Inexact Smooth Penalty for Convex Problems with Linear Constraints

Tatiana Tatarenko and Angelia Nedić

TU Darmstadt
Arizona State University at Tempe

November 15, 2021

Problem of Interest

- We want to *efficiently* solve the following convex problem

$$\text{minimize } f(x) \qquad \text{subject to } \langle a_i, x\rangle \le b_i, \; i = 1, \ldots, m, \tag{1}$$

  - The function $f : \mathbb{R}^n \to \mathbb{R}$ is strongly convex and has Lipschitz continuous gradients
  - The number $m \geq 2$ of inequalities is assumed to be large.
  - Such a problem arises in *machine learning* (training a classifier):
    - Each data point induces a constraint
    - $f(\cdot)$ corresponds to a "regularizer" that we choose (such as $\|\cdot\|^2$)
- When $f \equiv 0$, the randomized Kaczmarz algorithm can solve the resulting feasibility problem with a geometric rate [Strohmer and Vershynin 2008]
- Geometric rate results also exist for a general convex sets X_i admitting efficient projections [Nedić 2010, Richtarik and Necoara 2018]
- If max-type penalty is used, the resulting penalized problem is non-differentiable:

$$\text{minimize } f(x) + \frac{\gamma}{m}\sum_{i=1}^{m} \max\{0, \langle a_i, x\rangle - b_i\}$$

  - At best, the number k of iterations is in the order of $O(1/k)$ (subgradient method)
- Goal: to apply algorithm converging with a geometric rate, i.e. SAGA for example

Proposed Approach: Basic Idea

- Consider a penalized problem of the form

$$\text{minimize } f(x) + \frac{\gamma}{m}\sum_{i=1}^{m} h_\delta(x; a_i, b_i),$$

  where $h_\delta(x; a_i, b_i)$ is a *convex* penalty function associated with constraint $\langle a_i, x\rangle \le b_i$ and $\gamma > 0$ and $\delta > 0$ are penalty parameters
- Our choice of the penalty function is guided by the desire that **the resulting penalized objective is strongly convex and has Lipschitz gradients**
- So we *will choose* $h_\delta(x; a_i, b_i)$ *to have Lipschitz continuous gradients*
- Hence, SAGA or other fast incremental approaches can be applied for solving the penalized problem; however, the penalized problem is not necessarily equivalent to the original problem
- **Main goal:** Determine the parameters δ and γ, so that **the solution of the penalized problem is**
  - Feasible for the original problem and
  - Sub-optimal for the original problem with a desired pre-specified accuracy
- It turns out, we can accomplish this by choosing:
  - *Inexact penalty function that induces a positive penalty inside the feasible set*
- **The approach will work for the class of functions that have bounded level sets**

Penalized Problem

$$\min_x F_{\gamma\delta}(x), \qquad F_{\gamma\delta}(x) = f(x) + \frac{\gamma}{m}\sum_{i=1}^{m} h_\delta(x; a_i, b_i),$$

