A Smooth Inexact Penalty Reformulation of Convex Problems with Linear Constraints

Tatiana Tatarenko, Angelia Nedić

The School of Electrical, Computer and Energy Engineering at Arizona State University, Tempe, AZ 85281

Abstract

In this work, we consider a constrained convex problem with linear inequalities and provide an inexact penalty re-formulation of the problem. The novelty is in the choice of the penalty functions, which are smooth and can induce a non-zero penalty over some points in feasible region of the original constrained problem. The resulting unconstrained penalized problem is parametrized by two penalty parameters which control the slope and the curvature of the penalty function. With a suitable selection of these penalty parameters, we show that the solutions of the resulting penalized unconstrained problem are feasible for the original constrained problem, under some assumptions. Also, we establish that, with suitable choices of penalty parameters, the solutions of the penalized unconstrained problem can achieve a suboptimal value which is arbitrarily close to the optimal value of the original constrained problem. For the problems with a large number of linear inequality constraints, a particular advantage of such a smooth penalty-based reformulation is that it renders a penalized problem suitable for the implementation of fast incremental gradient methods, which require only one sample from the inequality constraints at each iteration. We consider applying SAGA proposed in [9] to solve the resulting penalized unconstrained problem.

Keywords: convex minimization, linear constraints, inexact penalty, incremental methods

1. Introduction

In this paper, we study the problem of minimizing a convex function $f : \mathbb{R}^n \to \mathbb{R}$ over a convex and closed set $X$ that is the intersection of finitely many convex and closed sets $X_i, i = 1, \ldots, m$ ($m \geq 2$ is large), i.e.,

$$\minimize f(x) \quad \text{subject to } x \in X = \cap_{i=1}^{m} X_i.$$  \hfill (1)

Throughout the paper, the function $f$ is assumed to be convex over $\mathbb{R}^n$. Optimization problems of the form (1) arise in many areas of research, such as digital filter settings in communication systems [1], energy consumption in Smart Grids [11], convex relaxations of various combinatorial optimization problems in machine learning applications [27, 42].

Our interest is in case when $m$ is large, which prohibits us from using projected gradient and augmented Lagrangian methods [3], which require either computation of the (Euclidean) projection or an estimation of the gradient for the sum of many functions, at each iteration. To reduce
the complexity, one may consider a method that operates on a single set $X_i$ from the constraint set collection $\{X_1, \ldots, X_m\}$ at each iteration. Algorithms using random constraint sampling for general convex optimization problems (1) have been first considered in [29] and were extended in [40] to a broader class of randomization over the sets of constraints. Moreover, the convergence rate analysis is performed in [40] to demonstrate that the feasibility error diminishes to zero at a rate $O(\log k/k)$, whereas the optimality error diminishes to zero with the rate of $O(1/\sqrt{k})$. For the general convex problems of type (1), the latter rate is optimal over the class of optimization methods based on noisy first-order information.

A special case of the problem (1) with $f \equiv 0$ is a feasibility problem, for which random sampling methods have been considered in [33] for the case of the sets given by convex inequalities, and in [8] for a more specialized case of linear matrix inequalities. In [28], a connection between the convergence properties of stochastic gradient methods and the existence of solutions for problem (1) has been studied, and a linear convergence rate has been established for some special cases of the constraint sets $X_i$ (such as those admitting easily computable Euclidean projections). Algorithms with the linear convergence to a solution of feasibility problems defined by a system of linear equations and inequalities have been considered in [22, 38]. An iterated randomized projection scheme for systems of linear equations is proposed in [38], which is a randomized variant of Kaczmarz’s method. This variant employs a single projection per each iteration and is shown to converge with the linear rate that does not depend on the number of equations, but instead, depends on the condition number associated with the linear system of equations.

A possible reformulation of problem (1) is through the use of the indicator functions of the constraint sets, resulting in the following unconstrained problem

$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^{m} \left\{ \frac{1}{m} f(x) + \chi_i(x) \right\},$$

where $\chi_i(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ is the indicator function of the set $X_i$ (taking value 0 at the points $x \in X_i$ and, otherwise, taking value $+\infty$). The advantage of this reformulation is that the objective function is the sum of convex functions and incremental methods can be employed that compute only a (sub)-gradient of one of the component functions at each iteration. The traditional incremental methods do not have memory, and their origin can be traced back to work of Kibardin [19]. They have been studied for smooth least-square problems [4, 5, 25], for training the neural networks [13, 14, 26], for smooth convex problems [37, 39] and for non-smooth convex problems [12, 16, 17, 20, 30, 31, 32, 41] (see [7] for a more comprehensive survey of these methods). These traditional memoryless incremental methods (randomized and deterministic), while simple to implement to solve problem (2), cannot achieve the optimal convergence rate even when $f$ is smooth and strongly convex. This is due to the non-smoothness of the indicator functions and the errors that are accumulated during the incremental processing of the functions in the sum.

Reformulation (2) has been considered in [21] as a departure point toward an exact penalty reformulation using the set-distance functions, thus yielding a penalized problem of the following form:

$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^{m} \left\{ \frac{1}{m} f(x) + \lambda h_P(x) \right\},$$

(3)
where

\[ h_p(x) = P(\text{dist}(x, X_1), \ldots, \text{dist}(x, X_m)), \]

with \( P \) being some norm in \( \mathbb{R}^m \) and \( \text{dist}(\cdot, Y) \) being the distance function to a set \( Y \). This exact penalty formulation has been motivated by a simple exact penalty model proposed in [6] (using only the set-distance functions) and a more general penalty model considered in [7]. In [21], a lower bound on the penalty level \( \lambda \) has been identified guaranteeing that the optimal solutions of the penalized problem are also optimal solutions of the original problem (2). However, the proposed approaches in [21] do not utilize incremental processing, but rather approaches where a full (sub)-gradient of the function objective in (3) is used.

Unlike [21], our objective in this paper is to consider a penalty-based reformulation of problem (1) (with linear constraints) that will allow us to take advantage of the penalized problem structure for the use of incremental methods. In order to achieve the optimal convergence rates, we would like to depart from the traditional incremental methods. In particular, we would like to have a penalty reformulation of problem (1) that will enable us to employ one of recently developed fast incremental algorithms. These algorithms are designed to solve optimization problems involving a large sum of functions [9, 18, 34] which arise in machine learning applications. Unlike the traditional incremental methods that are memoryless, these fast incremental algorithms require storage of the past (sub)-gradients. Typically, they require storing the same number \( m \) of the (sub)-gradients as the number \( m \) of the component functions in the objective. The stored information is effectively used to control the error due to the incremental processing of the functions, which in turn allows these algorithms to achieve optimal convergence rates. A drawback of the fast incremental algorithms, such as SAGA and its various modifications [2, 10, 18, 35, 24], is that they are not designed to efficiently handle a possibly large number of constraints. At most, these algorithms allow us to deal with so called composite optimization problems, where the composite term corresponds to a regularization function promoting some special properties of model parameters and has a simple structure for determining the proximal point [9].

Our focus is on problem (1) with linear constraints,

\[ X_i = \{ x \in \mathbb{R}^n \mid \langle a_i, x \rangle - b_i \leq 0 \}, \]

where \( a_i \in \mathbb{R}^n \) and \( b_i \in \mathbb{R} \) for all \( i = 1, \ldots, m \). Our objective is to develop a penalty model for this problem that will allow us to implement fast incremental methods [9, 18, 34] to solve the resulting unconstrained penalized problem. In order to do so, we will develop a smooth penalty framework motivated by the approach in [7], and provide the relations for the solutions of problem (5) and the solutions of the corresponding penalized problem. We consider a penalized reformulation of problem (5) in the following form:

\[
\begin{align*}
\text{minimize} & \quad f(x) + \frac{\gamma}{m} \sum_{i=1}^{m} h_\delta(x; a_i, b_i) \\
\text{subject to} & \quad x \in \mathbb{R}^n,
\end{align*}
\]

where the function \( h_\delta(x; a, b) \) is a smooth penalty function associated with a linear inequality constraint \( \langle a, x \rangle - b \leq 0 \), while \( \delta \geq 0 \) and \( \gamma > 0 \) are the penalty parameters. The penalty parameters
will control the slope and the curvature of the penalty function $\frac{\gamma}{m} \sum_{i=1}^{m} h_{\delta}(x; a_i, b_i)$.

The novelty is in the use of inexact smooth penalty function $h_{\delta}(x; a, b)$ that has Lipschitz continuous gradients, which are not related to the squared set-distance function, which is in contrast to the inexact distance-based smooth penalties considered in [36]. Also, this is contrast with the use of non-smooth exact penalty functions in [7]. A key property of our penalty framework is its accuracy guarantee, as follows: For a given accuracy $\delta^0 > 0$, we show that there exists a range of values for parameters $\delta$ and $\gamma$ such that any optimal solution of the penalized problem (4) is feasible for the original linearly constrained problem. Moreover, we provide estimates that characterize sub-optimality of the solutions of the penalized problem, i.e., we show that the solutions are located within the $\delta^0$-neighborhood of the solutions of the original constrained problem.

These properties of the penalized problem allow us to apply any fast incremental method [9, 18, 34]. We will employ SAGA to solve the smooth penalized problem to obtain a suboptimal point with the sublinear rate $O(1/k)$ in the case of smooth convex function $f$ and the linear rate $O(q^k)$, with $q < 1$, in the case of smooth strongly convex $f$.

The paper is organized as follows. In Section 2, we formulate the penalized problem, establish some properties of the chosen penalty function and provide some elementary relation between the penalized problem and the original constrained problem. In Section 3, we investigate the relation for the solutions of the original problem and its penalized variant. In Section 4 we consider applying an existing fast incremental method, namely SAGA, for solving the penalized problem. In Section 5, we provide some numerical results to illustrate the performance of SAGA for the penalized problem in comparison with a method that uses random projections, as proposed in [29]. We conclude the paper in Section 6.

