Handling Optimization Problems with a Big Number of Constraints

Konstantin Mishchenko [1] Peter Richtárik [1, 2, 3]

[1] KAUST, [2]University of Edinburgh, [3]Moscow Institute of Physics and Technology

Objective

$$\text{minimize} \quad \left[f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x)\right] \qquad (1)$$

$$\text{subject to} \quad x \in \mathcal{X} := \bigcap_{j=1}^{m} \mathcal{X}_j \subseteq \mathbb{R}^d \text{ (convex sets)}$$

Issue: both n and m can be BIG!

Motivation

Most modern stochastic methods for solving (1), such as SGD, SVRG and SAGA, have a version for constrained optimization of the form

$$x^{k+1} = \Pi_{\mathcal{X}}(x^k - \alpha^k g^k),$$

where g^k is a random approximation of $\nabla f(x^k)$, $\alpha^k > 0$ is a stepsize and $\Pi_{\mathcal{X}}(x)$ is the projection of x onto $\mathcal{X}$. While the cost of computing g^k can be extremely small (e.g., independent of n), cost of projecting onto $\mathcal{X}$ might be very large.

Our Reformulation

We turn the constrained problem (1) into a special unconstrained problem:

$$\min_{x\in\mathbb{R}^d} f(x) + \lambda h(x), \qquad (2)$$

where $\lambda > 0$ and $h(x) := \frac{1}{2m}\sum_{j=1}^{m} \|x - \Pi_{\mathcal{X}_j}(x)\|^2$.

Assumption on the Sets $\mathcal{X}_1, \dots, \mathcal{X}_m$

Sets $\{\mathcal{X}_j\}_{j=1}^m$ satisfy **linear regularity** property with some constant $\gamma > 0$, i.e., for all $x \in \mathbb{R}^d$

$$\frac{1}{m}\sum_j \|x - \Pi_{\mathcal{X}_j} x\|^2 \geqslant \gamma \|x - \Pi_{\mathcal{X}}(x)\|^2. \qquad (3)$$

Sufficient condition: $\cap_j ri\mathcal{X}_j$ is nonempty

Choice of λ

Let us denote

$$x_\lambda^* := \operatorname{argmin}_{x\in\mathbb{R}^d} f(x) + \lambda h(x).$$

Lemma. *If we increase λ:*

- $h_\lambda^* := h(x_\lambda^*) \searrow 0$,
- $f_\lambda^* := f(x_\lambda^*) \nearrow f^* := \min_{x\in\mathcal{X}} f(x)$,
- *$L_\lambda \to \infty$, where L_λ is the problem smoothness.*

Moreover, for smooth f it holds that

$$\frac{G}{L^2+\lambda^2} \leqslant h_\lambda^* \leqslant \frac{f^* - f_0^*}{\lambda},$$

$$f_\lambda^* \leqslant f^* - \frac{\lambda G}{L^2+\lambda^2}.$$

where $G := \inf_{x\in\mathcal{X}} \|\nabla f(x)\|^2/4$.

Feasible Solution

If we combine linear regularity to smoothness and convexity assumptions, we get

Theorem. *If $\lambda \geqslant \frac{L}{\gamma}$, then*

$$f(\Pi_{\mathcal{X}}(x_\lambda^*)) \leqslant f^* + \frac{2}{\gamma\lambda}(f^* - f(x_0^*)).$$

If, in addition, f is strongly convex,

$$f(\Pi_{\mathcal{X}}(x_\lambda^*)) \leqslant f^* + \frac{L}{2}\left(\frac{4L^2\|\nabla f(x^*)\|^2}{\mu^2\gamma^2\lambda^2} + \frac{1}{\lambda^2}\frac{1}{m}\sum_{j=1}^{m}\|g_m\|^2\right)$$

where g_j satisfy $\Pi_{\mathcal{X}_j}(x^* + g_j) = x^*$ and $\frac{1}{m}\sum_{j=1}^m g_j = \nabla f(x^*)$. **Hence, we obtain a good feasible solution by projecting onto $\mathcal{X}$ only once.**

