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 sets X_i, represented by linear inequalities, $i = 1, \ldots, m$ (where $m \geq 2$ is large), i.e.,

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in X = \cap_{i=1}^m X_i. \end{array} \tag{1}$$

Throughout the paper, the function f is assumed to be μ-*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 [7], convex relaxations of various combinatorial optimization problems in machine learning applications [17], [26].

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 [18] and were extended in [24] to a broader class of randomization over the sets of constraints. Moreover, the convergence rate analysis is performed in [24] 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

$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^m \left\{ \frac{1}{m} f(x) + \chi_i(x) \right\}, \tag{2}$$

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 [13]. They have been studied for smooth least-square problems [3], [16], for training the neural networks [9], [10], for smooth convex problems [21], [23] and for non-smooth convex problems [8], [11], [12], [25] (see [5] 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 [14] 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 [4] (using only the set-distance functions) and a more general penalty model considered in [5]. In [14], 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 [14] 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 [15], which are smooth and possess Lipschitz continuous gradients. In our previous work [22], 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@rmr.tu-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 $O(1/\sqrt{k})$ 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:

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

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

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

where

$$F_{\gamma\delta}(x) = f(x) + \frac{\gamma}{m}\sum_{i=1}^{m} h_\delta\left(x; a_i, b_i\right). \quad (5)$$

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 $h_\delta(\cdot; a, b)$ is given by[1]

$$h_\delta(x; a, b) = \begin{cases} \frac{\langle a,x\rangle - b}{\|a\|} & \text{if } \langle a, x\rangle - b > \delta, \\ \frac{(\langle a,x\rangle - b + \delta)^2}{4\delta\|a\|} & \text{if } -\delta \leq \langle a, x\rangle - b \leq \delta, \\ 0 & \text{if } \langle a, x\rangle - b < -\delta, \end{cases} \quad (6)$$

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

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

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

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

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

$$p_\delta(s) = \begin{cases} s & \text{if } s > \delta, \\ \frac{(s+\delta)^2}{4\delta} & \text{if } -\delta \leq s \leq \delta, \\ 0 & \text{if } s < -\delta, \end{cases} \quad (10)$$

[1]A version of the one-sided Huber losses [15].

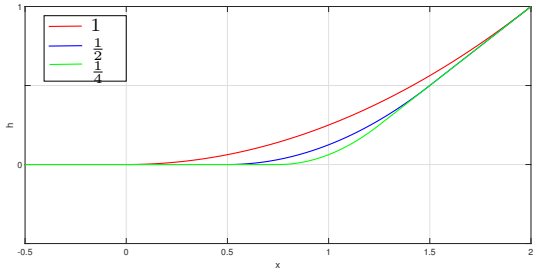

Fig. 1. Penalty functions $h_\delta(x; 1, 1)$ for the constraint $x - 1 \leq 0$, $x \in \mathbb{R}$, with $\delta \in \left\{\frac{1}{4}, \frac{1}{2}, 1\right\}$.

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

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

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 (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

$$p'_\delta(s) = \begin{cases} 1 & \text{if } s > \delta, \\ \frac{(s+\delta)}{2\delta} & \text{if } -\delta \leq s \leq \delta, \\ 0 & \text{if } s < -\delta, \end{cases} \quad (12)$$

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

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

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

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

$$\|\nabla h_\delta(x; a, b) - \nabla h_\delta(y; a, b)\| \leq \frac{\|a\|}{2\delta}\|x - y\| \quad (14)$$

for all $x, y \in \mathbb{R}^n$. In view of the definition of the penalty function $F_{\gamma\delta}$ 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 γ and δ 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 [15]. 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 [5]. 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 [20].

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

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 [22].