2. Penalized Problem and its Properties

We consider the following optimization problem:

\[
\begin{align*}
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad \langle a_i, x \rangle - b_i \leq 0, \ i = 1, \ldots, m,
\end{align*}
\]

(5)

where the vectors $a_i$, $i = 1, \ldots, m$, are nonzero. We will assume that the problem is feasible. Associated with problem (5), we consider a penalized problem

\[
\begin{align*}
\text{minimize} & \quad F_{\gamma \delta}(x) \\
\text{subject to} & \quad x \in \mathbb{R}^n,
\end{align*}
\]

(6)

where

\[
F_{\gamma \delta}(x) = f(x) + \frac{\gamma}{m} \sum_{i=1}^{m} h_{\delta}(x; a_i, b_i).
\]

(7)

Here, $\gamma > 0$ and $\delta \geq 0$ are penalty parameters. The vectors $a_i$ and scalars $b_i$ are the same as those characterizing the constraints in problem (5). For a given nonzero vector $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$, the
penalty function $h_\delta(\cdot; a, b)$ is given by (see also Figure 1)

$$h_\delta(x; a, b) = \begin{cases} \langle a, x \rangle - b, & \text{if } \langle a, x \rangle - b > \delta, \\ \frac{(\langle a, x \rangle - b + \delta)^2}{4\|a\|}, & \text{if } -\delta \leq \langle a, x \rangle - b \leq \delta, \\ 0, & \text{if } \langle a, x \rangle - b < -\delta. \end{cases}$$ \hspace{1cm} (8)

For any $\delta \geq 0$, the function $h_\delta(x; a, b)$ satisfies the following relations:

$$h_\delta(x; a, b) \geq 0 \quad \text{for all } x \in \mathbb{R}^n,$$ \hspace{1cm} (9)

$$h_\delta(x; a, b) \leq \frac{\delta}{4\|a\|}, \quad \text{when } \langle a, x \rangle \leq b,$$ \hspace{1cm} (10)

$$h_\delta(x; a, b) > \frac{\delta}{4\|a\|}, \quad \text{when } \langle a, x \rangle > b.$$ \hspace{1cm} (11)

Observe that $h_\delta(x; a, b)$ can be viewed as a composition of a scalar function

$$p_\delta(s) = \begin{cases} s, & \text{if } s > \delta, \\ \frac{(s+\delta)^2}{4\delta}, & \text{if } -\delta \leq s \leq -\delta, \\ 0, & \text{if } s < -\delta, \end{cases}$$ \hspace{1cm} (12)

with a linear function $x \mapsto \langle a, x \rangle - b$, which is scaled by $\frac{1}{\|a\|}$. In particular, we have

$$h_\delta(x; a, b) = \frac{1}{\|a\|} p_\delta(\langle a, x \rangle - b).$$ \hspace{1cm} (13)
The function $p_\delta(s)$ is convex on $\mathbb{R}$ for any $\delta \geq 0$. Thus, the function $h_\delta(x; a, b)$ is convex on $\mathbb{R}^n$, implying that the objective function (7) of the penalized problem (6) is convex over $\mathbb{R}^n$ for any $\delta \geq 0$ and $\gamma > 0$.

Furthermore, observe that the function $p_\delta(\cdot)$ is twice differentiable for any $\delta > 0$, with the second derivative given by

$$p''_\delta(s) = \begin{cases} \frac{1}{2\delta}, & \text{if } -\delta \leq s \leq \delta, \\ 0, & \text{if } s < -\delta \text{ or } s > \delta. \end{cases}$$

Thus, the function $p(s)$ has Lipschitz continuous derivatives with constant $\frac{1}{2\delta}$. Then, the function $h_\delta(\cdot; a, b)$ is differentiable for any $\delta > 0$ and its gradient is given by

$$\nabla h_\delta(x; a, b) = \frac{1}{\|a\|} p'_\delta(\langle a, x \rangle - b)a,$$

which is Lipschitz continuous with a constant $\frac{\|a\|}{2\delta}$,

$$\|\nabla h_\delta(x; a, b) - \nabla h_\delta(y; a, b)\| \leq \frac{\|a\|}{2\delta} \|x - y\| \quad \text{for all } x, y \in \mathbb{R}^n.$$

In view of the definition of the penalty function $F_{\gamma\delta}$ in (7) and relation (14), we can see that the magnitude of the “slope” of the penalty function is controlled by the parameter $\gamma > 0$, while the ratio of the parameters $\gamma$ and $\delta$ is controlling the “curvature” of the penalty function.

Our choice of the penalty function is motivated by a desire to have the minimizers of the penalized problem (6) being feasible for the original problem (5). Note that the penalty function proposed above is a version of the one-sided Huber losses. Originally, the Huber loss functions were introduced in applications of robust regression models to make them less sensitive to outliers in data in comparison with the squared error loss [23]. In contrast, we use this type of penalty function to smoothen the exact penalties based on the distance to the sets $X_i$ proposed in [7]. Furthermore, an appropriate choice of the parameter $\delta \geq 0$ allows us to overcome the limitation of the smooth penalties based on the squared distances to the sets $X_i$, which typically provide an infeasible solution (for the original problem), due to a small penalized value around an optimum lying close to the feasibility set boundary [36].

In what follows, we let $\Pi_Y[x]$ denote the (Euclidean) projection of a point $x$ on a convex closed set $Y$, so we have

$$\text{dist}(x, Y) = \|x - \Pi_Y[x]\|.$$

The following lemma provides some additional properties of the penalty function $h_\delta(x; a, b)$ that we will use later on. In fact, the lemma shows stronger results than what we will use, but the results may be of their own interest.

**Lemma 1.** Given a nonzero vector $a \in \mathbb{R}^n$ and a scalar $b \in \mathbb{R}$, consider the penalty function $h_\delta(x; a, b)$ defined in (8) with $\delta \geq 0$. Let $Y = \{ x | \langle a, x \rangle - b \leq 0 \}$. Then, we have for $\delta = 0$,

$$h_0(x; a, b) = \text{dist}(x, Y) \quad \text{for all } x \in \mathbb{R}^n,$$

and for any $0 < \delta \leq \delta'$,

$$h_\delta(x; a, b) \leq h_{\delta'}(x; a, b) \quad \text{for all } x \in \mathbb{R}^n.$$
Proof. Given a vector $x \in \mathbb{R}$, we have

$$\Pi_Y[x] = x - \frac{\max\langle a, x \rangle - b, 0}{\|a\|^2} a,$$

so that

$$\text{dist}(x, Y) = \|x - \Pi_Y[x]\| = \frac{\max\langle a, x \rangle - b, 0}{\|a\|}.$$  

If $\delta = 0$, then the last two cases in the definition of $h_\delta(x; a, b)$ reduce to $h_0(x; a, b) = 0$ when $\langle a, x \rangle - b \leq 0$, corresponding to $h_0(x; a, b) = \text{dist}(x, Y) = 0$ when $x \in Y$. When $x \notin Y$, we have $h_0(x; a, b) = \frac{\langle a, x \rangle - b}{\|a\|} = \text{dist}(x, Y)$.

To prove the monotonicity property, in view of relation (13), where $p_\delta(\cdot)$ is defined in (12), it suffices to show that the function $p_\delta(\cdot)$ has the monotonicity property, i.e., that we have for $0 < \delta \leq \delta'$,

$$p_\delta(s) \leq p_{\delta'}(s) \quad \text{for all } s \in \mathbb{R}.$$  

To show this let $0 \leq \delta \leq \delta'$. Note that, for $s < -\delta'$ and $s > \delta'$ the functions $p_\delta(\cdot)$ and $p_{\delta'}(\cdot)$ coincide, i.e.,

$$p_\delta(s) = p_{\delta'}(s) \quad \text{when } s < -\delta' \text{ or } s > \delta'.$$

When $-\delta' \leq s < -\delta$ we have

$$p_\delta(s) = 0 < p_{\delta'}(s).$$

Next, consider the case when $-\delta \leq s \leq \delta$. Let $s$ be fixed and we view the function $p_\delta(s)$ as a function of $\delta$. For the partial derivative with respect to $\delta$, we have

$$\frac{\partial p_\delta(s)}{\partial \delta} = \frac{1}{4} \frac{2(s + \delta) - (s + \delta)^2}{\delta^2} = \frac{1}{4} \frac{s^2 - \delta^2}{\delta^2} \geq 0,$$

where the inequality follows by $\delta \geq s$. Thus, $p_\delta(s)$ is non-decreasing in $\delta$, implying that $p_\delta(s) \leq p_{\delta'}(s)$. Since $s \in [-\delta, \delta]$ was arbitrary, it follows that

$$p_\delta(s) \leq p_{\delta'}(s) \quad \text{when } -\delta \leq s < \delta.$$

Finally, let $\delta < s \leq \delta'$, in which case we have

$$p_\delta(s) = s = \frac{4s\delta'}{4\delta'} \leq \frac{(s + \delta')^2}{4\delta'} = p_{\delta'}(s),$$

where the inequality is obtained by using $4st \leq (s + t)^2$ valid for any $s, t \in \mathbb{R}$.

In view of Lemma 1, for the function $F_{\gamma\delta}$ in (7) we obtain for any $\gamma > 0$ and any $\delta' \geq \delta \geq 0$,

$$F_{\gamma\delta}(x) \geq F_{\gamma\delta}(x) \geq f(x) + \sum_{i=1}^{m} \frac{\gamma}{m} \text{dist}(x, X_i) \geq f(x) \quad \text{for all } x \in \mathbb{R}^n.$$  

This relation implies an inclusion relation for the level sets of the functions $F_{\gamma\delta}$ and $f$, as given by the following corollary.
Corollary 1. For any $\gamma > 0$ and for any $t \in \mathbb{R}$, we have

$$\{ x \in \mathbb{R}^n \mid F_{\gamma \delta}(x) \leq t \} \subseteq \{ x \in \mathbb{R}^n \mid F_{\gamma \delta}(x) \leq t \} \subseteq \{ x \in \mathbb{R}^n \mid f(x) \leq t \} \quad \text{for all } \delta' \geq \delta \geq 0.$$ 

In particular, if the function $f$ has bounded level sets, then the functions $F_{\delta, \gamma}$ also have bounded level sets for any $\gamma > 0$ and $\delta \geq 0$.

While Corollary 1 shows some inclusion relations for the level sets of $F_{\gamma \delta}$ and $f$, for the same value $t$, it will be important in our analysis to identify a value of $t$ for which these level sets are nonempty. The following corollary shows that choosing $f(\hat{x})$, for any feasible $\hat{x}$, can be used to construct non-empty level sets.