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

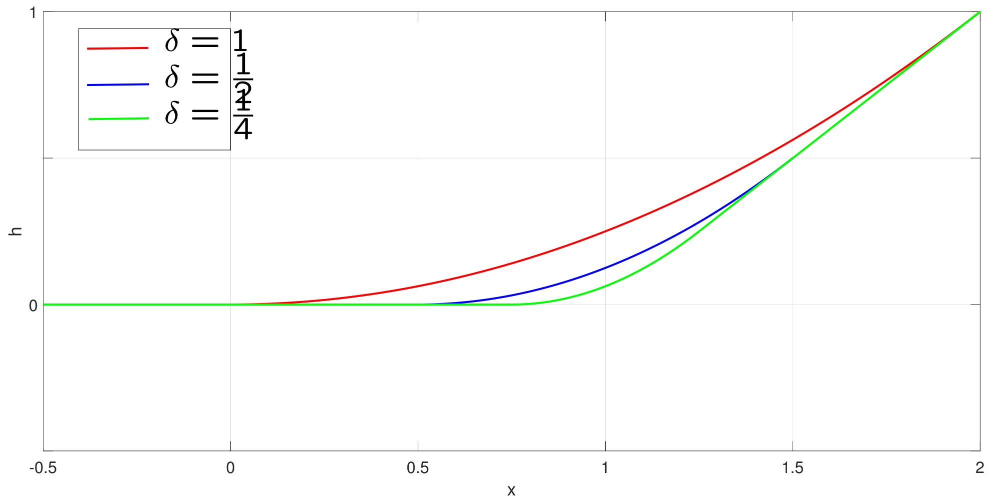

Figure 1: Penalty function $h_\delta(x; 1, 1)$ for the constraint $x \le 1$, $x \in \mathbb{R}$. The case $\delta = 0$ corresponds to $h_0(x; 1, 1) = \max\{0, x - 1\}$

Some properties of the penalty function $h_\delta(x; a, b)$:

- $h_0(x; a, b) = \mathrm{dist}(x, X_i), \qquad X_i = \{x \mid \langle a_i, x\rangle \le b_i\},\ i = 1, \dots, m.$
- If $0 < \delta \le \delta'$, then $\quad h_\delta(x; a, b) \le h_{\delta'}(x; a, b)\quad$ for all $x \in \mathbb{R}^n$.
- For $\delta > 0$, we have $\|\nabla h_\delta(x; a, b) - \nabla h_\delta(y; a, b)\| \le \frac{1}{2\delta}\|x - y\|$ for all $x, y \in \mathbb{R}^n$.
- If x satisfies $\langle a_i, x\rangle \le b_i$, then $\quad h_\delta(x; a_i, b_i) \le \frac{\delta}{4\|a_i\|}$.

An important consequence: nested level sets for any $\gamma > 0$, $\delta < \delta'$, and any $t \in \mathbb{R}$,

$$\{x \mid F_{\gamma\delta'}(x) \le t\} \subseteq \{x \mid F_{\gamma\delta}(x) \le t\} \subseteq \{x \mid f(x) \le t\}.$$

Thus, if f has bounded level sets, so does $F_{\gamma\delta}(x)$ for any $\gamma > 0$ and $\delta > 0$.

Assumption

- The feasible set $X = \{x \mid \langle a_i, x\rangle \le b_i,\ i = 1, \dots, m\}$ has a nonempty interior:

$$\text{there is } \hat{x} \text{ such that: } \langle a_i, \hat{x}\rangle \le b_i - \epsilon \quad \text{for all } i = 1, \dots, m.$$

We let $X_i = \{x \mid \langle a_i, x\rangle \le b_i\}$, so that $X = \cap_{i=1}^m X_i$.

Given a system of linear equations/inequalities, there is $\beta > 0$ such that

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

β is a Hoffman bound(O. Guler, A. Hoffman, and U. Rothblum 1992), and
$\mathrm{dist}(x, Y) = \|x - \Pi_Y[x]\|$ for a convex closed set $Y \subset \mathbb{R}^n$ and any $x \in \mathbb{R}^n$
For our affine set X, Hoffman bound β depends on the vectors a_i

Strongly Convex f

- Let x^* be the optimal solution of the original problem, i.e., $x^* = \mathrm{argmin}_{x\in X} f(x)$

- Let ϵ_a be the desired accuracy for solving the problem

Proposition 1 *Let X have nonempty interior, and let f be strongly convex with a constant $\mu > 0$. Let δ and γ be such that*

$$0 < \delta < \min\left\{\epsilon, \frac{16\alpha_{\min}^2}{\beta^2 m^2}\right\}, \qquad \gamma \le \frac{2\mu}{\delta}\,\epsilon_a\alpha_{\min}$$

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

where L is a Lipschitz constant for f over some suitably defined level set, $\alpha_{\min} = \min_i \|a_i\|$ and $\alpha_{\max} = \max_i \|a_i\|$. Then, the solution $x^*_{\gamma\delta}$ of the penalized problem $\min_x F_{\gamma\delta}(x)$ satisfies

$$x^*_{\gamma\delta} \in X, \qquad \|x^*_{\gamma\delta} - x^*\|^2 \le \epsilon_a.$$

