Second-order methods with global convergence in Convex Optimization

Nikita Doikov

UCLouvain, Belgium

June 3, 2022

The Goal: efficient **second-order** optimization methods with global convergence guarantees.

- The rate should be better than that of the **first-order** methods
- We analyse the complexity of the methods alongside suitable problem classes
- Implementable algorithms

Plan of the Talk

I. Intro

- Gradient methods
- Newton's method, classical approach

II. Modern techniques

- Cubic regularization
- Contracting-point methods
- Acceleration

III. Conclusions

$$\min_x f(x), \qquad x \in \mathbb{R}^n$$

f is *differentiable*; $\nabla f(x) \in \mathbb{R}^n$ — gradient of the function,

$$\left[\nabla f(x)\right]^{(i)} = \frac{\partial f(x)}{\partial x^{(i)}}, \qquad 1 \leq i \leq n$$

The Gradient Method

Iterate, for $k \geq 0$:

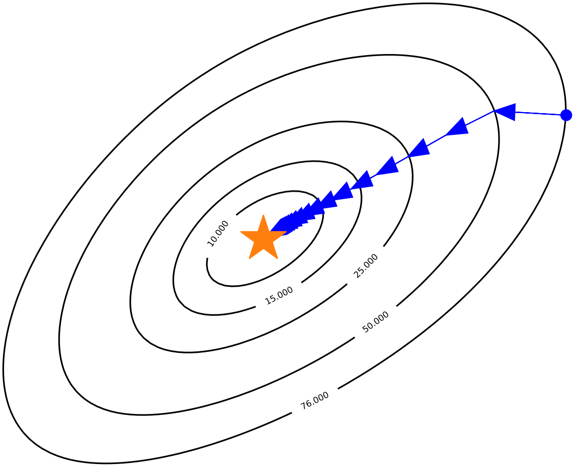
$$\boxed{x_{k+1} := x_k - \alpha_k \nabla f(x_k), \quad \text{for some} \quad \alpha_k > 0}$$

[Cauchy, 1847]

+ **Cheap iterations:** $\mathcal{O}(n)$
+ **Global convergence**
− **Slow rate:** $f(x_k) - f^* \leq \mathcal{O}(1/k)$

The Gradient Method: Trajectory

$$x_{k+1} \quad := \quad x_k - \alpha_k \nabla f(x_k)$$



$$\boxed{\min_{x \in \mathbb{R}^n} f(x)} \qquad (*)$$

f is convex and differentiable; the gradient is Lipschitz continuous; dimension n is big.

[Nemirovski-Yudin, 1979]: Any first-order method solving $(*)$ needs at least $\mathcal{O}(\frac{1}{\sqrt{\varepsilon}})$ iterations to solve the problem with ε accuracy:

$$f(\bar{x}) - f^* \quad \leq \quad \varepsilon.$$

- The Gradient Method: $\mathcal{O}(\frac{1}{\varepsilon})$ — not optimal
- The Fast Gradient Method: $\mathcal{O}(\frac{1}{\sqrt{\varepsilon}})$ [Nesterov, 1983] — optimal
- Better rates? — impossible for the first-order methods

$$\boxed{\min_{x \in \mathbb{R}^n} f(x)}$$

The Hessian $\nabla^2 f(x) \in \mathbb{R}^{n \times n}$ is the second-order information about the objective,

$$[\nabla^2 f(x)]^{(i,j)} = \frac{\partial^2 f(x)}{\partial x^{(i)} \partial x^{(j)}}, \qquad 1 \leq i, j \leq n$$

A full quadratic model of the objective, $f(y) \approx \Omega_2(x; y)$, where

$$\Omega_2(x; y) \stackrel{\text{def}}{=} f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{1}{2} \langle \nabla^2 f(x)(y - x), y - x \rangle.$$