**Corollary 2.** Let $\gamma > 0$ and $\delta \geq 0$ be arbitrary, and let $\hat{x}$ be a feasible point for the original problem (5). Then, for the scalar $t_{\gamma}(\hat{x})$ defined by

$$t_{\gamma}(\hat{x}) = f(\hat{x}) + \frac{\gamma \delta}{4 \min_{1 \leq i \leq m} \|a_i\|},$$

the level set $\{ x \in \mathbb{R}^n \mid F_{\gamma \delta}(x) \leq t_{\gamma}(\hat{x}) \}$ is nonempty and the solution set $X^{*}_{\gamma \delta}$ of the penalized problem (6) is contained in the level set $\{ x \in \mathbb{R}^n \mid f(x) \leq t_{\gamma}(\hat{x}) \}$.

**Proof.** Let $\gamma > 0$ and $\delta \geq 0$ be arbitrary, and $\hat{x}$ be any feasible point for the original problem. Since $\hat{x}$ is feasible, by relation (10), we have

$$h_{\delta}(\hat{x}; a_i, b_i) \leq \frac{\delta}{4\|a_i\|} \quad \text{for all } i = 1, \ldots, m.$$ 

Therefore,

$$F_{\gamma \delta}(\hat{x}) \leq f(\hat{x}) + \frac{\gamma \delta}{4m} \sum_{i=1}^{m} \frac{1}{\|a_i\|} \leq f(\hat{x}) + \frac{\gamma \delta}{4 \min_{1 \leq i \leq m} \|a_i\|} = t_{\gamma}(\hat{x}),$$

implying that $\hat{x}$ belongs to the level set $\{ x \in \mathbb{R}^n \mid F_{\gamma \delta}(x) \leq t_{\gamma}(\hat{x}) \}$. Noting that

$$X^{*}_{\gamma \delta} \subseteq \{ x \in \mathbb{R}^n \mid F_{\gamma \delta}(x) \leq t_{\gamma}(\hat{x}) \},$$

by Corollary 1, we obtain

$$X^{*}_{\gamma \delta} \subseteq \{ x \in \mathbb{R}^n \mid f(x) \leq t_{\gamma}(\hat{x}) \}.$$ 

In Corollary 2, the solution set $X^{*}_{\gamma \delta}$ of the penalized problem (6) may be empty. In the next section, we will consider the cases when the solution sets are nonempty for both the original and the penalized problems.
3. Relations for Penalized Problem and Original Problem Solutions

In what follows, we establish some important relations between the solutions of the penalized problem and the original problem. A key role in the analysis plays a special property of the linear constraint set, which is valid when the constraint set of problem (5) has a nonempty interior. To provide this property, we let $X_i$ be the set defined by the $i$th inequality in the constraint set of problem (5), i.e.,

$$X_i = \{ x \in \mathbb{R}^n \mid \langle a_i, x \rangle - b_i \leq 0 \},$$

and we define the set $X$ as the intersection of these sets

$$X = \cap_{i=1}^m X_i.$$

We make the following assumption on the interior of the set $X$.

**Assumption 1.** The interior of the set $X$ is not empty, i.e., there is a point $\bar{x}$ such that for some $\epsilon > 0$,

$$\langle a_j, \bar{x} \rangle - b_j \leq -\epsilon \quad \text{for all } j = 1, \ldots, m.$$

We next provide a lemma that will be important for our analysis of solution feasibility of the penalized problem. In this lemma and later on, we use the following notation

$$\alpha_{\min} = \min_{j=1,\ldots,m} \|a_j\|, \quad \alpha_{\max} = \max_{j=1,\ldots,m} \|a_j\|. \quad (16)$$

Moreover, conditions for solution feasibility of the penalized problem involve a constant $\beta$ from Hoffman’s lemma [15] stating that for the sets $X_i$

there exists $\beta = \beta(a_1, \ldots, a_m) > 0$ such that

$$\beta \sum_{i=1}^m \text{dist}(x, X_i) \geq \text{dist}(x, X) \quad \text{for all } x \in \mathbb{R}^n. \quad (17)$$

**Lemma 2.** Let Assumption 1 hold and let $\delta$ be a positive constant such that $\delta \leq \epsilon$, where $\epsilon$ is defined by Assumption 1. Then, for any $x \notin X$ there exists a feasible point $x_{in} \in X$ such that

(a) $h_\delta(x_{in}; a_j, b_j) = 0$ for all $j = 1, \ldots, m$,

(b) $\|x - x_{in}\| \leq \|x - \Pi_X[x]\| + \frac{\beta \epsilon \delta}{\alpha_{\min}}$,

where $\alpha_{\min}$ is defined in (16) and $\beta$ is Hoffman’s constant defined in (17).

**Proof.** Let $0 < \delta \leq \epsilon$ and consider the perturbed set $X_\delta$, which is obtained by perturbing the inequalities by amount of $\delta$ toward the interior of $X$ (see Figure 2), i.e.,

$$X_{\delta j} = \{ x \in \mathbb{R}^n \mid \langle a_j, x \rangle - b_j \leq -\delta \}, \quad X_\delta = \cap_{i=1}^m X_{\delta i}.$$

Assumption 1 and the condition $\delta \leq \epsilon$ imply that $X_\delta \neq \emptyset$.

Let us define

$$x_{in} = \Pi_{X_\delta}[x].$$
By the definition of $X_\delta$, we have $\langle a_j, \Pi_{X_\delta}[x] \rangle - b_j \leq -\delta$ for all $j = 1, \ldots, m$. Hence, taking into account the definition of the penalty functions $h_\delta(x_i; a_j, b_j)$, $j = 1, \ldots, m$, (see (8)), we obtain

$$h_\delta(x_i; a_j, b_j) = 0 \quad \text{for all } j = 1, \ldots, m,$$

thus showing the relation in part (a).

To estimate the distance $\|x - x_i\|$, let us consider an intermittent point $\Pi_{X_\delta}[\Pi_X[x]]$ obtained by projecting $x$ on $X$ and by projecting the resulting point on the set $X_\delta$. Since $x_i = \Pi_{X}[x]$ is the closest point in the set $X_\delta$ to $x$,

$$\|x - x_i\| \leq \|x - \Pi_{X_\delta}[\Pi_X[x]]\| \leq \|x - \Pi_X[x]\| + \|\Pi_X[x] - \Pi_{X_\delta}[\Pi_X[x]]\|.$$  

(18)

Next, note that the constant $\beta$ in Hoffmann’s lemma (see (17)) depends only on the vectors $a_i$, $i = 1, \ldots, m$ (not on the values $b_i$). Thus, Hoffmann’s result in (17) applies to the set $X_\delta$ with the same constant $\beta$ as it holds in respect to the set $X$, which implies that

$$\|\Pi_X[x] - \Pi_{X_\delta}[\Pi_X[x]]\| = \text{dist}(\Pi_X[x], X_\delta) \leq \beta \sum_{i=1}^{m} \text{dist}(\Pi_X[x], X_\delta_i).$$

Therefore, according to the definition of $X_{\delta_j}$, it follows that

$$\|\Pi_X[x] - \Pi_{X_\delta}[\Pi_X[x]]\| \leq \beta \sum_{j=1}^{m} \frac{\max(0, \langle a_j, \Pi_X[x] \rangle - b_j + \delta)}{\|a_j\|}.$$

Since $\Pi_X[x] \in X$, we have that $\langle a_j, \Pi_X[x] \rangle - b_j \leq 0$ for all $j$. Hence,

$$\|\Pi_X[x] - \Pi_{X_\delta}[\Pi_X[x]]\| \leq \beta \sum_{j=1}^{m} \frac{\delta}{\|a_j\|} \leq \frac{\beta m \delta}{\alpha_{\min}}.$$

From the preceding relation and relation (18) it follows that

$$\|x - x_i\| \leq \|x - \Pi_X[x]\| + \frac{\beta m \delta}{\alpha_{\min}},$$

thus establishing the result in part (b).
We next turn our attention to the solution sets of the problems. We let \( X^* \) and \( X^*_{y\delta} \) denote the solution sets of the original problem and the penalized problem, respectively, i.e.,
\[
X^* = \left\{ x \in X \mid f(x) = \min_{z \in X} f(z) \right\}, \quad X^*_{y\delta} = \left\{ x \in \mathbb{R}^n \mid F_{y\delta}(x) = \min_{z \in \mathbb{R}^n} F_{y\delta}(z) \right\}.
\]

In our main result establishing that \( X^*_{y\delta} \subset X \), under some conditions on the penalty parameters \( \gamma \) and \( \delta \), we will require that the function \( f \) has uniformly bounded subgradients over a suitably defined region. If the constraint set \( X \) is bounded, then the set \( X \) can be taken as such a region and an upper bound for the subgradient norms can be defined by
\[
L = \max\{||s|| \mid s \in \partial f(x), \ x \in X\},
\]
where \( \partial f(x) \) is the subdifferential set of \( f \) at \( x \). If \( X \) is unbounded, we identify a suitable region in the following lemma. In particular, the region should be large enough to contain the sets \( X^*_{y\delta} \) for a range of penalty values, and also the points \( x_{i\alpha} \) from Lemma 2(b) for each \( x \in X^*_{y\delta} \).

**Lemma 3.** Let Assumption 1 hold and let \( \delta \) be a positive constant such that \( \delta \leq \epsilon \), where \( \epsilon \) is defined by Assumption 1. Assume that \( f \) has bounded level sets. Then, for all \( \gamma > 0 \) and \( \delta > 0 \) satisfying \( \gamma \delta \leq c \) for some \( c > 0 \), there is a ball centered at the origin that contains all the points \( \Pi_X[x] \) with \( x \in X^*_{y\delta} \) and the points \( x_{i\alpha} \) satisfying Lemma 2(b) with \( x \in X^*_{y\delta} \). The radius of this ball depends on some feasible point \( \hat{x} \in X \), the given value of \( c \), the value \( \epsilon \) from Assumption 1, and the problem characteristics reflected in the constants \( \alpha_{\text{min}}, m \) and \( \beta \) from Hoffman’s result (see (17)).

**Proof.** Since \( f \) has bounded level sets, by Corollary 1, the functions \( F_{y\delta} \) also have bounded level sets for all \( \delta \geq 0 \) and \( \gamma \geq 0 \). Hence, the solution set \( X^* \) is nonempty and, also, the solution sets \( X^*_{y\delta} \) are nonempty for all \( \gamma > 0 \) and \( \delta > 0 \). We next employ Corollary 2 to construct a compact set that contains the optimal sets \( X^*_{y\delta} \) are nonempty for all \( \gamma > 0 \) and \( \delta > 0 \) for a range of values of these penalty parameters.

