Convex optimization based on global lower second-order models

Nikita Doikov Yurii Nesterov

UCLouvain, Belgium

NeurIPS 2020

Problem

Composite convex optimization problem:

$$\min_x F(x) \stackrel{\mathrm{def}}{=} f(x) + \psi(x)$$

- f is convex, differentiable.
- $\psi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is convex, simple.
- $\mathrm{dom}\,\psi$ is bounded. $D \stackrel{\mathrm{def}}{=} \mathrm{diam}\,(\mathrm{dom}\,\psi)$.

Example:

$$\psi(x) = \begin{cases} 0, & \|x\| \leq \frac{D}{2}, \\ +\infty, & \text{otherwise.} \end{cases}$$

$\Rightarrow$ The problem with ball-regularization:

$$\min_{\|x\| \leq \frac{D}{2}} f(x)$$

Review: Gradient Methods

Let ∇f be Lipschitz continuous: $\|\nabla f(y) - \nabla f(x)\|_* \leq L\|y - x\|$.

The Gradient Method:

$$x_{k+1} = \operatorname*{argmin}_y \Big\{ f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \tfrac{L}{2}\|y - x_k\|^2 + \psi(y) \Big\}.$$

- Global convergence: $F(x_k) - F^* \leq O(\frac{1}{k})$.

The Conditional Gradient Method [Frank-Wolfe, 1956]:

$$\begin{aligned} v_{k+1} &= \operatorname*{argmin}_y \Big\{ f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \psi(y) \Big\}, \\ x_{k+1} &= \gamma_k v_{k+1} + (1 - \gamma_k) x_k. \end{aligned}$$

- Set $\gamma_k = \frac{2}{k+2}$. Then $F(x_k) - F^* \leq O(\frac{1}{k})$.
  Note: Near-optimal for $\| \cdot \|_\infty$-balls [Guzmán-Nemirovski, 2015].

Review: Second-Order Methods

Let $\nabla^2 f$ be Lipschitz continuous: $\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L\|x - y\|$.

Newton Method:

$$x_{k+1} = \operatorname*{argmin}_y \Big\{ \langle \nabla f(x_k), y - x_k \rangle + \tfrac{1}{2}\langle \nabla^2 f(x_k)(y - x_k), y - x_k \rangle + \psi(y) \Big\}.$$

- Quadratic convergence (if $\nabla^2 f(x^*) \succ 0$ and x_0 close to x^*).
- No global convergence. A heuristic: use line-search in practice.

Newton Method with Cubic Regularization:

$$x_{k+1} = \operatorname*{argmin}_y \Big\{ \langle \nabla f(x_k), y - x_k \rangle + \tfrac{1}{2}\langle \nabla^2 f(x_k)(y - x_k), y - x_k \rangle + \tfrac{L}{6}\|y - x_k\|^3 + \psi(y) \Big\}.$$

- Global rate: $F(x_k) - F^* \leq O(\frac{1}{k^2})$ [Nesterov-Polyak, 2006].

Overview of the Contributions

New second-order algorithms with global convergence proofs.

- The methods are universal (no unknown parameters).
- Affine-invariant (the norm is not fixed).

Stochastic methods (basic and with the variance reduction).

Numerical experiments.

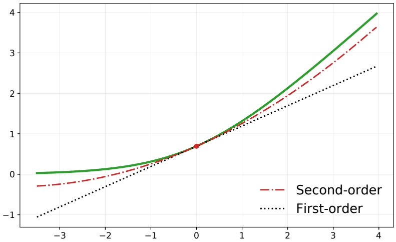
Second-Order Lower Model

1. f is convex: $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$.
2. $\nabla^2 f$ is Lipschitz continuous:

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \quad \leq \quad L\|x - y\|.$$

Convexity + Smoothness $\Rightarrow$ **tighter lower bound**: $\forall t \in [0, 1]$

$$f(y) \quad \geq \quad f(x) + \langle \nabla f(x), y - x \rangle + \frac{t}{2}\langle \nabla^2 f(x)(y - x), y - x \rangle - \frac{t^2 L\|y-x\|^3}{6}.$$



Contracting-Domain Newton Method:

$$
\begin{aligned}
v_{k+1} &= \underset{y}{\operatorname{argmin}}\Big\{ \langle \nabla f(x_k), y - x_k \rangle + \tfrac{\gamma_k}{2} \langle \nabla^2 f(x_k)(y - x_k), y - x_k \rangle \\
&\qquad + \psi(y) \Big\}, \\
x_{k+1} &= \gamma_k v_{k+1} + (1 - \gamma_k) x_k .
\end{aligned}
$$

Contracting-Domain Newton Method (reformulation):

$$x_{k+1} = \operatorname*{argmin}_y \Big\{ \langle \nabla f(x_k), y - x_k \rangle + \tfrac{1}{2} \langle \nabla^2 f(x_k)(y - x_k), y - x_k \rangle + \gamma_k \psi\big(x_k + \tfrac{1}{\gamma_k}(y - x_k)\big) \Big\}.$$

Regularization of quadratic model by the asymmetric trust region.