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 $\delta = 0$, $h_0(x; a, b) = \operatorname{dist}(x, Y)$ for all $x \in \mathbb{R}^n$ and for any $0 < \delta \leq \delta'$, $h_\delta(x; a, b) \leq h_{\delta'}(x; a, b)$ 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.,

$$F_k(x) = f(x) + \frac{\gamma_k}{m}\sum_{i=1}^{m} h_k(x; a_i, b_i), \qquad (15)$$

where we use h_k to denote the function h_{δ_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*

$$\|\nabla h_\delta(x; a, b)\| \leq 1 \qquad \text{for all } x \in \mathbb{R}^n.$$

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. Can be found in [19]. ∎

Our next lemma provides an upper bound on $\|x^*_{k+1} - x^*_k\|$, 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 $\gamma_{k+1} \geq \gamma_k$, $\delta_{k+1} \leq \delta_k$ 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 $x^*_k \neq x^*_{k+1}$ (for 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,

$$\nabla f(x^*_k) + \frac{\gamma_k}{m}\sum_{i=1}^{m}\nabla h_k(x^*_k; a_i, b_i) = 0,$$

$$\nabla f(x^*_{k+1}) + \frac{\gamma_{k+1}}{m}\sum_{i=1}^{m}\nabla h_{k+1}\left(x^*_{k+1}; a_i, b_i\right) = 0.$$

By subtracting the last relation from the preceding one, and by re-arranging the terms, we obtain

$$\begin{aligned}\nabla f(x^*_k) - \nabla f(x^*_{k+1}) =& \frac{\gamma_{k+1}}{m}\sum_{i=1}^{m}\nabla h_{k+1}\left(x^*_{k+1}; a_i, b_i\right)\\ &- \frac{\gamma_k}{m}\sum_{i=1}^{m}\nabla h_k(x^*_k; a_i, b_i).\end{aligned}$$

By adding and subtracting $\frac{\gamma_k}{m}\sum_{i=1}^{m}\nabla h_{k+1}\left(x^*_{k+1}; a_i, b_i\right)$, we have

$$\begin{aligned}&\nabla f(x^*_k) - \nabla f(x^*_{k+1}) = \frac{\gamma_{k+1} - \gamma_k}{m}\sum_{i=1}^{m}\nabla h_{k+1}\left(x^*_{k+1}; a_i, b_i\right)\\ &+ \frac{\gamma_k}{m}\sum_{i=1}^{m}\left(\nabla h_{k+1}\left(x^*_{k+1}; a_i, b_i\right) - \nabla h_k(x^*_k; a_i, b_i)\right).\end{aligned}$$

Hence,

$$\begin{aligned}&\langle\nabla f(x^*_k) - \nabla f(x^*_{k+1}), x^*_k - x^*_{k+1}\rangle\\ &= \frac{\gamma_{k+1} - \gamma_k}{m}\sum_{i=1}^{m}\langle\nabla h_{k+1}\left(x^*_{k+1}; a_i, b_i\right), x^*_k - x^*_{k+1}\rangle + \frac{\gamma_k}{m}\\ &\times\sum_{i=1}^{m}\langle\nabla h_{k+1}\left(x^*_{k+1}; a_i, b_i\right) - \nabla h_k(x^*_k; a_i, b_i), x^*_k - x^*_{k+1}\rangle.\end{aligned}$$

By the strong convexity of f, it follows that

$$\begin{aligned}&\mu\|x^*_k - x^*_{k+1}\|^2\\ &\leq \frac{\gamma_{k+1} - \gamma_k}{m}\sum_{i=1}^{m}\langle\nabla h_{k+1}\left(x^*_{k+1}; a_i, b_i\right), x^*_k - x^*_{k+1}\rangle + \frac{\gamma_k}{m}\\ &\times\sum_{i=1}^{m}\langle\nabla h_{k+1}\left(x^*_{k+1}; a_i, b_i\right) - \nabla h_k(x^*_k; a_i, b_i), x^*_k - x^*_{k+1}\rangle.\end{aligned}$$