To start, we choose some feasible point \( \hat{x} \in X \) and, by Corollary 2, we obtain
\[
X^*_{y\delta} \subseteq \{x \in \mathbb{R}^n \mid f(x) \leq t_{y\delta}(\hat{x})\} \quad \text{for all } \delta \geq 0 \text{ and } \gamma > 0,
\]
where
\[
t_{y\delta} = f(\hat{x}) + \frac{\gamma \delta}{4 \min_{1 \leq i \leq m} ||a_i||}.
\]
Under the assumption that \( \gamma \delta \leq c \) for some \( c > 0 \), we have \( t_{y\delta} \leq \hat{c} \), where \( \hat{c} = \frac{4 \alpha_{\text{min}}}{m} \) (see (16) for the definition of \( \alpha_{\text{min}} \)). Thus, we consider the level set
\[
\{x \in \mathbb{R}^n \mid f(x) \leq f(\hat{x}) + \hat{c}\},
\]
which is bounded by the assumption that \( f \) has bounded level sets. Furthermore,
\[
X^*_{y\delta} \subseteq \{x \in \mathbb{R}^n \mid f(x) \leq f(\hat{x}) + \hat{c}\} \quad \text{for all } \gamma > 0 \text{ and } \delta > 0 \text{ satisfying } \gamma \delta \leq c.
\]
Hence, these optimal sets are uniformly bounded, i.e., for some \( B_1(\hat{x}, \hat{c}) > 0 \),
\[
||x|| \leq B_1(\hat{x}, \hat{c}) \quad \text{for all } x \in X^*_{y\delta}, \text{ and all } \gamma > 0 \text{ and } \delta > 0 \text{ with } \gamma \delta \leq c.
\]
Since the projection operator is non-expansive, the projections of the points in the set $X_{\gamma \delta}^*$ on the set $X$ are also bounded, i.e., for some $B_2(\hat{x}, \hat{c}) > 0,$

$$\|\Pi_X[x]\| \leq B_2(\hat{x}, \hat{c}) \quad \text{for all } x \in X_{\gamma \delta}^*, \text{ and all } \gamma > 0 \text{ and } \delta > 0 \text{ with } \gamma \delta \leq c. \quad (19)$$

Finally, for each $x \in X_{\gamma \delta}^*$, consider a point $x_m$ as given in Lemma 2. Then, by Lemma 2(b) for each $x \in X_{\gamma \delta}^*$, it follows that

$$\|x_m\| \leq \|x_m - x\| + \|x - \Pi_X[x]\| + \frac{\beta m \delta}{\alpha_{\min}} + \|\Pi_X[x]\| + \frac{\beta m \epsilon}{\alpha_{\min}},$$

where we use assumption that $\delta \leq \epsilon$. Thus, for each $x \in X_{\gamma \delta}^*$, the point $x_m$ from Lemma 2(b) satisfies the following relation

$$\|x_m\| \leq B(\hat{x}, \hat{c}, \epsilon) \quad \text{for all } \gamma > 0 \text{ and } \delta > 0 \text{ with } \delta \leq \epsilon \text{ and } \gamma \delta \leq c,$$

where

$$B(\hat{x}, \hat{c}, \epsilon) = 2B_1(\hat{x}, \hat{c}) + B_2(\hat{x}, \hat{c}) + \frac{\beta m \epsilon}{\alpha_{\min}}.$$ 

In view of (19), the ball centered at the origin with the radius $B(\hat{x}, \hat{c}, \epsilon)$ also contains $\Pi_X[x]$ for all $x \in X_{\gamma \delta}^*$ and for all $\gamma > 0$ and $\delta > 0$, with $\delta \leq \epsilon$ and $\gamma \delta \leq c$. Since $\hat{c} = \frac{c}{4\alpha_{\min}}$, we see that the constant $B(\hat{x}, \hat{c}, \epsilon)$ depends on the choice of the feasible point $\hat{x} \in X$, the given value of $c$, the value $\epsilon$ from Assumption 1, and the problem characteristics reflected in the constants $\alpha_{\min}$, $m$ and $\beta$ from Hoffman’s result (see (17)).

In what follows, we will let $R(c, \epsilon)$ denote the radius of the ball identified in Lemma 3, and suppress the dependence on the other parameters. We define

$$L(c, \epsilon) = \max\{|\tau| : \tau \in \partial f(x), \|\tau\| \leq R(c, \epsilon)|. \quad (20)$$

With Lemma 2 and Lemma 3, we are ready to provide a key relation for the solutions of the penalized problem and the original problem. Specifically, we show that for sufficiently small values of the penalty $\delta$, the solutions of the penalized problem are feasible for the original problem.

**Proposition 1.** Let $\delta^0$ be a given accuracy parameter. Let Assumption 1 hold and assume that $f$ has bounded level sets. Let the parameters $\gamma$ and $\delta$ be chosen such that

$$0 < \delta < \min\left\{\epsilon, \frac{16\alpha_{\min}^2}{\beta^2 m^2}\right\}, \quad \gamma \delta \leq c, \quad \gamma \geq \Gamma,$$

with

$$\Gamma = \max\left\{L\left(\frac{1}{m\beta} - \frac{\sqrt{\delta}}{4\alpha_{\min}}\right)^{-1}, 4mL\alpha_{\max} \left(\frac{1}{\sqrt{\delta}} + \frac{\beta m}{\alpha_{\min}}\right)\right\},$$

where $c > 0$ is arbitrary, $\epsilon$ is the constant from Assumption 1, $\beta$ is the constant from Hoffman’s bound (see (17)), the scalars $\alpha_{\min}$ and $\alpha_{\max}$ are defined in (16), while $L = L(c, \epsilon)$ is defined by (20). Then, every point in the solution set $X_{\gamma \delta}^*$ of the penalized problem is feasible for the problem (5), namely $X_{\gamma \delta}^* \subset X$. 

12
Proof. Since $f$ has bounded level sets, the solution set $X^*$ and the solution sets $X^*_{\gamma \delta}$ are nonempty for all $\gamma > 0$ and $\delta > 0$. To arrive at a contradiction, let us assume that there exists some $\gamma$ and $\delta$ satisfying the conditions in the proposition and such that $X^*_{\gamma \delta} \nsubseteq X$. Thus, there exists a solution $x^*_{\gamma \delta} \in X^*_{\gamma \delta}$ and $x^*_{\gamma \delta} \notin X$. Define

$$\hat{x}^*_{\gamma \delta} = \Pi_X[x^*_{\gamma \delta}].$$

We consider two possibilities: $\|\hat{x}^*_{\gamma \delta} - x^*_{\gamma \delta}\| \geq \sqrt{\delta}$ and $\|\hat{x}^*_{\gamma \delta} - x^*_{\gamma \delta}\| < \sqrt{\delta}$.

Case 1: $\|\hat{x}^*_{\gamma \delta} - x^*_{\gamma \delta}\| \geq \sqrt{\delta}$. By Lemma 1 we have that $h_i(x; a_i, b_i) \geq \text{dist}(x, X_i)$ for all $i = 1, \ldots, m$. Thus, by the definition of the functions $F_{\gamma \delta}$, for any $x \in \mathbb{R}^n$ we can write

$$F_{\gamma \delta}(x) \geq f(x) + \frac{\gamma}{m} \sum_{i=1}^{m} \text{dist}(x, X_i).$$

Then, by Hoffman’s lemma (see (17)), for some $\beta > 0$ we have

$$F_{\gamma \delta}(x) \geq f(x) + \frac{\gamma}{m\beta} \text{dist}(x, X).$$

Letting $x = x^*_{\gamma \delta}$ in the preceding relation, we obtain

$$F_{\gamma \delta}(x^*_{\gamma \delta}) \geq f(x^*_{\gamma \delta}) + \frac{\gamma}{m\beta} \|\hat{x}^*_{\gamma \delta} - x^*_{\gamma \delta}\| + f(\hat{x}^*_{\gamma \delta}) - f(x^*_{\gamma \delta})$$

$$\geq \frac{\gamma}{m\beta} \|\hat{x}^*_{\gamma \delta} - x^*_{\gamma \delta}\| - L\|\hat{x}^*_{\gamma \delta} - x^*_{\gamma \delta}\| + f(\hat{x}^*_{\gamma \delta})$$

$$= \left(\frac{\gamma}{m\beta} - L\right) \|\hat{x}^*_{\gamma \delta} - x^*_{\gamma \delta}\| + F_{\gamma \delta}(\hat{x}^*_{\gamma \delta}) - \frac{\gamma}{m} \sum_{i=1}^{m} h_i(\hat{x}^*_{\gamma \delta}; a_i, b_i),$$

where in the second inequality we use the assumption that the norms of the subgradients in the subdifferential set $\partial f(x)$ are bounded by $L$ in a region containing the point $x = \hat{x}^*_{\gamma \delta}$ (see Lemma 3 and (20)). Taking into the account that $h_i(x; a_i, b_i) \leq \frac{\delta}{\text{dist}(x, X_i)}$ when $x \in X_i$ (see inequality (10) and the definition of the set $X_i$) and using $\hat{x}^*_{\gamma \delta} \in X \subseteq X_i$, we see that

$$F_{\gamma \delta}(x^*_{\gamma \delta}) \geq \left(\frac{\gamma}{m\beta} - L\right) \|x^*_{\gamma \delta} - \hat{x}^*_{\gamma \delta}\| + F_{\gamma \delta}(\hat{x}^*_{\gamma \delta}) - \frac{\gamma\delta}{4m} \sum_{i=1}^{m} \frac{1}{\text{dist}(x, X_i)}.$$ 

Note that the condition $\gamma \geq \Gamma$ and the definition of $\Gamma$ imply $\gamma \geq Lm\beta$. Using the relations $\gamma \geq Lm\beta$ and $\|\hat{x}^*_{\gamma \delta} - x^*_{\gamma \delta}\| \geq \sqrt{\delta}$, which we assumed, we further obtain

$$F_{\gamma \delta}(x^*_{\gamma \delta}) > \left(\frac{\gamma}{m\beta} - L\right) \sqrt{\delta} - \frac{\gamma\delta}{4\alpha_{\min}} + F_{\gamma \delta}(\hat{x}^*_{\gamma \delta}) \geq F_{\gamma \delta}(\hat{x}^*_{\gamma \delta}),$$

where the last inequality is obtained by using $\left(\frac{\gamma}{m\beta} - L\right) \sqrt{\delta} - \frac{\gamma\delta}{4\alpha_{\min}} \geq 0$, which is equivalent to

$$\frac{\gamma}{m\beta} - L - \frac{\gamma\sqrt{\delta}}{4\alpha_{\min}} \geq 0 \quad \iff \quad \gamma \left(\frac{1}{m\beta} - \frac{\sqrt{\delta}}{4\alpha_{\min}}\right) \geq L.$$
The last inequality holds due to the conditions that we imposed on the parameters \( \gamma \) and \( \delta \), namely, that \( \delta < \frac{16a_m^2}{\beta m^2} \) and \( \gamma \geq L \left( \frac{1}{m \bar{b}} - \frac{\delta^2}{4m \alpha_{\text{min}}} \right)^{-1} \). Thus, relation (21) implies that \( F_{\gamma \delta}(x_{\gamma \delta}^*) > F_{\gamma \delta}(\hat{x}_{\gamma \delta}^*) \), which contradicts the fact that \( \hat{x}_{\gamma \delta}^* \) is an unconstrained minimizer of \( F_{\gamma \delta} \).

