Convergence Rate of a Penalty Method for Strongly Convex Problems with Linear Constraints

Angelia Nedić and Tatiana Tatarenko

Abstract—We consider an optimization problem with strongly convex objective and linear inequalities constraints. To be able to deal with a large number of constraints we provide a penalty reformulation of the problem. As penalty functions we use a version of the one-sided Huber losses. The smoothness properties of these functions allow us to choose time-varying penalty parameters in such a way that the incremental procedure with the diminishing step-size converges to the exact solution with the rate $O(1/\sqrt{k})$. To the best of our knowledge, we present the first result on the convergence rate for the penalty-based gradient method, in which the penalty parameters vary with time.

I. 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 $X_i$, $i = 1, \ldots, m$ (where $m \geq 2$ is large), i.e.,

$$\begin{align*}
\min & \quad f(x) \\
\text{s. t.} & \quad x \in X = \cap_{i=1}^m X_i.
\end{align*}$$

(1)

Throughout the paper, the function $f$ is assumed to be $\mu$-strongly 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 [8], convex relaxations of various combinatorial optimization problems in machine learning applications [21], [32].

Our interest is in case when $m$ is large, which prohibits us from using projected gradient and augmented Lagrangian methods [2], 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 [22] and were extended in [30] to a broader class of randomization over the sets of constraints. Moreover, the convergence rate analysis is performed in [30] to demonstrate that the optimality error diminishes to zero with the rate of $O(1/\sqrt{k})$.

In this work, we present an alternative penalty-based approach to guarantee convergence to the optimum while processing a single set $X_i$ per iteration. A possible reformulation of the problem (1) is through the use of the indicator functions of the constraint sets, resulting in the following unconstrained problem

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

(2)

where $\chi_i(\cdot) : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ 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 [15]. They have been studied for smooth least-square problems [3], [4], [19], for training the neural networks [10], [11], [20], for smooth convex problems [27], [29] and for non-smooth convex problems [9], [13], [14], [16], [23]–[25], [31] (see [6] for a more comprehensive survey of these methods). However, no rate of convergence to the exact solution has been obtained for such procedures. Reformulation (2) has been considered in [17] as a departure point toward an exact penalty reformulation using the set-distance functions. This exact penalty formulation has been motivated by a simple exact penalty model proposed in [5] (using only the set-distance functions) and a more general penalty model considered in [6]. In [17], a lower bound on the penalty parameter has been identified guaranteeing that the optimal solutions of the penalized problem are also optimal solutions of the original problem (2). However, this bound depends on a so-called regularity constant for the constraint set, which might be difficult to estimate. Moreover, the proposed approaches in [17] do not utilize incremental processing, but rather primal-dual approaches where a full (sub)-gradient of the penalized function is used.

In contrast to the works mentioned above, this paper deals with a penalized reformulation of the problem (1), where the penalty parameter can be gradually increased to guarantee convergence of the incremental procedure to the exact solution. The corresponding penalty functions correspond to a version of the one-sided Huber losses [18], which are smooth and possess Lipschitz continuous gradients. In our previous work [28], we have demonstrated existence of the settings for this penalized reformulation under which the fast incremental algorithms can be applied to achieve convergence to a predefined feasible neighborhood of the optimum with a

A. Nedić (Angelia.Nedich@asu.edu) is with School of Electrical, Computer and Energy Engineering, Arizona State University, USA, and T. Tatarenko (tatiana.tatarenko@mr.ru-darmstadt.de) is with the Control Methods and Robotics Lab., Technical University Darmstadt, Darmstadt, Germany.
linear rate. However, to guarantee this convergence, we need to know some problem specific parameters. These parameters might be difficult to estimate in practice. That is why in this work, we study some new properties of these penalty functions which allow us to set up the time-dependent parameters of the reformulated unconstrained problem such that convergence to the exact optimum with the average rate $O(1/\sqrt{t})$ is guaranteed. To the best of our knowledge, this is the first result on the convergence rate for the penalty-based optimization with time-varying parameters.

II. PROBLEM FORMULATION AND ITS PENALTY-BASED REFORMULATION

We consider the following optimization problem:

$$\begin{align*}
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad \langle a_i, x \rangle - b_i \leq 0, \ i = 1, \ldots, m, \ (3)
\end{align*}$$

where the vectors $a_i$, $i = 1, \ldots, m$, are nonzero. We will assume that the problem is feasible. Associated with problem (3), we consider a penalized problem

$$\begin{align*}
\text{minimize} & \quad F_{\gamma}(x) \\
\text{subject to} & \quad x \in \mathbb{R}^n, \ (4)
\end{align*}$$

where

$$F_{\gamma}(x) = f(x) + \gamma \sum_{i=1}^{m} h_{\delta}(\langle a_i, x \rangle, b_i). \ (5)$$

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 (3). For a given nonzero vector $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$, the penalty function $h_{\delta}(\langle a, x \rangle)$ is given by

$$h_{\delta}(\langle a, x \rangle) = \begin{cases} 
\langle a, x \rangle - b & \text{if } \langle a, x \rangle - b > \delta, \\
\frac{\langle a, x \rangle - b}{\|a\|} + \delta & \text{if } -\delta \leq \langle a, x \rangle - b \leq \delta, \\
0 & \text{if } \langle a, x \rangle - b < -\delta,
\end{cases} \ (6)$$