By adding and subtracting $\nabla h_{k+1}\left(x_k^*; a_i, b_i\right)$ in the last terms, we obtain

$$\begin{aligned}
&\mu\|x_k^* - x_{k+1}^*\|^2 \\
&\leq \frac{\gamma_{k+1}-\gamma_k}{m}\sum_{i=1}^m \langle \nabla h_{k+1}\left(x_{k+1}^*; a_i, b_i\right), x_k^* - x_{k+1}^*\rangle + \frac{\gamma_k}{m} \\
&\times \sum_{i=1}^m \langle \nabla h_{k+1}\left(x_{k+1}^*; a_i, b_i\right) - \nabla h_{k+1}\left(x_k^*; a_i, b_i\right), x_k^* - x_{k+1}^*\rangle \\
&+ \frac{\gamma_k}{m}\sum_{i=1}^m \langle \nabla h_{k+1}\left(x_k^*; a_i, b_i\right) - \nabla h_k\left(x_k^*; a_i, b_i\right), x_k^* - x_{k+1}^*\rangle.
\end{aligned}$$

By the convexity of h_{k+1}, we have for all i,

$$\langle \nabla h_{k+1}\left(x_{k+1}^*; a_i, b_i\right) - \nabla h_{k+1}\left(x_k^*; a_i, b_i\right), x_k^* - x_{k+1}^*\rangle \leq 0,$$

implying that

$$\begin{aligned}
&\mu\|x_k^* - x_{k+1}^*\|^2 \\
&\leq \frac{\gamma_{k+1}-\gamma_k}{m}\sum_{i=1}^m \langle \nabla h_{k+1}\left(x_{k+1}^*; a_i, b_i\right), x_k^* - x_{k+1}^*\rangle \\
&+ \frac{\gamma_k}{m}\sum_{i=1}^m \langle \nabla h_{k+1}\left(x_k^*; a_i, b_i\right) - \nabla h_k\left(x_k^*; a_i, b_i\right), x_k^* - x_{k+1}^*\rangle.
\end{aligned}$$

Since $\gamma_{k+1} \geq \gamma_k > 0$, by using Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}
&\mu\|x_k^* - x_{k+1}^*\|^2 \\
&\leq \frac{\gamma_{k+1}-\gamma_k}{m}\sum_{i=1}^m \|\nabla h_{k+1}\left(x_{k+1}^*; a_i, b_i\right)\| \, \|x_k^* - x_{k+1}^*\| \\
&+ \frac{\gamma_k}{m}\sum_{i=1}^m \|\nabla h_{k+1}\left(x_k^*; a_i, b_i\right) - \nabla h_k\left(x_k^*; a_i, b_i\right)\| \\
&\qquad \times \|x_k^* - x_{k+1}^*\|.
\end{aligned}$$

By Lemma 2, we have that $\|\nabla h_{k+1}\left(x_{k+1}^*; a_i, b_i\right)\| \leq 1$ implying that

$$\begin{aligned}
&\mu\|x_k^* - x_{k+1}^*\|^2 \leq (\gamma_{k+1}-\gamma_k)\|x_k^* - x_{k+1}^*\| + \frac{\gamma_k}{m} \\
&\times \sum_{i=1}^m \|\nabla h_{k+1}\left(x_k^*; a_i, b_i\right) - \nabla h_k\left(x_k^*; a_i, b_i\right)\| \, \|x_k^* - x_{k+1}^*\|.
\end{aligned}$$

Since $\delta_{k+1} \leq \delta_k$, by Lemma 2 we have that for all i,

$$\max_{x\in\mathbb{R}^n} \|\nabla h_{\delta_k}(x; a_i, b_i) - \nabla h_{\delta_{k+1}}(x; a_i, b_i)\| \leq \frac{\delta_k - \delta_{k+1}}{2\delta_k}.$$

Hence,

$$\begin{aligned}
&\mu\|x_k^* - x_{k+1}^*\|^2 \\
&\leq (\gamma_{k+1}-\gamma_k)\|x_k^* - x_{k+1}^*\| + \gamma_k \frac{\delta_k - \delta_{k+1}}{2\delta_k}\|x_k^* - x_{k+1}^*\|.
\end{aligned}$$

Dividing by $\|x_k^* - x_{k+1}^*\|$, we obtain the result. ■

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. Can be found in [19]. ■

We next consider a set of conditions on parameters δ_k and γ_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.,

$$L = \max_{\|x\|\leq R} \|\nabla f(x)\|. \tag{16}$$

We have the following lemma.