**Case 2:** \( ||\hat{x}_{\gamma \delta}^* - x_{\gamma \delta}^*|| < \sqrt{\delta} \). Since \( x_{\gamma \delta}^* \notin X \), under Assumption 1 and the condition \( \delta \leq \epsilon \), we can apply Lemma 2 with \( x = x_{\gamma \delta}^* \). According to Lemma 2, there exists a feasible point \( x_{\gamma \delta}^* \in X \) such that

\[
h_\delta(x_{\gamma \delta}^*; a_i, b_i) = 0 \quad \text{for all } i = 1, \ldots, m,
\]

and

\[
||x_{\gamma \delta}^* - x_{\gamma \delta}|| \leq ||x_{\gamma \delta}^* - \Pi_X[x_{\gamma \delta}^*]|| + \frac{\beta m \delta}{\alpha_{\text{min}}},
\]

Using the point \( x_{\gamma \delta}^* \), we have

\[
F_{\gamma \delta}(x_{\gamma \delta}^*) - F_{\gamma \delta}(x_{\gamma \delta}^*) = f(x_{\gamma \delta}^*) - f(x_{\gamma \delta}^*) + \frac{\gamma}{m} \left( \sum_{i=1}^{m} h_\delta(x_{\gamma \delta}^*; a_i, b_i) - h_\delta(x_{\gamma \delta}^*; a_i, b_i) \right)
\geq -L||x_{\gamma \delta}^* - x_{\gamma \delta}|| + \frac{\gamma}{m} \sum_{i=1}^{m} h_\delta(x_{\gamma \delta}^*; a_i, b_i),
\]

where we use the assumption that \( f \) has bounded subgradients over the region containing the point \( x_{\gamma \delta}^* \) (see Lemma 3 and (20)) and relation (22). Since \( x_{\gamma \delta}^* \notin X \), there exists a constraint \( j \) that is violated at \( x_{\gamma \delta}^* \), i.e., we have

\[
\langle a_j, x_{\gamma \delta}^* \rangle - b_j > 0.
\]

For the violated constraint \( j \), by property (11), for the penalty function \( h_\delta(\cdot; a_j, b_j) \), we have

\[
h_\delta(x_{\gamma \delta}^*; a_j, b_j) > \frac{\delta}{4||a_j||} \geq \frac{\delta}{4\alpha_{\text{max}}}.
\]

Using (25) and the fact that the penalty functions are non-negative (see (9)), we obtain

\[
\sum_{i=1}^{m} h_\delta(x_{\gamma \delta}^*; a_i, b_i) \geq h_\delta(x_{\gamma \delta}^*; a_j, b_j) > \frac{\delta}{4\alpha_{\text{max}}}.
\]

Substituting estimate (26) in relation (24) we further obtain

\[
F_{\gamma \delta}(x_{\gamma \delta}^*) - F_{\gamma \delta}(x_{\gamma \delta}^*) > -L||x_{\gamma \delta}^* - x_{\gamma \delta}|| + \frac{\gamma \delta}{4m \alpha_{\text{max}}}
\geq -L \left( ||x_{\gamma \delta}^* - \Pi_X[x_{\gamma \delta}^*]|| + \frac{\beta m \delta}{\alpha_{\text{min}}} \right) + \frac{\gamma \delta}{4m \alpha_{\text{max}}},
\]

where the last inequality is obtained by using (23). Since \( \hat{x}_{\gamma \delta}^* = \Pi_X[x_{\gamma \delta}^*] \) and we work under the assumption that \( ||\hat{x}_{\gamma \delta}^* - x_{\gamma \delta}|| < \sqrt{\delta} \), from (27) we have

\[
F_{\gamma \delta}(x_{\gamma \delta}^*) - F_{\gamma \delta}(x_{\gamma \delta}^*) > -L \left( \sqrt{\delta} + \frac{\beta m \delta}{\alpha_{\text{min}}} \right) + \frac{\gamma \delta}{4m \alpha_{\text{max}}} \geq 0,
\]

(28)
where the last inequality is due to the conditions imposed on \( \delta \) and \( \gamma \), namely, the condition that 
\[
\gamma \geq 4mL\alpha_{\max} \left( \frac{1}{\sqrt{\delta}} + \frac{\beta m}{\sigma_{\min}} \right),
\]
which is equivalent to 
\[
\frac{\gamma}{4m\alpha_{\max}} - L \left( \frac{\sqrt{\delta}}{\sigma_{\min}} + \frac{\beta m}{\sigma_{\min}} \right) \geq 0.
\]
Hence, it follows that
\[
F_{\gamma \delta}(x^*_{\gamma \delta}) - F_{\gamma \delta}(x'_{\gamma \delta}) > 0,
\]
which contradicts the fact that \( x^*_{\gamma \delta} \) is an unconstrained minimizer of \( F_{\gamma \delta} \).

Let us note that, for the constant \( \Gamma \) in Proposition 1, we have \( \Gamma > 0 \) in view of the condition
\[
0 < \delta < \frac{16\alpha^2_{\min}}{\beta^2 m^2}.
\]
The condition \( \gamma \delta \leq c \) in Proposition 1 is imposed only to ensure the existence of the subgradient norm bound \( L(c, \epsilon) \). One way to think about the choices of \( \gamma \) and \( \delta \) that satisfy the conditions in Proposition 1 is as follows. We first select a penalty value \( \delta > 0 \) that satisfies \( \delta < \min \{ \epsilon, \frac{16\alpha^2_{\min}}{\beta^2 m^2} \} \).

Then, we choose a large \( c \) and constrain \( \gamma \) from above by \( \hat{\gamma} \). For the given \( c \), we determine an estimate \( \hat{L} \geq L(c, \epsilon) \). Having \( \hat{L} \), we compute \( \hat{\Gamma} \) by using \( \hat{L} \) instead of \( L \) in the definition of \( \Gamma \), and then impose the constraint \( \gamma \geq \hat{\Gamma} \).

Now, we provide a relation between optimal values for the original and penalized problems. We consider the cases when \( f \) is strongly convex and non-strongly convex, separately.

The following proposition establishes a key relation between solutions \( x^*_{\gamma \delta} \) and \( x^* \) for the case when \( f \) is strongly convex. In particular, the proposition provides a set of conditions on the parameters \( \delta \) and \( \gamma \) ensuring that the distance between \( x^*_{\gamma \delta} \) and \( x^* \) does not exceed a desired accuracy \( \delta^0 \), i.e., \( \|x^*_{\gamma \delta} - x^*\|^2 \leq \delta^0 \).

**Proposition 2.** Let \( \delta^0 \) be a given accuracy parameter. Let Assumption 1 hold and let \( f \) be strongly convex with a constant \( \mu_f > 0 \). Let the parameters \( \gamma \) and \( \delta \) be chosen such that

\[
0 < \delta < \min \left\{ \epsilon, \frac{16\alpha^2_{\min}}{\beta^2 m^2} \right\}, \quad \Gamma \leq \gamma \leq \frac{2\mu_f\alpha_{\min}\delta^0}{\delta},
\]

with

\[
\Gamma = \max \left\{ L \left( \frac{1}{m\beta} - \frac{\sqrt{\delta}}{4\alpha_{\min}} \right)^{-1}, \ 4mL\alpha_{\max} \left( \frac{1}{\sqrt{\delta}} + \frac{\beta m}{\sigma_{\min}} \right) \right\},
\]

where \( \epsilon \) is the constant from Assumption 1, \( \beta \) is the constant from Hoffman’s bound (see (17)), the scalars \( \alpha_{\min} \) and \( \alpha_{\max} \) are defined in (16), while \( L = L(c, \epsilon) \) is the bound on the subgradient norms as defined in (20) with \( c > 2\mu_f\alpha_{\min}\delta^0 \). Then, the original problem (5) and the penalized problem (6) have unique solutions, \( x^* \) and \( x^*_{\gamma \delta} \), respectively, which satisfy the following relation:

\[
\|x^*_{\gamma \delta} - x^*\|^2 \leq \delta^0.
\]

**Proof.** Since the function \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) is strongly convex with a constant \( \mu_f > 0 \), by the convexity of the penalty function \( h_{\delta \gamma} \), the penalized objective function \( F_{\gamma \delta} \) in (7) is also strongly convex with
the same strong convexity constant $\mu$, for any $\gamma \geq 0$. Hence, the original problem (5) and the penalized problem (6) have unique solutions, denoted respectively by $x^*$ and $x_{\gamma \delta}^*$.