(see Figure 1 for an illustration). For any $\delta \geq 0$, the function $h_{\delta}(\langle a, x \rangle)$ satisfies the following relations:

$$h_{\delta}(\langle a, x \rangle, b) \geq 0 \quad \text{for all } x \in \mathbb{R}^n, \ (7)$$

$$h_{\delta}(\langle a, x \rangle, b) \leq \frac{\delta}{4\|a\|} \quad \text{when } \langle a, x \rangle \leq b, \ (8)$$

$$h_{\delta}(\langle a, x \rangle, b) \geq \frac{\delta}{4\|a\|} \quad \text{when } \langle a, x \rangle > b. \ (9)$$

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

$$p_{\delta}(s) = \begin{cases} 
s & \text{if } s > \delta, \\
\frac{(s+\delta)^2}{\delta} & \text{if } -\delta \leq s \leq \delta, \\
0 & \text{if } s < -\delta,
\end{cases} \ (10)$$

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

$$h_{\delta}(\langle a, x \rangle) = \frac{1}{\|a\|} p_{\delta}(\langle a, x \rangle - b). \ (11)$$

The function $p_{\delta}(s)$ is convex on $\mathbb{R}$ for any $\delta \geq 0$. Thus, the function $h_{\delta}(\langle a, x \rangle, b)$ is convex on $\mathbb{R}^n$, implying that the objective function (5) of the penalized problem (4) is convex over $\mathbb{R}^n$ for any $\delta \geq 0$ and $\gamma > 0$.

Furthermore, the function $p_{\delta}(\cdot)$ is twice differentiable for any $\delta > 0$, with the first and second derivatives given by

$$p'_{\delta}(s) = \begin{cases} 
\frac{1}{\|a\|} & \text{if } s > \delta, \\
\frac{s+\delta}{\|a\|} & \text{if } -\delta \leq s \leq \delta, \\
0 & \text{if } s < -\delta,
\end{cases} \ (12)$$

$$p''_{\delta}(s) = \begin{cases} 
0 & \text{if } s > \delta, \\
\frac{1}{\|a\|} & \text{if } -\delta \leq s \leq \delta, \\
0 & \text{if } s < -\delta \text{ or } s > \delta.
\end{cases} \ (13)$$

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

$$\nabla h_{\delta}(\langle a, x \rangle, b) = \frac{1}{\|a\|} p'_\delta(\langle a, x \rangle - b)a, \ (13)$$

which is Lipschitz continuous with a constant $\frac{1}{\|a\|}$, i.e.,

$$\|\nabla h_{\delta}(\langle a, x \rangle, b) - \nabla h_{\delta}(\langle a, y \rangle, b)\| \leq \frac{|a|}{\|a\|} \|x - y\| \ (14)$$

for all $x, \ y \in \mathbb{R}^n$. In view of the definition of the penalty function $F_{\gamma}(\cdot)$ in (5) and relation (13), 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 (4) being feasible for the original problem (3). 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 [18]. 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 [6]. 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 [26].

In what follows, we let \( \Pi_Y[x] \) denote the (Euclidean) projection of a point \( x \) on a convex closed set \( Y \), i.e.,
\[
\text{dist}(x, Y) = \| x - \Pi_Y[x] \|.
\]

The following lemma and its corollary provide some additional properties of the penalty function \( h_\delta(x; a, b) \) that we will use later on. The proof can be found in [28].

**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 (6) with \( \delta \geq 0 \). Let \( Y = \{ x \mid \langle a, x \rangle - b \leq 0 \} \). 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 \( 0 < \delta \leq \delta'_0 \),
\[
h_\delta(x; a, b) \leq h_{\delta'_0}(x; a, b) \quad \text{for all } x \in \mathbb{R}^n.
\]

The following corollary shows that choosing \( f(\hat{x}) \), for any feasible \( \hat{x} \), can be used to construct non-empty level sets of \( F_{\gamma, \delta} \) and \( f \).

**Corollary 1.** Let \( \gamma > 0 \) and \( \delta \geq 0 \) be arbitrary, and let \( \hat{x} \) be a feasible point for the original problem (3). Then, for the scalar \( t_{\gamma, \delta}(\hat{x}) \) defined by
\[
t_{\gamma, \delta}(\hat{x}) = f(\hat{x}) + \frac{\gamma \delta}{4 \min_{1 \leq i \leq m} \| a_i \|},
\]
the level set \( \{ x \in \mathbb{R}^n \mid F_{\gamma, \delta}(x) \leq t_{\gamma, \delta}(\hat{x}) \} \) is nonempty and the solution set \( X^*_{\gamma, \delta} \) of the penalized problem (4) is contained in the level set \( \{ x \in \mathbb{R}^n \mid f(x) \leq t_{\gamma, \delta}(\hat{x}) \} \).

In the next section, we will consider the settings for the penalty parameters under which the incremental gradient-based procedure for the unconstrained problem (4) leads to the solution of the original constrained problem (3). Moreover, we will establish the convergence rate of this procedure.

### III. Penalized Optimization with Time-Varying Parameters

We consider sequences \( \{ \delta_k \} \) and \( \{ \gamma_k \} \) of positive scalars, and we denote the corresponding penalty function \( F_{\delta_k \gamma_k}(x) \) simply by \( F_k \), i.e.,
\[
F_k(x) = f(x) + \frac{\gamma_k}{m} \sum_{i=1}^m h_k(x; a_i, b_i), \tag{15}
\]
where we use \( h_k \) to denote the function \( h_{\delta_k} \). When \( f \) is strongly convex, each of these penalty functions has a unique solution, denoted by \( x^*_k \), and the original problem also has a unique solution \( x^* \in X \).

First, we derive an upper bound for the distance between \( x^*_k \) and \( x^*_{k+1} \). To provide such a bound, we use some properties of the gradients of \( h_k \), as given in the following lemma.

