\left\{\sup_{\mathbf{x}_1,\,\mathbf{x}_2\in\mathcal{X}_k}\left|\sum_{j=1}^{n}f(\mathbf{x}_1,W^j)/n\right.\right.$$
$$\left.\left.-\sum_{j=1}^{n}f(\mathbf{x}_2,W^j)/n\right|-2L_\mu\|\mathbf{x}_1-\mathbf{x}_2\|\le 0\right\},\quad k=1,\ldots,|V(t)|.$$

It is easily verifiable that $\mathcal{E}_2\subseteq\mathcal{E}_3(k)$ for any $k=1,\ldots,|V(t)|$. Conditioning on $\mathcal{E}_2$, we have that for any $k=1,\ldots,|V(t)|$:

$$\begin{aligned}
&\sup_{\mathbf{x}\in\mathcal{X}_k}\left|\frac{1}{n}\sum_{j=1}^{n} f(\mathbf{x},\ W^j) - F(\mathbf{x})\right| \\
&\leq \sup_{\mathbf{x}\in\mathcal{X}_k}\left|\frac{1}{n}\sum_{j=1}^{n} f(\mathbf{x},\ W^j) - \frac{1}{n}\sum_{j=1}^{n} f(\mathbf{x}^k,\ W^j)\right| + \left|F(\mathbf{x}) - F(\mathbf{x}^k)\right| \\
&\quad + \left|\frac{1}{n}\sum_{j=1}^{n} f(\mathbf{x}^k,\ W^j) - F(\mathbf{x}^k)\right| \\
&\overset{\mathcal{E}_2\subseteq\mathcal{E}_3(k)}{\leq} \sup_{\mathbf{x}\in\mathcal{X}_k} 2L_\mu \left\|\mathbf{x}-\mathbf{x}^k\right\| + \left|F(\mathbf{x}) - F(\mathbf{x}^k)\right| + \left|\frac{1}{n}\sum_{j=1}^{n} f(\mathbf{x}^k,\ W^j) - F(\mathbf{x}^k)\right| \\
&\overset{\text{Lemma 2}}{\leq} \sup_{\mathbf{x}\in\mathcal{X}_k} 2L_\mu \left\|\mathbf{x}-\mathbf{x}^k\right\| + L_\mu \left\|\mathbf{x}-\mathbf{x}^k\right\| + \left|\frac{1}{n}\sum_{j=1}^{n} f(\mathbf{x}^k,\ W^j) - F(\mathbf{x}^k)\right| \\
&= \frac{t}{2} + \left|\frac{1}{n}\sum_{j=1}^{n} f(\mathbf{x}^k,\ W^j) - F(\mathbf{x}^k)\right|, \qquad a.s.
\end{aligned}$$

Therefore, conditioning on the simultaneous occurrence of both $\mathcal{E}_1(t)$ and $\mathcal{E}_2$, we have

$$\begin{aligned}
&\sup_{\mathbf{x}\in\mathcal{X}\cap\{\mathbf{x}:\ \|\mathbf{x}\|_0\leq\tilde{p}\}}\left|\frac{1}{n}\sum_{j=1}^{n} f(\mathbf{x},\ W^j) - F(\mathbf{x})\right| \\
&\leq \max_{k=1,\ldots,|V(t)|}\sup_{\mathbf{x}\in\mathcal{X}_k}\left|\frac{1}{n}\sum_{j=1}^{n} f(\mathbf{x},\ W^j) - F(\mathbf{x})\right| \\
&\leq \frac{t}{2} + \max_{k=1,\ldots,|V(t)|}\left|\frac{1}{n}\sum_{j=1}^{n} f(\mathbf{x}^k,\ W^j) - F(\mathbf{x}^k)\right| \leq \frac{t}{2}+\frac{t}{2} = t, \quad a.s.
\end{aligned}$$

Now it suffices to bound the probability for $\mathcal{E}_1(t)$ and $\mathcal{E}_2$.

(i). To consider $\mathcal{E}_1(t)$, we know by union bound that

$$\begin{aligned}
&\mathbf{P}\left[\max_{\mathbf{y}\in V(t)}\left|\frac{1}{n}\sum_{j=1}^{n} f(\mathbf{y},\ W^j) - F(\mathbf{y})\right| > \frac{t}{2}\right] \\
&\leq \sum_{\mathbf{y}\in V(t)}\mathbf{P}\left[\left|\frac{1}{n}\sum_{j=1}^{n} f(\mathbf{y},\ W^j) - F(\mathbf{y})\right| > \frac{t}{2}\right]
\end{aligned}$$

Due to Assumption A.1, we may invoke the large deviation theorem on sub-gaussian i.i.d. random variables to obtain that, for any $t > 0$, it holds that

$\mathbf{P}\left[\left|\frac{1}{n}\sum_{j=1}^{n} f(\mathbf{y},\ W^j) - F(\mathbf{y})\right| \geq\ t\right] \leq\ 2\exp\left(-\frac{nt^2}{2\sigma^2}\right)$ for any $\mathbf{y} \in V(t)$. Therefore, we may continue as

$$\mathbf{P}[\mathcal{E}_1(t)] = \mathbf{P}\left[\max_{\mathbf{y}\in V(t)}\left|\frac{1}{n}\sum_{j=1}^{n} f(\mathbf{y},\ W^j) - F(\mathbf{y})\right| \leq t/2\right]$$
$$\geq 1 - 2|V(t)| \cdot \exp\left(-\frac{nt^2}{8\sigma^2}\right) \geq\ 1 - 2\left[\left(\frac{12\sqrt{\tilde{p}}RL_\mu}{t}\right)^{\tilde{p}}\binom{p}{\tilde{p}}\right]\exp\left(-\frac{nt^2}{8\sigma^2}\right) \tag{31}$$

(ii). To consider $\mathcal{E}_2$, we invoke Lemma 1 (in which we let $t\ :=\ L_\mu$ only within that lemma), we know that

$$\mathbf{P}\left[\mathcal{E}_2\right] \geq 1 - 2\exp\left(-\frac{nL_\mu^2}{2\sigma_L^2}\right) \tag{32}$$

Now, invoking both the De Morgan's Law and the union bound to combine all the above, we obtain the desired result. □

Lemma 4 *Suppose that Assumptions* A.1–A.3 *and Condition B hold. Let* $\hat{\varepsilon} \geq\ 0$ *and* $\mathbf{x}^* \in\ \mathcal{X}$ *be an* S^3*ONC solution. For any integer* $\tilde{p}\ :\ \tilde{p} \geq\ |\mathcal{S}|$ *and any scalars* $t\ >\ 0$, $\hat{\varepsilon} \geq 0$, *and* $\Gamma\ \geq\ 0$, *if* $F_{n,\lambda}(\mathbf{x}^*)\ \leq\ F_{n,\lambda}(\hat{\mathbf{x}}^{\min}) + \Gamma$ *almost surely,* $\frac{nt^2}{8\sigma^2} \geq \ln\left(\frac{12pRL_\mu}{t}\right)$, $a\lambda\ \leq\ 1$ *and*

$$P_\lambda(a\lambda) > \frac{\Gamma + 2t\sqrt{\tilde{p}+1} + \hat{\varepsilon}}{\tilde{p} - |\mathcal{S}| + 1}, \tag{33}$$

then $\|\mathbf{x}^*\|_0\ \leq\ \tilde{p}$ *with probability at least*

$$\mathbf{P}^*(t,\tilde{p}) :=\ 1 - 2\exp\left(-\frac{(\tilde{p}+1)nt^2}{8\sigma^2}\right)\cdot\frac{1}{1-\exp\left(-\frac{nt^2}{8\sigma^2}\right)}$$
$$-2p\exp\left(-\frac{nL_\mu^2}{2\sigma_L^2}\right) - 2\exp\left(-(\tilde{p}+1)\left[\frac{nt^2}{8\sigma^2} - \ln\left(\frac{12pRL_\mu}{t}\right)\right]\right)$$
$$\cdot\frac{1}{1-\exp\left(-\left[\frac{nt^2}{8\sigma^2}-\ln\left(\frac{12pRL_\mu}{t}\right)\right]\right)} \tag{34}$$

Proof If $\tilde{p}\ >\ p$, then $\|\mathbf{x}^*\|_0 \leq p < \tilde{p}$ with probability 1, while $\mathbf{P}^*(t,\tilde{p}) \leq 1$ for any $t\ >\ 0$ and $\tilde{p} \geq |\mathcal{S}|$. Thus the desired result holds if $\tilde{p}\ >\ p$. The rest of the proof then considers only the case where $\tilde{p} \leq p$.

For arbitrary integers $\tilde{p}: p \geq \tilde{p} \geq |\mathcal{S}|$ and $k: 1 \leq k \leq p - \tilde{p}$, consider the events

$$\mathcal{E}_a(\tilde{p}+k) := \{\|\mathbf{x}^*\|_0 = \tilde{p}+k\}; \quad \mathcal{E}_b := \{F_n(\hat{\mathbf{x}}^{\min}) - F_n(\mathbf{x}^*) \leq 2t\sqrt{\tilde{p}+k} + \hat{\varepsilon}\}$$

and

$$\mathcal{E}_c := \left\{ \sup_{\mathbf{x}\in\mathcal{X}:\, \|\mathbf{x}\|_0 \leq \tilde{p}+k} |F_n(\mathbf{x}) - F(\mathbf{x})| \leq t\sqrt{\tilde{p}+k} \right\}.$$

Firstly, we want to show that $\mathbf{P}[\mathcal{E}_a(\tilde{p}+k) \cap \mathcal{E}_b] = 0$. To this end, consider another two events

$$\mathbb{A} := \left\{\forall\, i : x_i^* \notin (0,\, a\lambda)\right\}$$
$$\mathbb{B} := \left\{F_{n,\lambda}(\mathbf{x}^*) \leq F_{n,\lambda}(\hat{\mathbf{x}}^{\min}) + \Gamma\right\}.$$

If we recall Property (iv) of P_λ and the assumption that $a\lambda \leq 1$, it holds that

$$\left.\begin{array}{l} \forall\, i : x_i^* \notin (0,\, a\lambda) \Longrightarrow \sum_{i=1}^p P_\lambda(x_i^*) = \|\mathbf{x}^*\|_0 P_\lambda(a\lambda) \\ F_{n,\lambda}(\mathbf{x}^*) \leq F_{n,\lambda}(\hat{\mathbf{x}}^{\min}) + \Gamma \end{array}\right\} \tag{35}$$

$$\Longrightarrow F_n(\mathbf{x}^*) + \|\mathbf{x}^*\|_0 P_\lambda(a\lambda) \leq F_n(\hat{\mathbf{x}}^{\min}) + |\mathcal{S}| P_\lambda(a\lambda) + \Gamma \tag{36}$$

Meanwhile,

$$\left.\begin{array}{l} (36) \\ \|\mathbf{x}^*\|_0 = \tilde{p}+k \\ F_n(\hat{\mathbf{x}}^{\min}) - F_n(\mathbf{x}^*) \leq 2t\sqrt{\tilde{p}+k} + \hat{\varepsilon} \end{array}\right\} \tag{37}$$

$$\Longrightarrow (\tilde{p}+k-|\mathcal{S}|) P_\lambda(a\lambda) \leq 2t\sqrt{\tilde{p}+k} + \hat{\varepsilon} + \Gamma \tag{38}$$

