Quasi-Newton and Second-Order Methods for Convex Optimization

Anton Rodomanov   Nikita Doikov   Yurii Nesterov

Quasi-Newton (QN) Methods

Problem: $\min_{x \in \mathbb{R}^n} f(x)$, where $f\colon \mathbb{R}^n \to \mathbb{R}$ is a smooth function.

QN methods: Choose $x_0 \in \mathbb{R}^n$, $H_0 \succ 0$ and iterate for $k \geq 0$:

- Set $x_{k+1} = x_k - \alpha_k H_k \nabla f(x_k)$ for some $\alpha_k \geq 0$.
- Update H_k into H_{k+1}.

Standard updating rules:

- (DFP) $H_{k+1} = H_k - \frac{H_k \gamma_k \gamma_k^T H_k}{\langle \gamma_k, H_k \gamma_k \rangle} + \frac{u_k u_k^T}{\langle \gamma_k, u_k \rangle}$,
- (BFGS) $H_{k+1} = H_k - \frac{H_k \gamma_k u_k^T + u_k \gamma_k^T H_k}{\langle \gamma_k, u_k \rangle} + \left(\frac{\langle \gamma_k, H_k \gamma_k \rangle}{\langle \gamma_k, u_k \rangle} + 1 \right) \frac{u_k u_k^T}{\langle \gamma_k, u_k \rangle}$,

where $u_k := x_{k+1} - x_k$, $\gamma_k := \nabla f(x_{k+1}) - \nabla f(x_k)$.

Main result: $\frac{\|\nabla f(x_{k+1})\|}{\|\nabla f(x_k)\|} \to 0$ (superlinear convergence).

Open question: Rate of convergence? (explicit nonasymptotic estimates)

New Results on QN Methods

Main assumptions: $\exists \mu, L > 0$, $\exists M \geq 0$, $\forall x, y, z, w \in \mathbb{R}^n$:

(1) $\mu I \preceq \nabla^2 f(x) \preceq LI$,   (2) $\nabla^2 f(x) - \nabla^2 f(y) \preceq M \|x - y\|_z \nabla^2 f(w)$,

where $\|h\|_z := \langle \nabla^2 f(z) h, h \rangle^{1/2}$ for $h \in \mathbb{R}^n$.

Note: (1) + (2) $\iff$ (1) + L_2-Lipschitz Hessian, but (2) is affine invariant.

Theorem: Let $H_{k+1} = (1 - \tau)\,\mathrm{BFGS}(H_k, u_k, \gamma_k) + \tau\,\mathrm{DFP}(H_k, u_k, \gamma_k)$, where $\tau \in [0, 1]$, and let $\lambda_k := \|\nabla f(x_k)\|_{x_k}^*$ for each $k \geq 0$. Suppose $H_0 = \frac{1}{L} I$ and

$$M \lambda_0 \leq \frac{\ln \frac{3}{2}}{(\frac{3}{2})^{\frac{3}{2}}} \max\left\{ \frac{1}{2Q}, \frac{1}{K_0 + 9} \right\}, \qquad K_0 := \lceil 8 n Q_\tau \ln(2Q) \rceil,$$

where $Q_\tau := (1 - \tau + \tau \frac{4}{9} Q^{-1})^{-1}$, $Q := \frac{L}{\mu}$ (condition number). Then, $\forall k \geq 1$:

$$\lambda_k \leq \left(1 - \frac{1}{2Q}\right)^k \sqrt{\tfrac{3}{2}}\,\lambda_0, \qquad \lambda_k \leq \left[\tfrac{5}{2} Q_\tau \left(\exp\left\{ \frac{13 n \ln(2Q)}{6k} \right\} - 1 \right) \right]^{k/2} \sqrt{\tfrac{3}{2} Q}\,\lambda_0.$$

Discussion:

- Global convergence for quadratic functions ($M = 0$).
- For BFGS ($\tau = 0$), the rate is
$$\left[\exp\left\{ \frac{n \ln Q}{k} \right\} - 1 \right]^k \lesssim \left(\frac{n \ln Q}{k} \right)^k, \qquad k \gtrsim n \ln Q.$$
Region of local convergence: $M \lambda_0 \lesssim \max\{Q^{-1}, [n \ln Q]^{-1}\}$.
- For DFP ($\tau = 1$), the rate is
$$\left[Q \left(\exp\left\{ \frac{n \ln Q}{k} \right\} - 1 \right) \right]^k \lesssim \left(\frac{n Q \ln Q}{k} \right)^k, \qquad k \gtrsim n Q \ln Q.$$
Region of local convergence: $M \lambda_0 \lesssim Q^{-1}$.