**Lemma 2.** Consider the function \( h_\delta(\cdot; a, b) \) as given in (6). Then, we have
\[
\| \nabla h_\delta(x; a, b) \| \leq 1 \quad \text{for all } x \in \mathbb{R}^n.
\]
If \( \delta_1 \geq \delta_2 \), then
\[
\max_{x \in \mathbb{R}^n} \| \nabla h_{\delta_1}(x; a, b) - \nabla h_{\delta_2}(x; a, b) \| \leq \frac{\delta_1 - \delta_2}{2 \delta_1}.
\]

**Proof.** In view of relation (13) we have that
\[
\| \nabla h_{\delta_1}(x; a, b) \| = | p_{\delta_1}(\langle a, x \rangle - b) |
\]
where \( p_{\delta} \) is given in (10). The first derivative of the function \( p_{\delta} \) (see (12)) is nonnegative and satisfies \( p_{\delta}(s) \leq 1 \). Hence,
\[
\| \nabla h_{\delta_1}(x; a, b) \| \leq 1 \quad \text{(see (13)).}
\]
Now, let \( \delta_1 \geq \delta_2 \). If \( \delta_1 = \delta_2 \), then the relation \( \max_{x \in \mathbb{R}^n} \| \nabla h_{\delta_1}(x; a, b) - \nabla h_{\delta_2}(x; a, b) \| \leq \frac{\delta_1 - \delta_2}{2 \delta_1} \) holds trivially, so assume that \( \delta_1 > \delta_2 \). By relation (13), we have
\[
\| \nabla h_{\delta_1}(x; a, b) - \nabla h_{\delta_2}(x; a, b) \| = | p_{\delta_1}(s) - p_{\delta_2}(s) |. \tag{16}
\]
By using the expression for \( p_{\delta}' \), as given in (12), we have that the difference \( p_{\delta_1}'(s) - p_{\delta_2}'(s) \) is given by
\[
p_{\delta_1}'(s) - p_{\delta_2}'(s) = \begin{cases} 0 & \text{if } s < -\delta_1 \text{ or } s \geq \delta_1, \\ \frac{s + \delta_1}{s - \delta_1} - 1 & \text{if } -\delta_1 \leq s < -\delta_2, \\ \frac{s - \delta_1}{s + \delta_1} & \text{if } s \leq \delta_2 < \delta_1. \end{cases}
\]
Hence,
\[
p_{\delta_1}'(s) - p_{\delta_2}'(s) = \begin{cases} 0 & \text{if } s < -\delta_1 \text{ or } s \geq \delta_1, \\ \frac{s + \delta_1}{s - \delta_1} - 1 & \text{if } -\delta_1 \leq s < -\delta_2, \\ \frac{s - \delta_1}{s + \delta_1} & \text{if } \delta_2 \leq s < \delta_1. \end{cases}
\]
For the cases \(-\delta_1 \leq s < -\delta_2 \) and \( \delta_2 \leq s \leq \delta_1 \), we have
\[
\max_{s \in [-\delta_2, \delta_2]} | p_{\delta_1}'(s) - p_{\delta_2}'(s) | = \frac{\delta_1 - \delta_2}{2 \delta_1},
\]
For the case \(-\delta_2 \leq s < \delta_2 \), we have
\[
\max_{s \in [-\delta_2, \delta_2]} | p_{\delta_1}'(s) - p_{\delta_2}'(s) | = \frac{\delta_1 - \delta_2}{2 \delta_1}.
\]
Therefore, we have
\[
\max_{s \in \mathbb{R}} | p_{\delta_1}'(s) - p_{\delta_2}'(s) | \leq \frac{\delta_1 - \delta_2}{2 \delta_1},
\]
which when combined with relation (16) implies the desired relation.

Our next lemma provides an upper bound on \( \| x^*_k - x^*_{k+1} \| \), which is critical for establishing the convergence of the method later on.

**Lemma 3.** Let \( f \) be strongly convex with a constant \( \mu > 0 \). Let \( \{ \gamma_k \} \) and \( \{ \delta_k \} \) be sequences of positive scalars, such that
\[
\gamma_{k+1} \geq \gamma_k, \quad \delta_{k+1} \leq \delta_k \quad \text{for all } k \geq 1.
\]
Then, we have for all $k \geq 1$
\[
\mu \|x_k^* - x_{k+1}^*\| \leq (\gamma_{k+1} - \gamma_k) + \gamma_k \frac{\delta_k - \delta_{k+1}}{2\delta_k}.
\]

Proof. Consider an arbitrary $k \geq 1$ and assume without loss of generality that $x_k^* \neq x_{k+1}^*$ (otherwise the stated relation holds trivially). The optimality conditions $\nabla F_k(x_k^*) = 0$ and $\nabla F_{k+1}(x_{k+1}^*) = 0$ yield, respectively,
\[
\nabla f(x_k^*) + \gamma_k \sum_{i=1}^{m} \nabla h_k(x_k^*; a_i, b_i) = 0, \\
\nabla f(x_{k+1}^*) + \gamma_{k+1} \sum_{i=1}^{m} \nabla h_{k+1}(x_{k+1}^*; a_i, b_i) = 0.
\]

By subtracting the last relation from the preceding one, and by re-arranging the terms, we obtain
\[
\nabla f(x_k^*) - \nabla f(x_{k+1}^*) = \frac{\gamma_{k+1}}{m} \sum_{i=1}^{m} \nabla h_{k+1}(x_{k+1}^*; a_i, b_i) \\
- \frac{\gamma_k}{m} \sum_{i=1}^{m} \nabla h_k(x_k^*; a_i, b_i).
\]