Proof: Main Steps

- When the vector $x^*_{\gamma\delta}$ is feasible i.e., $x^*_{\gamma\delta} \in X$, we have

$$f(x^*) \leq f(x^*_{\gamma\delta}). \tag{2}$$

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

$$h_\delta(x^*; a_i, b_i) - h_\delta(x^*_{\gamma\delta}; a_i, b_i) \leq \frac{\delta}{4\|a_i\|} \qquad \text{for all } i = 1, \ldots, m. \tag{3}$$

Using the relations (2) and (3), we obtain

$$\begin{aligned} F_{\gamma\delta}(x^*) - F_{\gamma\delta}(x^*_{\gamma\delta}) &= f(x^*) - f(x^*_{\gamma\delta}) + \frac{\gamma}{m}\left(\sum_{i=1}^m h_\delta(x^*; a_i, b_i) - h_\delta(x^*_{\gamma\delta}; a_i, b_i)\right) \\ &\leq \frac{\gamma\delta}{4\alpha_{\min}}. \end{aligned}$$

By the strong convexity of $F_{\gamma\delta}$ and the fact that $x^*_{\gamma\delta}$ is the minimum of $F_{\gamma\delta}(x)$, it follows that

$$\|x^* - x^*_{\gamma\delta}\|^2 \leq \frac{2}{\mu_f}(F_{\gamma\delta}(x^*) - F_{\gamma\delta}(x^*_{\gamma\delta})) \leq \frac{\gamma\delta}{2\mu_f\alpha_{\min}} \leq \epsilon_a, \tag{4}$$

where the last inequality in the preceding relation is due to the choice of $\gamma \leq \frac{2\mu_f}{\delta}\alpha_{\min}\epsilon_a$.

- Proving the feasibility of $x^*_{\gamma\delta}$ is much more involved. Define

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

We consider two possibilities: $\|\widehat{x}^*_{\gamma\delta} - x^*_{\gamma\delta}\| \geq \sqrt{\delta}$ and $\|\widehat{x}^*_{\gamma\delta} - x^*_{\gamma\delta}\| < \sqrt{\delta}$.
Case 1: $\|\widehat{x}^*_{\gamma\delta} - x^*_{\gamma\delta}\| > \sqrt{\delta}$. Recall that we have $h_\delta(x; a_i, b_i) \geq \mathrm{dist}(x, X_i)$ for all i. Thus, by the definition of the functions $F_{\gamma\delta}$, for any $x \in \mathbb{R}^n$ we can write

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

Then, by Hoffman's lemma, for some $\beta > 0$ we have

$$F_{\gamma\delta}(x) \geq f(x) + \frac{\gamma}{m\beta}\mathrm{dist}(x, X) \qquad \text{for all } x \in \mathbb{R}^n.$$

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

$$\begin{aligned}
F_{\gamma\delta}(x^*_{\gamma\delta}) &\geq f(x^*_{\gamma\delta}) + \frac{\gamma}{m\beta}\|\widehat{x}^*_{\gamma\delta} - x^*_{\gamma\delta}\| + f(\widehat{x}^*_{\gamma\delta}) - f(\widehat{x}^*_{\gamma\delta}) \\
&\geq \frac{\gamma}{m\beta}\|\widehat{x}^*_{\gamma\delta} - x^*_{\gamma\delta}\| - L\|\widehat{x}^*_{\gamma\delta} - x^*_{\gamma\delta}\| + f(\widehat{x}^*_{\gamma\delta}) \\
&= \left(\frac{\gamma}{m\beta} - L\right)\|\widehat{x}^*_{\gamma\delta} - x^*_{\gamma\delta}\| + F_{\gamma\delta}(\widehat{x}^*_{\gamma\delta}) - \frac{\gamma}{m}\sum_{i=1}^{m} h_\delta(\widehat{x}^*_{\gamma\delta}; a_i, b_i),
\end{aligned}$$

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 $x = \widehat{x}^*_{\gamma\delta}$ (requires a lemma to assert this). Taking into the account that $h_\delta(x; a_i, b_i) \leq \frac{\delta}{4\|a_i\|}$ when $x \in X_i$, and using $\widehat{x}^*_{\gamma\delta} \in X \subseteq X_i$, we see that