However, (38) contradicts with the assumed inequality (33), that is, the event {(38)} is a sub-event of the complement of the event {(33)}. Further noticing that {(33)} holds surely as per our assumption, therefore, {(38)}$= \emptyset$. Combining this with the observations that (35)⇒(36), and (37)⇒(38) as well as the definitions of $\mathbb{A}$, $\mathbb{B}$, $\mathcal{E}_a(\tilde{p}+k)$ and $\mathcal{E}_b$, we know that $\mathbb{A}\cap\mathbb{B}\cap\mathcal{E}_a(\tilde{p}+k)\cap\mathcal{E}_b = \emptyset$. Since $\mathbf{P}(\mathbb{A}\cap\mathbb{B}) = 1$ by assumption and by (28) with $a\lambda \leq 1$, it therefore obtains that

$$\begin{aligned} 1 &= \mathbf{P}[(\mathbb{A}\cap\mathbb{B}) \cup (\mathcal{E}_a(\tilde{p}+k)\cap\mathcal{E}_b)] \\ &= \mathbf{P}[\mathbb{A}\cap\mathbb{B}] + \mathbf{P}[\mathcal{E}_a(\tilde{p}+k)\cap\mathcal{E}_b] - \mathbf{P}[\mathbb{A}\cap\mathbb{B}\cap\mathcal{E}_a(\tilde{p}+k)\cap\mathcal{E}_b] \\ &= 1 + \mathbf{P}[\mathcal{E}_a(\tilde{p}+k)\cap\mathcal{E}_b] + 0 \\ &\Longrightarrow \mathbf{P}[\mathcal{E}_a(\tilde{p}+k)\cap\mathcal{E}_b] = 0. \end{aligned} \tag{39}$$

Secondly, we want to show that $\mathbf{P}[\bar{\mathcal{E}}_c] \geq \mathbf{P}[\mathcal{E}_c(\tilde{p}+k)]$, where $\bar{\mathcal{E}}_c$ is the complement of $\mathcal{E}_c$. To this end, consider one more event $\mathbb{C} := \{F(\mathbf{x}^{\min}) \leq F(\mathbf{x}^*)\}$, which satisfies that $\mathbf{P}[\mathbb{C}] = 1$ by the definition of $\mathbf{x}^{\min}$. We observe that, since $\|\hat{\mathbf{x}}^{\min}\|_0 = |\mathcal{S}|$,

$$\left.\begin{array}{l}\sup_{\mathbf{x}\in\mathcal{X}:\,\|\mathbf{x}\|_0\le\tilde{p}+k}|F_n(\mathbf{x})-F(\mathbf{x})|\le t\sqrt{\tilde{p}+k}\\ F(\mathbf{x}^{\min})\le F(\mathbf{x}^*)\\ \|\mathbf{x}^*\|_0=\tilde{p}+k\end{array}\right\}$$

$$\Longrightarrow\begin{cases}-F_n(\mathbf{x}^*)\le -F(\mathbf{x}^*)+t\sqrt{\tilde{p}+k}\\ F_n(\hat{\mathbf{x}}^{\min})\le F(\hat{\mathbf{x}}^{\min})+t\sqrt{\tilde{p}+k}\le F(\mathbf{x}^{\min})+t\sqrt{\tilde{p}+k}+\hat{\varepsilon}\\ \|\mathbf{x}^*\|_0=\tilde{p}+k\\ F(\mathbf{x}^{\min})\le F(\mathbf{x}^*)\end{cases}$$

which immediately leads to the simultaneous satisfaction of both $F_n(\hat{\mathbf{x}}^{\min})-F_n(\mathbf{x}^*)\le 2t\sqrt{\tilde{p}+k}+\hat{\varepsilon}$ and $\|\mathbf{x}^*\|_0=\tilde{p}+k$. Therefore, $\mathbb{C}\cap\mathcal{E}_c\cap\mathcal{E}_a(\tilde{p}+k)\subseteq\mathcal{E}_b\cap\mathcal{E}_a(\tilde{p}+k)$ and thus $\mathbf{P}[\mathbb{C}\cap\mathcal{E}_c\cap\mathcal{E}_a(\tilde{p}+k)]\le\mathbf{P}[\mathcal{E}_b\cap\mathcal{E}_a(\tilde{p}+k)]$. Since we have shown above that $\mathbf{P}[\mathcal{E}_b\cap\mathcal{E}_a(\tilde{p}+k)]=0$, we know that $\mathbf{P}[\mathbb{C}\cap\mathcal{E}_c\cap\mathcal{E}_a(\tilde{p}+k)]=0$. Further recall that we have also known that $\mathbf{P}(\mathbb{C})=1$. Therefore, by both the De Morgan's Law and the union bound, under the assumption of (33),

$$0\ge 1-\mathbf{P}[\bar{\mathcal{E}}_a(\tilde{p}+k)]-\mathbf{P}[\bar{\mathcal{E}}_c]-(1-\mathbf{P}(\mathbb{C}))\Longrightarrow\mathbf{P}[\bar{\mathcal{E}}_c]\ge\mathbf{P}[\mathcal{E}_a(\tilde{p}+k)],\quad(40)$$

where $\bar{\mathcal{E}}_a(\tilde{p}+k)$ and $\bar{\mathcal{E}}_c$ are complements of $\mathcal{E}_a(\tilde{p}+k)$ and $\mathcal{E}_c$.

Lastly, using the upper bound on $\mathbf{P}[\bar{\mathcal{E}}_c]$ provided by Lemma 3, we obtain

$$\begin{aligned}&\mathbf{P}[\mathcal{E}_a(\tilde{p}+k)]\\&\le 2\left\lceil\left(\frac{12RL_\mu\sqrt{\tilde{p}+k}}{t\sqrt{\tilde{p}+k}}\right)^{\tilde{p}+k}\binom{p}{\tilde{p}+k}\right\rceil\cdot\exp\left(-\frac{n(\tilde{p}+k)t^2}{8\sigma^2}\right)+2\exp\left(-\frac{nL_\mu^2}{2\sigma_L^2}\right)\\&\le 2\exp\left(-\frac{n(\tilde{p}+k)t^2}{8\sigma^2}+(\tilde{p}+k)\ln\left(\frac{12RL_\mu}{t}\right)+(\tilde{p}+k)\cdot\ln p\right)\\&\quad+2\exp\left(-\frac{n(\tilde{p}+k)t^2}{8\sigma^2}\right)+2\exp\left(-\frac{nL_\mu^2}{2\sigma_L^2}\right)\qquad(41)\\&=2\exp\left(-\frac{n(\tilde{p}+k)t^2}{8\sigma^2}+(\tilde{p}+k)\ln\left(\frac{12pRL_\mu}{t}\right)\right)\\&\quad+2\exp\left(-\frac{n(\tilde{p}+k)t^2}{8\sigma^2}\right)+2\exp\left(-\frac{nL_\mu^2}{2\sigma_L^2}\right).\qquad(42)\end{aligned}$$

To get (41) we make use of the facts that $\binom{p}{\tilde{p}+k}\le p^{\tilde{p}+k}$ and that $\lceil x\rceil\le x+1$ for any $x\ge 0$.

Notice that if $\|\mathbf{x}^*\|_0>\tilde{p}$, it must hold that $\|\mathbf{x}^*\|_0\in\{\tilde{p}+1,\,\ldots,\,p\}$ and that by the union bound:

$$\mathbf{P}\left[\left\{\|\mathbf{x}^*\|_0\in\{\tilde{p}+1,\,\ldots,\,p\}\right\}\right]\le\sum_{k=1}^{p-\tilde{p}}\mathbf{P}\left[\left\{\|\mathbf{x}^*\|_0=\tilde{p}+k\right\}\right].\qquad(43)$$

We therefore can find an upper bound to $\mathbf{P}[\{\|\mathbf{x}^*\|_0 \in \{\tilde{p}+1,\ \ldots,\ p\}\}]$ by invoking (42). That upper bound writes as

$$\begin{aligned}
\mathbf{P}\left[\left\{\|\mathbf{x}^*\|_0 \in \{\tilde{p}+1,\ \ldots,\ p\}\right\}\right] &\le \sum_{k=1}^{p-\tilde{p}}[\mathcal{E}_a(\tilde{p}+k)] \\
&\le \sum_{k=1}^{p-\tilde{p}} 2\exp\left(-\frac{(\tilde{p}+k)nt^2}{8\sigma^2}+(\tilde{p}+k)\ln\left(\frac{12pRL_\mu}{t}\right)\right) \\
&\quad + 2\sum_{k=1}^{p-\tilde{p}}\exp\left(-\frac{n(\tilde{p}+k)t^2}{8\sigma^2}\right)+2(p-\tilde{p})\exp\left(-\frac{nL_\mu^2}{2\sigma_L^2}\right) \\
&= 2\exp\left(-(\tilde{p}+1)\left[\frac{nt^2}{8\sigma^2}-\ln\left(\frac{12pRL_\mu}{t}\right)\right]\right) \\
&\quad \cdot\frac{1-\exp\left(-(p-\tilde{p})\left[\frac{nt^2}{8\sigma^2}-\ln\left(\frac{12pRL_\mu}{t}\right)\right]\right)}{1-\exp\left(-\left[\frac{nt^2}{8\sigma^2}-\ln\left(\frac{12pRL_\mu}{t}\right)\right]\right)}+2(p-\tilde{p})\exp\left(-\frac{nL_\mu^2}{2\sigma_L^2}\right) \\
&\quad + 2\exp\left(-\frac{(\tilde{p}+1)nt^2}{8\sigma^2}\right)\cdot\frac{1-\exp\left(-\frac{(p-\tilde{p})nt^2}{8\sigma^2}\right)}{1-\exp\left(-\frac{nt^2}{8\sigma^2}\right)} &(44)\\
&\le 1-\mathbf{P}^*(t,\tilde{p}), &(45)
\end{aligned}$$

where to achieve (44) we invoke the sum of a geometric series and to obtain (45) we make use of the assumptions that $\frac{nt^2}{8\sigma^2} \ge \ln\left(\frac{12pRL_\mu}{t}\right)$ and $\tilde{p} \le p$. The desired result then follows immediately. □

4.2.3 Proof of Proposition 1

For an arbitrary $\epsilon:\ 0<\epsilon\le 1$, denote that

$$\mathcal{E}_A := \left\{\left|F(\mathbf{x}^*)-F_n(\mathbf{x}^*)\right| \le \frac{\epsilon}{2}\right\};\ \ \mathcal{E}_B := \left\{\left|F(\hat{\mathbf{x}}^{\min})-F_n(\hat{\mathbf{x}}^{\min})\right| \le \frac{\epsilon}{2}\right\}. \qquad (46)$$

We examine the two parts of the proposition:

(i). For Part 1, according to (15), $0<\epsilon\le 1$, and $|\mathcal{S}|\ge 1$, as well as $a\le 1$, we obtain $n \ge N_1 \ge \sigma^2 = (\frac{\sigma^{2\delta}}{1})^{\frac{1}{\delta}} \ge (\frac{a\sigma^{2\delta}}{1})^{\frac{1}{\delta}}$. Combined with $0\le\rho\le\frac{1}{2}$, we know that $a\lambda = \frac{a\sigma^{2\delta}}{n^\delta|\mathcal{S}|^\rho} \le 1$. Conditioning on the event $\mathcal{E}_A\cap\mathcal{E}_B$, under the assumption that $F_{n,\lambda}(\mathbf{x}^*) \le\ F_{n,\lambda}(\hat{\mathbf{x}}^{\min})+\Gamma$ almost surely, it holds almost surely that

$$\begin{aligned}
F(\mathbf{x}^*)-F(\mathbf{x}^{\min})-\epsilon-\hat{\varepsilon} \le F(\mathbf{x}^*)-F(\hat{\mathbf{x}}^{\min})-\epsilon &\le F_n(\mathbf{x}^*)-F_n(\hat{\mathbf{x}}^{\min}) \\
&\le |\mathcal{S}|\cdot P_\lambda(a\lambda)+\Gamma = |\mathcal{S}|\cdot\frac{a\lambda^2}{2}+\Gamma
\end{aligned}$$

$$= \frac{a\sigma^{4\delta}}{2n^{2\delta}} |\mathcal{S}|^{1-2\rho} + \Gamma. \tag{47}$$

Since $a \leq 1$, if $n \geq N_1 \geq \sigma^2 \left(\frac{1}{\epsilon}\right)^{\frac{1}{2\delta}} |\mathcal{S}|^{\frac{1-2\rho}{2\delta}} \geq \sigma^2 \left(\frac{a}{\epsilon}\right)^{\frac{1}{2\delta}} |\mathcal{S}|^{\frac{1-2\rho}{2\delta}}$, then (47) implies that $F(\mathbf{x}^*) - F(\mathbf{x}^{\min}) \leq 2\epsilon + \hat{\varepsilon} + \Gamma$. Therefore, to show the first part of the proposition, it suffices to prove that there exists a problem-independent constant $c_2 > 0$ such that, if $n \geq N_1 \bigvee c_2 N^*(c_2)$ as in (15), then the event $\mathcal{E}_A \cap \mathcal{E}_B$ occurs with probability at least $1 - \alpha$, which will be shown soon afterwards.