By adding and subtracting $\frac{\gamma_k}{m} \sum_{i=1}^{m} \nabla h_k(x_k^*; a_i, b_i)$, we have
\[
\nabla f(x_k^*) - \nabla f(x_{k+1}^*) = \frac{\gamma_{k+1} - \gamma_k}{m} \sum_{i=1}^{m} \nabla h_{k+1}(x_{k+1}^*; a_i, b_i) \\
+ \frac{\gamma_k}{m} \sum_{i=1}^{m} \left( \nabla h_{k+1}(x_{k+1}^*; a_i, b_i) - \nabla h_k(x_k^*; a_i, b_i) \right).
\]

Hence,
\[
\langle \nabla f(x_k^*) - \nabla f(x_{k+1}^*), x_k^* - x_{k+1}^* \rangle \\
= \frac{\gamma_{k+1} - \gamma_k}{m} \sum_{i=1}^{m} \langle \nabla h_{k+1}(x_{k+1}^*; a_i, b_i), x_k^* - x_{k+1}^* \rangle + \frac{\gamma_k}{m} \sum_{i=1}^{m} \langle \nabla h_{k+1}(x_{k+1}^*; a_i, b_i) - \nabla h_k(x_k^*; a_i, b_i), x_k^* - x_{k+1}^* \rangle.
\]

By the strong convexity of $f$, it follows that
\[
\mu \|x_k^* - x_{k+1}^*\|^2 \\
\leq \frac{\gamma_{k+1} - \gamma_k}{m} \sum_{i=1}^{m} \langle \nabla h_{k+1}(x_{k+1}^*; a_i, b_i), x_k^* - x_{k+1}^* \rangle + \frac{\gamma_k}{m} \sum_{i=1}^{m} \langle \nabla h_{k+1}(x_{k+1}^*; a_i, b_i) - \nabla h_k(x_k^*; a_i, b_i), x_k^* - x_{k+1}^* \rangle.
\]

By adding and subtracting $\nabla h_{k+1}(x_{k+1}^*; a_i, b_i)$ in the last terms, we obtain
\[
\mu \|x_k^* - x_{k+1}^*\|^2 \\
\leq \frac{\gamma_{k+1} - \gamma_k}{m} \sum_{i=1}^{m} \langle \nabla h_{k+1}(x_{k+1}^*; a_i, b_i), x_k^* - x_{k+1}^* \rangle + \frac{\gamma_k}{m} \sum_{i=1}^{m} \langle \nabla h_{k+1}(x_{k+1}^*; a_i, b_i) - \nabla h_{k+1}(x_k^*; a_i, b_i), x_k^* - x_{k+1}^* \rangle \\
+ \frac{\gamma_k}{m} \sum_{i=1}^{m} \langle \nabla h_{k+1}(x_k^*; a_i, b_i) - \nabla h_k(x_k^*; a_i, b_i), x_k^* - x_{k+1}^* \rangle.
\]

By the convexity of $h_{k+1}$, we have for all $i$,
\[
\langle \nabla h_{k+1}(x_{k+1}^*; a_i, b_i) - \nabla h_{k+1}(x_k^*; a_i, b_i), x_k^* - x_{k+1}^* \rangle \leq 0,
\]
implying that
\[
\mu \|x_k^* - x_{k+1}^*\|^2 \\
\leq \frac{\gamma_{k+1} - \gamma_k}{m} \sum_{i=1}^{m} \langle \nabla h_{k+1}(x_{k+1}^*; a_i, b_i), x_k^* - x_{k+1}^* \rangle \\
+ \frac{\gamma_k}{m} \sum_{i=1}^{m} \langle \nabla h_{k+1}(x_{k+1}^*; a_i, b_i) - \nabla h_k(x_k^*; a_i, b_i), x_k^* - x_{k+1}^* \rangle.
\]

Since $\gamma_{k+1} \geq \gamma_k > 0$, by using Cauchy-Schwarz inequality, we obtain
\[
\mu \|x_k^* - x_{k+1}^*\|^2 \\
\leq \frac{\gamma_{k+1} - \gamma_k}{m} \sum_{i=1}^{m} \|\nabla h_{k+1}(x_{k+1}^*; a_i, b_i)\| \|x_k^* - x_{k+1}^*\| \\
+ \frac{\gamma_k}{m} \sum_{i=1}^{m} \|\nabla h_{k+1}(x_{k+1}^*; a_i, b_i) - \nabla h_k(x_k^*; a_i, b_i)\| \|x_k^* - x_{k+1}^*\|.
\]

By Lemma 2, we have that $\|\nabla h_{k+1}(x_{k+1}^*; a_i, b_i)\| \leq 1$ implying that
\[
\mu \|x_k^* - x_{k+1}^*\|^2 \leq (\gamma_{k+1} - \gamma_k)\|x_k^* - x_{k+1}^*\| + \frac{\gamma_k}{m} \sum_{i=1}^{m} \|\nabla h_{k+1}(x_{k+1}^*; a_i, b_i) - \nabla h_k(x_k^*; a_i, b_i)\| \|x_k^* - x_{k+1}^*\|.
\]

Since $\delta_{k+1} \leq \delta_k$, by Lemma 2 we have that for all $i$,
\[
\max_{x \in \mathbb{R}^n} \|\nabla h_{k}(x; a_i, b_i) - \nabla h_{k+1}(x; a_i, b_i)\| \leq \frac{\delta_k - \delta_{k+1}}{2\delta_k}.
\]
Hence,
\[
\mu \|x_k^* - x_{k+1}^*\|^2 \leq (\gamma_{k+1} - \gamma_k)\|x_k^* - x_{k+1}^*\| + \gamma_k \frac{\delta_k - \delta_{k+1}}{2\delta_k} \|x_k^* - x_{k+1}^*\|.
\]
Dividing by $\|x_k^* - x_{k+1}^*\|$, we obtain
\[
\mu \|x_k^* - x_{k+1}^*\| \leq (\gamma_{k+1} - \gamma_k) + \gamma_k \frac{\delta_k - \delta_{k+1}}{2\delta_k}.
\]

Our next result provides relations for the points $x_k^*$ and the optimal solution $x^*$ of the original problem.