By the relations $c > 2\mu f_{\text{min}}\delta^0$ and $\gamma \leq \frac{2\mu f_{\text{min}}\delta^0}{\delta}$, it follows that $\gamma \delta \leq c$. Thus, the conditions of Proposition 1 are satisfied. According to Proposition 1, the vector $x_{\gamma \delta}^*$ is feasible i.e., $x_{\gamma \delta}^* \in X$, implying that

$$f(x^*) \leq f(x_{\gamma \delta}^*).$$

Since the penalty functions are non-negative, we have $h_\delta(x^*; a_i, b_i) \geq 0$ for all $i = 1, \ldots, m$ (see (9)). The point $x^*$ is feasible but it may be penalized, in which case $h_\delta(x^*; a_i, b_i) \leq \frac{\delta}{4\|a_i\|}$ for all $i = 1, \ldots, m$ (see (10)). Therefore, we have

$$h_\delta(x^*; a_i, b_i) - h_\delta(x_{\gamma \delta}^*; a_i, b_i) \leq \frac{\delta}{4\|a_i\|} \quad \text{for all } i = 1, \ldots, m. \quad (30)$$

Using relations (33) and (30), we obtain

$$F_{\gamma \delta}(x^*) - F_{\gamma \delta}(x_{\gamma \delta}^*) = f(x^*) - f(x_{\gamma \delta}^*) + \frac{\gamma}{m} \left( \sum_{i=1}^{m} h_\delta(x^*; a_i, b_i) - h_\delta(x_{\gamma \delta}^*; a_i, b_i) \right) \leq \frac{\gamma \delta}{4\alpha_{\text{min}}}. \quad (31)$$

By the strong convexity of $F_{\gamma \delta}$, it follows that

$$\|x^* - x_{\gamma \delta}^*\|^2 \leq \frac{2}{\mu f} (F_{\gamma \delta}(x^*) - F_{\gamma \delta}(x_{\gamma \delta}^*)) \leq \frac{\gamma \delta}{2\mu f \alpha_{\text{min}}} \leq \delta^0, \quad (32)$$

where the last inequality in the preceding relation is due to the choice of $\gamma \leq \frac{2\mu f \alpha_{\text{min}} \delta^0}{\delta}$.

By slightly adapting the choices of $\delta$ and $\gamma$, we can provide an estimate for the function value $f(x_{\gamma \delta})$ at a solution $x_{\gamma \delta}$ of the penalized problem. For this, let us define

$$f^* = \min_{z \in X} f(z).$$

We have the following result for these optimal values.

**Proposition 3.** Let $\delta^0$ be a given accuracy parameter. Let Assumption 1 hold, and assume that $f$ is convex and has bounded level sets. Let the parameters $\gamma$ and $\delta$ be chosen such that

$$0 < \delta < \min \left\{ \epsilon, \frac{16\alpha_{\text{min}}^2}{\beta^2 m^2} \right\}, \quad \Gamma \leq \gamma \leq \frac{4\alpha_{\text{min}} \delta^0}{\delta},$$

with

$$\Gamma = \max \left\{ L \left( \frac{1}{m \beta} - \frac{\sqrt{\delta}}{4\alpha_{\text{min}}} \right)^{-1}, \frac{4mL \alpha_{\text{max}} \left( \frac{1}{\sqrt{\delta}} + \frac{\beta m}{\alpha_{\text{min}}} \right)}{4} \right\},$$

where $\epsilon$ is the constant from Assumption 1, $\beta$ is the constant from Hoffman’s bound (see (17)), the scalars $\alpha_{\text{min}}$ and $\alpha_{\text{max}}$ are defined in (16), while $L = L(c, \epsilon)$ is the bound on the subgradient norms as defined in (20) with $c > 4\alpha_{\text{min}} \delta^0$. Then, we have

$$0 \leq f(x_{\gamma \delta}^*) - f^* \leq \delta^0 \quad \text{for any } x_{\gamma \delta}^* \in X_{\gamma \delta}^*. \quad (29)$$
Proof. By the assumption that \( f \) has bounded level sets, the solution sets \( X^* \) and \( X^*_{\gamma \delta} \), for any \( \delta, \gamma \geq 0 \), are nonempty. In view of relations \( c > 4 \alpha \min \delta^0 \) and \( \gamma \leq \frac{4 \alpha \min \delta^0}{\delta} \), it follows that \( \gamma \delta \leq c \). Hence, all the conditions of Proposition 1 are satisfied. By Proposition 1, the solutions of the penalized problem are feasible, i.e., \( X^*_{\gamma \delta} \subseteq X \), implying that

\[
0 \leq f(x^*_\gamma \delta) - f^*.
\]  

(33)

Now, let \( x^*_\gamma \delta \in X^*_{\gamma \delta} \) and \( x^* \in X^* \) be arbitrary solutions, and consider the difference \( F_{\gamma \delta}(x^*_\gamma \delta) - F_{\gamma \delta}(x^*) \). By the definition of the functions \( F_{\gamma \delta} \) we have

\[
F_{\gamma \delta}(x^*_\gamma \delta) - F_{\gamma \delta}(x^*) = f(x^*_\gamma \delta) - f(x^*) + \frac{\gamma}{m} \sum_{i=1}^{m} \left( h_{\delta}(x^*_\gamma \delta; a_i, b_i) - h_{\delta}(x^*; a_i, b_i) \right).
\]

Since \( x^*_\gamma \delta \in X^*_{\gamma \delta} \), it follows that \( F_{\gamma \delta}(x^*_\gamma \delta) - F_{\gamma \delta}(x^*) \leq 0 \), thus implying that

\[
f(x^*_\gamma \delta) - f(x^*) \leq \frac{\gamma}{m} \sum_{i=1}^{m} \left( h_{\delta}(x^*_\gamma \delta; a_i, b_i) - h_{\delta}(x^*; a_i, b_i) \right).
\]

The functions \( h_{\delta}(:, a_i, b_i) \) are nonnegative, so it follows that

\[
f(x^*_\gamma \delta) - f(x^*) \leq \frac{\gamma}{m} \sum_{i=1}^{m} h_{\delta}(x^*; a_i, b_i).
\]

In view of the maximum penalty over feasible region (cf. (10)) and since \( x^* \in X \), we have that \( h_{\delta}(x^*; a_i, b_i) \leq \frac{\delta}{\|a_i\|} \) for all \( i = 1, \ldots, m \). Therefore,

\[
f(x^*_\gamma \delta) - f(x^*) \leq \frac{\gamma \delta}{4 \alpha \min}.\]

By the condition \( \gamma \leq \frac{4 \alpha \min \delta^0}{\delta} \), it follows that

\[
f(x^*_\gamma \delta) - f(x^*) \leq \delta^0.
\]

4. Applying SAGA to Penalized Problem

With the results presented in Proposition 2 and Proposition 3 in place, we can proceed with formulating fast incremental methods to find a solution of the penalized problem (6). Recently, many algorithms have been proposed to incrementally solve the following optimization problem of minimizing the average sum of functions:

\[
\min_{x \in \mathbb{R}^n} G(x), \quad G(x) = \frac{1}{N} \sum_{i=1}^{N} g_i(x).
\]  

(34)
Among these algorithms are, for example, SAG, SAGA, and SVRG [9, 10, 18], which leverage the idea to randomly sample the full gradient by processing only one function per iteration in a way to reduce the variance in the gradient estimation. Under the assumption of Lipschitz continuous gradients $\nabla g_i$, these algorithms possess the same asymptotic convergence rate to an optimal solution as the standard full gradient method requiring the full sum of the gradients $\nabla g_i$ at each iteration. More precisely, given an optimal choice of step size parameters, the aforementioned incremental methods approach an optimal solution with the convergence rate $O(q^t)$, $q \in (0, 1)$, in the case of strongly convex function $G$, and the convergence rate $O(1/t)$ in the case of non-strongly convex function $G$.

As an example of a fast incremental method, we will consider the SAGA algorithm\(^1\). The algorithm is summarized as follows.

**Algorithm 1 SAGA Algorithm**

0. Let $x^0 \in \mathbb{R}^n$ and $\nabla g_i(\phi^0_i)$ with $\phi^0_i = x^0$, $i = 1, \ldots, N$, be known.
1. Pick an index $j$ uniformly at random.
2. $\phi^{t+1}_j = x^t$.
3. $x^{t+1} = x^t - \alpha \left[ \nabla g_j(\phi^{t+1}_j) - \nabla g_j(\phi^t_j) + \frac{1}{N} \sum_{i=1}^{N} \nabla g_i(\phi^t_i) \right]$.

The main result for the SAGA algorithm is formulated in the following theorem, which is adapted from [9].

**Theorem 1. (\cite{9})**

(a) Let the functions $g_i$, $i = 1, \ldots, N$, be strongly convex with a parameter $\mu > 0$ and have Lipschitz continuous gradients with a constant $L_g > 0$. Let $x^*_g$ be the solution of the problem (34). Then, if the step size $\alpha = \frac{1}{2(\mu N + L_g)}$ is chosen in SAGA algorithm, then the following convergence rate result holds:

$$
E\|x^t - x^*_g\|^2 \leq O(q^t), \quad q = 1 - \frac{\mu}{2(\mu N + L_g)}.
$$

(b) Let the functions $g_i$, $i = 1, \ldots, N$, be non-strongly convex and have Lipschitz continuous gradients with a constant $L_g > 0$. Let $G^*$ be the optimal value of the problem (34) and $\bar{x} = \frac{1}{T} \sum_{k=1}^{T} x^k$. Then, if the step size $\alpha = \frac{1}{4L_g}$ is chosen, then the following convergence rate result is valid:

$$
E[G(\bar{x})] - G^* \leq O\left(\frac{4N}{T}\right).
$$

By applying Algorithm 1 to the penalized problem (6) under our consideration, namely by taking $g_i(x) = f_i(x) + \gamma h_{x_i}(x; a_i, b_i)$, we get the following incremental algorithm to find its solution.

---

\(^1\)The SAGA method in [9] is formulated for a composite objective function $G(x) = \frac{1}{N} \sum_{i=1}^{N} g_i(x) + h(x)$, where the proximal operator associated with the convex function $h$ is easy to evaluate. However, in our setting, it suffices to consider the case $h(x) = 0$. 

18
In terms of the original optimization problem \((5)\) the following result holds, as a direct consequence of Theorem 1, and Proposition 2 and Proposition 3.