Lemma 5. *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. Let L be given by (16). Then, for all k, we have*

$$\begin{aligned}
\frac{\mu}{2}\|x^* - x_k^*\|^2 + \frac{\mu}{2}\|x^* - p_k\|^2 + \left(\frac{\gamma_k}{4m\beta} - L\right)\operatorname{dist}(x_k^*, X) \\
\leq \frac{\gamma_k\delta_k}{4\alpha_{\min}},
\end{aligned}$$

where $p_k = \Pi_X[x_k^*]$ for all k, and $\alpha_{\min} = \min_{1\leq i\leq m}\|a_i\|$.

Proof. Can be found in [19]. ■

Lemma 5 indicates that, when $\gamma_k \to +\infty$, for all large enough k, we will have $\frac{\gamma_k}{4m\beta} > L$, implying that

$$\operatorname{dist}(x_k^*, X) \leq \frac{\gamma_k\delta_k}{4\alpha_{\min}\left(\frac{\gamma_k}{4m\beta} - L\right)} \approx O(\delta_k).$$

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 $\|x^* - x_k^*\|^2 \leq \frac{\gamma_k\delta_k}{2\mu\alpha_{\min}}$ for large enough k. 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 $O(\gamma_k\delta_k)$.

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. In the next section, we propose a more efficient single-loop algorithm, where at each iteration k we use the gradient of the penalty function F_k.

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 $k \geq 1$,

$$x_{k+1} = x_k - s_k[\nabla f(x_k) + \gamma_k \nabla h_k(x_k; a_{i_k}, b_{i_k})], \quad (17)$$

where x_1 is an initial point, $s_k > 0$ is a stepsize, and the index i_k is chosen uniformly at random. Note that $\nabla f(x_k) + \gamma_k \nabla h_k(x_k; a_{i_k}, b_{i_k})$ can be considered an unbiased estimation of $\nabla F_k(x_k)$, since by the choice of i_k for $k \geq 1$ we have $\mathbb{E}\left[\nabla f(x_k) + \gamma_k \nabla h_k(x; a_{i_k}, b_{i_k}) | \mathcal{F}_{k-1}\right] = \nabla F_k(x_k)$, where $\mathcal{F}_{k-1}$ is σ-algebra generated by the random variables $\{i_j,\, 1 \leq j \leq k-1\}$.

The idea behind the analysis of the method (17) is resting on a relation of the form $\mathbb{E}\left[\|x_{k+1} - x^*\|\right] \leq q_k \,\mathbb{E}\left[\|x_k - x^*\|\right] + r_k$ for some q_k and r_k and explores the conditions on q_k and r_k, for which the following Chung's lemma [6] 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 $u_{k+1} \leq \left(1 - \frac{a}{k^s}\right) u_k + O\left(\frac{b}{k^{s+t}}\right)$ 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 (17).

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 $\gamma_k = k^g$, $\delta_k = \frac{1}{k^d}$, where $g > 0$ and $d > 0$ are such that $\{\gamma_k \delta_k\}$ is nonincreasing. Consider the method (17) with the stepsize $s_k = \frac{1}{k^s}$ with $s > 0$. Then, as $k \to \infty$,*

$$\mathbb{E}\|x_k - x^*\|^2 = O\left(\frac{1}{k^{\min\{s-2g, 2-2s+2g\}}} + \frac{1}{k^{d-g}}\right).$$

In particular, when $s = 1$, $g = \frac{1}{4}$, and $d \geq \frac{3}{4}$ the iterates $\{x_k\}$ the method (17) converge to the solution x^* of the original problem (in expectation) and $\mathbb{E}\|x_k - x^*\|^2 = O\left(\frac{1}{k^{\frac{1}{2}}}\right)$.