Lemma 4. Let $f$ be strongly convex with a constant $\mu > 0$. Assume that the sequence $\{\delta_k\}$ and $\{\gamma_k\}$ are such that $\gamma_k > 0$, $\delta_k > 0$ and $\gamma_k \delta_k \leq c$ for all $k$. Then, the sequence $\{x_k\}$ of solutions to the corresponding penalized problems $\min_{x \in \mathbb{R}^n} F_k(x)$ is contained in the level set
\[
\left\{ x \in \mathbb{R}^n \mid f(x) \leq f(x^*) + \frac{c}{4\alpha_{\min}} \right\},
\]
where $x^*$ is the solution of the original problem and $\alpha_{\min} = \min_{1 \leq i \leq m} \|a_i\|$. In particular, the sequence $\{x_k\}$ is bounded.
Proof. By Corollary 1, where \( \hat{x} = x^* \), \( \delta = \delta_k \), \( \gamma = \gamma_k \), and \( t_{\gamma_0}(\hat{x}) \) is replaced with \( f(x^*) + \frac{\gamma_k \delta_k}{4\alpha_{\text{min}}} \), we obtain that \( f(x_k^*) \leq f(x^*) + \frac{\gamma_k \delta_k}{4\alpha_{\text{min}}} \). Since \( \gamma_k \delta_k \leq c \), it follows that \{\( x_k^* \)\} is contained in the level set \( \{ x \in \mathbb{R}^n \mid f(x) \leq f(x^*) + \frac{\gamma_k \delta_k}{4\alpha_{\text{min}}} \} \). Since \( f \) is strongly convex, its level sets are bounded, thus implying that \{\( x_k^* \)\} is bounded.

We next consider a set of conditions on parameters \( \delta_k \) and \( \gamma_k \) that will ensure that the sequence \{\( x_k^* \)\} converges to \( x^* \) as \( k \to \infty \). In what follows, we will use the projections of the points \( x_k^* \) on the feasible set, which we denote by \( p_k \), i.e., \( p_k = \Pi X [x_k^*] \). Under the assumptions of Lemma 4, the sequence \{\( x_k^* \)\} is bounded, and so is the sequence \{\( p_k \)\} of the projections of \( x_k^* \)'s on \( X \). Let \( R \) be large enough so that \{\( x_k \)\} \( \subset \mathbb{B}(0, R) \) and \{\( p_k \)\} \( \subset \mathbb{B}(0, R) \), where \( \mathbb{B}(0, R) \) denotes the ball centered at the origin with the radius \( R \). The subgradients of \( f(x) \) for \( x \in \mathbb{B}(0, R) \) are bounded, and let \( L \) be the maximum norm of the subgradients of \( f(x) \) over \( x \in \mathbb{B}(0, R) \), i.e.,

\[
L = \max \{ \| s \| \mid s \in \partial f(x), \| x \| \leq R \}. \tag{17}
\]

We have the following lemma.

**Lemma 5.** Let \( f \) be strongly convex with a constant \( \mu \geq 0 \). Assume that the sequence \{\( \delta_k \)\} and \{\( \gamma_k \)\} are such that \( \gamma_k > 0 \), \( \delta_k > 0 \) and \( \gamma_k \delta_k / c \) for all \( k \). Let \( L \) be given by (17). Then, for all \( k \), we have

\[
\frac{\mu}{2} \| x^* - x_k^* \|^2 + \frac{\mu}{2} \| x^* - p_k \|^2 + \left( \frac{\gamma_k}{4m\beta} - L \right) \text{dist}(x_k^*, X) \leq \frac{\gamma_k \delta_k}{4\alpha_{\text{min}}},
\]

where \( p_k = \Pi X [x_k^*] \) for all \( k \), and \( \alpha_{\text{min}} = \min_{1 \leq i \leq m} \| a_i \| \).

**Proof.** Since each \( F_k \) is strongly convex with the constant \( \mu \geq 0 \), by the optimality of \( x_k^* \) and by the definition of \( F_k \), we have

\[
\frac{\mu}{2} \| x^* - x_k^* \|^2 \leq f(x^*) - f(x_k^*) + \frac{\gamma_k}{m} \sum_{i=1}^m h_k(x^*; a_i, b_i) - \frac{\gamma_k}{m} \sum_{i=1}^m h_k(x_k^*; a_i, b_i). \tag{18}
\]

By adding and subtracting \( f(p_k) \), we have

\[
f(x^*) - f(x_k^*) = f(x^*) - f(p_k) + f(p_k) - f(x_k^*) \leq -\frac{\mu}{2} \| x^* - p_k \|^2 + f(p_k) - f(x_k^*),
\]

where in the last inequality is obtained using

\[
\frac{\mu}{2} \| p_k - x^* \|^2 + \langle \nabla f(x^*), p_k - x^* \rangle + f(x^*) \leq f(p_k),
\]

and the fact \( \langle \nabla f(x^*), p_k - x^* \rangle \geq 0 \), which holds since \( p_k \) is feasible and \( x^* \) is the optimal point. Since the subgradients of \( f \) are uniformly bounded at \( x_k^* \) and \( p_k \), we have

\[
f(x^*) - f(x_k^*) \leq \frac{\mu}{2} \| x^* - p_k \|^2 + L \| p_k - x_k^* \|,
\]