**Theorem 2.** Let Assumption 1 hold.

(a) Let the function \(f\) be strongly convex with a parameter \(\mu_f > 0\) and have Lipschitz continuous gradients with a constant \(L_f > 0\). Let \(x^*\) be the solution of the problem \((5)\). Assume that an accuracy level \(\delta^0\) is given, the penalty parameters \(\gamma\) and \(\delta\) are chosen to satisfy the conditions of Proposition 2, and the step size \(\alpha = \frac{1}{2(\mu_f m + L_f + \frac{\gamma\max_i}{\delta})}\) is selected. Then, the following convergence rate result is valid for the iterates of Algorithm 2:

\[
\mathbb{E}[\|x - \Pi_x[x]\|^2] \leq O\left(q_y^t\right), \quad \mathbb{E}[\|x - x^*\|^2] \leq O\left(q_y^t\right) + 2\delta^0, \quad q_y = 1 - \frac{\mu_f}{2(\mu_f m + L_f + \frac{\gamma\max_i}{\delta})}.
\]

(b) Let the function \(f\) be convex, have bounded level sets and have Lipschitz continuous gradients with a constant \(L_f > 0\). Let \(f^*\) be the optimal value problem \((5)\). Suppose that a desired accuracy level \(\delta^0\) is given, the penalty parameters \(\gamma\) and \(\delta\) are chosen to satisfy the conditions of Proposition 3, and the step size \(\alpha\) is given by \(\alpha = \frac{1}{3L_f + \frac{\gamma\max_i}{\delta}}\). Then, for the averages \(\bar{x} = \frac{1}{t}\sum_{k=1}^{t} x^k\) of the iterates \(x^k\) generated by Algorithm 2 the following holds:

\[
E[f(\bar{x})] - f^* \leq O\left(\frac{4m}{t}\right) + 2\delta^0 \quad \text{for all } t,
\]

and for any \(\varepsilon > 0\) there exists \(T > 0\) such that for all \(t > T\),

\[-\gamma\varepsilon \leq E[f(\bar{x})] - f^*.
\]

**Proof.** We apply Theorem 1 to the problem \((6)\) with the objective function

\[
F_{\gamma\delta} = \frac{1}{m}\sum_{i=1}^{m} g_i(x),
\]

where \(g_i(x) = f(x) + \gamma h_\delta (x; a_i, b_i), i = 1, \ldots, m\). Recall that, if the function \(f\) is strongly convex with a constant \(\mu_f > 0\), then the functions \(g_i(x)\) are strongly convex with the same constant \(\mu_f\). Moreover, the gradients of the functions \(g_i(x)\) are Lipschitz continuous with the constant \(L_f + \frac{\gamma\max_i}{\delta}\),

---

**Algorithm 2** SAGA-based Fast Incremental Method for Solving Penalized Problem

1. Pick an index \(j\) uniformly at random.
2. \(\phi_j = x^j\).
3. \(x^{j+1} = x^j - \alpha[\nabla f(\phi_j^j) + \gamma \nabla h_\delta (\phi_j^j; a_j, b_j) - \nabla f(\phi_j^j) - \gamma \nabla h_\delta (\phi_j^j; a_j, b_j)] + \frac{1}{m}\sum_{i=1}^{m} (\nabla f(\phi_i^j) + \gamma \nabla h_\delta (\phi_i^j; a_i, b_i)).\)
since each penalty function $h_i(x; a_i, b_i), i = 1, \ldots, m$, has the Lipschitz continuous gradient with the constant $\frac{\mu_i}{2\delta_i}$ (see (15)).

To obtain the result in part (a), let us notice that, according to Proposition 1 we have $x^*_{\gamma_0} \in X$. Hence,

$$\mathbb{E}[\|x' - \Pi_X[x']\|^2] \leq \mathbb{E}[\|x' - x^*_{\gamma_0}\|^2].$$

Thus, due to Theorem 1(a), Proposition 2, and the inequality $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, which is valid for any $a, b \in \mathbb{R}^n$, we conclude that

$$\mathbb{E}[\|x' - \Pi_X[x']\|^2] \leq O(\gamma'), \quad \mathbb{E}[\|x' - x^*\|^2] \leq O(\gamma') + 2\delta^0,$$

where $\gamma' = 1 - \frac{\mu_i}{2(\mu_i + \epsilon_i + \eta_i + \delta_i)}$.

To prove part (b), we consider $\mathbb{E}[F_{\gamma_0}(\bar{x})] - F_{\gamma_0}^*$, for which, according to Theorem 1(b), we have

$$0 \leq \mathbb{E}[F_{\gamma_0}(\bar{x})] - F_{\gamma_0}^* \leq O\left(\frac{4m}{t}\right). \quad (35)$$

By using the definition of the penalty function $F_{\gamma_0}$, for any $x^*_{\gamma_0} \in X_{\gamma_0}$ we can write

$$\mathbb{E}[F_{\gamma_0}(\bar{x})] - F_{\gamma_0}^* = \mathbb{E}[f(\bar{x})] - f(x^*_{\gamma_0}) + \frac{\gamma}{m} \sum_{i=1}^{m} \left[ \mathbb{E}[h_\delta(x^*_{\gamma_0}; a_i, b_i)] - h_\delta(x^*_{\gamma_0}; a_i, b_i) \right].$$

From the preceding relation, using the fact that the functions $h_\delta(\cdot; a_i, b_i)$ are nonnegative and using (35), we obtain

$$\mathbb{E}[f(\bar{x})] - f(x^*_{\gamma_0}) \leq O\left(\frac{4m}{t}\right) + \frac{\gamma}{m} \sum_{i=1}^{m} h_\delta(x^*_{\gamma_0}; a_i, b_i).$$

By adding and subtracting $f^*$ and re-arranging the terms, we further obtain

$$\mathbb{E}[f(\bar{x})] - f^* \leq O\left(\frac{4m}{t}\right) + f(x^*_{\gamma_0}) - f^* + \frac{\gamma}{m} \sum_{i=1}^{m} h_\delta(x^*_{\gamma_0}; a_i, b_i).$$

By Proposition 3, we have $f(x^*_{\gamma_0}) - f^* \leq \delta^0$, implying that

$$\mathbb{E}[f(\bar{x})] - f^* \leq O\left(\frac{4m}{t}\right) + \delta^0 + \frac{\gamma}{m} \sum_{i=1}^{m} h_\delta(x^*_{\gamma_0}; a_i, b_i).$$

By Proposition 1, the point $x^*_{\gamma_0}$ is feasible for the original problem, so that we have $h_\delta(x^*_{\gamma_0}; a_i, b_i) \leq \frac{\delta}{4\mu_i}$ for all $i = 1, \ldots, m$ (see (10)). Therefore, for all $t$ we have

$$\mathbb{E}[f(\bar{x})] - f^* \leq O\left(\frac{4m}{t}\right) + \delta^0 + \frac{\gamma\delta}{4\mu_{\min}} \leq O\left(\frac{4m}{t}\right) + 2\delta^0,$$

where the last inequality follows in view of the condition $\gamma \leq \frac{4\mu_{\min}\delta^0}{\delta}$ of Proposition 3.
Next, we provide a lower bound on $\mathbb{E}[f(\bar{x}')] - f^*$. We write

$$\mathbb{E}[f(\bar{x}')] - f^* = \mathbb{E}[f(\bar{x}')] - f(x^*_{\gamma \delta}) + f(x^*_{\gamma \delta}) - f^* \geq \mathbb{E}[f(\bar{x}')] - f(x^*_{\gamma \delta}),$$

where $x^*_{\gamma \delta}$ is an arbitrary solution of the penalized problem, i.e., $F(x^*_{\gamma \delta}) = F^*_{\gamma \delta}$, and the last inequality is obtained by using $f(x^*_{\gamma \delta}) - f^* \geq 0$ (see Proposition 3). By using the convexity of $f$, we further have

$$[\mathbb{E}[f(\bar{x}')] - f^* \geq f(\mathbb{E}[\bar{x}']) - f(x^*_{\gamma \delta}).]$$

By the definition of the penalty function, we have

$$f(\mathbb{E}[\bar{x}']) - f(x^*_{\gamma \delta}) = F_{\gamma \delta}(\mathbb{E}[\bar{x}']) - F^*_{\gamma \delta} - \gamma \sum_{i=1}^{m} \left\{ h_{\delta}(\mathbb{E}[\bar{x}']; a_i, b_i) - h_{\delta}(x^*_{\gamma \delta}; a_i, b_i) \right\}.$$ 

Hence, for any $x^*_{\gamma \delta} \in X^*_{\gamma \delta}$,

$$\mathbb{E}[f(\bar{x}')] - f^* \geq F_{\gamma \delta}(\mathbb{E}[\bar{x}']) - F^*_{\gamma \delta} + \gamma \sum_{i=1}^{m} \left\{ h_{\delta}(x^*_{\gamma \delta}; a_i, b_i) - h_{\delta}(\mathbb{E}[\bar{x}']; a_i, b_i) \right\}$$

$$\geq \gamma \sum_{i=1}^{m} \left\{ h_{\delta}(x^*_{\gamma \delta}; a_i, b_i) - h_{\delta}(\mathbb{E}[\bar{x}']; a_i, b_i) \right\},$$

where the last inequality follows by $F_{\gamma \delta}(\mathbb{E}[\bar{x}']) - F^*_{\gamma \delta} \geq 0$ since $x^*_{\gamma \delta}$ is a minimizer of $F_{\gamma \delta}$. The function $h_{\delta}(\cdot; a, b)$ has bounded gradient norms by 1 (see (12)–(14)), implying that for all $i = 1, \ldots, m$ and for all $x^*_{\gamma \delta} \in X^*_{\gamma \delta}$,

$$h_{\delta}(x^*_{\gamma \delta}; a_i, b_i) - h_{\delta}(\mathbb{E}[\bar{x}']; a_i, b_i) \geq -\|x^*_{\gamma \delta} - \mathbb{E}[\bar{x}']\|.$$ 

By choosing a particular solution $\Pi_{X^*_{\gamma \delta}}[\mathbb{E}[\bar{x}']]$, we have

$$\mathbb{E}[f(\bar{x}')] - f^* \geq -\gamma \|\Pi_{X^*_{\gamma \delta}}[\mathbb{E}[\bar{x}']] - \mathbb{E}[\bar{x}']\|. \tag{36}$$

According to Theorem 1(b), for Algorithm 2 there holds

$$\lim_{t \to \infty} \mathbb{E}F_{\gamma \delta}(\bar{x}') = F^*_{\gamma \delta}.$$ 

By the convexity of the function $F_{\gamma \delta}$ and the fact that $F^*_{\gamma \delta}$ is its unconstrained minimum value, it follows that

$$\lim_{t \to \infty} F_{\gamma \delta}(\mathbb{E}[\bar{x}']) = F^*_{\gamma \delta}.$$ 

Thus, any limit point of the sequence $\{\mathbb{E}[\bar{x}']\}$ belongs to the set of minimizers $X^*_{\gamma \delta}$ of the function $F_{\gamma \delta}$. Hence, for any given $\tilde{\epsilon} > 0$ there exists $T > 0$ such that for all $t > T$,

$$\|\mathbb{E}[\bar{x}'] - \Pi_{X^*_{\gamma \delta}}[\mathbb{E}[\bar{x}']]\| \leq \tilde{\epsilon}. \tag{37}$$

The result follows from (36) and (37).