(ii). For Part 2, according to (16), $0 < \epsilon \leq 1$, and $R \geq 1$, combined with $|\mathcal{S}| \geq 1$ and $0 \leq \rho \leq \frac{1}{2}$, we know that $n \geq N_2 \geq \sigma^2 \geq \left(\frac{a\sigma^{2\delta}}{1}\right)^{\frac{1}{\delta}} \Longrightarrow a\lambda = \frac{a\sigma^{2\delta}}{n^\delta |\mathcal{S}|^\rho} \leq 1$. Conditioning on the event $\mathcal{E}_A \cap \mathcal{E}_B$, under Assumption A.5, we obtain from (26) that

$$\begin{aligned} F(\mathbf{x}^*) - F(\mathbf{x}^{\min}) - \epsilon - \hat{\varepsilon} &\leq F(\mathbf{x}^*) - F(\hat{\mathbf{x}}^{\min}) - \epsilon \leq F_n(\mathbf{x}^*) - F_n(\hat{\mathbf{x}}^{\min}) \\ &\leq \lambda |\hat{\mathbf{x}}^{\min}| = \frac{\sigma^{2\delta}}{n^\delta} |\mathcal{S}|^{1-\rho} R \end{aligned} \tag{48}$$

Hence, if $n \geq N_2 \geq \left(\frac{|\mathcal{S}|^{1-\rho} R \sigma^{2\delta}}{\epsilon}\right)^{\frac{1}{\delta}}$, then (48) implies that $F(\mathbf{x}^*) - F(\mathbf{x}^{\min}) \leq 2\epsilon + \hat{\varepsilon}$. Therefore, to show the second part of the proposition, it also suffices to show that there exists a problem-independent constant $c_2 > 0$ such that, if $n \geq N_2 \bigvee c_2 N^*(c_2)$ as in (16), then the event $\mathcal{E}_A \cap \mathcal{E}_B$ occurs with probability at least $1 - \alpha$.

The following provides probability lower bound for the occurrence of $\mathcal{E}_A \cap \mathcal{E}_B$. Such a bound applies to both (i) and (ii) above.

We have shown above that $a\lambda \leq 1$ for both (i) and (ii), and we also have let Assumptions A.1–A.3 and Condition B hold. Under the assumption that $F_{n,\lambda}(\mathbf{x}^*) \leq F_{n,\lambda}(\hat{\mathbf{x}}^{\min}) + \Gamma$ almost surely, we may invoke Lemma 4, where we assume for now that

$$\frac{nt^2}{8\sigma^2} \geq \ln\left(\frac{12pRL_\mu}{t}\right) \tag{49}$$

which will be shown soon afterwards. It then follows that, for any integer $\tilde{p} \geq |\mathcal{S}|$ such that $\tilde{p} > |\mathcal{S}| + \frac{2t\sqrt{\tilde{p}+1} + \Gamma + \hat{\varepsilon}}{P_\lambda(a\lambda)} - 1 \Longleftrightarrow \sqrt{\tilde{p}+1} > \frac{t}{P_\lambda(a\lambda)} + \sqrt{\frac{t^2}{[P_\lambda(a\lambda)]^2} + |\mathcal{S}| + \frac{\Gamma + \hat{\varepsilon}}{P_\lambda(a\lambda)}}$, it holds that $\|\mathbf{x}^*\|_0 \leq \tilde{p}$ with probability at least $\mathbf{P}^*(t, \tilde{p})$ as defined in (34). Further notice that, since $\|\hat{\mathbf{x}}^{\min}\|_0 = |\mathcal{S}|$, for any $\tilde{p} \geq |\mathcal{S}|$ it holds that $\mathcal{E}_A \cap \mathcal{E}_B \supseteq \left\{\sup_{\mathbf{x} \in \mathcal{X}:\, \|\mathbf{x}\|_0 \leq \tilde{p} \bigwedge p} \left|\frac{1}{n}\sum_{j=1}^n f(\mathbf{x}, W^j) - F(\mathbf{x})\right| \leq \epsilon/2\right\} \cap \{\|\mathbf{x}^*\|_0 \leq \tilde{p}\}$. Hence we may combine Lemma 3 (in which we let $t = \frac{\epsilon}{2}$ and rescale $\tilde{p}$ only within that lemma into $p \bigwedge \tilde{p}$), and Lemma 4 (in which we let $\tilde{p} = \left\lfloor \frac{4t^2}{[P_\lambda(a\lambda)]^2} + 4|\mathcal{S}| + \frac{4(\Gamma + \hat{\varepsilon})}{P_\lambda(a\lambda)} \right\rfloor$ here and we will also let $t = \frac{\sigma^{2\delta}}{n^\delta |\mathcal{S}|^\rho}$ soon afterwards) through both the De Morgan's Law and the union bound to obtain that $\mathcal{E}_A \cap \mathcal{E}_B$ occurs with probability at least

$$\begin{aligned}
\mathcal{P}^* &:= \Bigg[\mathbf{P}^*\left(t, \tilde{p}\right) - 2 \Bigg[\left(\frac{24\sqrt{\tilde{p}}RL_\mu}{\epsilon} \right)^{(p \bigwedge \tilde{p})} \binom{p}{p \bigwedge \tilde{p}} \Bigg] \\
&\quad \times \exp\left(-\frac{n\epsilon^2}{32\sigma^2} \right) - 2\exp\left(-\frac{nL_\mu^2}{2\sigma_L^2} \right) \Bigg]_{\tilde{p} = \left\lfloor \frac{16t^2}{a^2\lambda^4} + \frac{8(\Gamma+\hat{\varepsilon})}{a\lambda^2} + 4|\mathcal{S}| \right\rfloor} \\
&\geq 1 - 2\exp\left(-\frac{n\epsilon^2}{32\sigma^2} \right) - 2(p+1)\exp\left(-\frac{nL_\mu^2}{2\sigma_L^2} \right) \\
&\quad - 2\exp\left(-\frac{n\epsilon^2}{32\sigma^2} + \left[p \bigwedge \left\lfloor \frac{16t^2}{a^2\lambda^4} + \frac{8(\Gamma + \hat{\varepsilon})}{a\lambda^2} + 4|\mathcal{S}| \right\rfloor \right] \ln\left(\frac{24RL_\mu p^{3/2}}{\epsilon} \right) \right) \\
&\quad - \frac{2\exp\left(-\left[\left\lfloor \frac{16t^2}{a^2\lambda^4} + \frac{8(\Gamma+\hat{\varepsilon})}{a\lambda^2} + 4|\mathcal{S}| \right\rfloor + 1 \right] \left[\frac{nt^2}{8\sigma^2} - \ln\left(\frac{12pRL_\mu}{t} \right) \right] \right)}{1 - \exp\left(-\left[\frac{nt^2}{8\sigma^2} - \ln\left(\frac{12pRL_\mu}{t} \right) \right] \right)} \\
&\quad - 2\exp\left(-\frac{\left(\left\lfloor \frac{16t^2}{a^2\lambda^4} + \frac{8(\Gamma+\hat{\varepsilon})}{a\lambda^2} + 4|\mathcal{S}| \right\rfloor + 1 \right) nt^2}{8\sigma^2} \right) \cdot \frac{1}{1 - \exp\left(-\frac{nt^2}{8\sigma^2} \right)},
\end{aligned} \tag{50}$$

where we may plug in $t = \frac{\sigma^{2\delta}}{n^\delta |\mathcal{S}|^\rho}$ in the next.

Now we want to show the satisfaction of (49). Observe that, with $t = \lambda = \frac{\sigma^{2\delta}}{n^\delta |\mathcal{S}|^\rho}$, $\delta < \frac{1}{2}$, $\rho \leq \frac{1}{2}$, $4p^2 \geq n$ and $p \geq |\mathcal{S}| \geq 1$, we know that

$$\begin{aligned}
&\frac{nt^2}{8\sigma^2} - \ln\left(\frac{12pRL_\mu}{t} \right) \\
&= \frac{n^{1-2\delta}}{8\sigma^{2-4\delta}|\mathcal{S}|^{2\rho}} - \ln\left(\frac{12n^\delta |\mathcal{S}|^\rho pRL_\mu}{\sigma^{2\delta}} \right) \\
&\geq \frac{n^{1-2\delta}}{8\sigma^{2-4\delta}|\mathcal{S}|^{2\rho}} - \ln\left(\frac{24p^{5/2}RL_\mu}{\sigma^{2\delta}} \right) \\
&= \frac{n^{1-2\delta}}{16\sigma^{2-4\delta}|\mathcal{S}|^{2\rho}} + \frac{n^{1-2\delta}}{16\sigma^{2-4\delta}|\mathcal{S}|^{2\rho}} - \ln\left(\frac{24p^{5/2}RL_\mu}{\sigma^{2\delta}} \right)
\end{aligned} \tag{51}$$

Observe that, if $n \geq \left[12\sigma^{2-4\delta}|\mathcal{S}|^{2\rho} \bigvee 16\sigma^{2-4\delta}|\mathcal{S}|^{2\rho} \ln\left(\frac{24p^{5/2}RL_\mu}{\sigma^{2\delta}} \right) \right]^{1/(1-2\delta)}$, then $\frac{n^{1-2\delta}}{16\sigma^{2-4\delta}|\mathcal{S}|^{2\rho}} \geq \ln\left(\frac{24p^{5/2}RL_\mu}{\sigma^{2\delta}} \right) \bigvee \frac{12}{16}$. Therefore, we know that (51) $\geq \frac{n^{1-2\delta}}{16\sigma^{2-4\delta}|\mathcal{S}|^{2\rho}} \geq \frac{12}{16} \geq \ln 2$. This inequality implies (49).