Combining the preceding inequality with relation (18), we obtain

\[
\frac{\mu}{2} \| x^* - x_k^* \|^2 \leq -\frac{\mu}{2} \| x^* - p_k \|^2 + L \| p_k - x_k^* \| + \frac{\gamma_k \delta_k}{4\alpha_{\text{min}}}
\]

\[
+ \frac{\gamma_k}{m} \sum_{i=1}^m h_k(x^*; a_i, b_i) - \frac{\gamma_k}{m} \sum_{i=1}^m h_k(x_k^*; a_i, b_i).
\]

Since \( x^* \) is feasible, by relation (8) (where \( \delta = \delta_k \)) we have for all \( i = 1, \ldots, m \),

\[
h_k(x^*; a_i, b_i) \leq \frac{\delta_k}{4\| a_i \|} \leq \frac{\delta_k}{4\alpha_{\text{min}}}.
\]

Hence, it follows that

\[
\frac{\mu}{2} \| x^* - x_k^* \|^2 + \frac{\mu}{2} \| x^* - p_k \|^2 \leq L \| p_k - x^* \| + \frac{\gamma_k \delta_k}{4\alpha_{\text{min}}}
\]

\[
- \frac{\gamma_k}{4m} \sum_{i=1}^m \text{dist}(x_k^*, X_i).
\]

By Lemma 1, where \( \delta = \delta_k \), \( a_i = a_i \), \( b_i = b_i \), and \( Y = X_i \), we have

\[
h_k(x_k^*; a_i, b_i) \geq \frac{1}{4} \text{dist}(x_k^*, X_i) \quad \text{for all } i = 1, \ldots, m.
\]

Therefore,

\[
\frac{\mu}{2} \| x^* - x_k^* \|^2 + \frac{\mu}{2} \| x^* - p_k \|^2 \leq L \| p_k - x^* \| + \frac{\gamma_k \delta_k}{4\alpha_{\text{min}}}
\]

\[
- \frac{\gamma_k}{4m} \sum_{i=1}^m \text{dist}(x_k^*, X_i).
\]

By Hoffman’s lemma [12], we have that there exists \( \beta = (\beta_1, \ldots, \beta_m) > 0 \) for the sets \( X_i, i = 1, \ldots, m \), such that

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

Hence,

\[
\frac{\mu}{2} \| x^* - x_k^* \|^2 + \frac{\mu}{2} \| x^* - p_k \|^2 \leq L \| p_k - x^* \| + \frac{\gamma_k \delta_k}{4\alpha_{\text{min}}}
\]

\[
- \frac{\gamma_k}{4m} \text{dist}(x_k^*, X).
\]

Since \( p_k \) is the projection of \( x_k^* \) on \( X \), we have \( \| p_k - x_k^* \|= \text{dist}(x_k^*, X) \) implying that

\[
\frac{\mu}{2} \| x^* - x_k^* \|^2 + \frac{\mu}{2} \| x^* - p_k \|^2 \leq \left( L - \frac{\gamma_k}{4m\beta} \right) \text{dist}(x_k^*, X) + \frac{\gamma_k \delta_k}{4\alpha_{\text{min}}}.
\]

**Lemma 5** indicates that, when \( \gamma_k \to +\infty \), for all large enough \( k \), we will have \( \frac{\gamma_k \delta_k}{4\alpha_{\text{min}}} > 0 \), implying that

\[
\text{dist}(x_k^*, X) \leq \frac{\gamma_k \delta_k}{4\alpha_{\text{min}} \left( \frac{\gamma_k \delta_k}{4\alpha_{\text{min}}} - L \right)} = \frac{\gamma_k \delta_k}{\alpha_{\text{min}} \left( \gamma_k - 4m\beta L \right)} \approx O(\delta_k).
\]
Thus, if \( \delta_k \to 0 \), the distance of \( x_k^* \) to the feasible set \( X \) will go to 0 at the rate of \( O(\delta_k) \). Lemma 5 also indicates that, for large enough \( k \),

\[
\|x^*- x_k^*\|^2 \leq \frac{\gamma_k \delta_k}{2\mu \alpha_{\min}}.
\]

Thus, if \( \gamma_k \delta_k \to 0 \), then the points \( x_k^* \) approach the optimal solution \( x^* \) of the original problem, with the rate of \( O(\delta_k) \).

To summarize, Lemma 5 characterizes the behavior of the sequence \( \{x_k^*\} \) in terms of the penalty parameters \( \{\gamma_k\} \) and \( \{\delta_k\} \). It shows that under conditions \( \gamma_k \to \infty \), \( \delta_k \to 0 \) and \( \gamma_k \delta_k \to 0 \), we have \( \|x_k^*- x^*\| \to 0 \). Based on Lemma 5, one can construct a two-loop approach to compute the optimal point \( x^* \) of the original problem, where for every outer loop \( k \), we have an inner loop of iterations to compute \( x_k^* \). This, however, will be quite inefficient.