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

The conditions on γ imply that $\gamma \geq Lm\beta$. Using the relations $\gamma \geq Lm\beta$ and $\|\widehat{x}^*_{\gamma\delta} - x^*_{\gamma\delta}\| > \sqrt{\delta}$, we further obtain

$$F_{\gamma\delta}(x^*_{\gamma\delta}) > \left(\frac{\gamma}{m\beta} - L\right) \sqrt{\delta} - \frac{\gamma\delta}{4\alpha_{\text{min}}} + F_{\gamma\delta}(\widehat{x}^*_{\gamma\delta}) \geq F_{\gamma\delta}(\widehat{x}^*_{\gamma\delta}), \qquad (5)$$

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

$$\frac{\gamma}{m\beta} - L - \frac{\gamma\sqrt{\delta}}{4\alpha_{\text{min}}} \geq 0 \quad \Longleftrightarrow \quad \gamma\left(\frac{1}{m\beta} - \frac{\sqrt{\delta}}{4\alpha_{\text{min}}}\right) \geq L.$$

The last inequality holds by the conditions imposed on the parameters γ and δ. Thus, relation (5) implies that

$$F_{\gamma\delta}(x^*_{\gamma\delta}) > F_{\gamma\delta}(\widehat{x}^*_{\gamma\delta}),$$

which contradicts the fact that $x^*_{\gamma\delta}$ is an unconstrained minimizer of $F_{\gamma\delta}$.

Case 2: $\|\widehat{x}^*_{\gamma\delta} - x^*_{\gamma\delta}\| \leq \sqrt{\delta}$. Uses the interior-point assumption and a special construction. Specifically, it relies on the following result.

Lemma 1 *Let the interior-point assumption hold and let* δ *be such that* $0 < \delta \leq \epsilon$. *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) $\|x - x_{in}\| \leq \|x - \Pi_X[x]\| + \frac{\beta m \delta}{\min_{1 \leq i \leq m} \|a_i\|}$,

where β *is Hoffman's bound.*

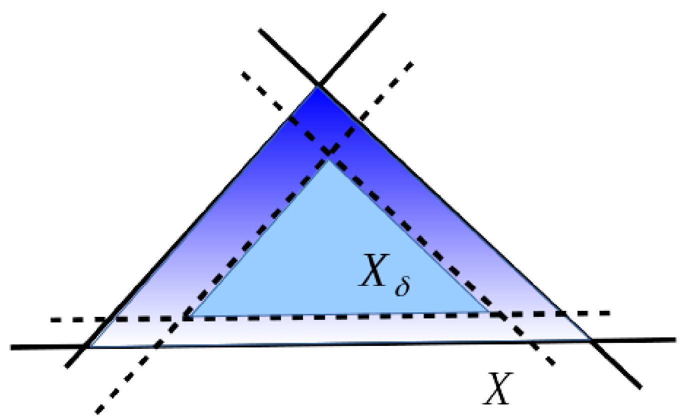

Figure 2: Illustration of the set X_δ**.**

Connecting with SAGA by Defazio et al. 2014

Algorithm 1 SAGA-based Fast Incremental Method for Solving Penalized Problem

- Let $x^0 \in \mathbb{R}^n$ and $\nabla f(\phi_i^0) + \gamma \nabla h_\delta\left(\phi_i^0; a_i, b_i\right)$ with $\phi_i^0 = x^0$ be available for $i = 1, \ldots, m$.

- Pick an index $j \in \{1, 2, \ldots, m\}$ uniformly at random.

- Set $\phi_j^{t+1} = x^t$ and store $\nabla f(\phi_j^{t+1}) + \gamma \nabla h_\delta\left(\phi_j^{t+1}; a_j, b_j\right)$.

- $x^{t+1} = x^t - \alpha[\nabla f(\phi_j^{t+1}) + \gamma \nabla h_\delta\left(\phi_j^{t+1}; a_j, b_j\right) - \nabla f(\phi_j^t)$
  $- \gamma \nabla h_\delta\left(\phi_j^t; a_j, b_j\right) + \frac{1}{m}\sum_{i=1}^m \left(\nabla f(\phi_i^t) + \gamma \nabla h_\delta\left(\phi_i^t; a_i, b_i\right)\right)]$.