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

$$\mathbb{E}\|x_{k+1} - x_k^*\|^2 \leq (1 - 2s_k\mu)\mathbb{E}\|x_k - x_k^*\|^2 + s_k^2 \mathbb{E}\|g_k(x_k)\|^2. \quad (18)$$

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

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

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

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

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 $\mathbb{E}\|x_k - x^*\|^2 \leq 2\mathbb{E}\|x_k - x_k^*\|^2 + 2\|x_k^* - x^*\|^2$, 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 for sufficiently large k we have

$$\|x_k^* - x^*\|^2 \leq \frac{\gamma_k \delta_k}{2\mu\alpha_{\min}}. \quad (19)$$

By combining the preceding two relations with relation (18) we obtain

$$\mathbb{E}\|x_{k+1} - x_k^*\|^2 \leq (1 - 2s_k\mu + 8L_f^2 s_k^2)\mathbb{E}\|x_k - x_k^*\|^2 + s_k^2 \left(\frac{4L_f^2 \gamma_k \delta_k}{\mu\alpha_{\min}} + 4\|\nabla f(x^*)\|^2 + 2\gamma_k^2\right).$$

We next consider $\|x_{k+1} - x_{k+1}^*\|^2$ for which we write

$$\|x_{k+1} - x_{k+1}^*\|^2 \leq (1 + s_k\mu)\|x_{k+1} - x_k^*\|^2 + (1 + s_k^{-1}\mu^{-1})\|x_k^* - x_{k+1}^*\|^2.$$

Combining the preceding two relations, we obtain

$$\begin{aligned}
&\mathbb{E}\|x_{k+1} - x_{k+1}^*\|^2 \\
&\leq (1 + s_k\mu)(1 - 2s_k\mu + 8L_f^2 s_k^2)\mathbb{E}\|x_k - x_k^*\|^2 \\
&+ (1 + s_k\mu)s_k^2 \left(\frac{2L_f^2 \gamma_k \delta_k}{\mu\alpha_{\min}} + 4\|\nabla f(x^*)\|^2 + 2\gamma_k^2\right) \\
&+ (1 + s_k^{-1}\mu^{-1})\|x_k^* - x_{k+1}^*\|^2. \quad (20)
\end{aligned}$$

Next we use Lemma 3 to upper bound $\|x_k^* - x_{k+1}^*\|^2$. Thus, we obtain for large enough k,

$$\begin{aligned}
&\mathbb{E}\|x_{k+1} - x_{k+1}^*\|^2 \\
&\leq (1 + s_k\mu)(1 - 2s_k\mu + 8L_f^2 s_k^2)\mathbb{E}\|x_k - x_k^*\|^2 \\
&+ (1 + s_k\mu)s_k^2 \left(\frac{2L_f^2 \gamma_k \delta_k}{\mu\alpha_{\min}} + 4\|\nabla f(x^*)\|^2 + 2\gamma_k^2\right) \\
&+ \frac{1 + s_k^{-1}\mu^{-1}}{\mu^2} \left(\gamma_{k+1} - \gamma_k + \gamma_k \frac{\delta_k - \delta_{k+1}}{2\delta_k}\right)^2. \quad (21)
\end{aligned}$$

The rest of the proof is verifying that Lemma 6 can be applied to the preceding inequality. Indeed, let

$$u_k = \mathbb{E}\|x_k - x_k^*\|^2, \qquad q_k = (1 + s_k\mu)(1 - 2s_k\mu + 8L_f^2 s_k^2),$$

$$r_k = (1+s_k\mu)s_k^2\left(\frac{2L_f^2\gamma_k\delta_k}{\mu\alpha_{\min}} + 4\|\nabla f(x^*)\|^2 + 2\gamma_k^2\right) + \frac{1+s_k^{-1}\mu^{-1}}{\mu^2}\left(\gamma_{k+1}-\gamma_k+\gamma_k\frac{\delta_k-\delta_{k+1}}{2\delta_k}\right)^2.$$