Let $\nabla^2 f$ be Lipschitz continuous: $\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L\|x - y\|$ (w.r.t. arbitrary norm).

Theorem 1. Set $\gamma_k = \frac{3}{k+3}$. Then

$$F(x_k) - F^* \;\; \leq \;\; O(\tfrac{LD^3}{k^2}).$$

Theorem 2. Let ψ be strongly convex with parameter $\mu > 0$.

- Set $\gamma_k = \frac{5}{k+5}$. Then

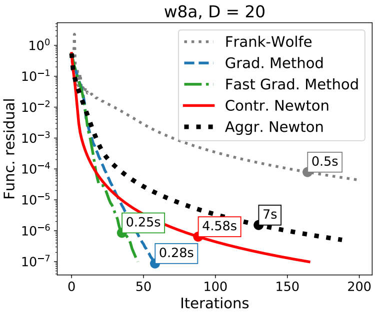
$$F(x_k) - F^* \;\; \leq \;\; O(\tfrac{LD}{\mu} \cdot \tfrac{LD^3}{k^4}).$$

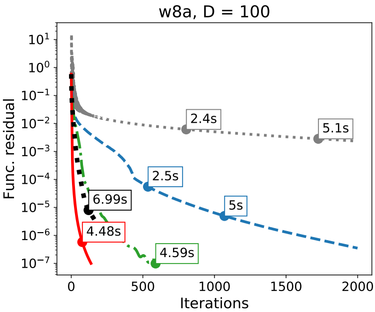
- Set $\gamma_k = \frac{1}{1+\omega}$, where $\omega \overset{\text{def}}{=} \left[\frac{LD}{2\mu}\right]^{\frac{1}{2}}$. Then

$$F(x_k) - F^* \;\; \leq \;\; \exp\left(-\tfrac{k-1}{1+\omega}\right)\tfrac{LD^3}{2}.$$

$$\min_{\|x\|_2 \le \frac{D}{2}} \sum_{i=1}^{M} f_i(x), \qquad f_i(x) \;=\; \log\big(1 + \exp\big(\langle a_i, x \rangle\big)\big).$$

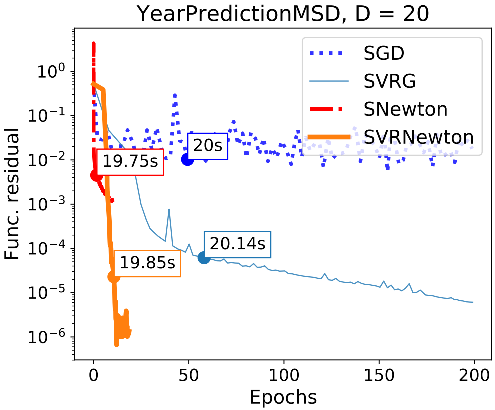
D plays the role of regularization parameter.





For bigger D the problem becomes more *ill-conditioned*.

Approximate $\nabla f(x)$, $\nabla^2 f(x)$ by stochastic estimates.



The problem with big dataset size ($M = 463715$) and small dimension ($n = 90$).

Conclusions

Second-order information helps in a case of

- ill-conditioning;
- small or moderate dimension (the subproblems are more expensive).

No need to tune stepsize.

Can be preferable for solving problems over the sets with a non-Euclidean geometry.

Follow Up Results

Nikita Doikov and Yurii Nesterov. "Affine-invariant contracting-point methods for Convex Optimization". In: *arXiv:2009.08894* (2020)

- General framework of Contracting-Point Methods.
- Contracting-Point Tensor Methods of order $p \geq 1$:

$$F(x_k) - F^* \quad \leq \quad O(\tfrac{1}{k^p}).$$

- Affine-invariant smoothness condition $\Rightarrow$ Affine-invariant analysis.

Thank you for your attention!