The above provides a lower bound on the probability for the event of interest. The rest of the proof seeks to simplify this bound. We have shown above that (51) $\geq$

$\frac{n^{1-2\delta}}{16\sigma^{2-4\delta}|\mathcal{S}|^{2\rho}} \geq \ln 2$. This inequality implies both $\exp(-\frac{nt^2}{8\sigma^2}) \leq 1/2$ and

$$\exp\left(-\left[\frac{nt^2}{8\sigma^2} - \ln\left(\frac{12pRL_\mu}{t}\right)\right]\right) \leq \frac{1}{2}.$$

Further observing $\frac{t^2}{\lambda^4} = \frac{n^{2\delta}|\mathcal{S}|^{2\rho}}{\sigma^{4\delta}}$, we may combine the above with (50) to obtain

$$\begin{aligned}
\mathcal{P}^* \geq\; & 1 - 2\exp\left(-\frac{16t^2}{a^2\lambda^4}\cdot\left[\frac{nt^2}{16\sigma^2} + \frac{nt^2}{16\sigma^2} - \ln\left(\frac{12pRL_\mu}{t}\right)\right]\right) \\
& \times \frac{1}{1-\exp\left(-\left[\frac{nt^2}{16\sigma^2} + \frac{nt^2}{16\sigma^2} - \ln\left(\frac{12pRL_\mu}{t}\right)\right]\right)} \\
& - 2\exp\left(-\frac{\left(\left\lfloor\frac{16t^2}{a^2\lambda^4} + \frac{8(\Gamma+\hat{\varepsilon})}{a\lambda^2} + 4|\mathcal{S}|\right\rfloor + 1\right)nt^2}{8\sigma^2}\right)\cdot\frac{1}{1-\exp\left(-\frac{nt^2}{8\sigma^2}\right)} \\
& - 2\exp\left(-\frac{n\epsilon^2}{32\sigma^2} + \left\lfloor\frac{16t^2}{a^2\lambda^4} + \frac{8(\Gamma+\hat{\varepsilon})}{a\lambda^2} + 4|\mathcal{S}|\right\rfloor\ln\left(\frac{24RL_\mu p^{3/2}}{\epsilon}\right)\right) \\
& - 2\exp\left(-\frac{n\epsilon^2}{32\sigma^2}\right) - 2(p+1)\exp\left(-\frac{nL_\mu^2}{2\sigma_L^2}\right) \\
\geq\; & 1 - 2\exp\left(-\frac{16t^2}{a^2\lambda^4}\cdot\frac{nt^2}{16\sigma^2}\right)\cdot\frac{1}{1-\frac{1}{2}} - 2\exp\left(-\frac{16t^2}{a^2\lambda^4}\cdot\frac{nt^2}{8\sigma^2}\right)\cdot\frac{1}{1-\frac{1}{2}} \\
& - 2\exp\left(-\frac{n\epsilon^2}{32\sigma^2} + \left\lfloor\frac{16t^2}{a^2\lambda^4} + \frac{8(\Gamma+\hat{\varepsilon})}{a\lambda^2} + 4|\mathcal{S}|\right\rfloor\ln\left(\frac{24RL_\mu p^{3/2}}{\epsilon}\right)\right) \\
& - 2\exp\left(-\frac{n\epsilon^2}{32\sigma^2}\right) - 2(p+1)\exp\left(-\frac{nL_\mu^2}{2\sigma_L^2}\right) \\
\geq\; & 1 - 8\exp\left(-\frac{n}{a^2\sigma^2}\right) - 2(p+1)\exp\left(-\frac{nL_\mu^2}{2\sigma_L^2}\right) - 2\exp\left(-\frac{n\epsilon^2}{32\sigma^2}\right) \\
& - 2\exp\left(-\frac{n\epsilon^2}{32\sigma^2} + \left\lfloor\frac{16n^{2\delta}|\mathcal{S}|^{2\rho}}{a^2\sigma^{4\delta}} + \frac{8(\Gamma+\hat{\varepsilon})|\mathcal{S}|^{2\rho}n^{2\delta}}{a\sigma^{4\delta}} + 4|\mathcal{S}|\right\rfloor\right. \\
& \left.\times\ln\left(\frac{24RL_\mu p^{3/2}}{\epsilon}\right)\right)
\end{aligned}$$

Combined with the above, it is easily verifiable that, if n is large enough to satisfy both $n \geq \sigma^2\left[12|\mathcal{S}|^{2\rho} \bigvee 16|\mathcal{S}|^{2\rho}\ln\left(\frac{24p^{5/2}RL_\mu}{\sigma^{2\delta}}\right)\right]^{1/(1-2\delta)}$ and

$$n \geq a^2\sigma^2\ln\frac{32}{\alpha} + \frac{2\sigma_L^2}{L_\mu^2}\ln\left(\frac{8(p+1)}{\alpha}\right) + \frac{\sigma^2}{\epsilon^2}\left(64\cdot\ln\frac{8}{\alpha} + 256\cdot|\mathcal{S}|\ln\left(\frac{24RL_\mu p^{3/2}}{\epsilon}\right)\right)$$

$$\bigvee \sigma^2 \left[\frac{64}{\epsilon^2}\left(\frac{16|\mathcal{S}|^{2\rho}}{a^2}+\frac{8(\Gamma+\hat{\varepsilon})|\mathcal{S}|^{2\rho}}{a}\right)\ln\frac{24RL_\mu p^{5/2}}{\epsilon}\right]^{\frac{1}{1-2\delta}},$$

then $\mathcal{P}^* \geq 1-\alpha$. Therefore, recalling that $a \leq 1$, $L_\mu \geq 1$, $p \geq |\mathcal{S}| \geq 1$ and $\epsilon \leq 1$, there exists a problem-independent constant $c_2 > 0$ such that the above stipulation of n is satisfied if

$$n \geq c_2^{\frac{1}{1-2\delta}} \sigma^2 |\mathcal{S}|^{\frac{2\rho}{1-2\delta}} \cdot \left(\frac{1+\Gamma+\hat{\varepsilon}}{a^2\epsilon^2}\ln\frac{24RL_\mu p}{\min\{\epsilon,\sigma^{2\delta}\}}\right)^{\frac{1}{1-2\delta}} \bigvee c_2 \cdot N^*(c_2). \tag{52}$$

Combining the above with (i) Eq. (47) and (ii) Eq. (48) yields the desired results for part 1 and part 2 of the proposition, respectively.

4.2.4 Proof of Proposition 2

For an arbitrary $\epsilon:\ 0<\epsilon\leq 1$, let us consider the events that

$$\mathcal{E}_A := \left\{\left|F(\mathbf{x}^*)-F_n(\mathbf{x}^*)\right| \leq \frac{\epsilon}{2}\right\};\quad \mathcal{E}_B := \left\{\left|F(\hat{\mathbf{x}}^{\min})-F_n(\hat{\mathbf{x}}^{\min})\right| \leq \frac{\epsilon}{2}\right\} \tag{53}$$

Conditioning on the event $\mathcal{E}_A \cap \mathcal{E}_B$, under Assumption A.5, we obtain from (25) that, almost surely,

$$\begin{aligned} F(\mathbf{x}^*)-F(\mathbf{x}^{\min})-\epsilon-\hat{\varepsilon} \leq& F(\mathbf{x}^*)-F(\hat{\mathbf{x}}^{\min})-\epsilon \\ \leq& F_n(\mathbf{x}^*)-F_n(\hat{\mathbf{x}}^{\min}) \leq \lambda\sum_{i\in\mathcal{S}}|\hat{x}_i^{\min}-x_i^*| \\ =& \lambda\sqrt{|\mathcal{S}|}\sqrt{\sum_{i\in\mathcal{S}}\|\hat{x}_i^{\min}-x_i^*\|^2} \end{aligned} \tag{54}$$

Further invoking (17), which immediately leads to $F(\mathbf{x})-F(\mathbf{x}^{\min}) \geq \frac{\mathcal{U}_{\mathcal{H}}}{2}\|\mathbf{x}-\mathbf{x}^{\min}\|^2$ for all $\mathbf{x}\in\mathcal{X}$, we may continue the above as, almost surely (conditioning on $\mathcal{E}_A\cap\mathcal{E}_B$),

$$\begin{aligned} &F(\mathbf{x}^*)-F(\mathbf{x}^{\min})-\epsilon-\hat{\varepsilon} \\ &\leq \lambda\sqrt{|\mathcal{S}|}\sqrt{\sum_{i\in\mathcal{S}}\|\hat{x}_i^{\min}-x_i^*\|^2} \leq \lambda\sqrt{|\mathcal{S}|}\cdot\|\mathbf{x}^*-\hat{\mathbf{x}}^{\min}\| \\ &\leq \lambda\sqrt{|\mathcal{S}|}\cdot\|\mathbf{x}^*-\mathbf{x}^{\min}\|+\lambda\sqrt{|\mathcal{S}|}\cdot\|\hat{\mathbf{x}}^{\min}-\mathbf{x}^{\min}\| \\ &\leq \lambda\sqrt{|\mathcal{S}|}\cdot\sqrt{\frac{2}{\mathcal{U}_{\mathcal{H}}}(F(\mathbf{x}^*)-F(\mathbf{x}^{\min}))}+\lambda\sqrt{|\mathcal{S}|}\cdot\sqrt{\frac{2}{\mathcal{U}_{\mathcal{H}}}(F(\hat{\mathbf{x}}^{\min})-F(\mathbf{x}^{\min}))} \\ &\leq \lambda\sqrt{|\mathcal{S}|}\cdot\sqrt{\frac{2}{\mathcal{U}_{\mathcal{H}}}(F(\mathbf{x}^*)-F(\mathbf{x}^{\min}))}+\lambda\sqrt{|\mathcal{S}|}\cdot\sqrt{\frac{2\hat{\varepsilon}}{\mathcal{U}_{\mathcal{H}}}}. \end{aligned}$$

Solving the inequality for $\sqrt{F(\mathbf{x}^*) - F(\mathbf{x}^{\min})}$, we have, almost surely (conditioning on $\mathcal{E}_A \cap \mathcal{E}_B$),

$$\sqrt{F(\mathbf{x}^*) - F(\mathbf{x}^{\min})} \leq \frac{\lambda\sqrt{\frac{2|\mathcal{S}|}{\mathcal{U}_{\mathcal{H}}}} + \sqrt{\frac{2\lambda^2|\mathcal{S}|}{\mathcal{U}_{\mathcal{H}}} + 4(\hat{\varepsilon} + \epsilon) + 4\lambda\sqrt{|\mathcal{S}|} \cdot \sqrt{\frac{2\hat{\varepsilon}}{\mathcal{U}_{\mathcal{H}}}}}}{2} \tag{55}$$

Therefore, combined with $\lambda = \frac{\sigma^{2\delta}}{n^\delta |\mathcal{S}|^\rho}$, we know that

$$\begin{aligned} &F(\mathbf{x}^*) - F(\mathbf{x}^{\min}) \\ &\leq \left(\sqrt{\frac{\sigma^{4\delta}|\mathcal{S}|^{1-2\rho}}{2\mathcal{U}_{\mathcal{H}} n^{2\delta}}} + \sqrt{\frac{\sigma^{4\delta}|\mathcal{S}|^{1-2\rho}}{2\mathcal{U}_{\mathcal{H}} n^{2\delta}} + \sqrt{\frac{2\sigma^{4\delta}\hat{\varepsilon}|\mathcal{S}|^{1-2\rho}}{n^{2\delta}\mathcal{U}_{\mathcal{H}}}} + (\hat{\varepsilon} + \epsilon)}\right)^2 \end{aligned}$$

almost surely (conditioning on $\mathcal{E}_A \cap \mathcal{E}_B$).