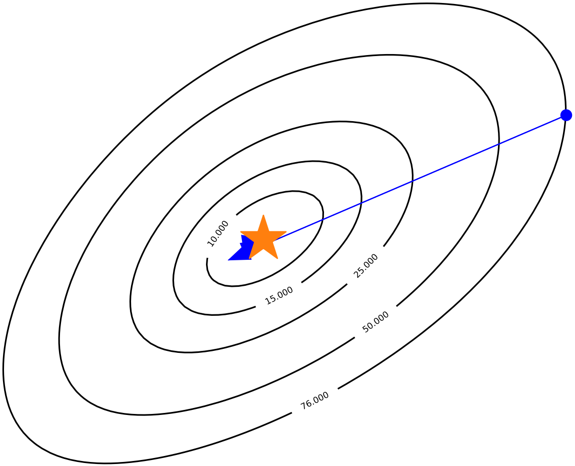
Newton's Method. Iterate, for $k \geq 0$:

$$\boxed{\begin{aligned} x_{k+1} &:= \operatorname*{argmin}_{y \in \mathbb{R}^n} \Omega_2(x_k; y) \\ &= x_k - \nabla^2 f(x_k)^{-1} \nabla f(x_k) \end{aligned}}$$

[Newton, 1669; Raphson, 1690; Fine-Bennett, 1916; Kantorovich, 1948]

Newton's Method: Trajectory

$$x_{k+1} \quad := \quad x_k - \nabla^2 f(x_k)^{-1} \nabla f(x_k)$$



$$x_{k+1} \quad := \quad x_k - \nabla^2 f(x_k)^{-1} \nabla f(x_k)$$

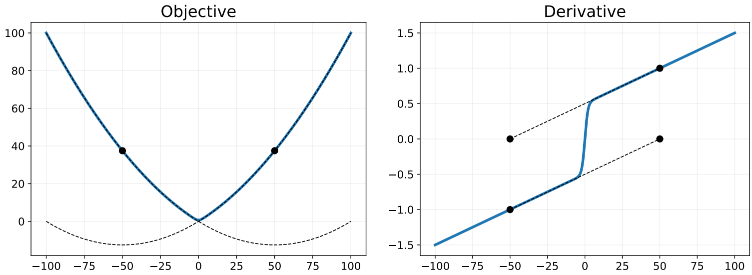
- Solving a linear system requires $\mathcal{O}(n^3)$ per iteration
- Fast **local** convergence:

$$\mathcal{O}(\log \log \tfrac{1}{\varepsilon})$$

  iterations to find an ε-solution, **when in the neighbourhood of the optimum**
- Global convergence — ?

$$\min_{x \in \mathbb{R}} \left\{ f(x) \quad := \quad \log(1 + \exp(x)) - \tfrac{1}{2}x + \tfrac{\mu}{2}x^2 \right\}, \qquad \mu \; := \; 10^{-2}.$$

- The objective is smooth and strongly convex; $x^* = 0$.



The method oscillates between two points!

How to Fix the Newton Method?

- **Damped Newton step** [Kantorovich, 1948]

$$x_{k+1} = x_k - \alpha_k \nabla^2 f(x_k)^{-1} \nabla f(x_k), \qquad \alpha_k \in (0, 1]$$

- **Quadratic regularization**
  [Levenberg, 1944; Marquardt, 1963]

$$x_{k+1} = x_k - \bigl(\nabla^2 f(x_k) + \alpha_k I\bigr)^{-1} \nabla f(x_k)$$

- **Trust-region approach**
  [Goldfeld-Quandt-Trotter, 1966; Conn-Gould-Toint, 2000]

$$x_{k+1} = \operatorname*{argmin}_{\|y - x_k\| \leq \Delta_k} \Omega_2(x_k; y)$$

Works well in practice. Difficult to establish good global rates

Plan of the Talk

I. Intro

- Gradient methods
- Newton's method, classical approach

II. Modern techniques

- Cubic regularization
- Contracting-point methods
- Acceleration

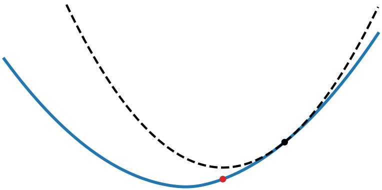
III. Conclusions

The Gradient Method: a Modern View

$\boxed{\min_{x \in \mathbb{R}^n} f(x)}$ **Assumption:** gradient is Lipschitz continuous

$$\|\nabla f(y) - \nabla f(x)\| \quad \leq \quad L_1 \|y - x\|, \qquad \forall x, y \in \mathbb{R}^n$$

$$\Rightarrow \qquad f(y) \quad \leq \quad f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{L_1}{2} \|y - x\|^2$$



The Gradient Step minimizes **the model of the objective**:

$$\begin{aligned} x_{k+1} &= \operatorname*{argmin}_{y \in \mathbb{R}^n} \Big[f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \tfrac{L_1}{2} \|y - x_k\|^2 \Big] \\ &= x_k - \tfrac{1}{L_1} \nabla f(x_k) \end{aligned}$$

$\boxed{\min_{x \in \mathbb{R}^n} f(x)}$ **New assumption:** Hessian is Lipschitz continuous

$$\|\nabla^2 f(y) - \nabla^2 f(x)\| \quad \leq \quad L_2 \|y - x\|, \qquad \forall x, y \in \mathbb{R}^n$$

$$\Rightarrow \qquad f(y) \quad \leq \quad \Omega_2(x; y) + \tfrac{L_2}{6}\|y - x\|^3,$$

where Ω_2 is the second-order Taylor approximation of f.

Newton method with cubic regularization:

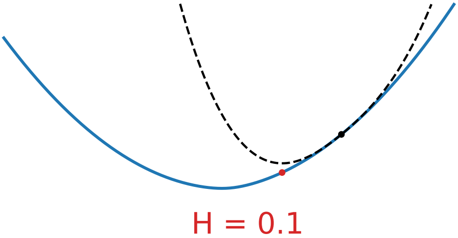
$$\begin{aligned} x_{k+1} \quad &= \quad \operatorname*{argmin}_{y \in \mathbb{R}^n} \Big[M_H(x_k; y) \ \stackrel{\mathrm{def}}{=} \ \Omega_2(x_k; y) + \tfrac{H}{6}\|y - x_k\|^3 \Big] \\ &= \quad x_k \ - \ \Big(\nabla^2 f(x_k) + \tfrac{H\|x_{k+1} - x_k\| I}{2} \Big)^{-1} \nabla f(x_k) \end{aligned}$$

Theorem. Set $H := L_2$. Then, $f(x_k) - f^* \leq \mathcal{O}(1/k^2)$
[Nesterov-Polyak, 2006]

Iterate, for $k \geq 0$:

$$\boxed{\begin{array}{rcl} x_{k+1} & := & \underset{y \in \mathbb{R}^n}{\operatorname{argmin}} \Big[M_H(x_k; y) \Big] \\ & = & x_k - \left(\nabla^2 f(x_k) + \frac{H r_k}{2} I \right)^{-1} \nabla f(x_k) \end{array}}$$

- $H := 0 \quad \Rightarrow \quad$ the classical Newton's Method
- **Constant choice** $H := L_2$
- **Adaptive strategy** [Nesterov-Polyak, 2006; Cartis-Gould-Toint, 2011; Grapiglia-Nesterov, 2017] ensures $f(x_{k+1}) \leq M_H(x_k; x_{k+1})$



Let $T_H(x) \stackrel{\mathrm{def}}{=} \operatorname*{argmin}_{y \in \mathbb{R}^n} \Big[M_H(x; y) \Big]$, $M_H^*(x) \stackrel{\mathrm{def}}{=} \min_{y \in \mathbb{R}^n} \Big[M_H(x; y) \Big]$

Algorithm 1 Adaptive Cubic Newton

Initialization: Choose $x_0 \in \mathbb{R}^n$, $H_0 > 0$.

Iterations: $k \geq 0$.

1: Find minimum integer $i_k \geq 0$ s.t. it holds

$$f(T_{H_k 2^{i_k}}(x_k)) \leq M^*_{H_k 2^{i_k}}(x_k).$$

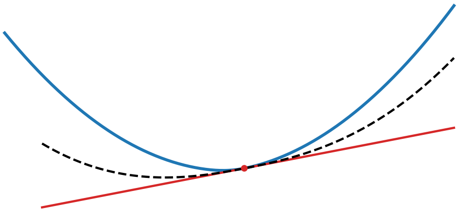
2: Make the Cubic step $x_{k+1} := T_{H_k 2^{i_k}}(x_k)$.

3: Set $H_{k+1} := H_k 2^{i_k - 1}$.

Uniformly Convex Functions

f is called uniformly convex of degree $q \geq 2$ with $\sigma_q > 0$, iff

$$f(y) \quad \geq \quad f(x) + \langle \nabla f(x), y - x \rangle + \frac{\sigma_q}{q} \|y - x\|^q, \quad \forall x, y$$



- Example: $f(x) = \frac{1}{q} \|x - x_0\|^q$ is uniformly convex of degree q with constant $\sigma_q = 2^{2-q}$
- Sum of convex and uniformly convex functions gives uniformly convex
- Strongly convex functions: $q = 2$. **GM:** global linear rate.

Second-order methods — ?

Global Lower Model

- Let $\nabla^2 f$ be Lipschitz continuous
- Let f be uniformly convex of degree $q = 3$

Theorem. We have a *global lower bound*, $\forall x, y$:

$$
\begin{aligned}
f(y) \quad \geq \quad & f(x) + \langle \nabla f(x), y - x \rangle \\
& + \tfrac{\gamma}{2} \langle \nabla^2 f(x)(y - x), y - x \rangle + \tfrac{\gamma^2 L_2}{6} \|y - x\|^3
\end{aligned}
$$

where $\gamma \geq \frac{1}{2} \min\{1, \frac{2}{3}\sqrt{\frac{\sigma_3}{L_2}}\}$ is a **condition number**.

Recall the global *upper model*:

$$
f(y) \quad \leq \quad f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{1}{2} \langle \nabla^2 f(x)(y - x), y - x \rangle + \tfrac{L_2 \|y - x\|^3}{6}
$$

Cubic Newton, with $H := L_2$:

$$\boxed{x_{k+1} \quad = \quad \operatorname*{argmin}_{y \in \mathbb{R}^n} \Big[\Omega_2(x_k; y) + \frac{H}{6}\|y - x_k\|^3 \Big]}$$

Theorem. $f(x_{k+1}) - f^* \le (1-\gamma)(f(x_k) - f^*)$.
Proof:

$$\gamma f^* + (1-\gamma) f(x_k)$$

$$\overset{\text{lower}}{\geq} \quad f(x_k) + \gamma \langle \nabla f(x_k), x^* - x_k \rangle$$

$$+ \frac{\gamma^2}{2}\langle \nabla^2 f(x_k)(x^* - x_k), x^* - x_k \rangle + \frac{\gamma^3 L_2}{6}\|x^* - x\|^3$$

Denote $y := x_k + \gamma(x^* - x_k)$. Then $y - x_k = \gamma(x^* - x_k)$.