Important Result

Let $x_{\lambda,\varepsilon}$ be such that $f(x_{\lambda,\varepsilon}) + \lambda h(x_{\lambda,\varepsilon}) \leqslant f(x_\lambda^*) + \lambda h(x_\lambda^*) + \varepsilon$, where f is convex and smooth. We obtain:

Infeasible solution $x_{\lambda,\varepsilon}$ satisfying $f(x_{\lambda,\varepsilon}) \leqslant f^*$ and $h(x_{\lambda,\varepsilon}) \leqslant \delta$ if $\lambda \sim \frac{1}{\sqrt{\delta}}$ and $\varepsilon \sim \sqrt{\delta}$.

Feasible solution $\Pi_{\mathcal{X}}(x_{\lambda,\varepsilon})$ satisfying $f(\Pi_{\mathcal{X}}(x_{\lambda,\varepsilon})) \leqslant \delta$ if $\lambda \sim \frac{1}{\delta}$ and $\varepsilon \sim \delta$.

If f is strongly convex, then we only need $\lambda \sim \frac{1}{\sqrt{\delta}}$ and $\varepsilon \sim \sqrt{\delta}$.

Summary of Solution Properties

Lower bound	Quantity	Upper bound
$\Omega\left(\frac{\lambda}{L^2+\lambda^2}\right)^*$	$f^* - (f_\lambda^* + \lambda h_\lambda^*)$	$O\left(\frac{1}{\lambda}\right)^*$
$\Omega\left(\frac{\lambda}{L^2+\lambda^2}\right)^*$	$f^* - f_\lambda^*$	$O\left(\frac{1}{\lambda}\right)^\dagger$
$\Omega\left(\frac{1}{L^2+\lambda^2}\right)^*$	h_λ^*	$O\left(\frac{1}{\lambda}\right)$ $O\left(\frac{1}{\lambda^2}\right)^\dagger$
$\Omega\left(\frac{1}{L^2+\lambda^2}\right)^*$	$\|x_\lambda^* - x^*\|^2$	$O(\frac{1}{\lambda})^\ddagger$ $O(\frac{1}{\lambda^2})^{*,\ddagger}$
$\Omega\left(\frac{1}{L^2+\lambda^2}\right)^*$	$\|x_\lambda^* - \Pi_{\mathcal{X}}(x_\lambda^*)\|^2$	$O\left(\frac{1}{\lambda^2}\right)^\dagger$
0	$f(\Pi_{\mathcal{X}}(x_\lambda^*)) - f^*$	$O\left(\frac{1}{\lambda}\right)^{*,\S}$ $O(\frac{1}{\lambda^2})^{*,\ddagger}$
0	$f(\Pi_{\mathcal{X}}(x_{\lambda,\varepsilon}) - f^*$	$O\left(\frac{1}{\lambda} + \varepsilon\right)^{*,\S}$ $O(\frac{1}{\lambda^2} + \varepsilon)^{*,\S,\ddagger}$

Table 1:Lower and upper bounds for different measures of solution's quality. Superscripts mean assumptions used to prove the bound: * - smoothnes, † - convex-

Algorithms

Algorithm 1: SGD applied to the reformulation.

for $k = 1, \dots$ **do**
  Sample i and j;
  $\omega_k = \frac{1}{2L+\theta k}$;
  $x^{k+1} = x^k - \omega_k(\nabla f_i(x^k) + \nabla h_j(x^k))$;
end

Theorem. The complexity of Algorithm 1 is $O\big((C(\Pi_{\mathcal{X}_j}) + C(\nabla f)\delta^{-1}\big)$.