Notice that if $n \geq \sigma^2 \left(\frac{8|\mathcal{S}|^{1-2\rho}}{\mathcal{U}_{\mathcal{H}}\epsilon}\right)^{\frac{1}{2\delta}} \bigvee \sigma^2 \left(\frac{8\hat{\varepsilon}|\mathcal{S}|^{1-2\rho}}{\mathcal{U}_{\mathcal{H}}\epsilon^2}\right)^{\frac{1}{2\delta}} \bigvee \sigma^2$, then the following three inequalities hold: (a) $a\lambda = a\frac{\sigma^{2\delta}}{n^\delta|\mathcal{S}|^\rho} \leq 1$; (b) $\frac{\sigma^{4\delta}|\mathcal{S}|^{1-2\rho}}{2\mathcal{U}_{\mathcal{H}} n^{2\delta}} \leq \frac{\epsilon}{16}$; and (c) $\sqrt{\frac{2\sigma^{4\delta}\hat{\varepsilon}|\mathcal{S}|^{1-2\rho}}{n^{2\delta}\mathcal{U}_{\mathcal{H}}}} \leq \frac{\varepsilon}{2}$. Thus,

$$\begin{aligned} &\left(\sqrt{\frac{\sigma^{4\delta}|\mathcal{S}|^{1-2\rho}}{2\mathcal{U}_{\mathcal{H}} n^{2\delta}}} + \sqrt{\frac{\sigma^{4\delta}|\mathcal{S}|^{1-2\rho}}{2\mathcal{U}_{\mathcal{H}} n^{2\delta}} + \sqrt{\frac{2\sigma^{4\delta}\hat{\varepsilon}|\mathcal{S}|^{1-2\rho}}{n^{2\delta}\mathcal{U}_{\mathcal{H}}}} + (\hat{\varepsilon} + \epsilon)}\right)^2 \\ &\leq \left(\frac{\sqrt{\epsilon}}{4} + \sqrt{\frac{25\epsilon}{16} + \hat{\varepsilon}}\right)^2 = \frac{26\epsilon}{16} + \hat{\varepsilon} + \sqrt{\frac{25\epsilon^2}{64} + \frac{\epsilon\hat{\varepsilon}}{4}} \\ &\leq \frac{26\epsilon}{16} + \hat{\varepsilon} + \sqrt{\frac{25\epsilon^2}{64} + \frac{\epsilon\hat{\varepsilon}}{4} + \frac{\hat{\varepsilon}^2}{25}} \\ &= \left(\frac{26}{16} + \frac{5}{8}\right)\epsilon + \left(1 + \frac{1}{5}\right)\hat{\varepsilon} = \frac{9}{4}\epsilon + \frac{6}{5}\hat{\varepsilon} \end{aligned} \tag{56}$$

Hence, if $n \geq \sigma^2|\mathcal{S}|^{\frac{1-2\rho}{2\delta}} \left[\left(\frac{8}{\mathcal{U}_{\mathcal{H}}\epsilon}\right)^{\frac{1}{2\delta}} + \left(\frac{8\hat{\varepsilon}}{\mathcal{U}_{\mathcal{H}}\epsilon^2}\right)^{\frac{1}{2\delta}}\right] \bigvee \sigma^2$, then (55) implies that $F(\mathbf{x}^*) - F(\mathbf{x}^{\min}) \leq 3\epsilon + 3\hat{\varepsilon}$ almost surely (conditioning on $\mathcal{E}_A \cap \mathcal{E}_B$). Therefore, to achieve the desired result of the proposition, it suffices to show that, if n additionally satisfies

$$n \geq c_3^{\frac{1}{1-2\delta}} \cdot \sigma^2|\mathcal{S}|^{\frac{2\rho}{1-2\delta}} \cdot \left(\frac{1 + \Gamma + \hat{\varepsilon}}{a^2\epsilon^2} \ln \frac{24RL_\mu p}{\min\{\epsilon, \sigma^{2\delta}\}}\right)^{\frac{1}{1-2\delta}} \bigvee c_3 \cdot N^*(c_3)$$

for some universal constant $c_3 > 0$, then the event $\mathcal{E}_A \cap \mathcal{E}_B$ occurs with probability at least $1-\alpha$, which can be shown by the same argument as in the proof for Proposition 1 as in Sect. 4.2.3 in showing (52). Further noticing that we can let $c_3 \geq 2$ to further satisfy that $c_3 N^*(c_3) \geq \frac{2\sigma^2}{\epsilon^2} \ln \frac{2}{\alpha} \geq \sigma^2$ (since $\alpha \leq 1$ and $\epsilon \leq 1$), we then have the desired result. □

4.2.5 Proof of Theorem 2

We first want to show that, if $\lambda = \frac{\sigma^{2\delta}}{n^\delta |\mathcal{S}|^\rho}$, then $F_{n,\lambda}(\mathbf{0}) - F_{n,\lambda}(\hat{\mathbf{x}}^{\min}) \leq 2L_\mu R\sqrt{|\mathcal{S}|}$ with a lower bounded probability. To this end, we observe that $\|\mathbf{0} - \hat{\mathbf{x}}^{\min}\| = \|\hat{\mathbf{x}}^{\min}\| \leq R\sqrt{|\mathcal{S}|}$. This combined with Lemma 1 (where we let $t = L_{\mu,s}$ in that lemma) in Sect. 4.1, we know that

$$|F_n(\mathbf{0}) - F_n(\hat{\mathbf{x}}^{\min})| \leq 2L_{\mu,s} R\sqrt{|\mathcal{S}|}, \tag{57}$$

with probability at least $1 - 2\exp(-\frac{n(L_{\mu,s})^2}{2\sigma_L^2})$. Furthermore, since $F_n(\mathbf{0}) = F_{n,\lambda}(\mathbf{0})$ and $F_{n,\lambda}(\hat{\mathbf{x}}^{\min}) = F_n(\hat{\mathbf{x}}^{\min}) + \sum_{i=1}^p P_\lambda(\hat{x}_i^{\min})$, we have that

$$\begin{aligned} F_{n,\lambda}(\mathbf{0}) - F_{n,\lambda}(\hat{\mathbf{x}}^{\min}) &= F_n(\mathbf{0}) - F_n(\hat{\mathbf{x}}^{\min}) - \sum_{i=1}^p P_\lambda(\hat{x}_i^{\min}) \\ &\leq F_n(\mathbf{0}) - F_n(\hat{\mathbf{x}}^{\min}) \leq 2L_{\mu,s} R\sqrt{|\mathcal{S}|} \end{aligned} \tag{58}$$

with a lower bounded probability $1 - 2\exp(-\frac{nL_{\mu,s}^2}{2\sigma_L^2})$.

Then, we may invoke both the De Morgan's Law and the union bound to combine the above with Part 2 of Proposition 1, where we let $\delta = \frac{1}{4}$ and $\Gamma = 2L_{\mu,s} R\sqrt{|\mathcal{S}|}$. As a result, there exists a problem-independent constant $\tilde{c}_5 > 0$ such that, if

$$\begin{aligned} n \geq \sigma^2 \cdot |\mathcal{S}|^{4-4\rho} \left(\frac{R}{\epsilon}\right)^4 \bigvee \tilde{c}_5 \cdot N^*(c_5) \\ \bigvee \tilde{c}_5 \cdot \sigma^2 |\mathcal{S}|^{4\rho} \cdot \left(\frac{1 + 2L_{\mu,s} R\sqrt{|\mathcal{S}|} + \hat{\varepsilon}}{a^2 \epsilon^2} \ln \frac{\tilde{c}_5 R L_\mu p}{\min\{\epsilon, \sigma^{1/2}\}}\right)^2 \end{aligned} \tag{59}$$

then $F(\mathbf{x}^*) - F(\mathbf{x}^{\min}) \leq 2\epsilon + \hat{\varepsilon}$ with probability lower bounded by $1 - \alpha - 2\exp(-\frac{n(L_{\mu,s})^2}{2\sigma_L^2})$. Recall again that $a \leq 1$. Then, inequality (59) holds with $2\exp(-\frac{n(L_{\mu,s})^2}{2\sigma_L^2}) \leq \alpha$, if $\rho = 3/8$ and if n is large enough to satisfy both of the following inequalities

$$n \geq 2\sigma_L^2 \cdot \ln \frac{2}{\alpha} \geq \frac{2\sigma_L^2}{L_{\mu,s}^2} \ln \frac{2}{\alpha} \tag{60}$$

where the last inequality is due to $L_{\mu,s} \geq 1$.

$$n \geq \sigma^2 \cdot |\mathcal{S}|^{5/2} \left(\frac{R}{\epsilon}\right)^4 \bigvee \tilde{c}_5 \cdot \sigma^2 |\mathcal{S}|^{5/2} \cdot \left(\frac{1+2L_{\mu,s}R+\hat{\varepsilon}}{a^2\epsilon^2} \ln \frac{\tilde{c}_5 R L_\mu p}{\min\{\epsilon, \sigma^{1/2}\}}\right)^2 \bigvee \tilde{c}_5 \cdot N^*(c_5). \tag{61}$$

The above immediately leads to the desired result by observing that $\tilde{c}_5 N^*(\tilde{c}_5) \geq 2\sigma_L^2 \cdot \ln \frac{2}{\alpha}$ if $\tilde{c}_5 \geq 2$. □

4.2.6 Proof of Theorem 3

Following the same argument as in the proof for Theorem 2, we have $F_{n,\lambda}(\mathbf{0}) \leq F_{n,\lambda}(\hat{\mathbf{x}}^{\min}) + 2L_{\mu,s}R\sqrt{|\mathcal{S}|}$ with lower-bounded probability $1 - 2\exp(-\frac{nL_{\mu,s}^2}{2\sigma_L^2})$. We may invoke both the De Morgan's Law and the union bound to combine the above with Proposition 2, where we let $\delta = \frac{1}{6}$, $\rho = 1/4$ and $\Gamma = 2L_{\mu,s}R\sqrt{|\mathcal{S}|}$. As a result, $F(\mathbf{x}^*) - F(\mathbf{x}^{\min}) \leq 3(\epsilon + \hat{\varepsilon})$ with probability lower bounded by $1-\alpha-2\exp(-\frac{nL_{\mu,s}^2}{2\sigma_L^2})$, for n satisfying

$$n \geq \tilde{c}_6 |\mathcal{S}|^{3/2}\sigma^2 \left[\left(\frac{1}{\mathcal{U}_{\mathcal{H}}\epsilon}\right)^3 + \left(\frac{\hat{\varepsilon}}{\mathcal{U}_{\mathcal{H}}\epsilon^2}\right)^3\right] \bigvee \tilde{c}_6 N^*(c_6) \bigvee \tilde{c}_6 \sigma^2 |\mathcal{S}|^{3/4} \cdot \left(\frac{1+2L_{\mu,s}R\sqrt{|\mathcal{S}|}+\hat{\varepsilon}}{a^2\epsilon^2} \ln \frac{\tilde{c}_6 R L_\mu p}{\min\{\epsilon, \sigma^{1/3}\}}\right)^{\frac{3}{2}} \tag{62}$$

Therefore, since $a \leq 1$ and $L_{\mu,s} \geq 1$, if one stipulates both

$$n \geq 2\sigma_L^2 \cdot \ln \frac{2}{\alpha} \geq \frac{2\sigma_L^2}{L_{\mu,s}^2} \ln \frac{2}{\alpha} \Longrightarrow 2\exp\left(-\frac{nL_{\mu,s}^2}{2\sigma_L^2}\right) \leq \alpha$$

and, for some problem-independent $\tilde{c}_6 > 0$,

$$n \geq \tilde{c}_6 \sigma^2 |\mathcal{S}|^{3/2} \left[\left(\frac{1}{\mathcal{U}_{\mathcal{H}}\epsilon}\right)^3 + \left(\frac{\hat{\varepsilon}}{\mathcal{U}_{\mathcal{H}}\epsilon^2}\right)^3\right] \bigvee \tilde{c}_6 \sigma^2 |\mathcal{S}|^{3/2} \cdot \left(\frac{1+2L_{\mu,s}R+\hat{\varepsilon}}{a^2\epsilon^2} \ln \frac{\tilde{c}_6 R L_\mu p}{\min\{\epsilon, \sigma^{1/3}\}}\right)^{\frac{3}{2}} \bigvee \tilde{c}_6 N^*(\tilde{c}_6),$$

we know that $F(\mathbf{x}^*) - F(\mathbf{x}^{\min}) \leq 3(\epsilon + \hat{\varepsilon})$ with probability lower bounded by $1 - 2\alpha$. This immediately leads to the desired result by further noticing that $\tilde{c}_6 N^*(\tilde{c}_6) \geq 2\sigma_L^2 \cdot \ln \frac{2}{\alpha}$ if $\tilde{c}_6 \geq 2$. □

4.2.7 Proof of Theorem 4

Consider again

$$\mathcal{E}_A := \left\{ \left| F(\mathbf{x}^*) - F_n(\mathbf{x}^*) \right| \leq \frac{\epsilon}{2} \right\}; \quad \text{and} \quad \mathcal{E}_B := \left\{ \left| F(\hat{\mathbf{x}}^{\min}) - F_n(\hat{\mathbf{x}}^{\min}) \right| \leq \frac{\epsilon}{2} \right\}.$$

Following the same steps as in the proof for Proposition 2, it obtains that (55) holds almost surely conditioning on $\mathcal{E}_A \cap \mathcal{E}_B$. When $\hat{\varepsilon} = 0$ and $\lambda = \frac{\sigma^{2\delta}}{n^\delta |\mathcal{S}|^\rho}$ with $\rho = \frac{1}{4}$ and $\delta = 0$, (55) immediately yields:

$$F(\mathbf{x}^*) - F(\mathbf{x}^{\min}) = F(\mathbf{x}^*) - F(\hat{\mathbf{x}}^{\min}) \leq \left(\sqrt{\frac{|\mathcal{S}|^{1/2}}{2\mathcal{U}_\mathcal{H}}} + \sqrt{\frac{|\mathcal{S}|^{1/2}}{2\mathcal{U}_\mathcal{H}} + \epsilon} \right)^2$$

almost surely conditioning on $\mathcal{E}_A \cap \mathcal{E}_B$. Since it is assumed that F is differentiable and strongly convex as in (17) with constant $\mathcal{U}_\mathcal{H}$, we know that $F(\mathbf{x}) - F(\mathbf{x}^{\min}) \geq \frac{\mathcal{U}_\mathcal{H}}{2} \|\mathbf{x} - \mathbf{x}^{\min}\|^2$ for all $\mathbf{x} \in \mathcal{X}$ and that $\hat{\mathbf{x}}^{\min} = \mathbf{x}^{\min}$ (because we have let $\hat{\varepsilon} = 0$). Therefore,

$$\begin{aligned}
&\frac{\mathcal{U}_\mathcal{H}}{2} \|\mathbf{x}^* - \hat{\mathbf{x}}^{\min}\|^2 \leq \left(\sqrt{\frac{|\mathcal{S}|^{1/2}}{2\mathcal{U}_\mathcal{H}}} + \sqrt{\frac{|\mathcal{S}|^{1/2}}{2\mathcal{U}_\mathcal{H}} + \epsilon} \right)^2 \\
&\overset{0<\epsilon\leq 1}{\leq} \left(\sqrt{\frac{|\mathcal{S}|^{1/2}}{2\mathcal{U}_\mathcal{H}}} + \sqrt{\frac{|\mathcal{S}|^{1/2}}{2\mathcal{U}_\mathcal{H}} + 1} \right)^2 \\
&\quad \Longrightarrow \min_{i\in\mathcal{S}} \hat{x}_i^{\min} - \min_{i\in\mathcal{S}} x_i^* \leq \|\mathbf{x}^* - \hat{\mathbf{x}}^{\min}\| \leq \sqrt{\frac{2}{\mathcal{U}_\mathcal{H}}} \cdot \left(\sqrt{\frac{|\mathcal{S}|^{1/2}}{2\mathcal{U}_\mathcal{H}}} + \sqrt{\frac{|\mathcal{S}|^{1/2}}{2\mathcal{U}_\mathcal{H}} + 1} \right)
\end{aligned}$$

almost surely conditioning on $\mathcal{E}_A \cap \mathcal{E}_B$, where we have made use of the assumption that $\mathbf{x}^*,\ \hat{\mathbf{x}}^{\min} \in \mathcal{X} \subseteq \Re_+^p$. Therefore, if

$$\min_{i\in\mathcal{S}} \hat{x}_i^{\min} > \frac{|\mathcal{S}|^{1/4} + \sqrt{|\mathcal{S}|^{1/2} + 2\mathcal{U}_\mathcal{H}}}{\mathcal{U}_\mathcal{H}},$$

it holds that $\min_{i\in\mathcal{S}} x_i^* > 0$ almost surely conditioning on $\mathcal{E}_A \cap \mathcal{E}_B$. Further invoking (27) with $a\lambda = \frac{a}{|\mathcal{S}|^{1/4}} \leq 1$, we know that $\min_{i\in\mathcal{S}} x_i^* \geq a\lambda$, and thus $P_\lambda'(x_i^*) = 0$ for all $i \in \mathcal{S}$ and $P_\lambda'(x_i^*) \geq 0$ for all $i = 1, \ldots, p$ due to Properties (iii) and (iv) of MCP in Sect. 4.1. If we recall (25) and the fact that $\hat{x}_i^{\min} = 0$ for all $i \notin \mathcal{S}$, conditioning on $\mathcal{E}_A \cap \mathcal{E}_B$,

$$F_n(\mathbf{x}^*) - F_n(\hat{\mathbf{x}}^{\min}) \leq \sum_{i=1}^p P_\lambda'\left(x_i^*\right)\left(\hat{x}_i^{\min} - x_i^*\right) \leq \sum_{i=1}^p P_\lambda'\left(x_i^*\right) \hat{x}_i^{\min}$$

$$= \sum_{i \in \mathcal{S}} P_\lambda'(x_i^*) \hat{x}_i^{\min} + \sum_{i \notin \mathcal{S}} P_\lambda'(x_i^*) \hat{x}_i^{\min} = 0, \quad a.s.$$

The above inequality yields that $F(\mathbf{x}^*) - F(\hat{\mathbf{x}}^{\min}) \leq \epsilon$ almost surely conditioning on $\mathcal{E}_A \cap \mathcal{E}_B$.

Now, to achieve the desired result of the theorem, it suffices to show that, if n satisfies

$$n \geq c_7 \cdot \sigma^2 |\mathcal{S}| \cdot \left(\frac{1 + L_{\mu,s} R}{a^2 \epsilon^2} \ln \frac{24 R L_\mu p}{\epsilon} \right) \bigvee c_7 \cdot N^*(c_7) \tag{63}$$

for some universal constant $c_7 > 0$, then the event $\mathcal{E}_A \cap \mathcal{E}_B$ occurs with probability at least $1 - 2\alpha$. To this end, notice that $\hat{\varepsilon} = 0$. We may use the same argument as in the proof for Proposition 1 in Sect. 4.2.3 in showing (52) and obtain that $\mathbf{P}[\mathcal{E}_A \cap \mathcal{E}_B] \geq 1 - \alpha$ if

$$n \geq \hat{c}_7 \cdot \sigma^2 |\mathcal{S}|^{1/2} \cdot \left(\frac{1 + \Gamma}{a^2 \epsilon^2} \ln \frac{24 R L_\mu p}{\epsilon} \right) \bigvee \hat{c}_7 \cdot N^*(\hat{c}_7) \tag{64}$$

for some universal constant $\hat{c}_7 > 0$.

Recall the assumption that $F_{n,\lambda}(\mathbf{x}^*) \leq F_{n,\lambda}(\mathbf{0})$ almost surely. Since $F_{n,\lambda}(\mathbf{0}) \leq F_{n,\lambda}(\hat{\mathbf{x}}^{\min}) + 2L_{\mu,s} R \sqrt{|\mathcal{S}|}$ with lower-bounded probability $1 - 2\exp(-\frac{nL_{\mu,s}^2}{2\sigma_L^2})$ (to see this, we can repeat the steps in showing (57) in Sect. 4.2.5), we may let $\Gamma = 2L_{\mu,s} R \sqrt{|\mathcal{S}|}$. It is then easily verifiable from (64) that there exists such a problem-independent constant $c_7 > 0$ such that, if (63) holds, then $2\exp(-\frac{nL_{\mu,s}^2}{2\sigma_L^2}) \leq \alpha$ and the desired result holds. □

5 Some discussions on solution schemes for RSAA

This section discusses two classes of solution techniques to ensure the desired S^{3}ONC solutions: local schemes (in Sect. 5.1) and a global technique (in Sect. 5.2).

5.1 Local optimization for RSAA

The S^{3}ONC is weaker than the second-order KKT condition. Therefore, any algorithm that guarantees the second-order KKT condition can satisfy the stipulations made by Part 2 of Proposition 1 and those by Proposition 2. Furthermore, among those algorithms, any descent algorithm that guarantees the second-order KKT condition can ensure the conditions as in Theorems 2–4, if initialized with an all-zero solution.

Algorithms that ensure the second-order KKT condition have been discussed by much literature. For instance, [4,7,19,25,26] provide algorithms with different convergence and complexity results. In particular, one of these algorithms, the interior point algorithm (IPA) presented by [4], is a descent, and fully polynomial-time approximation scheme (FPTAS) for a local solution that satisfies the desired second-order necessary condition, when $\mathcal{X}$ consists of a set of box constraints. In the special case where (6) is a quadratic program, [26] proposes a potential reduction (PR) algorithm and shows its convergence to a second-order KKT solution.

To facilitate the solution schemes we may reformulate the objective function into a twice continuously differentiable function. Specifically, according to [15], we have the following equivalence

$$P_\lambda(x) = \min_{\eta\in[0,a\lambda]} \frac{1}{2a}\eta^2 - \frac{1}{a}\eta x + \lambda x,$$

for which the optimizer admits a closed form:

$$\eta^{\min}(x) := \begin{cases} x & \text{if } 0 \le x \le a\lambda; \\ a\lambda & \text{if } x > a\lambda. \end{cases} \tag{65}$$

Therefore, we have the equivalence between the original regularized problem $\min_{\mathbf{x}\in\mathcal{X}} F_n(\mathbf{x}) + \sum_{i=1}^p P_\lambda(x_i)$ and an optimization problem with additional dummy variables:

$$\min_{\mathbf{x}\in\mathcal{X},\eta=(\eta_i)\in[0,a\lambda]^p} G_n(\mathbf{x}) := F_n(\mathbf{x}) + \sum_{i=1}^p \left(\frac{1}{2a}\eta_i^2 - \frac{1}{a}\eta_i x_i + \lambda x_i\right) \tag{66}$$

where η is the vector of dummy variables. Notice that Problem (66) is convex in η.

One can show that the second-order KKT condition to the reformulated program (66) implies the S^{3}ONC of (6). To see this, observe that, at a second-order KKT point $(\mathbf{x}^*, \eta^*)$ the first-order KKT condition also holds. Due to the convexity of (66) in η, it holds that $\eta^* = \eta^{\min}(x^*)$. Also by the definition of the second-order KKT condition, we know that

$$d^\top \begin{bmatrix} \nabla^2 F_n(x^*) & -\frac{1}{a}I \\ -\frac{1}{a}I & \frac{1}{a}I \end{bmatrix} d \ge 0, \text{ for all } d \text{ in the critical set.} \tag{67}$$

To check if S^{3}ONC is satisfied, we only need to consider the case where $x_i \in (0, \min\{1, a\lambda\})$. According to (65), it holds that $\eta_i^* \in (0, \min\{1, a\lambda\})$. As an immediate result, (67) implies that the submatrix $\begin{bmatrix} \frac{\partial^2 F_n(x^*)}{\partial x_i^2} & -1/a \\ -1/a & 1/a \end{bmatrix}$ is positive semi-definite. Invoking Schur complement condition, it obtains that $0 \le \frac{\partial^2 F_n(x^*)}{\partial x_i^2} - \frac{1}{a} = \left.\frac{\partial^2[F_n(\mathbf{x})+\sum_{i=1}^p P_\lambda(x_i)]}{(\partial x_i)^2}\right|_{\mathbf{x}=\mathbf{x}^*}$, where the last identity is immediate from the definition of P_λ for $x_i \in (0, \min\{1, a\lambda\})$. By its definition, the S^{3}ONC holds.