Applying SAGA to the Penalized Problem

Proposition 2 *Let the interior-point assumption hold, and let the function f be strongly convex with a parameter $\mu > 0$ and have Lipschitz continuous gradients with a constant $L_f > 0$. Let $x^* = \operatorname{argmin}_{x \in X} f(x)$, and assume that an accuracy level ϵ_a is given. Consider Algorithm SAGA applied to the penalized problem:*

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

where the penalty parameters γ and δ are chosen to satisfy the conditions of Proposition 1, and the step size is given by

$$\alpha = \frac{1}{2(\mu m + L_f + \frac{\gamma \alpha_{\max}}{2\delta})},$$

with $\alpha_{\max} = \max_i \|a_i\|$.
Then, the following convergence rate result is valid for the iterates of the algorithm:

$$\mathbb{E}[\|x^t - \Pi_X[x^t]\|^2] \le O\left(q_\gamma^t\right), \qquad \mathbb{E}[\|x^t - x^*\|]^2 \le O\left(q_\gamma^t\right) + 2\epsilon_a,$$
$$q_\gamma = 1 - \frac{\mu}{2(\mu m + L_f + \frac{\gamma \alpha_{\max}}{2\delta})}.$$

Proof: Immediately from Proposition 1 and a rate result from Defazio et al. 2014.

Convergence Result for Merely Convex Function f

Proposition 3 *Let the interior-point assumption hold. Let f be convex function with bounded level sets and Lipschitz continuous gradients with a constant L_f, and let $f^* = \min x \in X f(x)$. Let ϵ_a a desired accuracy. Consider the iterates x_k produced by SAGA method applied to the penalized problem:*

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

where γ and δ are chosen such that the conditions of Proposition 1 are satisfied, and the stepsize is given by

$$\alpha = \frac{1}{3(L_f + \gamma\alpha_{\max}/(2\delta))}$$

Then, for the iterate averages

$$\bar{x}_k = \frac{1}{k} \sum_{i=0}^{k-1} x_i$$

the following hold relation holds

$$\mathbb{E}[f(\bar{x}_k)] - f^* \leq O(1/k) + 2\epsilon_a \qquad \text{for all } k,$$

Moreover, there exists some $T > 0$ such that for all k,

$$-\gamma\epsilon_a \leq \mathbb{E}[f(\bar{x}_k)] - f^*$$

Simulation Results

$$f(x) = \frac{1}{2}\|\Phi x - x^0\|^2 \qquad Ax \le b, \qquad A \in \mathbb{R}^{m\times n}, \quad n = 30$$

where x^0 is chosen randomly according to a zero-mean normal distribution with covariance matrix $\Sigma = 10I$. Constraint set X is constructed so that the optimal point x^0 is on its boundary. The SAGA algorithm applied to the penalized problem is compared to an algorithm that, at time t, selects one constraint X_{ω_t} randomly from the collection $X_1, \dots, X_m$, and performs a gradient update*

$$x^{t+1} = \Pi_{X_{\omega_t}}[x^t - \alpha_t \nabla f(x^t)],$$

where α_t is a diminishing stepsize (not summable, but square summable).
The full gradient method is also simulated.

The algorithms are run for $1,000$ iterations, and the relative error is plotted

$$\frac{\|x^t - x^*\|}{\|x^*\|}$$

along the iterate sequence.

*Nedić 2011

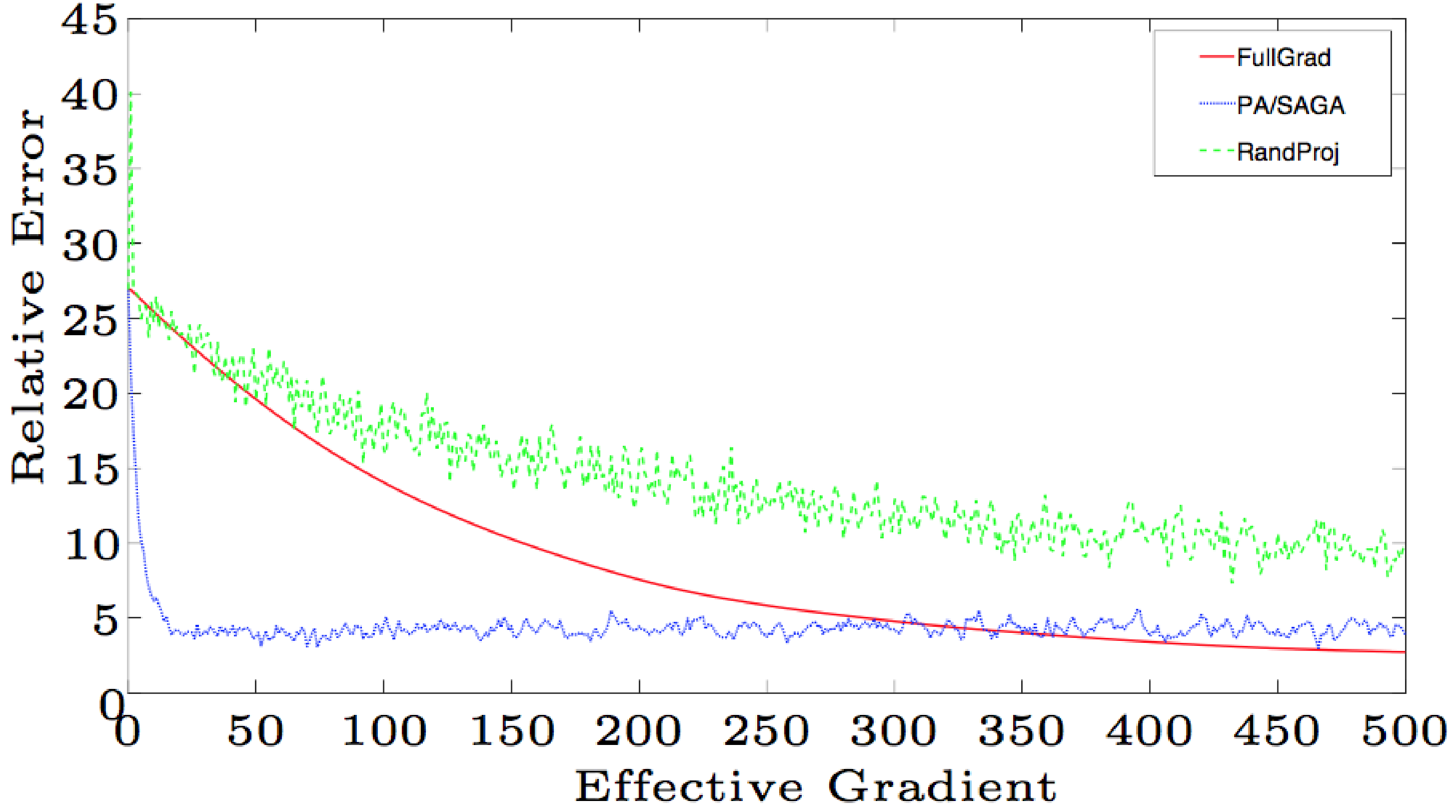

Figure 3: Full-Gradient, PA/SAGA (our penalized approach) and Random Projections, $m = 1500$.

Summary and Extensions

- 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.

- 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.

- We have a convergence result for our penalty approach also when f has bounded level sets (not necessarily strongly convex)

- Difficulty: our results rely on the availability of Hoffman constant β,

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

  which is hard to (upper) estimate; see recent work by J.Pena, J. Vera L.F. Zuluaga *New characterizations of Hoffman constants for systems of linear constraints*, Mathematical Programming 187, p. 79–109, 2021, arxiv https://arxiv.org/pdf/1905.02894.pdf

- As a remedy, we have considered an approach where we vary the penalty parameter γ with time at each iteration and show that such a method converges to the optimal point

- Details are in our paper:
  T. Tatarenko and AN *A Smooth Inexact Penalty Reformulation of Convex Problems with Linear Constraints*, SIAM J. on Optim., 31 (3), 2141–2170, 2021