Convex optimization based on global lower second-order models

Nikita Doikov nikita.doikov@uclouvain.be

Yurii Nesterov yurii.nesterov@uclouvain.be

UCLouvain University

Problem and Motivation

Composite optimization problem

$$\min_x F(x) \stackrel{\text{def}}{=} f(x) + \psi(x) \qquad (1)$$

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

Second-order optimization schemes are popular for solving ill-conditioned problems. However, for the classical Newton method we have only *local* convergence (the starting point is assumed to be sufficiently close to the optimum).

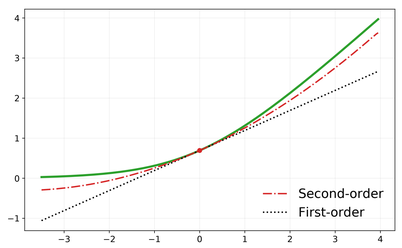
In this work, we develop new second-order algorithms equipped with the *global* complexity guarantees (which are significantly better, than their first-order counterparts).
It is important that our methods are affine-invariant and parameter-free.

Second-Order Lower Model of Objective Function

- Convexity of f implies global lower bound:
$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle.$$
- Let $\nabla^2 f$ be Lipschitz continuous:
$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L \|x - y\|.$$

Convexity + Smoothness $\Rightarrow$ tighter lower bound:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{t}{2} \langle \nabla^2 f(x)(y - x), y - x \rangle - \frac{t^2 L \|y - x\|^3}{6}, \qquad \forall t \in [0, 1].$$



Example

Let ψ be an indicator of a ball (in arbitrary norm):

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

Then, (1) becomes the problem with ball-regularization:

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

- D is a regularization parameter.

Contracting-Domain Newton Method

Algorithm for solving (1):

Initialization: Choose $x_0 \in \operatorname{dom} \psi$.
Iterations: $k \geq 0$.

1: Pick up $\gamma_k \in (0, 1]$.

2: Compute

$$v_{k+1} \in \operatorname*{Argmin}_y \Big\{ \langle \nabla f(x_k), y - x_k \rangle + \tfrac{\gamma_k}{2} \langle \nabla^2 f(x_k)(y - x_k), y - x_k \rangle + \psi(y) \Big\}.$$

3: Set $x_{k+1} := x_k + \gamma_k (v_{k+1} - x_k)$.

- Interpretation: Conditional gradient method (Frank-Wolfe algorithm) with second-order global lower model.

Global Convergence

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

$$F(x_k) - F^* \leq \mathcal{O}(1/k^2).$$

- This is much faster than the $\mathcal{O}(1/k)$-rate of Frank-Wolfe.
- The same rate as for Cubically regularized Newton method. However, the new algorithm is **affine-invariant** and **universal** (it does not depend on norms and parameters of the problem class).

NB: in Classical Newton Method: $\gamma_k \equiv 1$ (no global convergence).

Stochastic Optimization

We can use randomized unbiased estimators $g_k \approx \nabla f(x_k)$, $H_k \approx \nabla^2 f(x_k)$ instead of the true gradients and the Hessians. To keep the fast global convergence, we need to increase batch size over the iterations.

Theorem. Let $m_k^g = \mathcal{O}(k^4)$ (batch size for the gradients on iteration k), and $m_k^H = \mathcal{O}(k^2)$ (batch size for the Hessians). Then, for the stochastic method we have

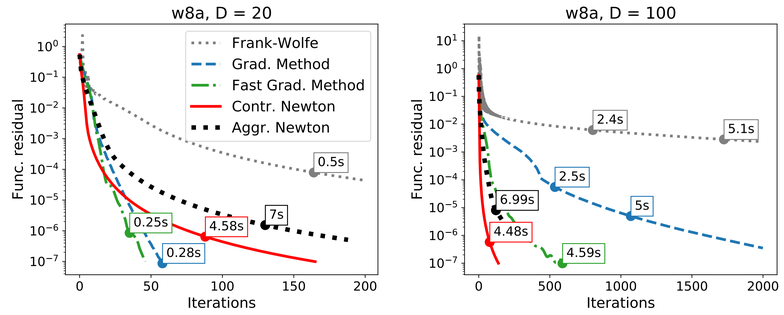
$$\mathbb{E}\big[F(x_k) - F^*\big] \leq \mathcal{O}(1/k^2). \qquad (2)$$

Theorem. For finite-sum minimization we can use **variance reduction** for the gradients. Then it is enough to set $m_k^g = m_k^H = \mathcal{O}(k^2)$ to have (2).

Experiments

Minimizing Logistic Regression with ℓ_2 ball-regularization.

- Exact methods:



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

- Stochastic methods: