

NEW PENALIZED STOCHASTIC GRADIENT METHODS FOR LINEARLY CONSTRAINED STRONGLY CONVEX OPTIMIZATION

MENG LI*, PAUL GRIGAS*, AND ALPER ATAMTÜRK*

Abstract. For minimizing a strongly convex objective function subject to linear inequality constraints, we consider a penalty approach that allows one to utilize stochastic methods for problems with a large number of constraints and/or objective function terms. We provide upper bounds on the distance between the solutions to the original constrained problem and the penalty reformulations, guaranteeing the convergence of the proposed approach. We give a nested accelerated stochastic gradient method and propose a novel way for updating the smoothness parameter of the penalty function and the step-size. The proposed algorithm requires at most $\tilde{O}(1/\sqrt{\epsilon})$ expected stochastic gradient iterations to produce a solution within an expected distance of ϵ to the optimal solution of the original problem, which is the best complexity for this problem class to the best of our knowledge. We also show how to query an approximate dual solution after stochastically solving the penalty reformulations, leading to results on the convergence of the duality gap. Moreover, the nested structure of the algorithm and upper bounds on the distance to the optimal solutions allows one to safely eliminate constraints that are inactive at an optimal solution throughout the algorithm, which leads to improved complexity results. Finally, we present computational results that demonstrate the effectiveness and robustness of our algorithm.

Key words. Convex optimization, linear constraints, penalty method, stochastic gradient, duality gap

AMS subject classifications. 90C30, 90C25, 65K05

1. Introduction. Consider the convex optimization problem

$$(1.1) \quad \begin{aligned} \min_{x \in \mathbb{R}^n} \quad & F(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} f_i(x) + \psi(x) \\ \text{s.t.} \quad & x \in X_i, \quad i = 1, \dots, m, \end{aligned}$$

where $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$, $i = 1, \dots, \ell$, are convex and smooth functions, $\psi : \mathbb{R}^n \rightarrow \mathbb{R}$ is a convex function for which one can build a proximal mapping, and $X_i \subseteq \mathbb{R}^n$ is a nonempty closed convex set for $i = 1, \dots, m$. Problems of the form (1.1) arise in many contexts, including predictive control [10], portfolio optimization [28], and others. In machine learning, some examples include isotonic regression [7], convex regression [37, 25], and strong convex relaxations of sparse combinatorial problems such as signal estimation [5] and sparse regression [2, 3, 21]. We are particularly interested in problems where the number of objective terms ℓ and/or the number of constraints m are very large.

In this paper, we consider the case where the objective function F is μ -strongly convex for $\mu > 0$, and the feasible region is defined by a set of linear inequalities, i.e., $X_i := \{x \in \mathbb{R}^n : a_i^T x \leq b_i\}$, $i = 1, \dots, m$. Let $A = (a_1, \dots, a_m)^T \in \mathbb{R}^{m \times n}$ and $b = (b_1, \dots, b_m)^T \in \mathbb{R}^m$. Then, our problem of interest is

$$(1.2) \quad \begin{aligned} \min_{x \in \mathbb{R}^n} \quad & F(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} f_i(x) + \psi(x) \\ \text{s.t.} \quad & a_i^T x \leq b_i, \quad i = 1, \dots, m. \end{aligned}$$

*Department of Industrial Engineering and Operations Research, University of California-Berkeley, Berkeley, CA (meng_li@berkeley.edu, pgrigas@berkeley.edu, atamturk@berkeley.edu).

Our objective herein is to devise a penalty-based stochastic first-order approach for solving problem (1.2). The main advantage of the penalty function approach is that the resulting penalized reformulation is an unconstrained, smooth, convex optimization problem whose objective involves a finite sum over the components of objectives and the penalty functions. Such problems with a finite sum structure are particularly amenable to stochastic (proximal) gradient methods (see, e.g., [11] and the references therein) and variants that are able to exploit the finite sum structure to achieve faster convergence such as the stochastic variance reduced gradient method (SVRG) [22], the stochastic average gradient methods (SAG, SAGA) [36, 15], and related methods.

In this work, we construct a penalty reformulation of (1.2) based on using the *softplus* penalty function, which arises from a smoothing [32] of the $\max(0, \cdot)$ function. We construct novel estimates of the 2-norm distance between solutions of the penalty reformulation and the original constrained problem by analyzing the smooth solution trajectories and results on C^∞ approximations of convex (and strongly convex) functions due to Azagra [6]. Then, based on these estimation results, we analyze the complexity of solving the penalty reformulations with various methods, most notably the accelerated stochastic methods that exploit the finite-sum structure. Moreover, using the structure of the proposed nested penalty method and our novel bounds, we are able to obtain approximate dual solutions and we can safely eliminate inactive constraints throughout the course of the algorithm.

There are several existing approaches for solving problems of the form (1.1). Classical approaches include interior point and projected gradient methods, as well as the augmented Lagrangian [9] and ADMM methods [12]. These approaches do not scale well when ℓ and m are very large, either because they require projecting onto or otherwise directly working with the intersection of all of the constraint sets or because they require calculating the gradient of a function that is the sum of many components. One approach for addressing the case where m is very large is the gradient descent method with projections onto randomly sampled constraint sets X_i [30, 39].

Penalty methods have been considered in several previous works on similar problems. Nedich and Tatarenko [38] consider problem (1.2) (without finite-sum structure in the objective), use the one-sided Huber penalty to construct the penalty reformulations, and apply SAGA to solve the unconstrained penalty reformulation. An unfortunate deficiency of this approach is that the smoothness parameter of the penalty function as well as the weight parameter on the penalty function must satisfy a complicated set of relations in order to apply a linear convergence rate inherited by the SAGA algorithm. In a follow up paper [31], they present an incremental gradient method that dynamically updates the parameters and achieves an asymptotic convergence rate of $O(1/\sqrt{k})$, where k is the iteration counter and each iteration involves working with a single constraint. Instead, we use the softplus function as our smooth approximation, which leads to improved estimation results (Theorem 2.4). This improved estimation result, as well as our refined analysis, ultimately leads to an improved $\tilde{O}(\frac{1}{\sqrt{\epsilon}})$ complexity bound in terms of the expected number of incremental steps (i.e., stochastic gradient and constraint evaluations) required to find a solution within a 2-norm distance of ϵ to the optimal solution.

Lan and Monteriro [23] consider convex problems with conic constraints $Ax - b \in \mathcal{K}^*$, where \mathcal{K}^* is the dual cone of some closed convex cone \mathcal{K} , and use the quadratic penalty $[d_{\mathcal{K}^*}(Ax - b)]^2 = \min_{u \in \mathcal{K}^*} \|u + b - Ax\|_2^2$ to relax the constraints. Compared

with their setting and methods, (i) we consider the case when ℓ and m are large and focus on the complexity with respect to incremental steps (i.e., calls of f_i , a_i , and gradients thereof) and use stochastic methods; (ii) we use the softplus function, which is close to $\max(a_i^T x - b_i, 0)$ instead of a quadratic function to penalize the constraints. Fercoq et al. [16] consider the objective to be an expectation of random smooth convex functions and the random constraints to be held almost surely, which can be thought as (1.2) when m tends to infinity, and use a similar quadratic penalty as in [23]. In addition to using a different penalty function, we apply the catalyst SVRG (or SAG, SAGA) method [26] which exploits the finite sum structure of the penalty reformulation to obtain a $\tilde{O}(1/\sqrt{\epsilon})$ expected complexity bound in terms of distance to the optimal solution. On the other hand, Fercoq et al. [16] demonstrate a $\tilde{O}(1/\epsilon)$ bound in terms of the objective function gap and the constraint violations (for the strongly convex case). Using the sum of squared distance functions $\frac{1}{2m} \sum_{i=1}^m d_{\mathcal{X}_i}(x)^2$ to penalize the constraints, Mishchenko and Richtárik [29] obtain $\tilde{O}(\frac{1}{\epsilon})$ complexity, and their algorithm and results cover nonlinear constraints as well. Compared with their result, we have a better convergence rate in the linear case, and we do not require a global Hoffman-type assumption.

Contributions. This paper has three main contributions. First, we propose to use the softplus function to penalize the constraints, and we show that this penalty function reformulation also arises by applying Nesterov’s smoothing technique [32] to the Lagrangian of the original problem. The penalty reformulation has a finite sum structure that allows one to employ stochastic gradient methods, and further, the relation with the Lagrangian enables us to query an approximately optimal dual solution after obtaining an approximately optimal primal solution. Second, we use ordinary differential equations techniques to estimate the 2-norm distance between solutions to the penalty reformulations and the solution to (1.2), which leads to novel and stronger estimation bounds as compared to existing approaches. Third, we analyze the complexity of accelerated stochastic methods to obtain approximate primal and dual solutions with our penalty method and compare it with the existing lower bounds on such problems. In addition, we propose to use a nested algorithm to solve the penalty reformulations, which ensures limiting convergence can help to determine the parameters during implementations. Moreover, based on the nested structure and upper bounds on the distance to the optimal solution, we design a screening procedure to effectively eliminate constraints that are inactive at the optimal solution throughout the algorithm, which leads to improved complexity results. We present the results of numerical experiments performed on quadratic programming problem and a SVM problems, to demonstrate the effectiveness of the proposed algorithm as compared to existing penalty function approaches.

Organization. This paper is organized as follows. In section 2, we introduce the penalty reformulation for the constrained problem (1.2) and study its properties. Our nested algorithm based on the penalty reformulation and its complexity analysis is presented in section 3. In section 4, we show how to query dual solutions and present results on the convergence of the duality gap. In section 5, we show how to incorporate the screening procedure to eliminate inactive constraints into our algorithm and improve the complexity. We present our numerical results to demonstrate the effectiveness of our proposed algorithm in section 6. We conclude with a few final remarks in section 7.

2. Penalty Function Reformulation and Key Properties. In this section, we examine several useful theoretical properties of a smooth penalized reformulation of problem (1.2). To penalize the constraints, we use the *softplus* function,

$$(2.1) \quad p_\delta(t) := \delta \log(1 + \exp(t/\delta)), \quad (\delta, t) \in [0, +\infty) \times \mathbb{R},$$

where $\delta \geq 0$ is a parameter controlling the smoothness of p_δ . Function p_δ is used as a loss function in many contexts, including, logistic regression. The softplus penalty may be viewed as an instance of Nesterov's smoothing technique [32] applied to the hinge function $\max(0, t)$ since it holds that

$$(2.2) \quad p_\delta(t) = \max_{r \in [0, 1]} \{tr - \delta[r \log(r) + (1 - r) \log(1 - r)]\}.$$

The above conjugate representation will be useful in studying the Lagrangian of (1.2) and developing duality gap results in section 4. Figure 1 shows the softplus function

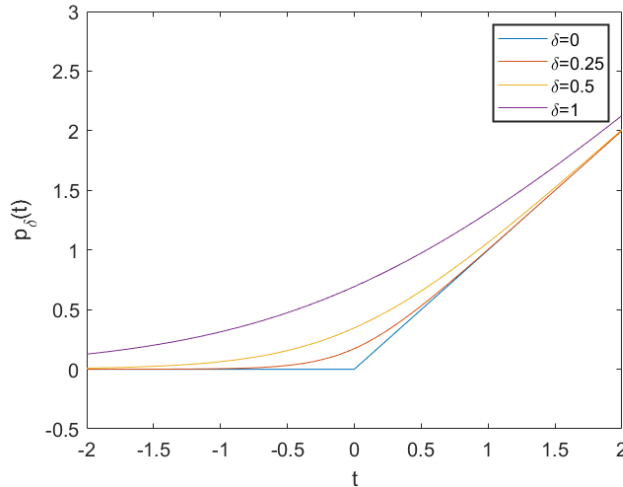


FIG. 1. *Softplus* function p_δ for different values of δ .

p_δ for different values of δ and highlights how δ controls the trade-off between smoothness of p_δ and closeness of the approximation to $p_0(t) = \max(0, t)$. Proposition 2.1 formalizes this intuition and provides additional properties of p_δ .

PROPOSITION 2.1. *The softplus function p_δ satisfies the following properties:*

1. $p_0(t) = \max(0, t)$ for all $t \in \mathbb{R}$;
2. p_δ is differentiable and convex for all $\delta > 0$;
3. $p'_\delta(t) \in [0, 1]$ for all $\delta > 0$ and $t \in \mathbb{R}$;
4. $|p'_\delta(t_1) - p'_\delta(t_2)| \leq \frac{1}{4\delta} |t_1 - t_2|$ for all $\delta > 0$ and $t_1, t_2 \in \mathbb{R}$;
5. $0 \leq p_{\delta_2}(t) - p_{\delta_1}(t) \leq p_{\delta_2}(0) - p_{\delta_1}(0) = (\delta_2 - \delta_1) \log 2$ for all $\delta_2 \geq \delta_1 \geq 0$ and $t \in \mathbb{R}$.

The proof of Proposition 2.1 is included for completeness in Appendix A.

Let us now consider the penalty reformulation for (1.2)

$$(2.3) \quad \min_{x \in \mathbb{R}^n} F_{\xi, \delta}(x) := F(x) + \xi \sum_{i=1}^m p_\delta(a_i^T x - b_i),$$

where $\xi \geq 0$ is the penalty parameter and $\delta \geq 0$ is the smoothness parameter of the softplus function. Let $x_{\xi, \delta}^*$ denote the unique optimal solution of (2.3) for given $\xi \geq 0$ and $\delta \geq 0$. We are interested in studying the relationship between the optimal solution x^* for (1.2) and $x_{\xi, \delta}^*$ as we vary the parameters ξ and δ . Towards this goal, throughout the paper, the following assumptions are made concerning problem (1.2).

Assumption 2.2. Problem (1.2) satisfies the following properties:

1. The objective function F is globally μ -strongly convex for some $\mu > 0$, i.e., $F - \frac{\mu}{2} \|\cdot\|_2^2$ is a convex function;
2. The component functions f_i , $i = 1, \dots, \ell$, are convex and globally L -smooth for some $L \geq 0$, i.e., $\|\nabla f_i(x) - \nabla f_i(y)\|_2 \leq L\|x - y\|_2$ for all $x, y \in \mathbb{R}^n$;
3. The proximal term ψ is a proper convex and closed function;
4. The feasible region $\{x \in \mathbb{R}^n : Ax \leq b\}$ is nonempty, and the rows of A are normalized so that $\|a_i\|_2 = 1$ for $i = 1, \dots, m$.

Let us also consider the dual problem of (1.2),

$$(2.4) \quad \max_{\lambda \geq 0} G(\lambda) := -F^*(-A^T \lambda) - b^T \lambda,$$

where $F^*(y) = \sup_x \{y^T x - F(x)\}$ is the conjugate function of F . Note that strong convexity of F ensures that F^* is a smooth function. Let $\Lambda := \{\lambda : \lambda \geq 0\}$ denote the feasible region of (2.4) and let Λ^* denote the set of dual optimal solutions of (2.4). The following lemma shows that under Assumption 2.2, for a large enough $\xi \geq 0$, the solution of the penalty reformulation (2.3) at $\delta = 0$, $x_{\xi, 0}^*$, equals the solution of the original problem (1.2), x^* .

LEMMA 2.3. *Define $\bar{\xi} := \inf_{\lambda^* \in \Lambda^*} \|\lambda^*\|_\infty$. Then, $\bar{\xi}$ is finite and it holds that $x_{\xi, 0}^* = x^*$ for all $\xi \geq \bar{\xi}$.*

Proof. By the Lagrangian necessary conditions, there exists $\lambda^* \in \mathbb{R}_+^m$ such that

$$0 \in \partial F(x^*) + \sum_{i=1}^m \lambda_i^* a_i,$$

and $a_i^T x^* - b_i = 0$ for all i with $\lambda_i^* > 0$. Hence, Λ^* is non-empty and $\bar{\xi}$ is finite. Given $\bar{\xi} = \inf_{\lambda^* \in \Lambda^*} \|\lambda^*\|_\infty$, since $\partial_t p_0(0) = [0, 1]$, when $\xi \geq \bar{\xi}$ it holds that

$$0 \in \partial F(x^*) + \xi \sum_{i=1}^m \partial_x p_0(a_i^T x^* - b_i) = \partial_x F_{\xi, 0}(x^*).$$

Thus, x^* is the unique solution of (2.3). \square

Tatarenko and Nedich [38] present a similar bound on their penalty parameter that is based on Hoffman's bound for the polyhedral feasible region, and Mishchenko and Richtárik [29] present a bound on their penalty parameter based on an assumption that is close to the definition of Hoffman's constant. In both cases, estimating the relevant Hoffman-type constant can be difficult in practice. On the other hand, the bound presented in Lemma 2.3 is based on the infinity norm of a dual optimal solution, which may possibly be estimated in an adaptive manner. One may also expect the lower bound in Lemma 2.3 to be a less strict requirement because it is based on a local property of optimal solutions as compared to the Hoffman-type bounds that must hold globally. Finally, in section 6, we observe that using conservatively large

values of ξ , e.g., values much larger than the infinity norm of a dual optimal solution, does not have a significant effect on the performance of our algorithms.

Our next objective is to bound the 2-norm distance between $x_{\xi,\delta}^*$ and $x_{\xi,0}^*$. [Theorem 2.4](#) gives three such upper bounds on $\|x_{\xi,\delta}^* - x_{\xi,0}^*\|_2$. Throughout, let s_{\max} be the maximum singular value of the matrix A , and let s_{\min} be the minimum positive singular value of A .

THEOREM 2.4. *The following upper bounds on $\|x_{\xi,\delta}^* - x_{\xi,0}^*\|_2$ hold.*

1. *For any $\xi > 0$ and any $\delta \geq 0$, it holds that*

$$\|x_{\xi,\delta}^* - x_{\xi,0}^*\|_2 \leq \sqrt{\frac{2m\xi\delta \log 2}{\mu}}.$$

2. *For any $\xi > 0$ and any $\delta \in [0, \frac{\xi}{\mu} \exp(-2)]$, it holds that*

$$\|x_{\xi,\delta}^* - x_{\xi,0}^*\|_2 \leq m\delta \log\left(\frac{\xi}{\mu\delta}\right).$$

3. *For any $\xi > 0$ and any $\delta \in [0, \frac{s_{\max} \max(s_{\min}, 1)\xi}{\mu} \exp(-1)]$, it holds that*

$$\|x_{\xi,\delta}^* - x_{\xi,0}^*\|_2 \leq \frac{4\sqrt{m}}{s_{\min}} \delta \log\left(\frac{s_{\max} \max(s_{\min}, 1)\xi}{\mu\delta}\right).$$

First, note that based on the penalty reformulation construction, and the 5th property in [Proposition 2.1](#),

$$F_{\xi,\delta}(x) - F_{\xi,0}(x) \in [0, m\xi\delta \log 2], \quad \forall x \in \mathbb{R}^n.$$

Then, the first claim follows from the μ -strong convexity of F . Before presenting the proof of the remaining two claims, we introduce the following notation for residual vectors:

$$\begin{aligned} z_{\xi,\delta} &= A(x_{\xi,\delta}^* - x_{\xi,0}^*), \\ \tilde{z}_{\xi,\delta} &= Ax_{\xi,\delta}^* - b. \end{aligned}$$

Furthermore, the following simple result will be used often in the proofs.

LEMMA 2.5. *Consider the scalar function $\phi_{c,d} : [0, \infty) \rightarrow [0, \infty)$ defined as*

$$\phi_{c,d}(\theta) := \frac{\theta}{d + c \exp(\theta)} \text{ for } \theta \in [0, \infty) \text{ for constants } c, d > 0.$$

Whenever $\log(\frac{d}{c}) \geq 2$, it holds $\phi_{c,d}(\theta) \leq (\log(\frac{d}{c}) - 1)/d$ for all $\theta \in [0, \infty)$.

The proof of [Lemma 2.5](#) is included for completeness in [Appendix A](#).

Proof of Claim 2 in [Theorem 2.4](#). First, note it is sufficient to prove the proposition for the case when F is twice-continuous differentiable. Otherwise, for any $\epsilon > 0$, by [Corollary 1.3](#) in [\[6\]](#), one can construct \tilde{F} to be infinitely differentiable and an ϵ -approximation of F , i.e., $F(x) - \epsilon \leq \tilde{F}(x) \leq F(x)$. Let $\tilde{x}_{\xi,\delta}^*$ be the solution to $\min_x \tilde{F}_{\xi,\delta}(x) = \tilde{F}(x) + \xi \sum_{i=1}^m p_\delta(a_i^T x - b_i)$. Then, for $\delta > 0$,

$$\tilde{F}_{\xi,\delta}(x_{\xi,\delta}^*) \leq F_{\xi,\delta}(x_{\xi,\delta}^*) \leq F_{\xi,\delta}(\tilde{x}_{\xi,\delta}^*) \leq \tilde{F}_{\xi,\delta}(\tilde{x}_{\xi,\delta}^*) + \epsilon,$$

which means $\|\tilde{x}_{\xi,\delta}^* - x_{\xi,\delta}^*\|_2 \leq \frac{2\epsilon}{\mu}$, for all $\delta > 0$. Similarly, we can construct an ϵ -approximation of $\tilde{F}_{\xi,0}$, $\hat{F}_{\xi,0}$, and the corresponding solution $\hat{x}_{\xi,0}$. Then, $\|\hat{x}_{\xi,0}^* - \tilde{x}_{\xi,0}^*\|_2^2 \leq \frac{2\epsilon}{\mu}$ and $\|\hat{x}_{\xi,0}^* - x_{\xi,0}^*\|_2^2 \leq \frac{4\epsilon}{\mu}$. Hence,

$$\|x_{\xi,\delta}^* - x_{\xi,0}^*\|_2 \leq \|\tilde{x}_{\xi,\delta}^* - \tilde{x}_{\xi,0}^*\|_2 + (2 + 2\sqrt{2})\sqrt{\frac{\epsilon}{\mu}}.$$

Because the choice of ϵ is arbitrary, we may assume F to be twice-continuous differentiable without loss of generality.

For simplicity of notation, in the proof, we let $\xi > 0$ to be a constant and drop it from $x_{\xi,\delta}^*$, $z_{\xi,\delta}$, and $\tilde{z}_{\xi,\delta}$. Now, consider x_δ^* as a function of δ . We have $\nabla F_{\xi,\delta}(x_\delta^*) = 0$. Then,

$$\nabla^2 F_{\xi,\delta}(x_\delta^*) \frac{dx_\delta^*}{d\delta} + \frac{\partial}{\partial \delta} \nabla F_{\xi,\delta}(x_\delta^*) = 0.$$

And since

$$\begin{aligned} \nabla F_{\xi,\delta}(x) &= \nabla F(x) + \xi \sum_{i=1}^m \nabla p_\delta(a_i^T x - b_i) \\ &= \nabla F(x) + \xi \sum_{i=1}^m a_i \frac{\exp((a_i^T x - b_i)/\delta)}{1 + \exp((a_i^T x - b_i)/\delta)}. \end{aligned}$$

Then,

$$\frac{\partial}{\partial \delta} \nabla F_{\xi,\delta}(x) = \xi \sum_{i=1}^m a_i \frac{-\frac{a_i^T x - b_i}{\delta^2} \exp((a_i^T x - b_i)/\delta)}{(1 + \exp((a_i^T x - b_i)/\delta))^2},$$

$$(2.5) \quad \nabla^2 F_{\xi,\delta}(x) = \nabla^2 F(x) + \xi \sum_{i=1}^m \frac{a_i a_i^T \exp((a_i^T x - b_i)/\delta)}{(1 + \exp((a_i^T x - b_i)/\delta))^2},$$

we have

$$(2.6) \quad \begin{aligned} \frac{dx_\delta^*}{d\delta} &= \left(\delta \nabla^2 F(x_\delta^*) + \xi \sum_{i=1}^m \frac{a_i a_i^T \exp((a_i^T x_\delta^* - b_i)/\delta)}{(1 + \exp((a_i^T x_\delta^* - b_i)/\delta))^2} \right)^{-1} \\ &\quad \left(\xi \sum_{i=1}^m a_i \frac{a_i^T x_\delta^* - b_i}{\delta} \frac{\exp((a_i^T x_\delta^* - b_i)/\delta)}{(1 + \exp((a_i^T x_\delta^* - b_i)/\delta))^2} \right). \end{aligned}$$

In other words,

$$\begin{aligned} \left\| \frac{dx_\delta^*}{d\delta} \right\|_2 &\leq \left\| \left(\mu \delta I + \xi \sum_{i=1}^m \frac{a_i a_i^T \exp(\tilde{z}_{\delta,i}/\delta)}{(1 + \exp(\tilde{z}_{\delta,i}/\delta))^2} \right)^{-1} \xi \sum_{i=1}^m a_i \frac{\tilde{z}_{\delta,i}}{\delta} \frac{\exp(\tilde{z}_{\delta,i}/\delta)}{(1 + \exp(\tilde{z}_{\delta,i}/\delta))^2} \right\|_2 \\ &\leq \sum_{i=1}^m \xi \left\| \left(\mu \delta I + \xi \frac{a_i a_i^T \exp(\tilde{z}_{\delta,i}/\delta)}{(1 + \exp(\tilde{z}_{\delta,i}/\delta))^2} \right)^{-1} a_i \right\|_2 \frac{\tilde{z}_{\delta,i}}{\delta} \frac{\exp(|\tilde{z}_{\delta,i}|/\delta)}{(1 + \exp(\tilde{z}_{\delta,i}/\delta))^2}, \end{aligned}$$

where by Woodbury's identity, and the fact $\|a_i\|_2 = 1$

$$\begin{aligned} & \left\| \left(\mu\delta I + \xi \frac{a_i a_i^T \exp(\tilde{z}_{\delta,i}/\delta)}{(1 + \exp(\tilde{z}_{\delta,i}/\delta))^2} \right)^{-1} a_i \right\|_2 \\ &= \left\| \left(\frac{1}{\mu\delta} I - \frac{1}{\mu^2 \delta^2} a_i \left(\frac{a_i^T a_i}{\mu\delta} + \frac{(1 + \exp(\tilde{z}_{\delta,i}/\delta))^2}{\xi \exp(\tilde{z}_{\delta,i}/\delta)} \right)^{-1} a_i^T \right) a_i \right\|_2 \\ &= \frac{1}{\mu\delta} \frac{(1 + \exp(\tilde{z}_{\delta,i}/\delta))^2}{\xi \exp(\tilde{z}_{\delta,i}/\delta)} \left(\frac{1}{\mu\delta} + \frac{(1 + \exp(\tilde{z}_{\delta,i}/\delta))^2}{\xi \exp(\tilde{z}_{\delta,i}/\delta)} \right)^{-1}. \end{aligned}$$

Hence, it holds

$$\left\| \frac{dx_\delta^*}{d\delta} \right\|_2 \leq \sum_{i=1}^m \frac{|\tilde{z}_{\delta,i}|/\delta}{1 + \frac{\mu\delta}{\xi} \frac{(1 + \exp(\tilde{z}_{\delta,i}/\delta))^2}{\exp(\tilde{z}_{\delta,i}/\delta)}} \leq \frac{|\tilde{z}_{\delta,i}|/\delta}{1 + \frac{\mu\delta}{\xi} \exp(|\tilde{z}_{\delta,i}|/\delta)}.$$

Then, by [Lemma 2.5](#),

$$\left\| \frac{dx_\delta^*}{d\delta} \right\|_2 \leq m \left(\log \left(\frac{\xi}{\mu\delta} \right) - 1 \right), \quad \forall \delta \leq \frac{\xi}{\mu} \exp(-2). \quad \square$$

Hence, whenever $\delta \in [0, \frac{\xi}{\mu} \exp(-2)]$, it holds $\|x_\delta^* - x_0^*\|_2 \leq m\delta \log \left(\frac{\xi}{\mu\delta} \right)$.

Before presenting the proof for Claim 3 in [Theorem 2.4](#), we first show the following proposition on the estimation of $\|z_{\xi,\delta}\|_2$, which then leads to an upper bounds on the violation of constraints at $x_{\xi,\delta}^*$.

PROPOSITION 2.6. *For any $\xi > 0$ and $\delta \in [0, \frac{s_{\max}^2 \xi}{\mu} \exp(-2)]$, it holds that*

$$\|z_{\xi,\delta}\|_2 \leq \sqrt{m}\delta \log \left(\frac{s_{\max}^2 \xi}{\mu\delta} \right).$$

Proof. As before, it suffices to prove the result for the case when $F(x)$ is twice-continuously differentiable. For simplicity of notation, we let $\xi > 0$ to be a constant and use the simpler notations x_δ^* , z_δ , \tilde{z}_δ in the proof. Letting

$$\Pi_{\tilde{z}_\delta, \delta} = \text{diag} \left(\frac{\exp(\tilde{z}_{\delta,1}/\delta)}{(1 + \exp(\tilde{z}_{\delta,1}/\delta))^2}, \dots, \frac{\exp(\tilde{z}_{\delta,m}/\delta)}{(1 + \exp(\tilde{z}_{\delta,m}/\delta))^2} \right),$$

we can express (2.6) as

$$(2.7) \quad \frac{dx_\delta^*}{d\delta} = \xi(\delta\nabla^2 F(x_\delta^*) + \xi A^T \Pi_{\tilde{z}_\delta, \delta} A)^{-1} A^T \Pi_{\tilde{z}_\delta, \delta} \frac{\tilde{z}_\delta}{\delta}.$$

Then,

$$\frac{dz_\delta}{d\delta} = \xi A(\delta\nabla^2 F(x_\delta^*) + \xi A^T \Pi_{\tilde{z}_\delta, \delta} A)^{-1} A^T \Pi_{\tilde{z}_\delta, \delta} \frac{\tilde{z}_\delta}{\delta}.$$

By Woodbury's identity,

$$\begin{aligned} & (\delta\nabla^2 F(x_\delta^*) + \xi A^T \Pi_{\tilde{z}_\delta, \delta} A)^{-1} = (\delta\nabla^2 F(x_\delta^*))^{-1} \\ & - (\delta\nabla^2 F(x_\delta^*))^{-1} A^T (A(\delta\nabla^2 F(x_\delta^*))^{-1} A^T + \xi^{-1} \Pi_{\tilde{z}_\delta, \delta}^{-1})^{-1} A (\delta\nabla^2 F(x_\delta^*))^{-1}; \end{aligned}$$

so

$$\begin{aligned} \frac{dz_\delta}{d\delta} &= A(\delta \nabla^2 F(x_\delta^*))^{-1} A^T (A(\delta \nabla^2 F(x_\delta^*))^{-1} A^T + \xi^{-1} \Pi_{\tilde{z}_\delta, \delta}^{-1})^{-1} \frac{\tilde{z}_\delta}{\delta} \\ &= A \left(\frac{\nabla^2 F(x_\delta^*)}{\mu} \right)^{-1} A^T \left(A \left(\frac{\nabla^2 F(x_\delta^*)}{\mu} \right)^{-1} A^T + \frac{\mu\delta}{\xi} \Pi_{\tilde{z}_\delta, \delta}^{-1} \right)^{-1} \tilde{z}_\delta / \delta. \end{aligned}$$

Then, because $A \left(\frac{\nabla^2 F(x_\delta^*)}{\mu} \right)^{-1} A^T \preceq s_{\max}^2 I$, we have

$$\left\| \frac{dz_\delta}{d\delta} \right\|_2 \leq \left\| \left(\frac{|\tilde{z}_{\delta,1}|/\delta}{1 + \frac{\mu\delta}{s_{\max}^2 \xi} \exp(|\tilde{z}_{\delta,1}|/\delta)}, \dots, \frac{|\tilde{z}_{\delta,m}|/\delta}{1 + \frac{\mu\delta}{s_{\max}^2 \xi} \exp(|\tilde{z}_{\delta,m}|/\delta)} \right) \right\|_2.$$

By [Lemma 2.5](#), for $\delta \in [0, \frac{s_{\max}^2 \xi}{\mu} \exp(-2)]$, it holds

$$\frac{|\tilde{z}_{\delta,i}|/\delta}{1 + \frac{\mu\delta}{s_{\max}^2 \xi} \exp(|\tilde{z}_{\delta,i}|/\delta)} \leq \log \left(\frac{s_{\max}^2 \xi}{\mu\delta} \right) - 1.$$

Hence, $\left\| \frac{dz_\delta}{d\delta} \right\|_2 \leq \sqrt{m} \left(\log \left(\frac{s_{\max}^2 \xi}{\mu\delta} \right) - 1 \right)$, and $\|z_\delta\|_2 \leq \sqrt{m} \delta \log \left(\frac{s_{\max}^2 \xi}{\mu\delta} \right)$ follows. \square

When $\xi \geq \bar{\xi}$, by [Lemma 2.3](#), $x_{\xi, \delta}^* = x^*$ and $\tilde{z}_{\xi, 0} = Ax^* - b \leq 0$. Then, by definition, $\|(\tilde{z}_{\xi, \delta})^+\|_2 \leq \|z_{\xi, \delta}\|_2$, and [Proposition 2.6](#) shows that the violation of constrains for the optimal solution to the penalty reformulation $\|(Ax_{\xi, \delta}^* - b)^+\|$, is bounded by $\tilde{O}(\sqrt{m}\delta)$.

Proof of Claim 3 in [Theorem 2.4](#). Similarly, we may assume that $F(x)$ is twice-continuously differentiable. For simplicity of notation, we again let $\xi > 0$ to be a constant and use the simplified notations x_δ^* , z_δ , \tilde{z}_δ in the proof. Based on the proof of [Proposition 2.6](#), it holds

$$\left\| \frac{dx_\delta}{d\delta} \right\|_2 \leq \left\| A^T \left(AA^T + \frac{\mu\delta}{\xi} \Pi_{\tilde{z}_\delta, \delta}^{-1} \right)^{-1} \frac{\tilde{z}_\delta}{\delta} \right\|_2.$$

Let $\hat{z}_\delta = (\hat{z}_{\delta,1}, \dots, \hat{z}_{\delta,1})^T$, where

$$\hat{z}_{\delta,i} = \begin{cases} \tilde{z}_{\delta,i} & \text{if } |\tilde{z}_{\delta,i}| \geq \delta \log \left(\frac{s_{\max} \max(s_{\text{pmin}}, 1) \xi}{\mu\delta} \right) \\ 0 & \text{o.w.} \end{cases}$$

Then,

$$\begin{aligned} \left\| \frac{dx_\delta}{d\delta} \right\|_2 &\leq \left\| A^T (AA^T + \frac{\mu\delta}{\xi} \Pi_{\tilde{z}_\delta, \delta}^{-1})^{-1} \frac{\hat{z}_\delta}{\delta} \right\|_2 + \left\| A^T \left(AA^T + \frac{\mu\delta}{\xi} \Pi_{\tilde{z}_\delta, \delta}^{-1} \right)^{-1} \frac{\tilde{z}_\delta - \hat{z}_\delta}{\delta} \right\|_2 \\ &\leq \left\| A^T \Pi_{\tilde{z}_\delta, \delta} \frac{\hat{z}_\delta}{\delta} \right\|_2 + \left\| A^T \left(AA^T + \frac{\mu\delta}{\xi} I \right)^{-1} \frac{\tilde{z}_\delta - \hat{z}_\delta}{\delta} \right\|_2 \\ &\leq s_{\max} \left\| \frac{\xi}{\mu\delta} \Pi_{\tilde{z}_\delta, \delta} \frac{\hat{z}_\delta}{\delta} \right\|_2 + \left\| \frac{\tilde{z}_\delta - \hat{z}_\delta}{\delta} \right\|_2 / s_{\text{pmin}}. \end{aligned}$$

By the definition of \hat{z}_δ , for $\delta \in [0, \frac{s_{\max} \max(s_{\text{pmin}}, 1) \xi}{\mu} \exp(-1)]$,

$$\begin{aligned} \left\| \frac{\xi}{\mu\delta} \Pi_{\tilde{z}_\delta, \delta} \frac{\hat{z}_\delta}{\delta} \right\|_2 &\leq \frac{\sqrt{m}}{s_{\text{pmin}} s_{\max}}, \\ \left\| \frac{\tilde{z}_\delta - \hat{z}_\delta}{\delta} \right\|_2 &\leq \sqrt{m} \log \left(\frac{s_{\max} \max(s_{\text{pmin}}, 1) \xi}{\mu\delta} \right). \end{aligned}$$

Hence, the claim holds. \square

3. Nested Penalty Method and Complexity Analysis. In this section, we present a generic algorithmic framework for solving the original constrained optimization problem (1.2) via algorithms that solve the unconstrained penalty reformulation (2.3). Our primary algorithm is a dynamic nested method (Algorithm 3.1), which solves a sequence of penalty reformulation problems as the smoothing parameter δ_k tends to zero at an appropriate rate.

We first present an algorithmic framework that uses a generic unconstrained convex optimization method \mathcal{M} to solve the penalty reformulation problems. We present a corresponding complexity result expressed in terms of the complexity of \mathcal{M} and our complexity analysis utilizes the results established in section 2 concerning bounds between the solutions of the original constrained problem and the penalty formulation. We are primarily interested in the use of fast stochastic gradient methods to solve the penalty reformulations; therefore, we then specialize our analysis to the exemplary case where \mathcal{M} is the proximal SVRG method with catalyst acceleration [26]. In this case, we show that our algorithm requires at most $\tilde{O}(1/\sqrt{\epsilon})$ total stochastic gradient evaluations to achieve a solution within a 2-norm distance of ϵ to the optimal solution x^* to (1.2). While the complexity analysis is focused primarily on the use of the proximal SVRG method with catalyst acceleration [26] as a black box to solve the penalty reformulations (2.3), it is not strongly tied to the choice of this method; other accelerated methods for finite sum problems, such as Katyusha [1] and the RPDG method [24], may be utilized as well.

Recall that the objective function of the penalty reformulation (2.3) decomposes into the sum of three terms: (i) the average of ℓ smooth components of the objective function, (ii) the possibly non-smooth proximal term ψ , and (iii) the weighted sum of smooth penalty functions p_δ applied to the constraints. Let $\tilde{F}_{\xi,\delta}(x) := \frac{1}{\ell} \sum_{i=1}^{\ell} f_i(x) + \xi \sum_{i=1}^m p_\delta(a_i^T x - b_i)$ denote the smooth terms so that the objective function $F_{\xi,\delta}$ of (2.3) decomposes as $F_{\xi,\delta}(x) = \tilde{F}_{\xi,\delta}(x) + \psi(x)$. In our generic algorithm, we will utilize the proximal operator and generalized gradient mapping induced by the proximal objective component ψ . Recall that the scaled proximal operator of ψ with parameter $\alpha > 0$ is defined by

$$\text{prox}_\psi(\bar{x}; \alpha) := \arg \min_{x \in \mathbb{R}^n} \left\{ \psi(x) + \frac{1}{2\alpha} \|x - \bar{x}\|_2^2 \right\} \quad \text{for all } \bar{x} \in \mathbb{R}^n.$$

The above scaled proximal operator $\text{prox}_\psi(\cdot; \alpha)$ is equivalent to the standard proximal operator of the function $\alpha\psi$. The scaled proximal operator $\text{prox}_\psi(\cdot; \alpha)$ induces a (generalized) gradient mapping g_ψ that can be evaluated on any smooth convex function $\tilde{F} : \mathbb{R}^n \rightarrow \mathbb{R}$. For any $\bar{x} \in \mathbb{R}^n$ and $\alpha > 0$, the gradient mapping induced by ψ is defined by $g_\psi(\bar{x}; \tilde{F}, \alpha) := \frac{1}{\alpha}(\bar{x} - \tilde{x})$, where

$$\begin{aligned} \tilde{x} &:= \text{prox}_\psi(\bar{x} - \alpha \nabla \tilde{F}(\bar{x}); \alpha) \\ &= \arg \min_{x \in \mathbb{R}^n} \left\{ \tilde{F}(\bar{x}) + \nabla \tilde{F}(\bar{x})^T (x - \bar{x}) + \frac{1}{2\alpha} \|x - \bar{x}\|_2^2 + \psi(x) \right\}. \end{aligned}$$

Note that the above notion of gradient mapping based on the proximal operator of the function ψ is considered by Lin et al. [27] and is a generalization of the gradient mapping introduced by Nesterov [34]. The intuition of the gradient mapping is that, when the function \tilde{F} is smooth and strongly convex, the squared norm of the gradient mapping is directly related to the suboptimality gap, as formalized by the following Lemma.

LEMMA 3.1. *Let $\psi : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$ be a proper, convex, and closed proximal function with gradient mapping g_ψ , and let $\tilde{F} : \mathbb{R}^n \rightarrow \mathbb{R}$ be convex and L -smooth for some $L \geq 0$. Consider $F := \tilde{F} + \psi$, and let $F^* := \min_{x \in \mathbb{R}^n} F(x)$. Suppose F is μ -strongly convex for some $0 \leq \mu \leq L$. Then, for any $\alpha \in (0, 1/2L]$ and for any $\bar{x} \in \mathbb{R}^n$ with $\tilde{x} := \text{prox}_\psi(\bar{x} - \alpha \nabla \tilde{F}(\bar{x}); \alpha)$, the following inequalities hold:*

1. $\|g_\psi(\bar{x}; \tilde{F}, \alpha)\|_2^2 \leq \frac{2}{\alpha}(F(\bar{x}) - F^*)$;
2. $\|g_\psi(\tilde{x}; \tilde{F}, \alpha)\|_2^2 \geq 2\mu(F(\tilde{x}) - F^*)$.

Lemma 3.1 is an extension of Theorem 2.2.13 of Nesterov [34], which covers the case when $\psi(x)$ is an indicator function of a convex set. Part (2.) of Lemma 3.1 is the same as Lemma 2 of [27]. The proof of Lemma 3.1 is included, for completeness, in Appendix A.

We presume herein that the unconstrained convex optimization method \mathcal{M} , which is applied to the penalty reformulations (2.3) in our generic algorithm, is a randomized iterative method, such as a stochastic first-order method. That is, whenever we apply \mathcal{M} to minimize a convex function $F : \mathbb{R}^n \rightarrow \mathbb{R}$, the iterate sequence $\{x_k\}$ generated by \mathcal{M} is a stochastic process. For any convex function $F : \mathbb{R}^n \rightarrow \mathbb{R}$ such that $F^* := \inf_{x \in \mathbb{R}^n} F(x) > -\infty$ and for any $\epsilon > 0$, define the stopping time

$$(3.1) \quad T_{\mathcal{M}, F}(\epsilon; x_0) := \inf\{k \geq 0 : F(x_k) - F^* \leq \epsilon\},$$

where $\{x_k\}$ is the sequence generated by \mathcal{M} initialized at x_0 and applied to minimize F . Notice that $T_{\mathcal{M}, F}(\epsilon; x_0)$ is a random variable when \mathcal{M} is a randomized method. We assume throughout that $\mathbb{E}[T_{\mathcal{M}, F}(\epsilon; x_0)] < \infty$ for any $\epsilon > 0$. In the case when the convex function F decomposes as $F = \tilde{F} + \psi$ as in the conditions of Lemma 3.1, we can also naturally define, for any fixed $\alpha > 0$ and $\epsilon' > 0$, a stopping time in terms of the gradient mapping:

$$(3.2) \quad \tilde{T}_{\mathcal{M}, F}(\epsilon'; x_0, \alpha) := \inf\{k \geq 0 : \|g_\psi(x_k; \tilde{F}, \alpha)\|_2 \leq \epsilon'\}.$$

Items (2.) and (3.) of Lemma 3.1 clearly imply that these two distinct notions of stopping times can be related to each other. The main advantage of the stopping time (3.2) over (3.1) is that (3.2) provides an implementable stopping criterion even when F^* is unknown, as long as the gradient mapping is computable. The previously defined stopping times will be useful in the analysis of our nested algorithm, which involves several stages where \mathcal{M} is applied separately at each stage.

We are now ready to present our generic nested penalty method in Algorithm 3.1. Algorithm 3.1 requires specifying a value for the penalty parameter $\xi > 0$ and a value for the “smoothness scaling parameter” $\eta > 1$ which controls the rate at which we shrink the smoothness parameter δ towards zero. The unconstrained convex optimization method \mathcal{M} as well as its total number of iterations at each stage τ_t are both currently left as generic. Note that, after finding an approximate solution \hat{x}_t using method \mathcal{M} , an additional computation is performed to compute the proximal operator as required in the calculation of the gradient mapping at \hat{x}_t .

Our main result concerning the generic Algorithm 3.1 is presented in Lemma 3.2, which shows that the sequence $\{\hat{x}_t\}$ converges linearly to the optimal solution x^* of (1.2) when the subproblems are solved to a sufficiently small accuracy.

LEMMA 3.2. *Consider applying the generic nested penalty method (Algorithm 3.1) using penalty parameter $\xi \geq \tilde{\xi}$ and initial smoothness parameter $\delta_0 \leq \frac{\tilde{\xi}}{\mu} \exp(-2)$. Let*

Algorithm 3.1 Generic Nested Penalty Method

Parameters: Penalty parameter $\xi > 0$, smoothness scaling parameter $\eta > 1$, and generic unconstrained convex optimization method \mathcal{M} .

Initialize: $\hat{x}_{-1} \in \mathbb{R}^n$, $\delta_0 > 0$.

for $t = 0, 1, \dots, T$ **do**

1. Apply method \mathcal{M} on the penalty reformulation problem (2.3), with penalty parameter ξ and smoothness parameter δ_t , initialized at \tilde{x}_{t-1} and for $\tau_t \geq 1$ total iterations, to obtain an approximate solution

$$\hat{x}_t \approx \arg \min_{x \in \mathbb{R}^n} \left\{ F_{\xi, \delta_t}(x) := F(x) + \xi \sum_{i=1}^m p_{\delta_t}(a_i^T x - b_i) \right\}.$$

2. Choose step-size $\alpha_t > 0$ and set $\tilde{x}_t \leftarrow \text{prox}_{\psi}(\hat{x}_t - \alpha_t \nabla \tilde{F}_{\xi, \delta_t}(\hat{x}_t); \alpha_t)$.
3. Update $\delta_{t+1} \leftarrow \delta_t / \eta$.

end for

$\{\epsilon_t\}$ be an accuracy sequence for the penalty reformulation subproblems satisfying

$$0 < \epsilon_t \leq \frac{1}{2} m \mu \delta_t^2 \log^2 \left(\frac{\xi s_{\max}}{\mu \delta_t} \right) \text{ for } t \geq 0,$$

and suppose that stopping criterion (3.2) is used with step-size $0 < \alpha_t \leq 1/(L + \frac{\xi s_{\max}^2}{4\delta_t})$, i.e., $\tau_t = \tilde{T}_{\mathcal{M}, F_{\xi, \delta_t}}(\sqrt{2\mu\epsilon_t}; \tilde{x}_{t-1}, \alpha_t)$ for all $t \geq 0$. Then, for any $t \geq 0$, it holds that:

$$\begin{aligned} \|\tilde{x}_t - x^*\|_2 &\leq \left(\min \left\{ m, \frac{4\sqrt{m}}{s_{\text{pmin}}} \right\} + \sqrt{m} \right) \delta_t \log \left(\frac{\xi s_{\max}}{\mu \delta_t} \right) \\ (3.3) \quad &= \frac{\left(\min \left\{ m, \frac{4\sqrt{m}}{s_{\text{pmin}}} \right\} + \sqrt{m} \right) \delta_0}{\eta^t} \left[t \log(\eta) + \log \left(\frac{\xi s_{\max}}{\mu \delta_0} \right) \right]. \end{aligned}$$

If instead the number of inner iterations τ_t is chosen so that $F_{\xi, \delta_t}(\hat{x}_t) - F_{\xi, \delta_t}^* \leq \epsilon_t$ for all $t \geq 0$, then (3.3) also holds.

Proof. Note that, when $\tau_t = \tilde{T}_{\mathcal{M}, F_{\xi, \delta_t}}(\sqrt{2\mu\epsilon_t}; \tilde{x}_{t-1}, \alpha_t)$, item (2.) of Lemma 3.1 implies $F_{\xi, \delta_t}(\tilde{x}_t) - F_{\xi, \delta_t}^* \leq \epsilon_t$. Furthermore, when it is assumed that $F_{\xi, \delta_t}(\hat{x}_t) - F_{\xi, \delta_t}^* \leq \epsilon_t$, then item (1.) of Lemma 3.1 implies $F_{\xi, \delta_t}(\tilde{x}_t) - F_{\xi, \delta_t}^* \leq F_{\xi, \delta_t}(\hat{x}_t) - F_{\xi, \delta_t}^* \leq \epsilon_t$. Thus, under both cases we have that $F_{\xi, \delta_t}(\tilde{x}_t) - F_{\xi, \delta_t}^* \leq \epsilon_t$. Furthermore, it holds that

$$\|\tilde{x}_t - x^*\|_2 \leq \|\tilde{x}_t - x_{\xi, \delta_t}^*\|_2 + \|x_{\xi, \delta_t}^* - x^*\|_2 \leq \sqrt{\frac{2\epsilon_t}{\mu}} + \|x_{\xi, \delta_t}^* - x^*\|_2.$$

Then, the result holds by μ -strong convexity of F_{ξ, δ_t} , and parts 2–3 of Theorem 2.4. \square

Note that Lemma 3.2 requires availability of the constants $\bar{\xi}$, μ , and s_{\max} and in order to set the penalty parameter ξ and the sequences of accuracy parameters ϵ_t and step-sizes α_t . As mentioned previously, using a conservatively large value of $\xi \geq \bar{\xi}$ is not prohibitive in practice, as observed in the numerical experiments presented in section 6. Likewise, depending on the information available about the original problem (1.2), one can replace μ with an available lower bound and s_{\max} with an available upper bound (e.g., $s_{\max} \leq \sqrt{m}$ by part (4.) of Assumption 2.2)

and the results of [Lemma 3.2](#) will carry through. Using the stopping criterion [\(3.2\)](#) requires an additional calculation of the gradient mapping at each iteration of the method \mathcal{M} . The stochastic accelerated methods we consider, including the catalyst accelerated SVRG, Katyusha, and RPDG, all require periodic cyclic calculations of the full gradients and therefore this stopping criterion can be implemented without any additional complexity. (Note that, in the case of catalyst accelerated SVRG for example, we consider each “outer loop iteration” to be one iteration of \mathcal{M} .)

We are now ready to specialize our complexity results to the exemplary case of using the proximal SVRG method with catalyst acceleration [\[26\]](#) in place of the generic method \mathcal{M} in [Algorithm 3.1](#). We state the overall complexity in terms of the number of “incremental steps,” which we define as any calculation of: (i) the objective components $f_i(x)$ and their gradients $\nabla f_i(x)$, (ii) the penalty function values $p_\delta(a_i^T x - b_i)$ and their gradients, and (iii) the proximal operator $\text{prox}_\psi(\bar{x}; \alpha)$. In the remainder of the text, \tilde{O} notation is used to hide universal constants and poly-logarithmic dependencies on problem parameters and $1/\epsilon$.

PROPOSITION 3.3. *Consider applying the nested penalty method ([Algorithm 3.1](#)), with penalty parameter $\xi \geq \bar{\xi}$ and initial smoothness parameter $\delta_0 = \frac{\xi}{\mu} \exp(-2)$, and using proximal SVRG with catalyst acceleration [\[26, 27\]](#) to solve the penalty reformulation subproblems. Let $\{\epsilon_t\}$ be an accuracy sequence for the subproblems satisfying*

$$\epsilon_t = \min \left\{ \frac{1}{2} m \mu \delta_t^2 \log^2 \left(\frac{\xi s_{\max}}{\mu \delta_t} \right), m \xi \delta_t \log \left(\frac{\xi s_{\max}}{\mu \delta_t} \right) \right\} \text{ for } t \geq 0,$$

and suppose that the stopping criterion based on the gradient mapping [\(3.2\)](#) is used with step-size $\alpha_t = 1/(L + \frac{\xi s_{\max}^2}{4\delta_t})$, i.e., $\tau_t = \tilde{T}_{\mathcal{M}, F_{\xi, \delta_t}}(\sqrt{2\mu\epsilon_t}; \tilde{x}_{t-1}, \alpha_t)$ for all $t \geq 0$. Then, for any desired accuracy $\epsilon \in (0, \frac{m\xi}{\mu} \exp(-2)]$, the expected number of incremental steps required to satisfy $\|\tilde{x}_T - x^*\|_2 \leq \epsilon$ is upper bounded by

$$(3.4) \quad \tilde{O} \left(\ell + \sqrt{\frac{(\ell+m)L}{\mu}} + m \sqrt{\frac{(\ell+m)\xi}{\mu\epsilon}} \right).$$

Proof. Suppose we use the sequence of δ : $\delta_0 > \dots > \delta_T$, where $\delta_0 = \frac{\xi}{\mu} \exp(-2)$, $\delta_{t-1}/\delta_t = \eta > 1$, $t = 1, \dots, T$, and T is the smallest positive integer such that inequality

$$2m\delta_T \log \left(\frac{\xi s_{\max}}{\mu \delta_T} \right) \leq \epsilon$$

holds. Since $\frac{1}{w} > 2 \log(\frac{1}{w})$ for all $w > 0$, we have

$$\frac{w}{2 \log(\frac{1}{w})} \log \left(\frac{2 \log(\frac{1}{w})}{w} \right) \leq w, \quad \forall w \in (0, 1).$$

Letting $w = \frac{\mu\epsilon}{2m\xi s_{\max}}$ in the inequality, we see that

$$2m\delta \log \left(\frac{\xi s_{\max}}{\mu \delta} \right) \leq \epsilon, \quad \forall \delta \in \left(0, \frac{\epsilon}{4m} \left(\log \left(\frac{2m\xi s_{\max}}{\mu\epsilon} \right) \right)^{-1} \right].$$

Hence,

$$\delta_T \geq \frac{\epsilon}{4m\eta} \left(\log \left(\frac{2m\xi s_{\max}}{\mu\epsilon} \right) \right)^{-1}.$$

Then, by [Lemma 3.2](#), $\|\tilde{x}_T - x^*\|_2 \leq 2m\delta_T \log\left(\frac{\xi s_{\max}}{\mu\delta_T}\right) \leq \epsilon$ holds.

Since we start minimizing $F_{\xi, \delta_{t+1}}(x)$, $t = 0, \dots, T-1$, at \tilde{x}_t , by [Lemma 3.2](#)

$$\begin{aligned} & F_{\xi, \delta_{t+1}}(\tilde{x}_t) - F_{\xi, \delta_{t+1}}(x_{\xi, \delta_{t+1}}^*) \\ & \leq (F_{\xi, \delta_{t+1}}(\tilde{x}_t) - F_{\xi, \delta_t}(x_{\xi, \delta_t}^*)) + (F_{\xi, \delta_t}(x_{\xi, \delta_t}^*) - F_{\xi, \delta_{t+1}}(x_{\xi, \delta_{t+1}}^*)) \\ & \leq m\xi\delta_t \log\left(\frac{\xi s_{\max}}{\mu\delta_{t+1}}\right) + m\xi(\delta_t - \delta_{t+1}) \log 2 \\ & \leq (2\eta - 1)m\xi\delta_{t+1} \log\left(\frac{\xi s_{\max}}{\mu\delta_{t+1}}\right). \end{aligned}$$

For $t = 0, 1, \dots, T$, we consider $\frac{\ell+m}{\ell}f_i(x)$, $i = 1, \dots, \ell$, and $(\ell+m)\xi p_{\delta_t}(a_j^T x - b_j)$, $j = 1, \dots, m$ to be the $\ell+m$ components. Then, the problems are μ -strongly convex, and consist of ℓ components that are $\frac{\ell+m}{\ell}L$ -smooth, and m components that are L_t -smooth, respectively, where

$$L_t = \frac{(\ell+m)\xi}{4\delta_t}.$$

Then, since $\tau_t = \tilde{T}_{\mathcal{M}, F_{\xi, \delta_t}}(\sqrt{2\mu\epsilon_t}; \tilde{x}_{t-1}, \alpha_t)$, by [Lemma 3.1](#), we have

$$\tau_t \leq T_{\mathcal{M}, F_{\xi, \delta_t}}\left(\frac{L + \frac{\xi s_{\max}}{4\delta_t}}{\mu}\epsilon_t; \tilde{x}_{t-1}\right)$$

and, by [Lemma A.1](#), we have the following bounds
(3.5)

$$\begin{aligned} \mathbb{E}\tau_0 &= \tilde{O}\left(\ell + m + \sqrt{\frac{(m+\ell)L + mL_0}{\mu}}\right) = \tilde{O}\left(\ell + \sqrt{\frac{(\ell+m)L}{\mu}} + \sqrt{(\ell+m)m}\right) \\ \mathbb{E}\tau_t &= \tilde{O}\left(\ell + m + \sqrt{\frac{(m+\ell)L + mL_t}{\mu}}\right) \\ &= \tilde{O}\left(\ell + \sqrt{\frac{(\ell+m)L}{\mu}} + \sqrt{\frac{(\ell+m)m\xi}{\mu\delta_t}}\right), \quad t = 1, \dots, T. \end{aligned}$$

Since $\delta_T \geq \frac{\epsilon}{4m\eta} \left(\log\left(\frac{2m\xi s_{\max}}{\mu\epsilon}\right)\right)^{-1}$,

$$T \leq \frac{\log\left(\frac{4\eta m\xi}{\mu\epsilon} \log\left(\frac{2m\xi s_{\max}}{\mu\epsilon}\right)\right)}{\log(\eta)}.$$

Therefore, $\mathbb{E}\sum_{t=1}^T \tau_t$ is bounded by

$$\begin{aligned} & \tilde{O}\left(T\left(\ell + \sqrt{\frac{(\ell+m)L}{\mu}}\right) + \sqrt{\frac{(\ell+m)m\xi}{\mu\delta_1} \frac{\eta^{\frac{T}{2}} - 1}{\sqrt{\eta} - 1}}\right) \\ &= \tilde{O}\left(\ell + \sqrt{\frac{(\ell+m)L}{\mu}} + \frac{\eta}{\sqrt{\eta} - 1} m \sqrt{\frac{(\ell+m)\xi}{\mu\epsilon}}\right). \end{aligned}$$

Then, we can take η to be a constant greater than 1, (or $\eta = 4$ which minimizes $\frac{\eta}{\sqrt{\eta-1}}$), combine it with the estimation of τ_0 , and within

$$\tilde{O} \left(\ell + \sqrt{\frac{(\ell+m)L}{\mu}} + m\sqrt{\frac{(\ell+m)\xi}{\mu\epsilon}} \right),$$

expected number of incremental steps, we will reach \tilde{x}_T satisfying $\|\tilde{x}_T - x^*\|_2 \leq \epsilon$. \square

Remark 3.4. It is possible to slightly improve the result of [Proposition 3.3](#) by considering a smaller value of the accuracy parameter ϵ_T for *only* the last inner loop iteration. In particular, by the third claim in [Theorem 2.4](#), since δ_T satisfies $2m\delta_T \log\left(\frac{\xi s_{\max}}{\mu\delta_T}\right) \leq \epsilon$, we will have $\|x_{\xi, \delta_T}^* - x^*\|_2 \leq \frac{2\epsilon}{\sqrt{m s_{\min}}}$. Since the dependence on ϵ_T is logarithmic in the complexity result, if we use

$$\epsilon'_T = \frac{m\mu\delta_T^2}{2s_{\min}} \log^2 \left(\frac{\xi s_{\max}}{\mu\delta_T} \right)$$

for the last inner loop, we obtain $\|\tilde{x}_T - x^*\|_2 \leq \frac{4\epsilon}{\sqrt{m s_{\min}}}$ within similar bound of $\tilde{O} \left(\ell + \sqrt{\frac{(\ell+m)L}{\mu}} + m\sqrt{\frac{(\ell+m)\xi}{\mu\epsilon}} \right)$ incremental steps. Therefore, by using the accuracy ϵ'_T in the last inner loop iteration, we obtain $\|\tilde{x}_T - x^*\|_2 \leq \min\{\epsilon, \frac{4\epsilon}{\sqrt{m s_{\min}}}\}$ within at most $\tilde{O} \left(\ell + \sqrt{\frac{(\ell+m)L}{\mu}} + m\sqrt{\frac{(\ell+m)\xi}{\mu\epsilon}} \right)$ expected number of incremental steps. When $s_{\min} \geq \frac{4}{\sqrt{m}}$, the bound achieved with the new ϵ'_T value will be smaller. For example, if the rows of A are generated from i.i.d. sub-Gaussian distributions, the minimal nonzero singular value will be at the constant level, and this is expected to be the case.

As a point of comparison, suppose that we instead use the accelerated proximal full-gradient method [\[8, 33\]](#) in place of \mathcal{M} in [Algorithm 3.1](#). Then, we have the following complexity result.

PROPOSITION 3.5. *Consider applying the nested penalty method ([Algorithm 3.1](#)), with penalty parameter $\xi \geq \bar{\xi}$, and using the accelerated proximal full-gradient method [\[8, 33\]](#) to solve the penalty reformulation subproblems. Let $\{\epsilon_t\}$ be an accuracy sequence for the subproblems satisfying*

$$0 < \epsilon_t \leq \min \left\{ \frac{1}{2} m\mu\delta_t^2 \log^2 \left(\frac{\xi s_{\max}}{\mu\delta_t} \right), m\xi\delta_t \log \left(\frac{\xi s_{\max}}{\mu\delta_t} \right) \right\} \text{ for } t \geq 0,$$

and suppose that the stopping criterion [\(3.2\)](#) is used with step-size $0 < \alpha_t \leq 1/(L + \frac{\xi s_{\max}^2}{4\delta_t})$, i.e., $\tau_t = \tilde{T}_{\mathcal{M}, F_{\xi, \delta_t}}(\sqrt{2\mu\epsilon_t}; \tilde{x}_{t-1}, \alpha_t)$ for all $t \geq 0$. Then, for any desired accuracy $\epsilon \in (0, \frac{m\xi}{\mu} \exp(-2)]$, the number of incremental steps required to satisfy $\|\tilde{x}_T - x^*\|_2 \leq \epsilon$ is upper bounded by

$$\tilde{O} \left((\ell+m) \left(\sqrt{\frac{L}{\mu}} + \sqrt{\frac{m\xi s_{\max}^2}{\mu\epsilon}} \right) \right).$$

Proof. The sequence of δ_t is the same as in the proof of [Proposition 3.3](#). We only need to modify the estimated number of iterations [\(3.5\)](#). Because $\tilde{F}_{\xi, \delta_t}$ is $\left(L + \frac{\xi s_{\max}^2}{4\delta_t} \right)$ -

smooth, by Theorem 6 in [33], the number of iterations will be

$$(3.6) \quad \begin{aligned} \tau_0 &= \tilde{O} \left(\sqrt{\frac{L}{\mu}} + s_{\max} \right). \\ \tau_t &= \tilde{O} \left(\sqrt{\frac{L}{\mu}} + \sqrt{\frac{\xi s_{\max}^2}{\mu \delta_t}} \right), \quad t = 1, \dots, T. \end{aligned}$$

Similar to the proof of Proposition 3.3, we get a solution \tilde{x}_T satisfying $\|\tilde{x}_T - x^*\|_2 \leq \epsilon$ within $\tilde{O} \left(\sqrt{\frac{L}{\mu}} + \sqrt{\frac{m\xi s_{\max}^2}{\mu\epsilon}} \right)$ iterations, i.e., $\tilde{O} \left((\ell + m) \left(\sqrt{\frac{L}{\mu}} + \sqrt{\frac{m\xi s_{\max}^2}{\mu\epsilon}} \right) \right)$ incremental steps. \square

Remark 3.6. Analogous to Remark 3.4, for the accelerated full-gradient method, by using a slightly smaller accuracy parameter in the final inner loop iteration, we can also achieve the bound $\|\tilde{x}_T - x^*\|_2 \leq \frac{4\epsilon}{\sqrt{m s_{\text{pmin}}}}$ within $\tilde{O} \left((\ell + m) \left(\sqrt{\frac{L}{\mu}} + \sqrt{\frac{m\xi s_{\max}^2}{\mu\epsilon}} \right) \right)$ expected incremental steps.

The stochastic method based on the catalyst acceleration requires at least $O \left(\min \left(\sqrt{\frac{L}{\mu}}, \sqrt{\ell + m}, \sqrt{\frac{\ell + m}{m}} s_{\max} \right) \right)$ fewer evaluations on the components of the objective and the constraints and their gradients, compared to the accelerated full-gradient method of Proposition 3.5, which indicates the advantage of application stochastic methods.

As another point of comparison, we can also obtain a complexity result for a static version of Algorithm 3.1 that uses a single value of δ_ϵ , which is determined based on the required accuracy ϵ . The following corollary follows immediately from Lemma 3.2 and Proposition 3.3.

COROLLARY 3.7. *For any desired accuracy $\epsilon \in (0, \frac{m\xi}{\mu} \exp(-2)]$ and penalty parameter $\xi \geq \bar{\xi}$, consider applying proximal SVRG with catalyst acceleration [26, 27] to solve a penalty reformulation subproblem (2.3) with penalty parameter*

$$\delta_\epsilon = \frac{\epsilon}{4m} \left(\log \left(\frac{2m\xi s_{\max}}{\mu\epsilon} \right) \right)^{-1}.$$

Then, the expected number of incremental steps required to satisfy $\|\tilde{x} - x^\|_2 \leq \min \left\{ \epsilon, \frac{4\epsilon}{\sqrt{m s_{\text{pmin}}}} \right\}$ for some \tilde{x} is upper bounded by*

$$(3.7) \quad \tilde{O} \left(\ell + \sqrt{\frac{(\ell + m)L}{\mu}} + m \sqrt{\frac{(\ell + m)\xi}{\mu\epsilon}} \right).$$

One can also obtain a result similar to Proposition 3.5 for a static version that uses the accelerated proximal gradient method. It is instructive to mention that the dynamic nested method, as presented in Algorithm 3.1, has several advantages over the static strategy of Corollary 3.7. First, in Corollary 3.7, δ depends directly on problem parameters and the desired accuracy $\epsilon > 0$, and thus Corollary 3.7 only provides a solution up to but no better than an error of magnitude ϵ . On the other hand, Algorithm 3.1 and Proposition 3.3 provide a dynamic strategy for shrinking δ and thereby guarantee that the sequence $\{\tilde{x}_t\}$ eventually converges to x^* . Second, the

nested structure has the advantage that it enables the use of the screening procedure that we develop in [section 5](#) and performs better in practice as we observe in [section 6](#). Finally, our proofs reveal that the complexity bound in [Proposition 3.3](#) is indeed better than that of [Corollary 3.7](#) by a factor of $O(\log \frac{1}{\epsilon})$, which is hidden behind the \tilde{O} notation. In practice, this improvement is more obvious, because the nested algorithm can address problems with relatively small condition number (due to larger δ) compared with setting $\delta = \tilde{O}(\frac{\epsilon}{m})$ instead at the beginning.

Remark 3.8. When $\epsilon > 0$ is small enough, the third term in [\(3.4\)](#) will be dominant and superior than the complexity for the deterministic method. Ouyang and Xu [\[35\]](#) establish lower bounds on the complexity for first-order methods for the affinely constrained convex problem:

$$(3.8) \quad \begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{s.t.} \quad & a_i^T x = b_i, \quad i = 1, \dots, m. \end{aligned}$$

Theorem 8 in [\[35\]](#) states, for $t = O(m)$, there are instances of f, a_i, b_i in [\(3.8\)](#), such that the iterate obtained from a deterministic first-order method, \bar{x}_t , will have an error $\|\bar{x}_t - x^*\|_2 \geq \Omega(\frac{\|A\|_2 \|\lambda^*\|_2}{\mu t})$. This result indicates a lower bound $\Omega(\min(m^2, m \frac{\|A\|_2 \|\lambda^*\|_2}{\mu \epsilon}))$ of calls on the $(a_i^T x - b_i)$ and their gradients to have a 2-norm error ϵ of the solution for a deterministic first-order algorithm. Compared with this result, notice for any problem with affine constraints $Ax = b$, one can scale the rows of A and the entries of b to make the norms of rows equal to 1; the entries of the dual solution will be scaled respectively. Then, the third term of [\(3.7\)](#) (for $\ell = O(m)$) becomes $\tilde{O}(m \sqrt{\frac{m \|\lambda^*\|_\infty}{\mu \epsilon}})$, when $\xi = \|\lambda^*\|_\infty$, which is better than the deterministic methods when $\epsilon \leq O(\frac{\|A\|_2^2 \|\lambda^*\|_2^2}{m \mu \|\lambda^*\|_\infty})$, considering $\|A\|_2 \geq 1$ and $\|\lambda^*\|_2 \geq \|\lambda^*\|_\infty$.

4. Convergence of the Duality Gap. In this section, we first show that given an approximate solution to the penalty reformulation problem [\(2.3\)](#), we can efficiently obtain an approximate solution to the dual problem [\(2.4\)](#). Then, we extend the results developed in [section 3](#) to derive the complexity of obtaining both primal and dual solutions using stochastic methods, again using the proximal SVRG with catalyst acceleration as our exemplary method.

Toward establishing a result on the duality gap for [\(1.2\)](#) and [\(2.4\)](#), for fixed values of the parameters $\xi, \delta > 0$, we define the following Lagrangian functions:

$$(4.1) \quad \begin{aligned} L(x, \lambda) &:= F(x) + \lambda^T (Ax - b), \\ L_{\xi, \delta}(x, \lambda) &:= F(x) + \lambda^T (Ax - b) - \delta \pi_\xi(\lambda), \end{aligned}$$

for any $x \in \mathbb{R}^n$ and $\lambda \in [0, \xi]^m$ and where

$$\pi_\xi(\lambda) := \sum_{i=1}^m \lambda_i \log \lambda_i + \sum_{i=1}^m (\xi - \lambda_i) \log(\xi - \lambda_i) - m \xi \log(\xi).$$

Note that, when $\lambda \in [0, \xi]^m$, we have $\pi_\xi(\lambda) \in [-m \xi \log 2, 0]$ and thus $L_{\xi, \delta}(x, \lambda) - L(x, \lambda) \in [0, m \xi \delta \log 2]$. Following [Lemma 2.3](#), if the penalty parameter ξ is large enough, i.e., $\xi \geq \tilde{\xi}$ then there exists a dual optimal solution λ^* such that $\xi \geq \tilde{\xi} \geq \|\lambda^*\|_\infty$ and the assumption that $\lambda \in [0, \xi]^m$ is without loss of generality. Now, we

define the following functions:

$$\begin{aligned}
G(\lambda) &:= \min_x L(x, \lambda) = -F^*(-A^T \lambda) - b^T \lambda, \\
G_{\xi, \delta}(\lambda) &:= \min_x L_{\xi, \delta}(x, \lambda) = G(\lambda) - \delta \pi_{\xi}(\lambda), \\
\Phi(x) &:= \max_{\lambda \in [0, \xi]^m} L(x, \lambda), \\
\Phi_{\xi, \delta}(x) &:= \max_{\lambda \in [0, \xi]^m} L_{\xi, \delta}(x, \lambda),
\end{aligned} \tag{4.2}$$

for any $x \in \mathbb{R}^n$ and $\lambda \in [0, \xi]^m$. Notice also that, for any $x \in \mathbb{R}^n$, we have that

$$\begin{aligned}
&\arg \max_{\lambda \in [0, \xi]^m} L_{\xi, \delta}(x, \lambda) = \\
&\xi \left(\frac{\exp((a_1^T x - b_1)/\delta)}{1 + \exp((a_1^T x - b_1)/\delta)}, \dots, \frac{\exp((a_m^T x - b_m)/\delta)}{1 + \exp((a_m^T x - b_m)/\delta)} \right)^T,
\end{aligned} \tag{4.3}$$

and hence

$$\Phi_{\xi, \delta}(x) = \max_{\lambda \in [0, \xi]^m} L_{\xi, \delta}(x, \lambda) = F_{\xi, \delta}(x),$$

for all $x \in \mathbb{R}^n$, which follows from (2.2). In other words, our penalty reformulation with the softplus penalty can be considered as a version of Nesterov's smoothing technique inside the Lagrangian function. With this observation, we have the following proposition, which demonstrates how to construct an approximately optimal dual solution from an approximately optimal primal solution. Recall that $x_{\xi, \delta}^*$ denotes the unique optimal solution of (2.3) for given $\xi \geq 0$ and $\delta \geq 0$ and Λ^* denotes the set of dual optimal solutions of (2.4).

PROPOSITION 4.1. *Let $\epsilon > 0$ be given, and suppose that $\xi \geq \bar{\xi} = \inf_{\lambda^* \in \Lambda^*} \|\lambda^*\|_{\infty}$ and $0 < \delta \leq \frac{\epsilon}{2m \log 2}$. For a given $\hat{x}_{\xi, \delta}$, define $\hat{\lambda}_{\xi, \delta}$ by*

$$\hat{\lambda}_{\xi, \delta} = \xi \left(\frac{\exp((a_1^T \hat{x}_{\xi, \delta} - b_1)/\delta)}{1 + \exp((a_1^T \hat{x}_{\xi, \delta} - b_1)/\delta)}, \dots, \frac{\exp((a_m^T \hat{x}_{\xi, \delta} - b_m)/\delta)}{1 + \exp((a_m^T \hat{x}_{\xi, \delta} - b_m)/\delta)} \right)^T.$$

If $\hat{x}_{\xi, \delta}$ satisfies $\|\hat{x}_{\xi, \delta} - x_{\xi, \delta}^\|_2 \leq \sqrt{\frac{\delta \epsilon}{m}} \min(1, \frac{4\mu\delta}{m})$, then it holds $G(\lambda^*) - G(\hat{\lambda}_{\xi, \delta}) \leq \xi \epsilon$. In addition, if $F_{\xi, \delta}(\hat{x}_{\xi, \delta}) - F_{\xi, \delta}(x_{\xi, \delta}^*) \leq \xi \epsilon$, then it holds $G(\lambda^*) - G(\hat{\lambda}_{\xi, \delta}) \leq F_{\xi, 0}(\hat{x}_{\xi, \delta}) - G(\hat{\lambda}_{\xi, \delta}) \leq 2\xi \epsilon$.*

Proof. Letting $\lambda_{\xi, \delta}^* = \arg \max_{\lambda \in [0, \xi]^m} G_{\xi, \delta}(\lambda)$ and $\lambda^* = \arg \max_{\lambda \in [0, \xi]^m} G(\lambda)$, we see that (x^*, λ^*) and $(x_{\delta}^*, \lambda_{\delta}^*)$ are the saddle pairs of $L(x, \lambda)$ and $L_{\xi, \delta}(x, \lambda)$, respectively. Then,

$$G(\lambda^*) \leq G_{\xi, \delta}(\lambda^*) \leq G_{\xi, \delta}(\lambda_{\xi, \delta}^*) \leq G(\lambda_{\xi, \delta}^*) + m\xi\delta \log 2 \leq G(\lambda^*) + m\xi\delta \log 2.$$

Hence, when $\delta \leq \frac{\epsilon}{2m \log 2}$, we have $G(\lambda^*) - G(\lambda_{\xi, \delta}^*) \leq \frac{\xi \epsilon}{2}$.

Suppose we obtain an approximate solution $\hat{x}_{\xi, \delta}$, and the corresponding $\hat{\lambda}_{\xi, \delta} = \arg \max_{\lambda \in [0, \xi]^m} L_{\xi, \delta}(\hat{x}_{\xi, \delta}, \lambda)$. Then, because $p_{\delta}'' \in [0, \frac{1}{4\delta}]$, we have

$$\|\hat{\lambda}_{\xi, \delta} - \lambda_{\xi, \delta}^*\|_2 \leq \frac{\sqrt{m}}{4\delta} \|\hat{x}_{\xi, \delta} - x_{\xi, \delta}^*\|_2.$$

Because $F(x)$ is μ -strongly convex, $F^*(y)$ is $\frac{1}{\mu}$ -smooth. Since $\|A\|_2 \leq \sqrt{m}$,

$$\|\nabla G(\hat{\lambda}_{\xi,\delta}) - \nabla G(\lambda_{\xi,\delta}^*)\|_2 \leq \frac{m}{\mu} \|\hat{\lambda}_{\xi,\delta} - \lambda_{\xi,\delta}^*\|_2.$$

And then,

$$\begin{aligned} & \|\nabla G_{\xi,\delta}(\hat{\lambda}_{\xi,\delta})\|_2 = \|\nabla G_{\xi,\delta}(\hat{\lambda}_{\xi,\delta}) - \nabla G_{\xi,\delta}(\lambda_{\xi,\delta}^*)\|_2 \\ & \leq \|\nabla G(\hat{\lambda}_{\xi,\delta}) - \nabla G(\lambda_{\xi,\delta}^*)\|_2 + \delta \|\nabla \pi_{\xi}(\hat{\lambda}_{\xi,\delta}) - \nabla \pi_{\xi}(\lambda_{\xi,\delta}^*)\|_2 \\ & \leq \frac{m}{\mu} \|\hat{\lambda}_{\xi,\delta} - \lambda_{\xi,\delta}^*\|_2 + \|A(\hat{x}_{\xi,\delta} - x_{\xi,\delta}^*)\|_2 \\ & \leq \left(\frac{m\sqrt{m}}{4\mu\delta} + \sqrt{m} \right) \|\hat{x}_{\xi,\delta} - x_{\xi,\delta}^*\|_2. \end{aligned}$$

Since $G(\lambda)$ is concave, $G_{\xi,\delta}(\lambda)$ is $\frac{4\delta}{\xi}$ -strongly concave. Therefore,

$$G_{\xi,\delta}(\lambda_{\xi,\delta}^*) - G_{\xi,\delta}(\hat{\lambda}_{\xi,\delta}) \leq \frac{\xi}{8\delta} \|\nabla G_{\xi,\delta}(\hat{\lambda}_{\xi,\delta})\|_2^2,$$

which indicates that when $\|\hat{x}_{\xi,\delta} - x_{\xi,\delta}^*\|_2 \leq \sqrt{\frac{\delta\epsilon}{m}} \min(1, \frac{4\mu\delta}{m})$,

$$G_{\xi,\delta}(\lambda_{\xi,\delta}^*) - G_{\xi,\delta}(\hat{\lambda}_{\xi,\delta}) \leq \frac{\xi\epsilon}{2},$$

and then, when additionally $\delta \leq \frac{\epsilon}{2m \log 2}$,

$$\begin{aligned} G(\lambda^*) & \geq G(\hat{\lambda}_{\xi,\delta}) \geq G_{\xi,\delta}(\hat{\lambda}_{\xi,\delta}) - \frac{\xi\epsilon}{2} \\ & \geq G_{\xi,\delta}(\lambda_{\xi,\delta}^*) - \xi\epsilon \geq G_{\xi,\delta}(\lambda^*) - \xi\epsilon \geq G(\lambda^*) - \xi\epsilon. \end{aligned}$$

In other words, $\lambda_{\xi,\delta,\epsilon}$ is a $(\xi\epsilon)$ -solution to the dual problem (2.4).

Now when $F_{\xi,\delta}(\hat{x}_{\xi,\delta}) - F_{\xi,\delta}(x_{\xi,\delta}^*) \leq \xi\epsilon$,

$$\begin{aligned} G(\lambda^*) & = F(x^*) = F_{\xi,0}(x^*) \leq F_{\xi,0}(\hat{x}_{\xi,\delta}) \leq F_{\xi,0}(\hat{x}_{\xi,\delta}) \\ & \leq F_{\xi,\delta}(\hat{x}_{\xi,\delta}) \leq G_{\xi,\delta}(\lambda_{\xi,\delta}^*) + \xi\epsilon \leq G(\lambda_{\xi,\delta,\epsilon}) + 2\xi\epsilon, \end{aligned}$$

and then $G(\lambda^*) - G(\hat{\lambda}_{\xi,\delta}) \leq F_{\xi,0}(\hat{x}_{\xi,\delta}) - G(\hat{\lambda}_{\xi,\delta}) \leq 2\xi\epsilon$. \square

Notice that the gap $F_{\xi,0}(\hat{x}_{\xi,\delta}) - G(\hat{\lambda}_{\xi,\delta})$, which we bound in [Proposition 4.1](#), is the duality gap between the penalized problem $\min_x F_{\xi,0}(x)$, and its dual

$$\begin{aligned} \max_{\lambda} \quad & G(\lambda) := -F^*(-A^T \lambda) - b^T \lambda, \\ \text{s.t.} \quad & 0 \leq \lambda \leq \xi, \end{aligned}$$

which have the same optimal costs with the original primal and dual problems. Hence, we report it as the duality gap in the numerical experiments in [Section 6](#). In addition, since the duality gap $F_{\xi,\delta}(\hat{x}_{\xi,\delta}) - G_{\xi,\delta}(\hat{\lambda}_{\xi,\delta})$ upper bounds $F_{\xi,\delta}(\hat{x}_{\xi,\delta}) - F_{\xi,\delta}^*$, we can use this gap as the stopping criterion (e.g., instead of the norm of the gradient mapping) in [Algorithm 3.1](#). One can then develop results analogous to [Lemma 3.2](#) and [Proposition 3.3](#) that use this duality gap in place of the norm of the gradient mapping.

We are now ready to state our main theorem on the primal and dual convergence of the nested penalty method (Algorithm 3.1) using proximal SVRG with catalyst acceleration [26, 27], which combines the results of Proposition 3.3 and Proposition 4.1. Note that, given the sequence $\{\tilde{x}_t\}$ output by Algorithm 3.1, we define a sequence of dual solutions $\{\tilde{\lambda}_t\}$ by

$$\tilde{\lambda}_t := \xi \left(\frac{\exp((a_1^T \tilde{x}_t - b_1)/\delta_t)}{1 + \exp((a_1^T \tilde{x}_t - b_1)/\delta_t)}, \dots, \frac{\exp((a_m^T \tilde{x}_t - b_m)/\delta_t)}{1 + \exp((a_m^T \tilde{x}_t - b_m)/\delta_t)} \right)^T.$$

THEOREM 4.2. *Consider applying the nested penalty method (Algorithm 3.1), with penalty parameter $\xi \geq \bar{\xi}$, and using proximal SVRG with catalyst acceleration [26, 27] to solve the penalty reformulation subproblems. Let $\{\epsilon_t\}$ be an accuracy sequence for the subproblems satisfying*

$$\epsilon_t = \min \left\{ \frac{1}{2} m \mu \delta_t^2 \log^2 \left(\frac{\xi s_{\max}}{\mu \delta_t} \right), m \xi \delta_t \log \left(\frac{\xi s_{\max}}{\mu \delta_t} \right) \right\} \text{ for } t \geq 0,$$

and suppose that the stopping criterion (3.2) is used with step-size $\alpha_t = 1/(L + \frac{\xi s_{\max}^2}{4\delta_t})$, i.e., $\tau_t = \tilde{T}_{\mathcal{M}, F_{\xi, \delta_t}}(\sqrt{2\mu\epsilon_t}; \tilde{x}_{t-1}, \alpha_t)$ for all $t \geq 0$. Then, for any desired accuracy $\epsilon \in (0, \frac{m\xi}{\mu} \exp(-2)]$, the expected number of incremental steps required to satisfy $\|\tilde{x}_T - x^*\|_2 \leq \epsilon$ and $G(\lambda^*) - G(\tilde{\lambda}_T) \leq F_{\xi, 0}(\tilde{x}_T) - G(\tilde{\lambda}_T) \leq \xi\epsilon$ is upper bounded by

$$\tilde{O} \left(\ell + \sqrt{\frac{(\ell + m)L}{\mu}} + m \sqrt{\frac{(\ell + m)\xi}{\mu\epsilon}} \right).$$

Remark 4.3. Analogous to Remark 3.4, by using a smaller accuracy parameter in the final inner loop iteration, in Theorem 4.2, one can also achieve the bound $\|\tilde{x}_T - x^*\|_2 \leq \frac{4\epsilon}{\sqrt{m s_{\min}}} \text{ within } \tilde{O} \left((\ell + m) \left(\sqrt{\frac{L}{\mu}} + \sqrt{\frac{m\xi s_{\max}^2}{\mu\epsilon}} \right) \right)$ expected incremental steps.

5. Screening Procedure for the Nested Penalty Method. In this section, we describe an additional enhancement to the dynamic nested penalty method (Algorithm 3.1) that uses a *safe* screening procedure to remove constraints that are not tight. In particular, the KKT conditions of problem (1.2) imply that if the i -th constraint is not tight, i.e., it is not active at x^* , $a_i^T x^* - b_i < 0$, then the corresponding dual variable $\lambda_i^* = 0$ and removing this constraint from (1.2) will not change the optimal solution x^* . Recall that Proposition 2.6 provides a bound on the norm of $z_{\xi, \delta} = A(x_{\xi, \delta}^* - x_{\xi, 0}^*)$, i.e., the distance between the slack variables at $x_{\xi, \delta}^*$ and $x_{\xi, 0}^*$. Based on this bound, if $\xi \geq \bar{\xi}$ and if the slack $b_i - a_i^T x_{\xi, \delta}^*$ is large enough, then one can safely conclude that $a_i^T x^* - b_i < 0$ and hence the i -th constraint can be dropped from the problem. One can also apply a similar procedure to the approximate solutions \tilde{x}_t to the penalty reformulation problems (2.3) computed at each outer loop iteration of the nested method (Algorithm 3.1). We refer to dropping such constraints as a screening procedure, and we prove that it is *safe*, i.e., does not remove constraints that are tight at an optimal solution, by combining a result similar to Proposition 2.6 with a bound on the distance between the slack variables at the approximate penalty solution \tilde{x}_t and the exact solution x_{ξ, δ_t}^* . Note that safe screening procedures to eliminate redundant variables and constraints are prevalent in machine learning literature (see, e.g., [17, 4]).

Let us now formally introduce these ideas by considering the following version of the original problem (1.2) based on a subset of the constraint indices $\hat{S} \subseteq S^F := \{1, \dots, m\}$:

$$(5.1) \quad \begin{aligned} \min_{x \in \mathbb{R}^n} \quad & F(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} f_i(x) + \psi(x) \\ \text{s.t.} \quad & a_i^T x \leq b_i, \quad i \in \hat{S}, \end{aligned}$$

as well as the corresponding the penalty reformulation:

$$(5.2) \quad \min_{x \in \mathbb{R}^n} F_{\hat{S}, \xi, \delta}(x) := F(x) + \xi \sum_{i \in \hat{S}} p_{\delta}(a_i^T x - b_i).$$

Let $x_{\hat{S}, \xi, \delta}^*$ denote the unique optimal solution of (5.2) for a given $\xi \geq 0$, $\delta \geq 0$. Furthermore, let $S^A := \{i \in S^F : a_i^T x^* - b_i = 0\}$ denote the set of indices of active (tight) constraints at the optimal solution x^* of (1.2). Also, for $\Delta \geq 0$, we let $S_{\Delta} := \{i \in S^F : a_i^T x^* - b_i \in [-\Delta, 0]\}$ and $m_{\Delta} := \text{card}(S_{\Delta})$. Notice, $S^A \subseteq S_{\Delta} \subseteq S^F$ and $m^A := \text{card}(S^A) \leq m_{\Delta} \leq m$ for all $\Delta \geq 0$. Finally, let A_S denote the submatrix consisting of the rows of A with indices in S . Then, the Lemma below follows from applying Lemma 2.3 on (5.1) and (5.2).

LEMMA 5.1. *If $\xi \geq \bar{\xi}$ and $S^A \subseteq \hat{S} \subseteq S^F$, then it holds that $x_{\hat{S}, \xi, 0}^* = x^*$.*

Similarly, we obtain Proposition 5.2 as a corollary of Proposition 2.6.

PROPOSITION 5.2. *For $\xi > 0$ and $\delta \in [0, \frac{s_{\max}^2(A_S)\xi}{\mu} \exp(-2)]$, it holds that*

$$\|z_{S, \xi, \delta}\|_2 = \|A_S(x_{\hat{S}, \xi, \delta}^* - x_{\hat{S}, \xi, 0}^*)\|_2 \leq \sqrt{\text{card}(S)} \delta \log \left(\frac{s_{\max}^2(A_S)\xi}{\mu\delta} \right).$$

Algorithm 5.1 presents the nested penalty method that includes the screening procedure at every outer loop iteration. Lemma 5.3 presents a convergence results for Algorithm 5.1 and verifies that the screening procedure safely removes only constraints that are guaranteed to not be tight at the optimal solution x^* .

LEMMA 5.3. *Consider applying the generic nested penalty method (Algorithm 5.1) using penalty parameter $\xi \geq \bar{\xi}$ and initial smoothness parameter $\delta_0 \leq \frac{\xi}{\mu} \exp(-2)$. Let $\{\epsilon_t\}$ be an accuracy sequence for the penalty reformulation subproblems satisfying*

$$0 < \epsilon_t \leq \frac{1}{2} \hat{m}_t \mu \delta_t^2 \log^2 \left(\frac{\xi s_{\max}}{\mu \delta_t} \right) \quad \text{for } t \geq 0,$$

and suppose that the stopping criterion (3.2) is used with step-size $0 < \alpha_t \leq 1/(L + \frac{\xi s_{\max}^2}{4\delta_t})$, i.e., $\tau_t = \tilde{T}_{\mathcal{M}, F_{\xi, \delta_t}}(\sqrt{2\mu\epsilon_t}; \tilde{x}_{t-1}, \alpha_t)$ for all $t \geq 0$. Then, for any $t \geq 0$, it holds that:

$$(5.3) \quad \begin{aligned} \|\tilde{x}_t - x^*\|_2 &\leq \left(\min \left\{ \hat{m}_t, \frac{4\sqrt{\hat{m}_t}}{s_{\text{pmin}}} \right\} + \sqrt{\hat{m}_t} \right) \delta_t \log \left(\frac{\xi s_{\max}}{\mu \delta_t} \right) \\ &= \frac{\left(\min \left\{ \hat{m}_t, \frac{4\sqrt{\hat{m}_t}}{s_{\text{pmin}}} \right\} + \sqrt{\hat{m}_t} \right) \delta_0}{\eta^t} \left[t \log(\eta) + \log \left(\frac{\xi s_{\max}}{\mu \delta_0} \right) \right]. \end{aligned}$$

In addition, it holds that $S^A \subseteq \hat{S}_{t+1} \subseteq S_{\hat{\Delta}_t} \subseteq S_{\Delta_t}$, where

$$\hat{\Delta}_t = 4\sqrt{\hat{m}_t} \delta_t \log \left(\frac{\hat{m}_t \xi}{\mu \delta_t} \right), \quad \Delta_t = 4\sqrt{m} \delta_t \log \left(\frac{m \xi}{\mu \delta_t} \right).$$

Algorithm 5.1 Generic Nested Penalty Method with Screening Procedure

Parameters: Penalty parameter $\xi > 0$, smoothness scaling parameter $\eta > 1$, and generic unconstrained convex optimization method \mathcal{M} .

Initialize: $\hat{x}_{-1} \in \mathbb{R}^n$, $\delta_0 > 0$, $\hat{S}_0 = S^F$, $\hat{m}_0 = m$.

for $t = 0, 1, \dots, T$ **do**

1. Apply method \mathcal{M} on the penalty reformulation problem (5.2), initialized at \tilde{x}_{t-1} and for $\tau_t \geq 1$ total iterations, to obtain an approximate solution

$$\hat{x}_t \approx \arg \min_{x \in \mathbb{R}^n} \left\{ F_{\hat{S}_t, \xi, \delta_t}(x) := F(x) + \xi \sum_{i \in \hat{S}_t} p_{\delta_t}(a_i^T x - b_i) \right\}.$$

2. Choose step-size $\alpha_t > 0$ and set $\tilde{x}_t \leftarrow \text{prox}_{\psi}(\hat{x}_t - \alpha_t \nabla \tilde{F}_{\hat{S}_t, \xi, \delta_t}(\hat{x}_t); \alpha_t)$.
3. Screening procedure: set $\hat{S}_{t+1} \leftarrow \{i \in \hat{S}_t : a_i^T \tilde{x}_t - b_i \geq -2\sqrt{\hat{m}_t} \delta_t \log(\frac{\hat{m}_t \xi}{\mu \delta_t})\}$ and update $\hat{m}_{t+1} \leftarrow \text{card}(\hat{S}_{t+1})$.
4. Update $\delta_{t+1} \leftarrow \delta_t / \eta$.

end for

Proof. We only need to prove the second part as the first part is completely analogous to the proof of Lemma 3.2. Suppose $S^A \subseteq \hat{S}_t$, then by Proposition 5.2, for all $i \in \hat{S}_t$, we have

$$|a_i^T x_{\hat{S}_t, \xi, \delta_t}^* - a_i^T x^*| \leq \|A_{\hat{S}_t}(x_{\hat{S}_t, \xi, \delta_t}^* - x^*)\|_2 \leq \sqrt{\hat{m}_t} \delta_t \log \left(\frac{s_{\max}^2(A_{\hat{S}_t}) \xi}{\mu \delta_t} \right).$$

Also, by μ -strongly convexity of $F_{\hat{S}_t, \xi, \delta_t}$, $\|\tilde{x}_t - x_{\hat{S}_t, \xi, \delta_t}^*\|_2 \leq \sqrt{\hat{m}_t} \delta_t \log \left(\frac{\hat{m}_t \xi}{\mu \delta_t} \right)$. And then, since $\|a_i\|_2 = 1$ (Assumption 2.2),

$$|a_i^T \tilde{x}_t - a_i^T x^*| \leq 2\sqrt{\hat{m}_t} \delta_t \log \left(\frac{\hat{m}_t \xi}{\mu \delta_t} \right),$$

which means $S^A \subseteq \hat{S}_{t+1} \subseteq S_{\hat{\Delta}_t} \subseteq S_{\Delta_t}$. By induction, since $S^A \subseteq \hat{S}_0 = S^F$, the result holds for all $t = 0, \dots, T-1$. \square

As mentioned, the first part of Lemma 5.3 states the convergence of the nested method, as a corollary of Lemma 3.2. The second part shows the safety of the screening procedure, as a result of the choice of ϵ_t and Proposition 5.2.

Similar to Proposition 4.1, we can also obtain a dual solution when solving the penalty reformulation (5.2).

PROPOSITION 5.4. *Let $\epsilon > 0$ be given, and suppose that $\xi \geq \bar{\xi} = \inf_{\lambda^* \in \Lambda^*} \|\lambda^*\|_\infty$, $S^A \subseteq \hat{S} \subseteq S^F$, $\hat{m} = \text{card}(\hat{S})$, and $0 < \delta \leq \frac{\epsilon}{2\hat{m} \log 2}$. For a given $\hat{x}_{\hat{S}, \xi, \delta}$, define $\hat{\lambda}_{\hat{S}, \xi, \delta}$ by*

$$\hat{\lambda}_{\hat{S}, \xi, \delta, i} = \begin{cases} \xi \frac{\exp((a_i^T \hat{x}_{\hat{S}, \xi, \delta} - b_i)/\delta)}{1 + \exp((a_i^T \hat{x}_{\hat{S}, \xi, \delta} - b_i)/\delta)}, & i \in \hat{S} \\ 0, & i \in S^F \setminus \hat{S}. \end{cases}$$

If $\hat{x}_{\hat{S}, \xi, \delta}$ satisfies $\|\hat{x}_{\hat{S}, \xi, \delta} - x_{\hat{S}, \xi, \delta}^\|_2 \leq \sqrt{\frac{\delta \epsilon}{\hat{m}}} \min(1, \frac{4\mu \delta}{\hat{m}})$, then it holds that $G(\lambda^*) -$*

$G(\hat{\lambda}_{\hat{S},\xi,\delta}) \leq \xi\epsilon$. In addition, if $F_{\hat{S},\xi,\delta}(\hat{x}_{\hat{S},\xi,\delta}) - F_{\hat{S},\xi,\delta}(x_{\hat{S},\xi,\delta}^*) \leq \xi\epsilon$, then it holds that $G(\lambda^*) - G(\hat{\lambda}_{\hat{S},\xi,\delta}) \leq F_{\hat{S},\xi,0}(\hat{x}_{\hat{S},\xi,\delta}) - G(\hat{\lambda}_{\hat{S},\xi,\delta}) \leq 2\xi\epsilon$.

Let $\varsigma := \min_{i \in S^F \setminus S^A} (b_i - a_i^T x^*)$ denote the smallest slackness of inactive constraints at x^* , which naturally arises as part of the complexity analysis of our screening procedure. In particular, we obtain the following result by combining [Proposition 3.3](#), [Lemma 5.3](#), and [Proposition 5.4](#).

PROPOSITION 5.5. *Consider applying the nested penalty method with screening procedure ([Algorithm 5.1](#)), with penalty parameter $\xi \geq \bar{\xi}$ and initial smoothness parameter $\delta_0 = \frac{\xi}{\mu} \exp(-2)$, and using proximal SVRG with catalyst acceleration [[26](#), [27](#)] to solve the penalty reformulation subproblems. Let $\{\epsilon_t\}$ be an accuracy sequence for the subproblems satisfying*

$$\epsilon_t = \min \left\{ \frac{1}{2} \hat{m}_t \mu \delta_t^2 \log^2 \left(\frac{\xi s_{\max}}{\mu \delta_t} \right), \hat{m}_t \xi \delta_t \log \left(\frac{\xi s_{\max}}{\mu \delta_t} \right) \right\} \text{ for } t \geq 0,$$

and suppose that the stopping criterion based on the gradient mapping (3.2) is used with step-size $\alpha_t = 1/(L + \frac{\xi s_{\max}^2}{4\delta_t})$, i.e., $\tau_t = \tilde{T}_{\mathcal{M}, F_{\xi,\delta_t}}(\sqrt{2\mu\epsilon_t}; \tilde{x}_{t-1}, \alpha_t)$ for all $t \geq 0$. Then, for any desired accuracy $\epsilon \in (0, \frac{m\xi}{\mu} \exp(-2)]$, the expected number of incremental steps required to satisfy $\|\tilde{x}_T - x^*\|_2 \leq \epsilon$ and $G(\lambda^*) - G(\tilde{\lambda}_T) \leq F_{S^A, \xi, 0}(\tilde{x}_T) - G(\tilde{\lambda}_T) \leq \xi\epsilon$ is upper bounded by

$$\tilde{O} \left(\ell + \sqrt{\frac{(\ell+m)L}{\mu}} + \sqrt{\frac{m(\ell+m)\xi}{\mu\varsigma}} + m^A \sqrt{\frac{(\ell+m^A)\xi}{\mu\epsilon}} \right).$$

Proof. Suppose we use the sequence of δ : $\delta_0 > \dots > \delta_T$, where $\delta_0 = \frac{\xi}{\mu} \exp(-2)$, $\delta_{t-1}/\delta_t = \eta > 1$, $t = 1, \dots, T$, and T is the smallest positive integer such that the following inequality holds

$$2\hat{m}_T \delta_T \log \left(\frac{\xi \hat{m}_T}{\mu \delta_T} \right) \leq \epsilon.$$

Then, the estimation of the complexity is similar with the proof for [Proposition 3.3](#). Notice

$$\begin{aligned} (5.4) \quad \mathbb{E}\tau_0 &= \tilde{O} \left(\ell + m + \sqrt{\frac{(m+\ell)L + mL_0}{\mu}} \right) = \tilde{O} \left(\ell + \sqrt{\frac{(\ell+m)L}{\mu}} + \sqrt{(\ell+m)m} \right). \\ \mathbb{E}\tau_t &= \tilde{O} \left(\ell + \hat{m}_t + \sqrt{\frac{(\hat{m}_t+\ell)L + \hat{m}_t L_t}{\mu}} \right) \\ &= \tilde{O} \left(\ell + \sqrt{\frac{(\ell+\hat{m}_t)L}{\mu}} + \sqrt{\frac{(\ell+\hat{m}_t)\hat{m}_t \xi}{\mu \delta_t}} \right), \quad t = 1, \dots, T. \end{aligned}$$

By [Lemma 5.3](#), when $4\sqrt{m}\delta_t \log \left(\frac{m\xi}{\mu\delta_t} \right) \leq \varsigma$, $\hat{S}_{t+1} = S^A$. Hence,

$$\mathbb{E}\tau_t \leq \tilde{O} \left(\ell + \sqrt{\frac{(\ell+m)L}{\mu}} + \max \left(\sqrt{\frac{m(\ell+m)\xi}{\mu\varsigma}}, m^A \sqrt{\frac{(\ell+m^A)\xi}{\mu\epsilon}} \right) \right).$$

The total complexity follows from the summation of $T \leq O(\log(\frac{m\xi}{\mu\epsilon}))$ estimations. \square

As compared to [Proposition 3.3](#), [Proposition 5.5](#) introduces an additional term, which is $\tilde{O}\left(\sqrt{\frac{m(\ell+m)\xi}{\mu\varsigma}}\right)$, to the complexity bound, and which naturally depends inversely proportional to the square root of ς . On the other hand, the last term of the complexity bound of [Proposition 3.3](#) is $\tilde{O}\left(m\sqrt{\frac{(\ell+m)\xi}{\mu\epsilon}}\right)$, whereas [Proposition 5.5](#) replaces m in this term with m^A , the number of *active* constraints at x^* , which can be much smaller. In particular, when t goes to infinity, \hat{S}_t will converge to S^A . In other words, in the long run, only the tight constraints will be maintained in the penalty reformulations in [Algorithm 5.1](#). Furthermore, if $m^A \neq 0$, the $\tilde{O}\left(m^A\sqrt{\frac{(\ell+m^A)\xi}{\mu\epsilon}}\right)$ term in [Proposition 5.5](#) will dominate the complexity for small enough ϵ , indicating a substantial improvement due to the incorporation of the screening procedure, especially when $m^A \ll m$.

6. Numerical Results. In this section, we present the our numerical experiments performed to test the empirical effectiveness and safety of the proposed algorithms. The first problem used for the experiments is quadratic programming, i.e., the minimization of a convex quadratic function over linear inequality constraints; the second one is a support vector machine problem. We mainly focus on three algorithms for solving the problems: **SASC-SGD**: the SASC algorithm in [\[16\]](#); **Nested-SGD**: [Algorithm 3.1](#) with SGD for solving the sub-problems; **Nested-SGDM**: [Algorithm 3.1](#) with the momentum method (see [\(B.1\)](#) in [Appendix B](#)) for solving the sub-problems. For the simplicity of implementation, we use SGD instead of SVRG for most of the experiments, and the momentum method instead of the catalyst accelerated SVRG. We will also compare their performance with the parameter updating scheme in [\[31\]](#), and the static penalty method **Static-SGDM**. We do not to include the comparison with algorithms in [\[29\]](#) because it has a similar penalty structure with [\[16\]](#), while the latter contains more information about implementation. For the second experiment, we also include comparisons with **Nested-SVRG**: [Algorithm 3.1](#) with SVRG solving the sub-problems, and **Nested-SVRG-Screening**: [Algorithm 5.1](#) with SVRG solving the sub-problems, to verify the safety and efficiency of the screening procedure in [section 5](#). Further details about experiments are given in [Appendix B](#).

6.1. Quadratic Programming. We consider the following QP problem

$$(6.1) \quad \begin{aligned} \min_{x \in \mathbb{R}^n} \quad & \frac{1}{2\ell} \sum_{i=1}^{\ell} (\phi_i^T x - y_i)^2 + \frac{w}{2} \|x\|_2^2 \\ \text{s.t.} \quad & a_j^T x \leq b_j, \quad j = 1, \dots, m, \end{aligned}$$

where $\Phi = (\phi_1, \dots, \phi_\ell)^T \in \mathbb{R}^{\ell \times n}$, $y = (y_1, \dots, y_\ell)^T \in \mathbb{R}^\ell$, $A \in (a_1, \dots, a_m)^T \in \mathbb{R}^{m \times n}$, and $b = (b_1, \dots, b_m)^T \in \mathbb{R}^m$ are generated randomly. Namely, the entries y , b_0 are all generated i.i.d. from the standard normal distribution. We then let b be the absolute value vector of b_0 to ensure the feasible region is nonempty. The entries of A are drawn from i.i.d. $\text{Normal}(0,1)$, and then normalized so that $\|a_i\|_2 = 1$, $i = 1, \dots, m$. We let $\ell = m = n = 100$, and the regularization parameter $w = 0.1$. All algorithms are initialized at point $x_0 = 0$.

For this experiment, we compare the relative error of the solutions found by the algorithms, i.e., $\frac{\|x - x_{\text{cvx}}\|_2}{\|x_{\text{cvx}}\|_2}$, where x_{cvx} is the solution obtained from CVX [\[14, 19\]](#) calling [gurobi](#) [\[20\]](#). We first select the best parameters for the three methods based on 20

simulations (see Appendix B). Figure 2 shows the box plot of the relative errors for the three methods after $1e7$ iterations. We could see that `Nested-SGDM` has the fastest convergence after $1e7$ iterations, and then it's `Nested-SGD`. Figure 3 shows a similar box plot for `Static-SGDM`, where δ is chosen as the initial, the last, and their geometric average in `Nested-SGDM`.

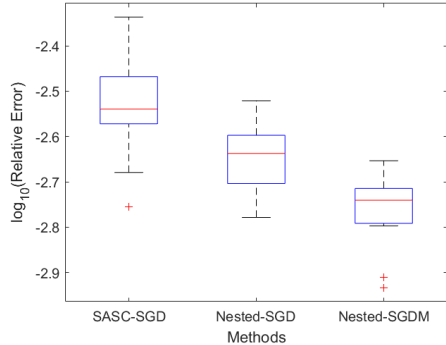


FIG. 2. Relative errors of 3 methods after $1e7$ iterations in 20 simulations.

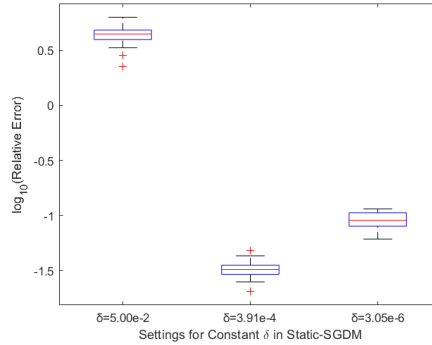


FIG. 3. Relative errors of `Static-SGDM` after $1e7$ iterations in 20 simulations.

Figure 4 shows the convergence of the relative error for the three methods with the best parameter settings and `Static-SGDM` (using the geometric average δ), respectively. In Figure 4 we see that using the momentum method in the nested algorithm to solve the sub-problems is more efficient compared to SGD. Algorithm 3.1 performs slightly better with SGD compared to using SASC to solve the sub-problems. For this run, even though $\|\lambda_{\text{cvx}}\|_\infty = 0.027$, we use $\xi = 1$, which indicates that, in practice, ξ does not need to be a close upper bound of $\|\lambda^*\|_\infty$. It is also shown in Appendix B that the `Nested-SGD` and `Nested-SGDM` are more robust than `SASC-SGD` when the parameters are tuned. Note that `Static-SGDM` does not converge because the accuracy of the penalty reformulation is directly proportional to δ . On the other hand, if we set δ in `Static-SGDM` to be the last δ in `Nested-SGDM`, as shown in Figure 3, its performance gets worse due to the large condition number, which is inversely proportional to δ . These results confirm the advantage of the nested algorithm.

The availability of duality gap is yet another advantage of our approach. Figure 5 shows the convergence of the minimal duality gap for the three algorithms. During iterations, whereas the primal solutions are not necessarily feasible, the dual solutions obtained by (4.3) are feasible. Hence, we use $F_{\xi,0}(x_t) - G(\lambda_t)$ as the duality gap (see section 4) and plot $\min_{1 \leq s \leq t} [F_{\xi,0}(x_s) - G(\lambda_s)]$ with respect to t in Figure 5. In this figure it is clear that the proposed nested algorithms perform better than the `Static-SGDM`.

6.2. Support Vector Machine. In this subsection, we consider the following hard margin support vector machine (SVM) problem for classification:

$$(6.2) \quad \begin{aligned} \min_{x \in \mathbb{R}^n} \quad & \frac{1}{2} \|x\|_2^2 \\ \text{s.t.} \quad & b_i \langle a_i, x \rangle \geq 1, \quad i = 1, \dots, m, \end{aligned}$$

where $\{a_1, \dots, a_m\} \subset \mathbb{R}^n$ and $\{b_1, \dots, b_m\} \subset \{-1, 1\}^n$ are the features and labels of the observations. For the experiment, we use the `mushrooms` dataset of libsvm database

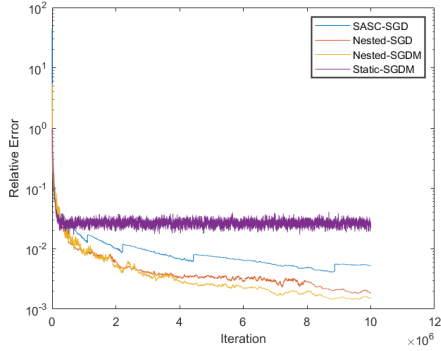


FIG. 4. Convergence of relative error.

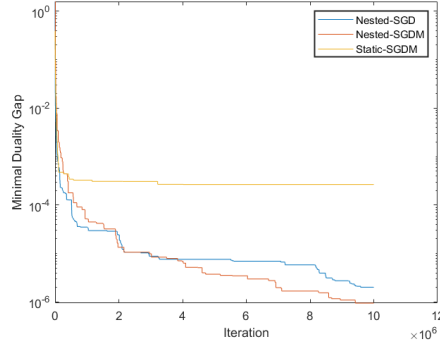


FIG. 5. Convergence of duality gap.

[13], with 8,124 observations and 112 features (the features are normalized before training). Features in `mushrooms` are separable, and we use CVX calling `gurobi` to produce an approximate solution to (6.2), x_{CVX} .

Figure 6 shows the convergence of relative errors of the solutions, i.e., $\frac{\|x - x_{\text{CVX}}\|_2}{\|x_{\text{CVX}}\|_2}$, for the four algorithms considered in subsection 6.1, whereas Figure 7 shows the convergence of the duality gaps. In these figures we observe that, consistent with the results in subsection 6.1, `Nested-SGDM` converges the fastest, followed next by `Nested-SGD`.

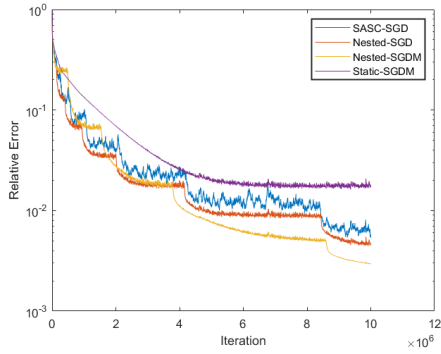


FIG. 6. Convergence of relative errors.

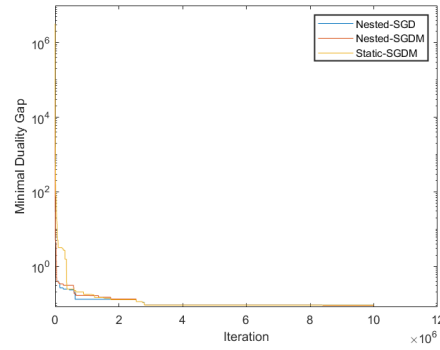


FIG. 7. Convergence of duality gaps.

To test the safety and effectiveness of the screening procedure, in Figure 8, we compare algorithms `SASC-SGD`, `Nested-SGD`, `Nested-SVRG`, `Nested-SVRG-Screening`. In this figure, we observe that even though without screening `Nested-SVRG` performs worse than `Nested-SGD`, with screening it outperforms `Nested-SGD`. The reason for the performance improvement is that we dynamically drop a significant amount of redundant constraints in Algorithm 5.1, which saves iterations that query these constraints. In other words, the screening procedure in Algorithm 5.1 is effective in detecting and eliminating redundant constraints. Details on the constraints that are dropped and kept in `Nested-SVRG-Screening` are presented in Appendix B.

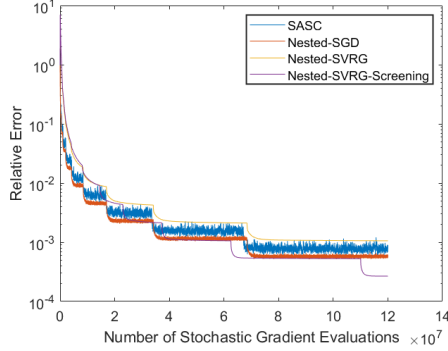


FIG. 8. Relative errors during iterations

7. Conclusions. We give a nested accelerated stochastic method for penalized reformulations of strongly convex function minimization subject to linear constraints. The algorithm generates a primal solution within a distance of ϵ to the optimum and an $O(\epsilon)$ sub-optimal dual solution in an expected $\tilde{O}(1/\sqrt{\epsilon})$ stochastic first-order iterations. We also design the screening procedure for our algorithm, which can effectively drop inactive constraints. Computational results on synthetic quadratic programming instances and support vector machine instances demonstrate the effectiveness and robustness of the nested algorithm.

Appendix A. Lemmas and Missing Proofs.

A.1. Proof of Proposition 2.1.

Proof. Observe that

$$\begin{aligned} \lim_{\delta \rightarrow 0^+} \delta \log(1 + \exp(t/\delta)) &= t, \quad t \in (0, \infty), \\ \lim_{\delta \rightarrow 0^+} \delta \log(1 + \exp(t/\delta)) &= 0, \quad t \in (-\infty, 0]. \end{aligned}$$

Hence, we may define $p_0(t) := \max(0, t)$. Then, $p_\delta(t)$ is continuous with respect to δ , for $\delta \in [0, \infty)$. For $\delta > 0$, we have

$$\begin{aligned} p'_\delta(t) &= \frac{\exp(t/\delta)}{1 + \exp(t/\delta)} \in [0, 1], \\ p''_\delta(t) &= \frac{1}{\delta} \frac{\exp(t/\delta)}{(1 + \exp(t/\delta))^2} \in [0, \frac{1}{4\delta}]. \end{aligned}$$

For $0 < \delta_1 \leq \delta_2$, $p_{\delta_1}(t) \leq p_{\delta_2}(t)$

$$p'_{\delta_2}(t) - p'_{\delta_1}(t) \begin{cases} \geq 0, & t < 0, \\ = 0, & t = 0, \\ \leq 0, & t > 0. \end{cases}$$

Then, $0 \leq p_{\delta_2}(t) - p_{\delta_1}(t) \leq p_{\delta_2}(0) - p_{\delta_1}(0) = (\delta_2 - \delta_1) \log 2$. Because $p_\delta(t)$ is continuous with respect to δ for $\delta \in [0, \infty)$, letting $\delta_1 = 0$ the result holds. \square

A.2. Proof of Lemma 2.5.

Proof. Let $\tilde{\theta}_{c,d} = \arg \max_{\theta} \phi_{c,d}(\theta)$. Since

$$\frac{d\phi_{c,d}(\theta)}{d\theta} = \frac{1}{(d + c \exp(\theta))^2} ((d + c \exp(\theta)) - c\theta \exp(\theta)),$$

$\tilde{\theta}_{c,d}$ satisfies

$$\tilde{\theta}_{c,d} = 1 + \frac{d}{c} \exp(-\tilde{\theta}_{c,d}).$$

Moreover, $\frac{d\phi_{c,d}(\theta)}{d\theta} > 0$ for $\theta < \tilde{\theta}_{c,d}$ and $\frac{d\phi_{c,d}(\theta)}{d\theta} < 0$ for $\theta > \tilde{\theta}_{c,d}$. When $\log(\frac{d}{c}) \geq 2$,

$$\frac{d\phi_{c,d}(\theta)}{d\theta} \Big|_{\theta=\log(\frac{d}{c})} \leq 0.$$

Then,

$$\phi_{c,d}(\theta) \leq \phi_{c,d}(\tilde{\theta}_{c,d}) = (\tilde{\theta}_{c,d} - 1)/d \leq (\log(\frac{d}{c}) - 1)/d. \quad \square$$

A.3. Proof of Lemma 3.1.

Proof. 1. By the definition, $\tilde{x} = \arg \min_{y \in \mathbb{R}^n} \left[\psi(y) + \nabla \tilde{F}(\tilde{x})^T (y - \tilde{x}) + \frac{1}{2\alpha} \|y - \tilde{x}\|_2^2 \right]$. ■

Hence,

$$g_{\psi}(\tilde{x}; \tilde{F}, \alpha) - \nabla \tilde{F}(\tilde{x}) \in \partial \psi(\tilde{x}),$$

which means

$$g_{\psi}(\tilde{x}; \tilde{F}, \alpha) - \nabla \tilde{F}(\tilde{x}) + \nabla \tilde{F}(\tilde{x}) \in \partial F(\tilde{x}).$$

For any given x , by the μ -strong convexity,

$$(A.1) \quad F(x) \geq F(\tilde{x}) + (g_{\psi}(\tilde{x}; \tilde{F}, \alpha) - \nabla \tilde{F}(\tilde{x}) + \nabla \tilde{F}(\tilde{x}))^T (x - \tilde{x}) + \frac{\mu}{2} \|x - \tilde{x}\|_2^2.$$

Letting $x = \bar{x}$ in (A.1), we obtain

$$\begin{aligned} F(\bar{x}) &\geq F(\tilde{x}) + (g_{\psi}(\tilde{x}; \tilde{F}, \alpha) - \nabla \tilde{F}(\tilde{x}) + \nabla \tilde{F}(\tilde{x}))^T (\bar{x} - \tilde{x}) + \frac{\mu}{2} \|\bar{x} - \tilde{x}\|_2^2 \\ &= F(\tilde{x}) + \left(\frac{\mu\alpha^2}{2} + \alpha \right) \|g_{\psi}(\tilde{x}; \tilde{F}, \alpha)\|_2^2 - (\nabla \tilde{F}(\tilde{x}) - \nabla \tilde{F}(\tilde{x}))^T (\bar{x} - \tilde{x}) \\ &\geq F(\tilde{x}) + (\alpha - L\alpha^2) \|g_{\psi}(\tilde{x}; \tilde{F}, \alpha)\|_2^2 \\ &\geq F(\tilde{x}) + \frac{\alpha}{2} \|g_{\psi}(\tilde{x}; \tilde{F}, \alpha)\|_2^2. \end{aligned}$$

Therefore, $\|g_{\psi}(\tilde{x}; \tilde{F}, \alpha)\|_2^2 \leq \frac{2}{\alpha} (F(\bar{x}) - F(\tilde{x})) \leq \frac{2}{\alpha} (F(\bar{x}) - F^*)$.

2. Minimizing the right-hand side of (A.1), we find that $\forall y \in \mathbb{R}^n$,

$$\begin{aligned} F(y) &\geq \min_x \left[(g_{\psi}(\tilde{x}; \tilde{F}, \alpha) - \nabla \tilde{F}(\tilde{x}) + \nabla \tilde{F}(\tilde{x}))^T (x - \tilde{x}) + \frac{\mu}{2} \|x - \tilde{x}\|_2^2 \right] \\ &\geq F(\tilde{x}) - \frac{1}{2\mu} \|g_{\psi}(\tilde{x}; \tilde{F}, \alpha) - \nabla \tilde{F}(\tilde{x}) + \nabla \tilde{F}(\tilde{x})\|_2^2 \\ &= F(\tilde{x}) - \frac{1}{2\mu} \|g_{\psi}(\tilde{x}; \tilde{F}, \alpha)\|_2^2 + \frac{1}{\mu} g_{\psi}(\tilde{x}; \tilde{F}, \alpha)^T (\nabla \tilde{F}(\tilde{x}) - \nabla \tilde{F}(\tilde{x})) \\ &\quad - \frac{1}{2\mu} \|\nabla \tilde{F}(\tilde{x}) + \nabla \tilde{F}(\tilde{x})\|_2^2 \\ &\geq F(\tilde{x}) - \frac{1}{2\mu} \|g_{\psi}(\tilde{x}; \tilde{F}, \alpha)\|_2^2 + \left(\frac{1}{\mu\alpha L} - \frac{1}{2\mu} \right) \|\nabla \tilde{F}(\tilde{x}) + \nabla \tilde{F}(\tilde{x})\|_2^2, \end{aligned}$$

where the last inequality follows from the co-coercivity of $\nabla \tilde{F}$. Hence, $\|g_\psi(\bar{x}; \tilde{F}, \alpha)\|_2^2 \geq 2\mu(F(\bar{x}) - F^*)$. \square

A.4. Complexity of Proximal SVRG with Catalyst Acceleration. In this section, we review the main complexity result of the proximal SVRG method with catalyst acceleration [26] that is used in the analysis of [Algorithm 3.1](#), particularly in the proof of [Proposition 3.3](#). Consider the problem

$$(A.2) \quad \min_{x \in \mathbb{R}^n} \Phi(x) := \frac{1}{m} \sum_{i=1}^m \phi_i(x) + \psi(x),$$

where $\phi_i : \mathbb{R}^n \rightarrow \mathbb{R}$ are convex L_i -smooth functions, and $\psi : \mathbb{R}^n \rightarrow \mathbb{R}$ is a convex proximal function. Namely, it is assumed that the proximal operator of ψ , defined by

$$\text{prox}_{t\psi}(x) := \arg \min_u \psi(u) + \frac{1}{2t} \|x - u\|_2^2,$$

is computable for any $x \in \mathbb{R}^n$ and $t > 0$. Lin et al. [26] propose the catalyst acceleration technique and develop the following complexity result, in the strongly convex case, for the proximal SVRG method with catalyst acceleration. Note that this complexity result is stated in terms of the number of evaluations of the the gradients of the individual component functions ϕ_i and of the proximal operator $\text{prox}_{t\psi}$. In the result below, the \tilde{O} notation hides universal constants and logarithmic dependencies in μ , L and m .