Consider the coefficient q_k, for which we have for sufficiently large k, $q_k \geq 1-\frac{\mu}{2}s_k$, where in the last inequality we use the fact that $s_k \to 0$ as $k \to \infty$. For the coefficient r_k, since $\gamma_k\delta_k$ is nonincreasing and $\gamma_k \to \infty$ we have for large enough k,

$$r_k \approx O(s_k^2\gamma_k^2) + O\left(\frac{\left(\gamma_{k+1}-\gamma_k+\gamma_k\frac{\delta_k-\delta_{k+1}}{2\delta_k}\right)^2}{s_k}\right).$$

Next, taking into account the settings $s_k = \frac{1}{k^s}$, $\gamma_k = k^g$, $\delta = \frac{1}{k^d}$, we obtain

$$u_{k+1} \leq \left(1-\frac{\mu}{2}\frac{1}{k^s}\right)u_k + O\left(\frac{1}{k^{2s-2g}}\right) + O\left(\frac{1}{k^{2g-s}}\left[\left(1+\frac{1}{k}\right)^g - 1 + \frac{1-\left(1-\frac{1}{k+1}\right)^d}{2}\right]^2\right).$$

Due to the fact that $\left(1+\frac{1}{k}\right)^g = O\left(1+\frac{g}{k}\right)$ and $\left(1-\frac{1}{k+1}\right)^d = O\left(1-\frac{d}{k+1}\right)$, we conclude that

$$u_{k+1} \leq \left(1-\frac{\mu}{2}\frac{1}{k^s}\right)u_k + O\left(\frac{1}{k^{2s-2g}} + \frac{1}{k^{2-s+2g}}\right).$$

Next, we write $\|x_k - x^*\|^2 \leq 2\|x_k - x_k^*\|^2 + 2\|x_k^* - x^*\|^2$, which together with (19) implies

$$\mathbb{E}\|x_k - x^*\|^2 = O\left(\frac{1}{k^{\min\{s-2g,2-2s+2g\}}} + \frac{1}{k^{d-g}}\right).$$

By optimizing the parameters s, g, and d , we get $s=1$, $g=\frac{1}{4}$, and $d \geq \frac{3}{4}$. Under this setting $\mathbb{E}\|x_k - x^*\|^2 = O\left(\frac{1}{k^{\frac{1}{2}}}\right)$, and the iterates $\{x_k\}$ the method (17) 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. A hybrid incremental gradient method for least squares. *SIAM Journal on Optimization*, 7:913–926, 1997.

[4] D. P. Bertsekas. Incremental proximal methods for large scale convex optimization. *Mathematical Programming*, 129(2):163–195, 2011.

[5] 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.

[6] K. L. Chung. On a stochastic approximation method. *Ann. Math. Statist.*, 25(3):463–483, 09 1954.

[7] 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.

[8] 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.

[9] L. Grippo. A class of unconstrained minimization methods for neural network training. *Optimization Methods and Software*, 4:135–150, 1994.

[10] L. Grippo. Convergent on-line algorithms for supervised learning in neural networks. *IEEE Transactions on Neural Networks*, 11:1284–1299, 2000.

[11] 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.

[12] 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.

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

[14] 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.

[15] W. Li and J. Swetits. The linear l1 estimator and the huber m-estimator. *SIAM Journal on Optimization*, 8(2):457–475, 1998.

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

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

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

[19] A. Nedić and T. Tatarenko. Convergence rate of a penalty method for strongly convex problems with linear constraints. available on arxiv at https://arxiv.org/abs/2004.13417, 2020.

[20] 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.

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

[22] T. Tatarenko and A. Nedić. A smooth inexact penalty reformulation of convex problems with linear constraints. available on arxiv at https://arxiv.org/abs/1808.07749, 2018.

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

[24] M. Wang and D. P. Bertsekas. Incremental constraint projection methods for variational inequalities. *Mathematical Programming*, 150(2):321–363, 2015.

[25] 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.

[26] 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.