Ellipsoid Method (EM)

Problem: $\min_{x \in Q} f(x)$, where

- Q is a closed convex set in $\mathbb{R}^n$, represented by the Separation Oracle: for any $x \notin Q$, it returns $g_Q(x) \in \mathbb{R}^n \setminus \{0\}$:
$$\langle g_Q(x), x - y \rangle \geq 0, \quad \forall y \in Q.$$
- $f\colon \mathbb{R}^n \to \mathbb{R}$ is a convex (possibly nonsmooth) function.

Main assumption: $B(\bar{x}, r) \subseteq Q \subseteq B(x_0, R)$ for some $\bar{x}, x_0 \in \mathbb{R}^n$, $r, R > 0$.

Oracle: $\mathcal{G}(x) := f'(x)$ if $x \in Q$ and $\mathcal{G}(x) := g_Q(x)$ if $x \notin Q$.

EM [Yudin-Nemirovski, 1976]: Set $W_0 = R^2 I$. Iterate for $k \geq 0$:

$$x_{k+1} = x_k - \frac{1}{n+1} \frac{W_k g_k}{\langle g_k, W_k g_k \rangle^{1/2}}, \qquad W_{k+1} = \frac{n^2}{n^2 - 1} \left(W_k - \frac{2}{n+1} \frac{W_k g_k g_k^T W_k}{\langle g_k, W_k g_k \rangle} \right),$$

where $g_k := \mathcal{G}(x_k)$.

Complexity: EM finds $\bar{x} \in Q$ with $f(\bar{x}) - f^* \leq \epsilon$ in $O(n^2 \ln \frac{\alpha V}{\epsilon})$ iterations, where $\alpha := R/r$ is the asphericity of Q and $V := \max_Q f - \min_Q f$ is the variation of f on Q.

Main problem: Does not work when $n \to \infty$: $x_{k+1} = x_k$, $W_{k+1} = W_k$.

Note: The simplest subgradient method

$$x_{k+1} = x_k - h g_k, \qquad g_k := \mathcal{G}(x_k), \qquad k \geq 0,$$

does not have this problem. Its complexity $O(\frac{\alpha^2 V^2}{\epsilon^2})$ is independent of n.

Subgradient Ellipsoid Method

Idea: Combine Subgradient method (SM) with EM.

Algorithm: Set $\ell_0(x) := 0$, $\omega_0(x) := \frac{1}{2}\|x - x_0\|^2$. Iterate for $k \geq 0$:

1. Compute $g_k := \mathcal{G}(x_k)$ and $U_k := \max_{x \in \Omega_k \cap L_k^-} \langle g_k, x_k - x \rangle$, where
$$\Omega_k := \{x \colon \omega_k(x) \leq \tfrac{1}{2} R^2\}, \qquad L_k^- := \{x \colon \ell_k(x) \leq 0\}.$$
2. Choose coefficients $a_k, b_k \geq 0$ and update the functions
$$\ell_{k+1}(x) := \ell_k(x) + a_k \langle g_k, x - x_k \rangle,$$
$$\omega_{k+1}(x) := \omega_k(x) + \tfrac{1}{2} b_k (U_k - \langle g_k, x_k - x \rangle) \langle g_k, x - x_k \rangle.$$
3. Set $x_{k+1} := \operatorname{argmin}_{x \in \mathbb{R}^n} [\ell_{k+1}(x) + \omega_{k+1}(x)]$.

Geometry:

- Ω_k is an ellipsoid, while L_k^- is a halfspace.
- For all $k \geq 0$ and any solution x^* to our problem, we have
$$x^* \in \Omega_k \cap L_k^- \quad \text{and} \quad \{x \in \Omega_k \cap L_k^- \colon \langle g_k, x - x_k \rangle \leq 0\} \subseteq \Omega_{k+1} \cap L_{k+1}^-.$$

Choice of coefficients:

- $a_k > 0$, $b_k = 0 \implies$ SM.
- $a_k = 0$, $b_k > 0 \implies$ EM.

Iteration cost: $O(n^2)$.

Complexity: $\approx \min\{\frac{\alpha^2 V^2}{\epsilon^2}, n^2 \ln \frac{\alpha V}{\epsilon}\}$ ($\approx$ best of SM and EM).

Cubic Newton

- An issue of the classical Newton method with unit step size: it has **no global convergence** guarantees.

We consider Newton's method with Cubic Regularization [Nesterov-Polyak, 2006]:

$$x_{k+1} = \operatorname*{argmin}_{y \in \mathbb{R}^n} \left\{ \langle \nabla f(x_k), y - x_k \rangle + \tfrac{1}{2} \langle \nabla^2 f(x_k)(y - x_k), y - x_k \rangle + \tfrac{H \|y - x_k\|^3}{6} \right\}$$
$$= x_k - \left(\nabla^2 f(x_k) + \tfrac{H r_k}{2} I \right)^{-1} \nabla f(x_k).$$

- $H := 0 \ \Rightarrow$ the classical Newton method
- $H := L_2$ (Lipschitz constant for the Hessian) $\Rightarrow$ global convergence
- adaptive strategy for choosing H [Nesterov-Polyak, 2006, Cartis-Gould-Toint, 2011; Grapiglia-Nesterov, 2017]

Uniformly Convex Functions

The class of **nondegenerate** problems:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{\sigma}{q} \|y - x\|^q, \quad \forall x, y \in \mathbb{R}^n$$

- $q \geq 2$ is the degree of uniform convexity, $\sigma > 0$ is a parameter
- $q = 2$: strongly convex functions $\Rightarrow$ the Gradient method has fast global linear rate of convergence; Second-order methods — ?

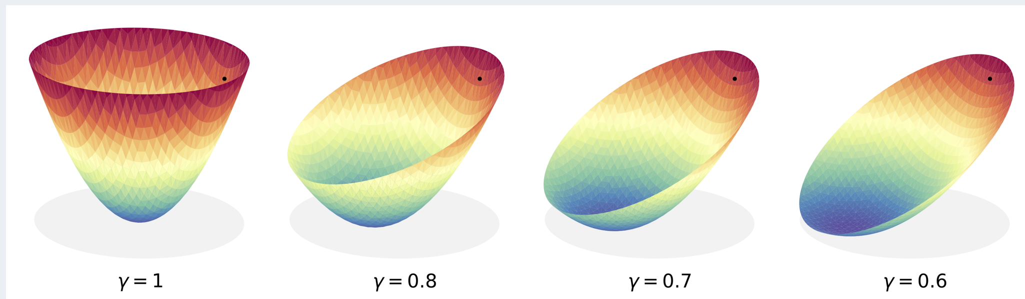
Theorem: The global complexity of the adaptive Cubic Newton for $q \in [2, 3]$ is

$$\mathcal{O}\left(\omega \log \tfrac{f(x_0) - f^*}{\varepsilon} \right)$$

iterations, where ω is a second-order condition number $\Rightarrow$ **Cubic Newton** is better than the **Gradient method**.

Contraction Technique

We consider contraction of the objective: $g(x) = f(\gamma x + (1 - \gamma)\bar{x})$, $\gamma \in [0, 1]$



- Smoothness properties of $g(\cdot)$ are better than that of $f(\cdot)$

New Contracting Newton method:

$$x_{k+1} = \operatorname*{argmin}_{y \in x_k + \gamma_k (Q - x_k)} \left\{ \langle \nabla f(x_k), y - x_k \rangle + \tfrac{1}{2} \langle \nabla f(x_k)(y - x_k), y - x_k \rangle \right\}$$

- $\gamma_k = 1$: the classical Newton method
- Interpretation: regularization of quadratic model by the asymmetric trust region

Theorem: Set $\gamma_k = \frac{3}{3+k}$. Global convergence: $f(x_k) - f^* \leq \mathcal{O}(\frac{1}{k^2})$