We emphasize that Algorithm 2 presented above is just an example of fast incremental methods which use the Lipschitz gradient property of the objective function and are, thus, applicable to the penalized optimization problem (6). Other methods with potentially better rate dependence on the problem’s parameters include [2, 9, 10, 18, 24, 35]. All these algorithms guarantee a fast convergence to a feasible point lying within some $\delta^0$-neighborhood of an optimal solution for a predefined accuracy parameter $\delta^0 > 0$.  

21
5. Simulation Results

To test the theoretic results presented above, we consider the problem of minimizing the quadratic function \( f(x) = \frac{1}{2} \| x - x^0 \|^2 \), \( x \in \mathbb{R}^3 \) subject to the set of linear inequality constraints. Here \( x^0 \in \mathbb{R}^3 \), is chosen at random from the normal distribution with the mean value 0 and variance 10. The set of linear constraints is chosen in such a way that its interior is not empty and the optimal solution is located on its boundary.

The run of two algorithms, namely the SAGA procedure for solving the problem based on the penalized function approach (PA/SAGA) from Algorithm 2 and the random projection algorithm (RandProj) from [29], are presented on Figures 3-6 for the number of inequality constraints \( m = 25, 100, 300, 500 \) respectively. As we can see, during the first 1000 iterations SAGA-based algorithm outperforms the random projection procedure by decreasing the relative error \( \frac{\| x_t - x^* \|}{\| x^* \|} \) faster. Moreover, the termination state \( x^T, T = 1000 \), in RandProj occurs to be non-feasible in around 16% of implementations, whereas for a specific setting of \( \delta \) and \( \gamma \) all implementations of PA/SAGA terminate at a feasible point \( x^T, T = 1000 \).

6. Conclusion

In this paper, we provided a novel penalty re-formulation for a convex minimization problem with linear constraints. The structure of the penalty functions that we used to penalize the linear
constraints, and the suitable choices of the penalty parameters render the penalized unconstrained problem with solutions that are feasible for the original constrained problem. In addition, with an additional constraint on the penalty parameters imposed by a desired accuracy level, the solutions of the penalized unconstrained problem are guaranteed to be arbitrarily close to the solution set of the original problem. An advantage of the proposed penalty reformulation is in the ability to employ fast incremental gradient methods, such as SAGA.

Acknowledgement

A. Nedić gratefully acknowledges a support of this work by the Office of Naval Research grant no. N00014-12-1-0998.

References

References

[1] J. W. Adams. FIR digital filters with least-squares stopbands subject to peak-gain constraints. IEEE Transactions on Circuits and Systems, 38(4):376–388, Apr 1991.
[2] Z. Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. In Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, pages 1200–1205, New York, NY, USA, 2017. ACM.
[3] D. Bertsekas. Constrained Optimization and Lagrange Multiplier Methods (Optimization and Neural Computation Series). Athena Scientific, 1 edition, 1996.
[4] D. P. Bertsekas. Incremental least squares methods and the extended Kalman filter. SIAM Journal on Optimization, 6:807–822, 1996.
[5] D. P. Bertsekas. A hybrid incremental gradient method for least squares. SIAM Journal on Optimization, 7:913–926, 1997.
[6] D. P. Bertsekas. Incremental proximal methods for large scale convex optimization. Mathematical Programming, 129(2):163–195, 2011.
[7] D. P. Bertsekas. Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. available on arxiv at https://arxiv.org/abs/1507.01030, 2015.
[8] G. Calafiore and B. T. Polyak. Stochastic algorithms for exact and approximate feasibility of robust Lmis. IEEE Transactions on Automatic Control, 40(11):1755–1759, 2001.
[9] A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, Advances in Neural Information Processing Systems 27, pages 1646–1654. Curran Associates, Inc., 2014.
[10] A. Defazio, J. Domke, and Caetano. Finito: A faster, permutable incremental gradient method for big data problems. In E. P. Xing and T. Jebara, editors, Proceedings of the 31st International Conference on Machine Learning, volume 32 of Proceedings of Machine Learning Research, pages 1125–1133, Beijing, China, 22–24 Jun 2014. PMLR.
[11] G. Dorini, P. Pinson, and H. Madsen. Chance-constrained optimization of demand response to price signals. IEEE Transactions on Smart Grid, 4(4):2072–2080, Dec 2013.
[12] M. Gaudioso, G. Giallombardo, and G. Miglionico. An incremental method for solving convex finite min-max problems. Mathematics of Operations Research, 31:173–187, 2006.
[13] L. Grippo. A class of unconstrained minimization methods for neural network training. Optimization Methods and Software, 4:135–150, 1994.
[14] L. Grippo. Convergent on-line algorithms for supervised learning in neural networks. IEEE Transactions on Neural Networks, 11:1284–1299, 2000.

[15] O. Guler, A. Hoffman, and U. Rothblum. Approximations to Solutions to Systems of Linear Inequalities. DIMACS technical report. DIMACS, Center for Discrete Mathematics and Theoretical Computer Science, 1992.

[16] E. S. Helou and A. R. De Pierro. Incremental subgradients for constrained convex optimization, a unified framework and new methods. SIAM Journal on Optimization, 20:1547–1572, 2009.

[17] B. Johansson, M. Rabi, and M. Johansson. A randomized incremental subgradient method for distributed optimization in networked systems. SIAM Journal on Optimization, 20:1157–1170, 2009.

[18] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, editors, NIPS, pages 315–323, 2013.

[19] V. M. Kibardin. Decomposition into functions in the minimization problem. Automation and Remote Control, 40:1311–1323, 1980.

[20] K. C. Kiwiel. Convergence of approximate and incremental subgradient methods for convex optimization. SIAM Journal on Optimization, 14:807–840, 2004.

[21] A. Kundu, F. Bach, and C. Bhattacharyya. Convex optimization over intersection of simple sets: improved convergence rate guarantees via an exact penalty approach. Available on arxiv at https://arxiv.org/abs/1710.06465, 2017.

[22] D. Leventhal and A. S. Lewis. Randomized methods for linear constraints: Convergence rates and conditioning. Mathematics of Operations Research, 35(3):641–654, 2010.

[23] W. Li and J. Swetits. The linear L1 estimator and the huber m-estimator. SIAM Journal on Optimization, 8(2):457–475, 1998.

[24] H. Lin, J. Mairal, and Z. Harchaoui. A universal catalyst for first-order optimization. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, Advances in Neural Information Processing Systems 28, pages 3384–3392. Curran Associates, Inc., 2015.

[25] Z. Q. Luo. On the convergence of the lms algorithm with adaptive learning rate for linear feedforward networks. Neural Computation, 3:226–245, 1991.

[26] Z. Q. Luo and P. Tseng. Analysis of an approximate gradient projection method with applications to the backpropagation algorithm. Optimization Methods and Software, 4:85–101, 1994.

[27] C. Mathieu and W. Schudy. Correlation clustering with noisy input. In SODA, pages 712–728. SIAM, 2010.

[28] A. Nedić. Random projection algorithms for convex set intersection problems. In Proceedings of the 49th IEEE Conference on Decision and Control, pages 7655–7660, 2010.

[29] A. Nedić. Random algorithms for convex minimization problems. Mathematical Programming, 129(2):225–253, Oct 2011.

[30] A. Nedić and D. P. Bertsekas. Convergence rate of the incremental subgradient algorithm. In S. Uryasev and P. M. Pardalos, editors, Stochastic Optimization: Algorithms and Applications, pages 263–304. Kluwer Academic Publishers, 2000.

[31] A. Nedić and D. P. Bertsekas. Incremental subgradient methods for nondifferentiable optimization. SIAM Journal on Optimization, 12:109–138, 2001.

[32] A. Nedić, D. P. Bertsekas, and V. Borkar. Distributed asynchronous incremental subgradient methods. In D. Butnariu, Y. Censor, and S. Reich, editors, Inherently Parallel Algorithms in Feasibility and Optimization and their Applications. Elsevier, Amsterdam, Netherlands, 2001.

[33] B. T. Polyak. Random algorithms for solving convex inequalities. In D. Butnariu, Y. Censor, and S. Reich, editors, Inherently Parallel Algorithms in Feasibility and Optimization and their Applications, pages 409–422. Elsevier, Amsterdam, Netherlands, 2001.

[34] M. Schmidt, N. Le Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. Math. Program., 162(1-2):83–112, Mar. 2017.

[35] S. Shalev-Shwartz and T. Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. In E. P. Xing and T. Jebara, editors, Proceedings of the 31st International Conference on Machine Learning, volume 32 of Proceedings of Machine Learning Research, pages 64–72. Beijing, China, 22–24 Jun 2014. PMLR.

[36] W. Siedlecki and J. Sklansky. Constrained genetic optimization via dynamic reward-penalty balancing and its
use in pattern recognition. In *Proceedings of the Third International Conference on Genetic Algorithms*, pages 141–150, San Francisco, CA, USA, 1989. Morgan Kaufmann Publishers Inc.

[37] M. V. Solodov. Incremental gradient algorithms with stepsizes bounded away from zero. *Comput. Opt. Appl.*, 11:28–35, 1998.

[38] T. Strohmer and R. Vershynin. A randomized kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262, Apr 2008.

[39] P. Tseng. An incremental gradient(-projection) method with momentum term and adaptive stepsize rule. *SIAM Journal on Optimization*, 8:506–531, 1998.

[40] M. Wang and D. P. Bertsekas. Stochastic first-order methods with random constraint projection. *SIAM Journal on Optimization*, 26(1):681–717, 2016.

[41] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo. Sparse reconstruction by separable approximation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3373–3376, 2008.

[42] M. Zaslavskiy, F. Bach, and J. P. Vert. A path following algorithm for the graph matching problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(12):2227–2242, Dec 2009.