The reformulated problem (66) then satisfies all the assumptions for some existing FPTASs that guarantee a second-order KKT point, such as the interior point method by [4].

5.2 Global optimization for RSAA

The global minimizer is a local minimizer, and, thus, also satisfies the S^{3}ONC. To compute this solution, the RSAA formulation can be equivalently formulated as a mixed integer program. Let Assumption A.3 hold and $a\lambda \leq 1$. This inequality is not restrictive as a and λ are user-specified parameters for P_λ. Then, as per (27), one can immediately rewrite the RSAA formulation into the following

$$\begin{aligned} \min \;\; & F_n(\mathbf{x}) + P_\lambda(a\lambda) \cdot \left(\mathbf{1}^\top \mathbf{z}_1 + \mathbf{1}^\top \mathbf{z}_2\right) \\ s.t. \;\; & \mathbf{x} \geq a\lambda \cdot \mathbf{z}_2 - \mathcal{M}\mathbf{z}_1; \quad \mathbf{x} \leq \mathcal{M} \cdot \mathbf{z}_2 \\ & -\mathbf{x} \geq a\lambda \cdot \mathbf{z}_1 - \mathcal{M}\mathbf{z}_2; \quad \mathbf{x} \geq -\mathcal{M} \cdot \mathbf{z}_1 \\ & \mathbf{x} \in \mathcal{X}; \quad \mathbf{z}_1, \mathbf{z}_2 \in \{0,\ 1\}^p. \end{aligned}$$

where $\mathcal{M}$ is a big-M and can be any scaler greater than $R + a\lambda$ in our case and where $P_\lambda(a\lambda) = \frac{a\lambda^2}{2}$. In particular, if Assumption A.5 holds, F_n is convex almost surely and the above formulation falls into the category of mixed integer convex programming, which admits numerical solvers to ensure global optimality. Liu et al. [15] presents MILP reformulations when F_n is a quadratic but not necessarily convex function.

6 Preliminary numerical results

This section presents a preliminary set of numerical experiments following similar setups with [12,16]. Specifically, we consider the following SP problem

$$\min\{\mathbb{E}[(\boldsymbol{\varrho}\mathbf{x} - \beta)^2] : \mathbf{x} \in [0,\ 5]^p\}, \tag{68}$$

where the relationship between $\boldsymbol{\varrho}$ and β is governed by $\beta = \boldsymbol{\varrho}\mathbf{x}^{\min} + \omega$ with $\mathbf{x}^{\min} = [3;\ 1.5;\ 0;\ 0;\ 2;\ \mathbf{0}_{p-5}]$. Let the ω be a standard normally distributed random variable; that is $\omega \sim \mathcal{N}(0,\ 1)$. Also assume that $\boldsymbol{\varrho} \sim \mathcal{N}_p(0, \Sigma)$, which is a p-variate normally distributed random variable with covariance matrix defined by $\Sigma = (\varsigma_{ij}) \in \Re^{p\times p}$ and $\varsigma_{ij} = 0.5^{|i-j|}$. It is easily verifiable that the optimal solution to the SP problem in (68) is $\mathbf{x}^{\min}$.

We compare the following approaches to solving (68) in problems with different choices of sample sizes and dimensions:

SAA: A global minimal solution to SAA in (3) computed using Mosek.

RSAA-local: An S^{3}ONC solution to RSAA in (6) generated by the PR algorithm as discussed in Sect. 5.1. The PR is initialized with an (approximate) all-zero solution. Our theories in Sect. 3 have predicted that such a local solution can approximate (68) globally.

RSAA-global: A global solution to RSAA in (6) solved with Mosek through the reformulation given in Sect. 5.2.

All experiments are conducted in Matlab on a computer with 2.2 GHz Intel Core i7 processor and 16 GB memory. Mosek is invoked via Matlab to generate solutions for SAA and RSAA-global. For both RSAA-local and RSAA-global, the parameters for FCP are fixed as $\lambda = 0.5$ and $a = 0.9$. We would also like to remark that, since the PR algorithm requires the starting point to be an interior point, we approximate the all-zero solution by $10^{-4} \cdot \mathbf{1}$ for the PR's initialization.

For every $(n,\ p)$ combination, we replicate each solution scheme five times with independently generated samples for each repetition. We report the average, maximal, and minimal suboptimality gaps as measured by $F(\cdot) - F(\mathbf{x}^{\min})$ in Tables 2 and 5. In Table 2, we fix the number of samples $n = 100$ and gradually increase p from 10 to 1500. From this table, we can observe a clear trend that the solution quality of SAA deteriorates dramatically. In contrast, the suboptimality gaps are well contained by the proposed RSAA, even if the RSAA is only solved locally (as shown in the "RSAA-local" column). When $p = 1400$, RSAA-global is noticeably better than RSAA-local, as the former has a smaller maximal suboptimality gap than the latter. Nonetheless, the two different types of solutions yield almost the same quality in approximating (68). Note that our theories in fact provide a sharper performance bound for RSAA-global than RSAA-local. Therefore, the closely similar numerical performance between RSAA-global and RSAA-local is an indication that our bounds for RSAA-local may not be tight enough for at least the special case in the numerical experiments.

Figure 1 shows the dependence between the suboptimality gap and p. Particularly, in Fig. 1a, the suboptimality gaps of SAA increase faster than linearly in p. In contrast, the suboptimality gaps for both RSAA-local and RSAA-global increase very slowly when p grows, as shown in both Fig. 1a, b.

Table 3 shows the sparsity of the solutions generated by the three different schemes. We can see from this table that SAA generates dense solutions in all the test instances, while both RSAA-local and RSAA-global can maintain sparsity in the output solutions.

Table 4 reports the computational time of the three different approaches. We notice that SAA is the most efficient among the three. RSAA-local incurs a noticeable increase in the computational efforts than SAA. Nonetheless, considering the substantial improvement generated by the RSAA-local in solution quality, we argue that the additional amount of computational cost is reasonable. RSAA-global is significantly slower than RSAA-local, even though the two have almost the same solution quality in our experiments.

We further compare the three approaches in problems that have different sample sizes n and a fixed number of dimensions $p = 100$. The comparison is presented in Table 5 and Fig. 2. By comparison, we see that the solution quality of both RSAA-local and RSAA-global increase rapidly with the growth of n. Their rates are significantly faster than SAA.

In summary, our numerical results verify our theoretical predictions that the RSAA is particularly effective when n is much smaller than than p. In such a case, RSAA may significantly improve solution quality over SAA.

Table 2 Comparison in solution quality measured by the suboptimality gaps for problems with different numbers of dimensions p and a fixed sample size $n = 100$

p	SAA			RSAA-local			RSAA-global		
	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min
10	0.13	0.22	4.79	0.04	0.11	0.00	0.04	0.11	0.00
30	0.466	0.617	0.31	0.04	0.06	0.02	0.04	0.06	0.02
50	1.05	1.25	0.76	0.05	0.09	0.00	0.05	0.09	0.00
70	2.42	4.09	1.55	0.03	0.05	0.01	0.03	0.05	0.01
90	11.8	17.4	8.91	0.04	0.06	0.02	0.04	0.06	0.02
200	366.56	488.31	279.27	0.02	0.06	0.01	0.02	0.06	0.01
300	1.25e3	1.57e3	1.04e3	0.02	0.04	0.00	0.02	0.04	0.00
400	2.48e3	2.74e3	2.18e3	0.03	0.07	0.01	0.03	0.07	0.01
500	3.40e3	3.75e3	3.00e3	0.03	0.06	0.00	0.03	0.06	0.00
600	4.89e3	5.18e3	4.35e3	0.02	0.04	0.01	0.02	0.04	0.01
700	6.21e3	6.41e3	5.75e3	0.02	0.04	0.00	0.02	0.04	0.00
800	7.96e3	8.54e3	7.34e3	0.02	0.03	0.01	0.02	0.03	0.01
900	9.92e3	1.06e4	9.44e3	0.04	0.10	0.01	0.04	0.10	0.01
1000	1.17e4	1.31e4	1.04e4	0.03	0.08	0.01	0.03	0.08	0.01
1100	1.32e4	1.43e4	1.19e4	0.03	0.08	0.01	0.03	0.08	0.01
1200	1.51e4	1.58e4	1.35e4	0.04	0.09	0.01	0.04	0.09	0.01
1300	1.73e4	1.85e4	1.59e4	0.01	0.03	0.00	0.01	0.03	0.00
1400	1.88e4	1.97e4	1.81e4	0.07	0.15	0.03	0.07	0.14	0.03
1500	2.18e4	2.34e4	2.10e4	0.03	0.08	0.01	0.03	0.08	0.01

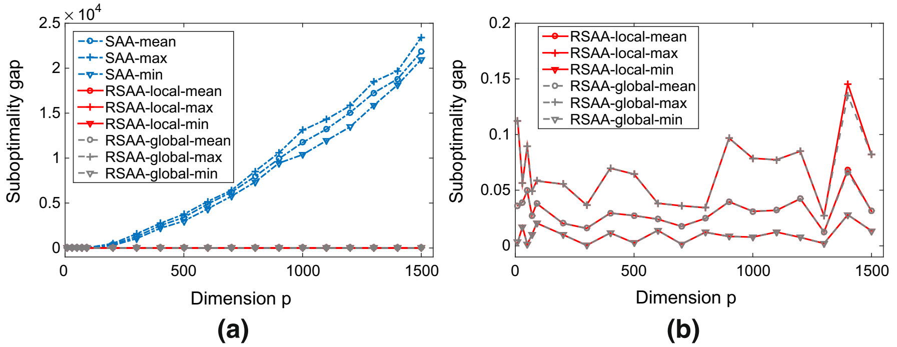

Fig. 1 Comparison of suboptimality gaps of solutions generated by SAA, local optimization of RSAA, and global optimization of RSAA when $n = 100$ and p increases from 10 to 1500. “SAA-mean”, “SAA-max”, and “SAA-min” are the average, maximal, and minimal suboptimality gaps of SAA out of the five replications, “RSAA-local-mean”, “RSAA-local-max”, and “RSAA-local-min” are the average, maximal, and minimal suboptimality gaps of RSAA-local, “RSAA-global-mean”, “RSAA-global-max”, and “RSAA-global-min” are the average, maximal, and minimal suboptimality gaps of RSAA-global

Table 3 The numbers of nonzeros in the solutions generated by SAA, RSAA-local, and RSAA-global, when $n = 100$

p	SAA			RSAA-local			RSAA-global		
	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min
10	10	10	10	3	3	3	3	3	3
30	30	30	30	3	3	3	3	3	3
50	50	50	50	3	3	3	3	3	3
70	70	70	70	3	3	3	3	3	3
90	90	90	90	3	3	3	3	3	3
200	200	200	200	3	3	3	3	3	3
300	300	300	300	3	3	3	3	3	3
400	400	400	400	3	3	3	3	3	3
500	500	500	500	3	3	3	3	3	3
600	600	600	600	3	3	3	3	3	3
700	700	700	700	3	3	3	3	3	3
800	800	800	800	3	3	3	3	3	3
900	900	900	900	3	3	3	3	3	3
1000	1000	1000	1000	3	3	3	3	3	3
1100	1100	1100	1100	3	3	3	3	3	3
1200	1200	1200	1200	3	3	3	3	3	3
1300	1300	1300	1300	3	3	3	3	3	3
1400	1400	1400	1400	3.8	6	3	3	3	3
1500	1500	1500	1500	3	3	3	3	3	3