Proof: continue

We have

$$
\begin{aligned}
&\gamma f^* + (1-\gamma) f(x_k) \\
&\quad \geq \quad f(x_k) + \langle \nabla f(x_k), y - x_k \rangle \\
&\qquad\quad + \tfrac{1}{2}\langle \nabla^2 f(x_k)(y - x_k), y - x_k \rangle + \tfrac{L_2}{6}\|y - x\|^3 \\
&\quad \overset{\text{method}}{\geq} \quad f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle \\
&\qquad\quad + \tfrac{1}{2}\langle \nabla^2 f(x_k)(x_{k+1} - x_k), x_{k+1} - x_k \rangle + \tfrac{L_2}{6}\|x_{k+1} - x_k\|^3 \\
&\quad \overset{\text{upper}}{\geq} \quad f(x_{k+1})
\end{aligned}
$$

Hence,

$$
f(x_{k+1}) - f^* \quad \leq \quad (1-\gamma)\big(f(x_k) - f^*\big).
$$

□

- Let Hessian be Hölder continuous of degree $\nu \in [0, 1]$:

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \quad \leq \quad \mathcal{H}_\nu \|x - y\|^\nu$$

- Let f be uniformly convex of degree $q = 2 + \nu$.

Condition number of degree ν:

$$\boxed{\gamma_\nu \quad \stackrel{\text{def}}{=} \quad \left(\frac{\sigma_{2+\nu}}{\mathcal{H}_\nu}\right)^{\frac{1}{1+\nu}}}$$

Theorem. [D-Nesterov, 2019]: The global complexity of the adaptive Cubic Newton is

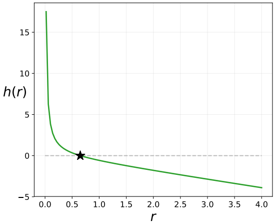
$$\mathcal{O}\left(\inf_{\nu \in [0,1]} \left[\gamma_\nu{}^{-1}\right] \log \frac{f(x_0) - f^*}{\varepsilon}\right)$$

Note: $\gamma_0{}^{-1} = \frac{\mathcal{H}_0}{\sigma_2} \leq \frac{L_1 - \sigma_2}{\sigma_2} < \frac{L_1}{\sigma_2} \quad \Rightarrow$ **Cubic Newton** is better than the **Gradient Method**

How to Compute Iteration?

$$\text{Cubic step:} \qquad x^{+} \;=\; x - \left(\nabla^2 f(x) + \tfrac{Hr^{+}}{2} I\right)^{-1} \nabla f(x),$$

where $r^{+} = \|x^{+} - x\|$ is the root of 1-D equation $h(r) = 0$:



We can apply any one-dimensional method (bisection, Newton, ...)

- $\tilde{\mathcal{O}}(1)$ matrix inversions, or one matrix factorization — $\mathcal{O}(n^3)$

Gradient regularization

[Ueda-Yamashita, 2014; Mishchenko, 2021; D-Nesterov, 2021]:

$$x^{+} \;=\; x - \left(\nabla^2 f(x) + \sqrt{\tfrac{H\|\nabla f(x)\|}{3}} I\right)^{-1} \nabla f(x)$$

- One matrix inversion; fast global rates

$$\boxed{\min_{x \in \mathbb{R}^n} f(x)}$$ where n is Huge

- even $\mathcal{O}(n)$ per iteration can be expensive

Idea: sample random coordinates $S \subset \{1, \ldots, n\}$ of size $\boldsymbol{\tau} = |S|$ and apply one step of the Cubic Newton along these coordinates

- The cost of each step is $\mathcal{O}(\boldsymbol{\tau}^3)$

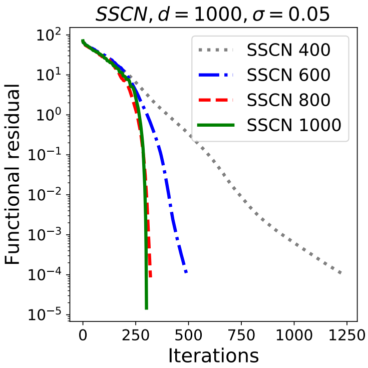
[D-Richtárik, 2018; Hanzely-D-Richtárik-Nesterov, 2020]