Algorithm 2: Nesterov's accelerated GD.

input : $x^0, \theta_0 = 0$
for $k = 1, \dots$ **do**
  $\theta_k = (1 + \sqrt{1 + 4\theta_{k-1}^2})/2$;
  $\eta_k = \frac{1-\theta_{k-1}}{\theta k}$;
  $y^{k+1} = x^k - \frac{1}{L+\lambda}(\nabla f(x^k) + x^k - \frac{1}{m}\sum_j \Pi_{\mathcal{X}_j}(x^k))$;
  $x^{k+1} = x^k - \omega_k(\nabla f_i(x^k) + \nabla h_j(x^k))$;
end

Theorem. The complexity of Algorithm 2 is

Numerical Results

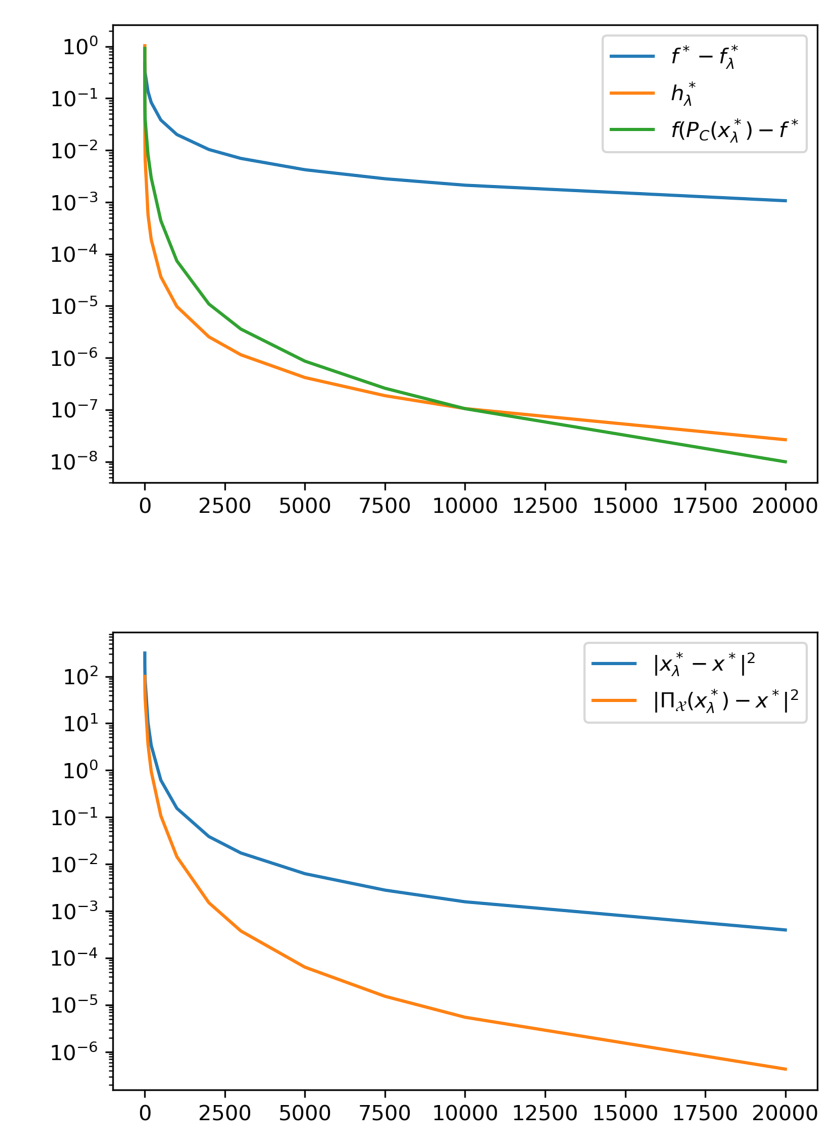

Figure 1:Horizontal axis: λ

References

[1] Ion Necoara, Peter Richtarik, and Andrei Patrascu. Randomized projection methods for convex feasibility problems: conditioning and convergence rates.

[2] Lam Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient.

[3] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course.*

Contact Information

- Web: konstmish.github.io
- Email: konstantin.mishchenko@kaust.edu.sa