Table 4 Comparison of the average computational time out of the five replications for problems with different dimensionality p and fixed sample size $n = 100$

p	SAA (s)	RSAA-local (s)	RSAA-global (s)	p	SAA (s)	RSAA-local (s)	RSAA-global (s)
10	3.19	1.71	9.77	700	3.42	20.92	241.68
30	3.21	4.08	13.22	800	3.38	34.13	1220.89
50	3.20	3.86	17.31	900	3.42	40.34	1425.75
70	3.17	4.46	30.28	1000	3.42	34.59	2693.44
90	3.13	8.55	27.31	1100	3.38	33.50	4014.09
200	3.06	19.03	7.21	1200	3.66	37.62	3686.88
300	3.13	15.82	45.60	1300	3.89	39.30	11658.30
400	3.35	14.02	157.64	1400	3.38	54.65	16927.54
500	3.33	19.34	134.08	1500	3.37	63.68	13463.53
600	3.40	20.92	240.10				

Table 5 Comparison in solution quality measured by the suboptimality gaps for problems with different sample sizes n and a fixed number of dimensions $p = 100$

n	SAA			RSAA-local			RSAA-global		
	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min
15	608.79	746.59	470.60	3.80	14.39	0.03	0.38	1.17	0.03
20	536.58	660.87	423.64	0.69	2.25	0.03	0.22	0.70	0.03
25	540.28	746.75	403.39	0.57	1.04	0.23	0.37	0.85	0.09
30	422.14	523.62	331.26	0.31	0.55	0.13	0.26	0.35	0.13
35	387.38	472.50	265.12	0.12	0.21	0.06	0.12	0.21	0.06
40	261.00	323.83	176.91	0.09	0.15	0.01	0.09	0.15	0.01
45	268.50	343.60	141.38	0.10	0.31	0.01	0.05	0.08	0.01
50	149.85	188.51	112.81	0.08	0.20	0.02	0.08	0.20	0.02
55	122.59	172.12	96.07	0.06	0.15	0.01	0.06	0.15	0.01
60	142.53	159.97	110.20	0.03	0.05	0.02	0.03	0.05	0.02
65	122.31	130.33	110.29	0.04	0.07	0.01	0.04	0.07	0.01
70	69.64	92.05	32.02	0.05	0.13	0.01	0.05	0.13	0.01
75	80.03	127.81	45.62	0.07	0.11	0.02	0.07	0.11	0.02
80	42.01	53.67	29.14	0.04	0.07	0.02	0.04	0.07	0.02
85	46.52	84.56	31.37	0.07	0.16	0.02	0.07	0.16	0.02
90	24.21	36.26	14.04	0.03	0.09	0.01	0.03	0.09	0.01
95	32.96	48.93	8.22	0.03	0.07	0.00	0.03	0.07	0.00
100	116.52	201.05	42.98	0.02	0.03	0.01	0.02	0.03	0.01
105	17.20	19.94	13.04	0.03	0.06	0.01	0.03	0.06	0.01
110	10.48	13.88	6.41	0.02	0.06	0.01	0.02	0.06	0.01

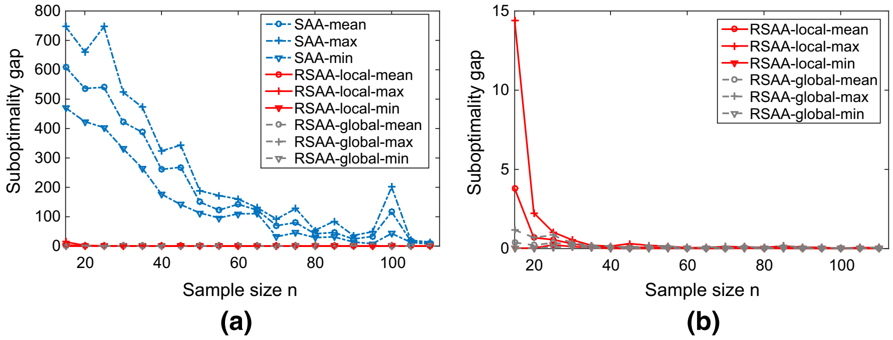

Fig. 2 Comparison of suboptimality gaps of solutions generated by SAA, local optimization of RSAA, and global optimization of RSAA when $p = 100$ and n increases from 15 to 110. "SAA-mean", "SAA-max", and "SAA-min" are the average, maximal, and minimal suboptimality gaps of SAA out of the five replications, "RSAA-local-mean", "RSAA-local-max", and "RSAA-local-min" are the average, maximal, and minimal suboptimality gaps of RSAA-local, "RSAA-global-mean", "RSAA-global-max", and "RSAA-global-min" are the average, maximal, and minimal suboptimality gaps of RSAA-global

7 Conclusion

This paper proposes the RSAA, a modification to the SAA by incorporating a regularization scheme called the FCP. This modification targets the high-dimensional SP problems with sparsity. We show that when the solution is sparse or can be approximated by a sparse solution, the regularization can significantly reduce the required number of samples in some high-dimensional SP applications: Compared to the conventional SAA approach that requires the sample size to grow polynomially in the number of dimensions, the RSAA stipulates number of samples that is only polylogarithmic in the dimensionality.

Although the incorporation of FCP renders the RSAA formulation nonconvex, we argue that any S^{3}ONC solution achieved by a decent algorithm starting at the all-zero vector is good enough to ensure the optimization performance of the local solution. The S^{3}ONC is a necessary condition (for local minimality) weaker than the second-order KKT condition. Numerical algorithms to ensure the second-order KKT condition are known from the literature. Furthermore, under some conditions on the feasible region, the S^{3}ONC solutions admit an FPTAS. We also discuss a mixed integer convex reformulation to the RSAA formulation that allows for exact, though exponential-time in the worst case, computation of the global solution. Our preliminary numerical experiments have verified our theoretical predictions.

A limitation of the current development is the assumption of coordinate-wise constraints. We would like to relax such an assumption in our future research.

Acknowledgements The authors are indebted to the referees and the associate editor for their valuable comments, which have significantly improved the paper. This research was partially supported by National Science Foundation Grant Nos. CMMI 1300638 and DMS 1512422, National Institute on Drug Abuse (NIDA) Grant Nos. P50 DA039838 and P50 DA036107, National Library of Medicine, T32 LM012415, National Institute of Allergy and Infectious Diseases, U19AI089672, Institute of CyberScience Seed Grant, Grace Woodward Grants for Collaborative Research in Engineering and Medicine, the Pennsylvania State University. This work was also partially supported by National Natural Science Foundation of China Grant Nos. 71528005 and 11690015. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF, the NIDA, the NIH, or the NNSFC. Part of the research was done when Liu was working as a postdoctoral research fellow at the Department of Radiation Oncology, Stanford University, Stanford, CA 94305, USA.

References

1. Agarwal, A., Negahban, S., Wainwright, M.J.: Stochastic optimization and sparsity statistical recovery: optimal algorithms for high dimensions. In: Advances in Neural Information Processing Systems, pp. 1538–1546 (2012)
2. Bach, F., Moulines, E.: Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In: Advances in Neural Information Processing Systems, pp. 451–459 (2011)
3. Bartlett, P.L., Jordan, M.I., McAuliffe, J.D.: Convexity, classification, and risk bounds. J. Am. Stat. Assoc. **101**, 138–156 (2006)
4. Bian, W., Chen, X., Ye, Y.: Complexity analysis of interior point algorithms for non-Lipschitz and non-convex minimization. Math. Prog. A **149**, 301–327 (2015)
5. Bickel, P.J., Ritov, Y., Tsybakov, A.B.: Simultaneous analysis of Lasso and Dantzig selector. Ann. Stat. **37**, 1705–1732 (2009)
6. Candés, E., Tao, T.: The Dantzig selector: statistical estimation when p is much larger than n. Ann. Stat. **35**, 2313–2351 (2007)

7. Cartis, C., Gould, N.I.M., Toint, P.I.: Adaptive cubic regularization methods for unconstrained optimization. Part I: motivation, convergence and numerical results. Math. Prog. A **127**, 245–295 (2011)
8. Chen, X., Ge, D., Wang, Z., Ye, Y.: Complexity of unconstrained L_2-L_p minimization. Math. Prog. A **143**, 371–383 (2014)
9. Dupačová, J., Wets, R.: Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems. Ann. Stat. **16**, 1517–1549 (1988)
10. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**, 1348–1360 (2001)
11. Fan, J., Lv, J.: Nonconcave penalized likelihood with NP-dimensionality. IEEE Trans. Inform. Theory **57**, 5467–5484 (2011)
12. Fan, J., Xue, L., Zou, H.: Strong oracle optimality of folded concave penalized estimation. Ann. Stat. **42**, 819–849 (2014)
13. Ge, D., Wang, Z., Ye, Y., Yin, H.: Strong NP-hardness result for regularized L_q-minimization problems with concave penalty functions. Cornell University Library (2015). arXiv:1501.00622v1
14. Kleywegt, A.J., Shapiro, A., Homem-de-Mello, T.: The sample average approximation method for stochastic discrete optimization. SIAM J. Optim. **12**, 479–502 (2001)
15. Liu, H., Yao, T., Li, R.: Global solutions for folded concave penalized non-convex learning. Ann. Stat. **44**, 629–659 (2016)
16. Liu, H., Yao, T., Li, R., Ye, Y.: Folded concave penalized sparse linear regression: sparsity, statistical performance, and algorithmic theory. Math. Program. (2017). https://doi.org/10.1007/s10107-017-1114-y
17. Loh, P.-L., Wainwright, M.J.: Regularized M-estimators with nonconvexity: statistical and algorithmic theory for local optima. J. Mach. Learn. Res. **16**, 559–616 (2015)
18. Negahban, S.N., Ravikumar, P., Wainwright, M.J., Yu, B.: A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. Stat. Sci. **27**, 538–557 (2012)
19. Nesterov, Y., Polyak, B.T.: Cubic regularization of Newton's method and its global performance. Math. Program. **108**, 177–205 (2006)
20. Shapiro, A.: Monte Carlo sampling methods. In: Ruszczynski, A., Shapiro, A. (eds.) Stochastic Programming, Handbook in OR and MS, vol. 10. North-Holland Publishing Company, Amsterdam (2003)
21. Shapiro, A., Dentcheva, D., Ruszczyński, A.: Lectures on Stochastic Programming Modeling and Theory. The Society for Industrial and Applied Mathematics and the Mathematical Programming Society, Philadelphia (2009)
22. Shapiro, A., Xu, H.: Stochastic mathematical programs with equilibrium constraints, modelling and sample average approximation. Optimization **57**(3), 395–418 (2008)
23. Wang, L., Kim, Y., Li, R.: Calibrating non-convex penalized regressioni in ultra-high dimension. Ann. Stat. **41**, 2505–2536 (2013)
24. Wang, Z., Liu, H., Zhang, T.: Optimal computational and statistical rates of convergence for sparse non-convex learning problems. Ann. Stat. **42**, 2164–2201 (2014)
25. Ye, Y.: On affine scaling algorithms for non-convex quadratic programming. Math. Prog. **56**, 285–300 (1992)
26. Ye, Y.: On the complexity of approximating a KKT point of quadratic programming. Math. Prog. **80**, 195–211 (1998)
27. Zhang, C.: Nearly unbiased variable selection under minimax concave penalty. Ann. Stat. **28**, 894–942 (2010)
28. Zhang, C.-H., Huang, J.: The sparsity and bias of the Lasso selection in high-dimensional linear regression. Ann. Stat. **36**, 1567–1594 (2008)
29. Zhang, C., Zhang, T.: A general theory of concave regularization for high dimensional sparse estimation problems. Stat. Sci. **27**, 576–593 (2012)