IV. CONVERGENCE RATE OF INCREMENTAL GRADIENT ALGORITHM

The results of Lemma 3 and Lemma 5 are useful for analyzing the convergence behavior of an incremental algorithm that, when the iterate \( x_k \) is available at iteration \( k \), uses only one randomly chosen constraint (indexed by \( i_k \)) to estimate the gradient \( \nabla F_k(x_k) \). This estimation is employed to construct \( x_{k+1} \), as opposed to determining \( x_k^* \) for each function \( F_k \). We illustrate this on a simple incremental gradient-based method, given by: for \( k \geq 1 \),

\[
x_{k+1} = x_k - s_k \nabla f(x_k) + \gamma_k \nabla h_k(x; a_{i_k}, b_{i_k}),
\]

where \( x_1 \) is an initial point, \( s_k \geq 0 \) is a stepsize, and the index \( i_k \) is chosen uniformly at random. Note that \( \nabla f(x_k) + \gamma_k \nabla h_k(x; a_{i_k}, b_{i_k}) \) is an unbiased estimation of \( \nabla F_k(x_k) \), since by the choice of \( i_k \) we have

\[
\mathbb{E} [\nabla f(x_k) + \gamma_k \nabla h_k(x; a_{i_k}, b_{i_k})] = \nabla F_k(x_k) \text{ for } k \geq 1.
\]

The idea behind the analysis of the method (19) is resting on a relation of the form \( \mathbb{E} [\|x_{k+1} - x^*\|] \leq q_k \mathbb{E} [\|x_k - x^*\|] + r_k \) for some \( q_k \) and \( r_k \) and explores the conditions on \( q_k \) and \( r_k \), for which the following Chung’s lemma [7] ensures the convergence of \( \|x_k - x^*\| \to 0 \), as \( k \to \infty \) with some definite convergence rate.

Lemma 6. Let \( \{u_k\} \) be a nonnegative scalar sequence and \( k_0 \) be such that

\[
u_{k+1} \leq \left(1 - \frac{a}{k^{s+t}}\right) \nu_k + O\left(\frac{b}{k^{s+t}}\right) \text{ for all } k > k_0,
\]

where \( 0 < s \leq 1 \), \( a > 0 \), \( b > 0 \), and \( t > 0 \). Then, we have

\[
u_k = O\left(\frac{1}{k^{t}}\right).
\]

With Lemma 3, Lemma 5, and Lemma 6 in place, we next establish a set of conditions on \( \{\gamma_k\} \) and \( \{\delta_k\} \) that ensure convergence of the iterates produced by the method (19).

**Proposition 1.** Let \( f \) be strongly convex with a constant \( \mu > 0 \) and have Lipschitz continuous gradients with a constant \( L_f \). Let the sequences \( \{\gamma_k\} \) and \( \{\delta_k\} \) satisfy

\[
\gamma_k = k^q, \; \delta_k = \frac{1}{k^d},
\]

where \( g > 0 \) and \( d > 0 \) are such that \( \{\gamma_k \delta_k\} \) is nonincreasing. Consider the method (19) with the stepsize \( s_k = \frac{1}{\delta_k} \) with \( s > 0 \). Then, as \( k \to \infty \),

\[
\mathbb{E}[\|x_k - x^*\|^2] = O\left(\frac{1}{k^{2(d-g)}}\right).
\]

In particular, when \( s = 1 \), \( g = \frac{1}{4} \), and \( d \geq \frac{3}{4} \) the iterates \( \{x_k\} \) the method (19) converge to the solution \( x^* \) of the original problem (in expectation) and

\[
\mathbb{E}[\|x_k - x^*\|^2] = O\left(\frac{1}{k^{2}}\right).
\]

**Proof.** For any \( k \geq 0 \), for the iterates of the method we have

\[
\|x_{k+1} - x_k^*\|^2 = \|x_k - x_k^*\|^2 - 2s_k \langle g_k(x_k), x_k - x_k^* \rangle + s_k^2 \|g_k(x_k)\|^2,
\]

where \( g_k(x_k) = \nabla f(x_k) + \gamma_k \nabla h_k(x; a_{i_k}, b_{i_k}) \). By the strong convexity of \( F_k \) and the fact \( \nabla F_k(x_k) = 0 \), it follows that

\[
\mathbb{E}[\|x_{k+1} - x_k^*\|^2] \leq (1 - 2s_k \mu) \mathbb{E}[\|x_k - x_k^*\|^2] + s_k^2 \mathbb{E}[\|g_k(x_k)\|^2].
\]

For \( \|g_k(x_k)\|^2 \) we write

\[
\mathbb{E}[\|g_k(x_k)\|^2] \leq 2E[\|\nabla f(x_k)\|^2] + 2E [\|\gamma_k \nabla h_k(x; a_{i_k}, b_{i_k})\|^2] \leq 2E \|\nabla f(x_k)\|^2 + 2 \gamma_k^2,
\]

where the last inequality is obtained by using the convexity of the squared-norm function and the fact that \( \|\nabla h_k(x; a_{i_k}, b_{i_k})\| \leq 1 \) for any \( x \) and \( i \) (see Lemma 2). We further estimate \( \mathbb{E}[\|\nabla f(x_k)\|^2] \) as follows:

\[
\mathbb{E}[\|\nabla f(x_k)\|^2] \leq 2E \|\nabla f(x_k) - \nabla f(x^*)\|^2 + 2\|\nabla f(x^*)\|^2 \\
\leq 2L_f^2 \mathbb{E}[\|x_k - x^*\|^2] + 2\|\nabla f(x^*)\|^2,
\]

where in the last inequality we use the Lipschitz gradient property of \( f \). Thus,

\[
\mathbb{E}[\|g_k(x_k)\|^2] \leq 4L_f^2 \mathbb{E}[\|x_k - x^*\|^2] + 4\|\nabla f(x^*)\|^2 + 2 \gamma_k^2.
\]