LEMMA A.1 (Lemma C.1 of [26]). *Consider applying the proximal SVRG method with catalyst acceleration to problem (A.2). Suppose that Φ is μ -strongly convex for some $\mu > 0$, and let the average Lipschitz constant of the gradients $\nabla \phi_i$ be defined by $L := \frac{1}{m} \sum_{i=1}^m L_i$. Then, the expectation of the number of evaluations of gradients $\nabla \phi_i$ and proximal mappings required to satisfy $\Phi(x_t) - \Phi^* \leq \epsilon$ is upper bound bounded by*

$$\tilde{O} \left(\left(m + \sqrt{\frac{mL}{\mu}} \right) \log \left(\frac{\Phi(x_0) - \Phi^*}{\epsilon} \right) \right),$$

where \tilde{O} hides universal constants and logarithmic dependencies in μ , L and m .

Appendix B. Details of Experiments.

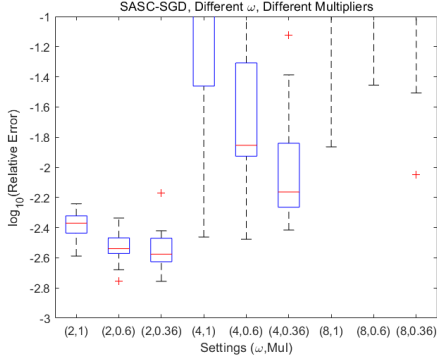
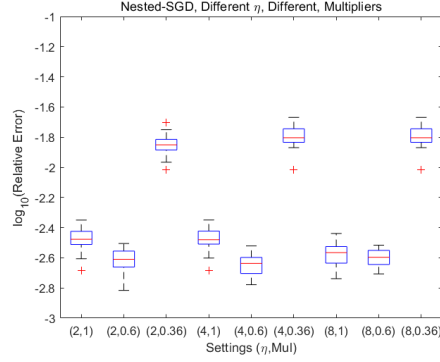
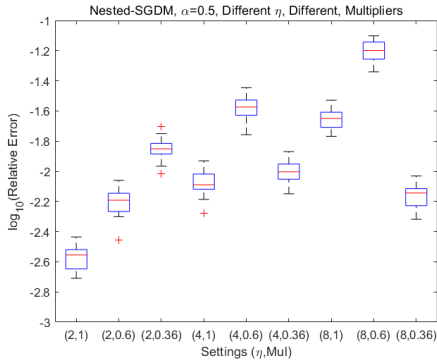
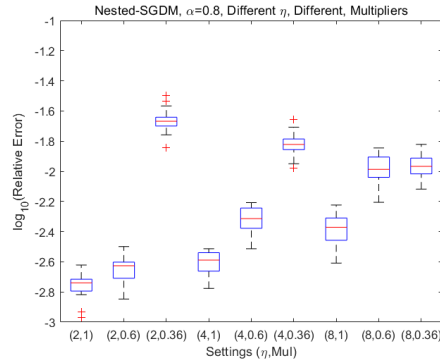
B.1. Momentum Method. For `Nested-SGDM`, we use the momentum method in [18]. The algorithm to solve the problem $\min_x \frac{1}{m} \sum_{i=1}^m \phi_i(x)$, is stated as

$$(B.1) \quad \begin{aligned} v &\leftarrow \alpha v - t \nabla_x \phi_i(x + \alpha v), \\ x &\leftarrow x + v, \end{aligned}$$

where t is the step size, and α is the ratio related to the momentum, which is often set to be $(0.5, 0.8, 0.9, 0.95)$. We use (B.1) as a simpler representative for the catalyst accelerated SVRG method.

B.2. Details for [subsection 6.1](#). For `SASC-SGD` and `Nested-SGD`, the parameters include: η (or ω), the ratios between 2 consecutive inner loops (see Algorithm 1 in [16] and [Algorithm 3.1](#)), from $\{2, 4, 8\}$; multipliers for the theoretical inner-loop iteration limit used in the code, from $\{1, 0.6, 0.36\}$. The theoretical inner-loop iteration limit for `Nested-SGD` is $\log(2\eta - 1) \left(\frac{L}{\mu} + \frac{m\xi}{4\mu\delta} \right)$, and that for `Nested-SGDM` is

$(2 \log(2\eta - 1) + \log(\frac{L}{\mu} + \frac{m\xi}{4\mu\delta}))\sqrt{m(\frac{L}{\mu} + \frac{m\xi}{4\mu\delta})}$. Similarly, the step sizes for the algorithms are $\frac{1}{L+\mu+\frac{m\xi}{4\delta}}$. For SASC, see the settings in Algorithm 1 in [16]. For *Nested-SGDM*, parameter α in (B.1) is taken as $\{0.5, 0.8, 0.9\}$. The following figures show the relative errors of the three methods with different parameters after $1e7$ iterations (starting from 0 point) for (6.1), in 20 simulations (including the generation of the problems and the stochastic sequence of indexes). For *Nested-SGD* and *Nested-SGDM*, $\xi = 1$ and $\delta_0 = 0.05$. We pick (2,0.36), (4,0.6), and (2,1,0.9) as the best parameter settings for SASC-SGD, *Nested-SGD*, and *Nested-SGDM*, considering the medium and largest errors (see Figure 9–Figure 13). We also see here that SASC-SGD is more likely to diverge for different parameter settings. The setting for the initial step size for SASC-SGD is $\frac{1}{10L}$, instead of $\frac{3}{4L}$ in Algorithm 1 in [16], which results in divergence. For *Static-SGDM*, $\xi = 1$ and the choices of δ have been shown. The duality gaps in Figure 5 are calculated every 1000 steps.

FIG. 9. Relative errors of *SASC-SGD*FIG. 10. Relative errors of *Nested-SGD*FIG. 11. Relative errors of *SASC, SGDM* ($\alpha = 0.5$)FIG. 12. Relative errors of *SASC, SGDM* ($\alpha = 0.8$)

B.3. Details for subsection 6.2. The starting point for all methods is $x_0 = 0$. For SASC-SGD, the parameter setting is the same with 6.3 in [16]. For *Nested-SVRG* and *Nested-SVRG-Screening*, the full gradients are calculated every $5m$ stochas-

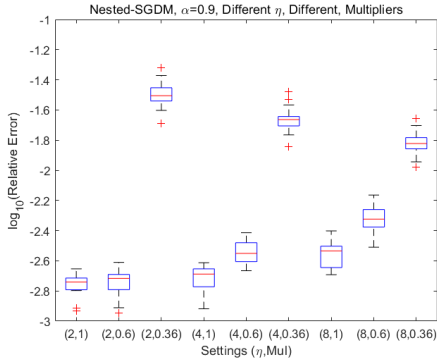
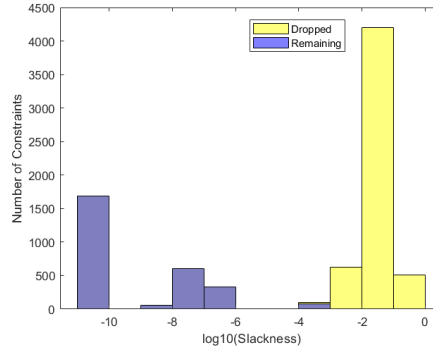

 FIG. 13. Relative errors of *SASC*, *SGDM* ($\alpha = 0.9$)


FIG. 14. Slackness of dropped and remaining constraints

tic gradient steps, where m is the numbers of constraints or remaining constraints. The numbers of iterations for these two algorithms are approximately $5/6$ of other algorithms to maintain the same number of stochastic gradient evaluations. For *Nested-SGD*, *Nested-SVRG*, *Nested-SVRG-Screening*, and *Nested-SGDM*, $\xi = 1$ and $\delta_0 = 0.005$. And other parameter settings are $(2, 0.3)$, $(2, 0.5)$, $(2, 0.5)$, and $(4, 0.5, 0.5)$ for *Nested-SGD*, *Nested-SVRG*, *Nested-SVRG-Screening*, and *Nested-SGDM*. Here, the theoretical inner-loop iteration numbers for *Nested-SGD*, *Nested-SVRG*, *Nested-SVRG-Screening* are $\log(2\eta - 1) \left(\frac{L}{\mu} + \frac{m\xi}{4\mu\delta} \right)$, and that for *Nested-SGDM* is $(2 \log(2\eta - 1) + \log(\frac{L}{\mu} + \frac{m\xi}{4\mu\delta})) \sqrt{m \left(\frac{L}{\mu} + \frac{m\xi}{4\mu\delta} \right)}$. Similarly, the step sizes for the *Nested-SGD*, *Nested-SVRG*, *Nested-SVRG-Screening* are $\frac{4}{L + \mu + \frac{m\xi}{4\delta}}$, and for *Nested-SGDM* they are $\frac{2}{L + \mu + \frac{m\xi}{4\delta}}$ (considering the momentum term) respectively. (We checked increasing the step sizes and (or) decreasing inner-loop iteration numbers for *SASC-SGD*, which turned out not to bring significantly better performance.) The duality gaps in [Figure 7](#) are calculated every 1000 steps.

In the experiment of [Figure 8](#), for simplicity, constraint dropping criterion is set as $a_i^T \tilde{x}_t - b_i < -2\sqrt{m_i}\delta_t$. As such, 2,777 out of 8,198 constraints are kept in the last inner loop of *Nested-SVRG-Screening*. In [Figure 14](#), the slackness of the constraints, $(b_i - a_i^T x_{cvx})$, $i = 1, \dots, m$ are compared (to include the constraints that $b_i - a_i^T x_{cvx} < 0$, the slackness values are set as $1e-11$). Constraints with slackness greater than $1e-3$ (5,536 out of 8,198) are dropped and all the constraints with slackness less than $1e-7$ (2,358 out of 8,198) are kept.

Acknowledgments. This research is supported, in part, by NSF AI Institute for Advances in Optimization Award 211253, DOD ONR grant 12951270, and NSF Awards CCF-1755705 and CMMI-1762744.

REFERENCES

- [1] Z. ALLEN-ZHU, *Katyusha: The first direct acceleration of stochastic gradient methods*, The Journal of Machine Learning Research, 18 (2017), pp. 8194–8244.
- [2] A. ATAMTÜRK AND A. GÓMEZ, *Strong formulations for quadratic optimization with M -matrices and indicator variables*, Mathematical Programming, 170 (1998), pp. 141–176.
- [3] A. ATAMTÜRK AND A. GÓMEZ, *Rank-one convexification for sparse regression*, arXiv preprint

- arXiv:1901.10334, (2019).
- [4] A. ATAMTÜRK AND A. GÓMEZ, *Safe screening rules for ℓ_0 -regression from perspective relaxations*, in Proceedings of the 37th International Conference on Machine Learning, vol. 119, PMLR, 2020, pp. 421–430.
 - [5] A. ATAMTÜRK, A. GÓMEZ, AND S. HAN, *Sparse and smooth signal estimation: Convexification of ℓ_0 formulations*, Journal of Machine Learning Research, 22 (2021), pp. 1–43.
 - [6] D. AZAGRA, *Global and fine approximation of convex functions*, Proceedings of the London Mathematical Society, 107 (2013), pp. 799–824.
 - [7] R. E. BARLOW AND H. D. BRUNK, *The isotonic regression problem and its dual*, Journal of the American Statistical Association, 67 (1972), pp. 140–147.
 - [8] A. BECK AND M. TEBoulLE, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM Journal on Imaging Sciences, 2 (2009), pp. 183–202.
 - [9] D. P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, 2014.
 - [10] F. BORRELLI, A. BEMPORAD, AND M. MORARI, *Predictive Control for Linear and Hybrid Systems*, Cambridge University Press, 2017.
 - [11] L. BOTTOU, F. E. CURTIS, AND J. NOCEDAL, *Optimization methods for large-scale machine learning*, SIAM Review, 60 (2018), pp. 223–311.
 - [12] S. BOYD, N. PARIKH, AND E. CHU, *Distributed Optimization and Statistical Learning Via the Alternating Direction Method of Multipliers*, Now Publishers Inc, 2011.
 - [13] C.-C. CHANG AND C.-J. LIN, *LIBSVM: A library for support vector machines*, ACM Transactions on Intelligent Systems and Technology, 2 (2011), pp. 27:1–27:27. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
 - [14] I. CVX RESEARCH, *CVX: Matlab software for disciplined convex programming, version 2.0*. <http://cvxr.com/cvx>, Aug. 2012.
 - [15] A. DEFazio, F. BACH, AND S. LACOSTE-JULIEN, *Saga: A fast incremental gradient method with support for non-strongly convex composite objectives*, in Advances in Neural Information Processing Systems, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, eds., vol. 27, 2014, pp. 1646–1654.
 - [16] O. FERCOQ, A. ALACAogLU, I. NECOARA, AND V. CEVHER, *Almost surely constrained convex optimization*, in International Conference on Machine Learning, PMLR, 2019, pp. 1910–1919.
 - [17] L. E. GHAoui, V. VIALlon, AND T. RABBANI, *Safe feature elimination for the lasso and sparse supervised learning problems*, arXiv preprint arXiv:1009.4219, (2010).
 - [18] I. GOODFELLOW, Y. BENGIO, AND A. COURVILLE, *Deep Learning*, MIT press, 2016.
 - [19] M. GRANT AND S. BOYD, *Graph implementations for nonsmooth convex programs*, in Recent Advances in Learning and Control, V. Blondel, S. Boyd, and H. Kimura, eds., Lecture Notes in Control and Information Sciences, Springer-Verlag Limited, 2008, pp. 95–110. http://stanford.edu/~boyd/graph_dcp.html.
 - [20] L. GUROBI OPTIMIZATION, *Gurobi optimizer reference manual*, 2021, <http://www.gurobi.com>.
 - [21] S. HAN, A. GÓMEZ, AND A. ATAMTÜRK, *2x2 convexifications for convex quadratic optimization with indicator variables*, arXiv preprint arXiv:2004.07448, (2020).
 - [22] R. JOHNSON AND T. ZHANG, *Accelerating stochastic gradient descent using predictive variance reduction*, in Advances in Neural Information Processing Systems, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, eds., vol. 26, 2013, pp. 315–323.
 - [23] G. LAN AND R. D. MONTEIRO, *Iteration-complexity of first-order penalty methods for convex programming*, Mathematical Programming, 138 (2013), pp. 115–139.
 - [24] G. LAN AND Y. ZHOU, *An optimal randomized incremental gradient method*, Mathematical Programming, 171 (2018), pp. 167–215.
 - [25] E. LIM AND P. W. GLYNN, *Consistency of multidimensional convex regression*, Operations Research, 60 (2012), pp. 196–208.
 - [26] H. LIN, J. MAIRAL, AND Z. HARCHAOUI, *A universal catalyst for first-order optimization*, arXiv preprint arXiv:1506.02186, (2015).
 - [27] H. LIN, J. MAIRAL, AND Z. HARCHAOUI, *Catalyst acceleration for first-order convex optimization: from theory to practice*, Journal of Machine Learning Research, 18 (2018), pp. 7854–7907.
 - [28] H. M. MARKOWITZ AND G. P. TODD, *Mean-variance Analysis in Portfolio Choice and Capital Markets*, vol. 66, John Wiley & Sons, 2000.
 - [29] K. MISHCHENKO AND P. RICHTÁRIK, *A stochastic penalty model for convex and nonconvex optimization with big constraints*, arXiv preprint arXiv:1810.13387, (2018).
 - [30] A. NEDIĆ, *Random algorithms for convex minimization problems*, Mathematical Programming, 129 (2011), pp. 225–253.

- [31] A. NEDIĆ AND T. TATARENKO, *Convergence rate of a penalty method for strongly convex problems with linear constraints*, in 2020 59th IEEE Conference on Decision and Control (CDC), IEEE, 2020, pp. 372–377.
- [32] Y. NESTEROV, *Smooth minimization of non-smooth functions*, *Mathematical Programming*, 103 (2005), pp. 127–152.
- [33] Y. NESTEROV, *Gradient methods for minimizing composite functions*, *Mathematical Programming*, 140 (2013), pp. 125–161.
- [34] Y. NESTEROV ET AL., *Lectures on convex optimization*, vol. 137, Springer, 2018.
- [35] Y. OUYANG AND Y. XU, *Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems*, *Mathematical Programming*, 185 (2021), pp. 1–35.
- [36] M. SCHMIDT, N. LE ROUX, AND F. BACH, *Minimizing finite sums with the stochastic average gradient*, *Mathematical Programming*, 162 (2017), pp. 83–112.
- [37] E. SEIJO, B. SEN, ET AL., *Nonparametric least squares estimation of a multivariate convex regression function*, *The Annals of Statistics*, 39 (2011), pp. 1633–1657.
- [38] T. TATARENKO AND A. NEDICH, *A smooth inexact penalty reformulation of convex problems with linear constraints*, arXiv preprint arXiv:1808.07749, (2018).
- [39] M. WANG AND D. P. BERTSEKAS, *Incremental constraint projection methods for variational inequalities*, *Mathematical Programming*, 150 (2015), pp. 321–363.