Further, we have

\[
\mathbb{E}[\|x_k - x^*\|^2] \leq 2E[\|x_k - x_k^*\|^2] + 2\|x_k^* - x^*\|^2,
\]

so that

\[
\mathbb{E}[\|g_k(x_k)\|^2] \leq 8L_f^2 \mathbb{E}[\|x_k - x_k^*\|^2] + 8L_f^2 \|x_k^* - x^*\|^2 + 4\|\nabla f(x^*)\|^2 + 2 \gamma_k^2.
\]

By Lemma 5 we have

\[
\|x_k - x^*\|^2 \leq \frac{\gamma_k \delta_k}{2\mu \alpha_{\min}}.
\]
By combining the preceding two relations with relation (20) we obtain
\[
E\|x_{k+1} - x_k\|^2 \leq (1 - 2s_k\mu + 8L_j^2s_k^2)E\|x_k - x_k^*\|^2 + 2s_k^2 \left( \frac{4L_j^2\gamma \delta_k}{\mu \omega_{\min}} + 4\|\nabla f(x^*)\|^2 + 2\gamma^2 \right).
\]

We next consider \(\|x_{k+1} - x_k^*\|^2\) for which we write
\[
\|x_{k+1} - x_k^*\|^2 \leq (1 + s_k\mu)\|x_{k+1} - x_k^*\|^2 + (1 + s_k^{-1}\mu^{-1})\|x_k^* - x_{k+1}^*\|^2.
\]

Combining the preceding two relations, we obtain
\[
E\|x_{k+1} - x_k^*\|^2 \leq (1 + s_k\mu)(1 - 2s_k\mu + 8L_j^2s_k^2)E\|x_k - x_k^*\|^2 + 2s_k \left( \frac{2L_j^2\gamma \delta_k}{\mu \omega_{\min}} + 4\|\nabla f(x^*)\|^2 + 2\gamma^2 \right)
\]
\[
+ (1 + s_k^{-1}\mu^{-1})\|x_k^* - x_{k+1}^*\|^2.
\]

Next we use Lemma 3 to upper bound \(\|x_k^* - x_{k+1}^*\|^2\). Thus, we obtain for large enough \(k\),
\[
E\|x_{k+1} - x_k^*\|^2 \leq (1 + s_k\mu)(1 - 2s_k\mu + 8L_j^2s_k^2)E\|x_k - x_k^*\|^2 + 2s_k \left( \frac{2L_j^2\gamma \delta_k}{\mu \omega_{\min}} + 4\|\nabla f(x^*)\|^2 + 2\gamma^2 \right)
\]
\[
+ (1 + s_k^{-1}\mu^{-1})\left( \gamma_{k+1} - \gamma_k + \gamma_k \frac{\delta_k - \delta_{k+1}}{2\delta_k} \right)^2.
\]

Next, taking into account the settings \(s_k = \frac{1}{\kappa^3}, \gamma = k^q, \delta = \frac{1}{\kappa^s}\), we obtain
\[
u_{k+1} \leq \left( 1 - \frac{\mu}{2k^s} \right) \nu_k + O \left( \frac{1}{k^{2s-2g}} \right)
\]
\[
+ O \left( \frac{1}{k^{2g-s}} \right) \left[ \left( 1 + \frac{1}{k} \right)^g - 1 + \left( 1 - \frac{1}{\kappa+1} \right)^d \right]^2.
\]

Due to the fact that \((1 + \frac{1}{k})^g \equiv 1 + \frac{g}{k} and \((1 - \frac{1}{\kappa+1})^d \equiv 1 - \frac{d}{\kappa+1}\), we conclude that
\[
u_{k+1} \leq \left( 1 - \frac{\mu}{2k^s} \right) \nu_k + O \left( \frac{1}{k^{2s-2g}} + \frac{1}{k^{2g-s}} \right).
\]

Next, we write
\[
\|x_k - x^*\|^2 \leq 2\|x_k - x_k^*\|^2 + 2\|x_k^* - x^*\|^2
\]
which together with (21) implies
\[
E\|x_k - x^*\|^2 = O \left( \frac{1}{k^{2s-2g}} + \frac{1}{k^{2g-s}} \right).
\]

By optimizing the parameters \(s, g, and d\), we get \(s = 1, g = \frac{1}{4}, and d \geq \frac{1}{4}\). Under this setting
\[
E\|x_k - x^*\|^2 = O \left( \frac{1}{k^{\frac{1}{2}}} \right),
\]
and the iterates \(\{x_k\}\) the method (19) converge, in the expectation, to the solution \(x^*\) of the original problem.

V. CONCLUSION

In this work we considered penalty reformulation of optimization problems with strongly convex objectives and linear constraints. We proposed using Huber losses as penalty functions. The properties of these functions allowed us to set up the penalty parameter and the step-size of the standard incremental gradient-based optimization procedure to guarantee convergence to the solution. Moreover, we provided the estimation of the convergence rate for this algorithm. In the future work, we will investigate applicability of accelerated incremental algorithms for the proposed penalty reformulation in the case of both strongly and non-strongly convex optimization.

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] D. P. Bertsekas. Constrained Optimization and Lagrange Multiplier Methods (Optimization and Neural Computation Series). Athena Scientific, 1 edition, 1996.
[3] D. P. Bertsekas. Incremental least squares methods and the extended Kalman filter. SIAM Journal on Optimization, 6:807–822, 1996.
[4] D. P. Bertsekas. A hybrid incremental gradient method for least squares. SIAM Journal on Optimization, 7:913–926, 1997.
[5] D. P. Bertsekas. Incremental proximal methods for large scale convex optimization. Mathematical Programming, 129(2):163–195, 2011.