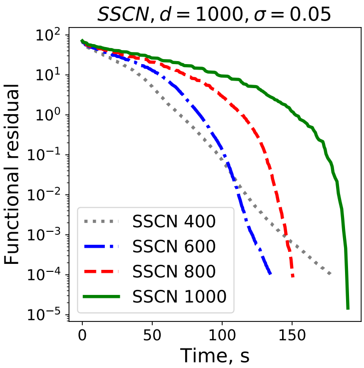
Theorem. The method converges globally, after k iterations:

$$\mathbb{E} f(x_k) - f^* \quad \leq \mathcal{O}\Big(\tfrac{n - \boldsymbol{\tau}}{\boldsymbol{\tau}} \cdot \tfrac{1}{k} + \big(\tfrac{n}{\boldsymbol{\tau}}\big)^2 \cdot \tfrac{1}{k^2} \Big)$$

Experiment

$$\min_{x \in \mathbb{R}^d} f(x) \quad = \quad \sigma \log\left(\sum_{i=1}^{m} e^{(\langle a_i, x\rangle - b_i)/\sigma} \right) \quad \approx \quad \max_{i=1}^{m} \langle a_i, x\rangle - b_i$$





Plan of the Talk

I. Intro

- Gradient methods
- Newton's method, classical approach

II. Modern techniques

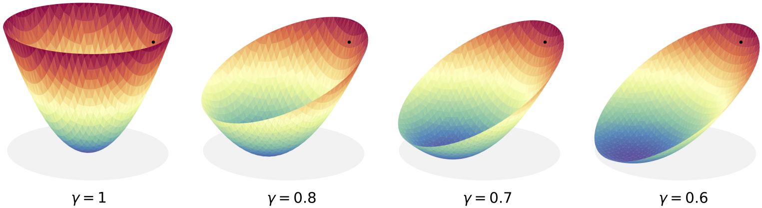
- Cubic regularization
- Contracting-point methods
- Acceleration

III. Conclusions

Contraction Technique

Let us consider contraction of the objective:

$$g(x) \quad := \quad f(\gamma x + (1-\gamma)\bar{x}), \qquad \gamma \in [0,1].$$



Note:

$$\nabla g(x) \quad = \quad \gamma \nabla f(\ldots),$$

$$\nabla^2 g(x) \quad = \quad \gamma^{\mathbf{2}} \nabla^2 f(\ldots),$$

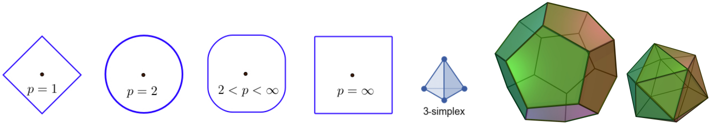
$$\ldots$$

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

Idea: use γ to balance the error of $g(x) \approx f(x)$ and smoothness

Contracting-Point Methods

$$\boxed{\min_{x \in Q} f(x)}$$ where $Q \subset \mathbb{R}^n$ is a bounded convex set, e.g.:



Conceptual Contracting-Point Method. Iterate, $k \geq 0$:

$$\boxed{\begin{array}{rcl} v_{k+1} & \approx & \operatorname*{argmin}_{v \in Q} f(\gamma_k v + (1 - \gamma_k) x_k), \\ x_{k+1} & = & \gamma_k v_{k+1} + (1 - \gamma_k) x_k \end{array}}$$

Approximate f by p-th order Taylor's polynomial.

- $p = 1$: The Conditional Gradient Method [Frank-Wolfe, 1956]
- $p = 2$: Contracting Newton [D-Nesterov, 2020]

Contracting Newton

$$\boxed{\min_{x \in Q} f(x)}$$

Iterate, for $k \geq 0$:

$$\boxed{x_{k+1} \quad := \quad \operatorname*{argmin}_{y} \Big\{ \ \Omega_2(x_k; y) \ : \ y \in x_k + \gamma_k (Q - x_k) \ \Big\}}$$

where Ω_2 is the second-order approximation of f

- $\gamma_k = 1$: The classical Newton's Method
- Interpretation: regularization of quadratic model by the assymmetric trust region

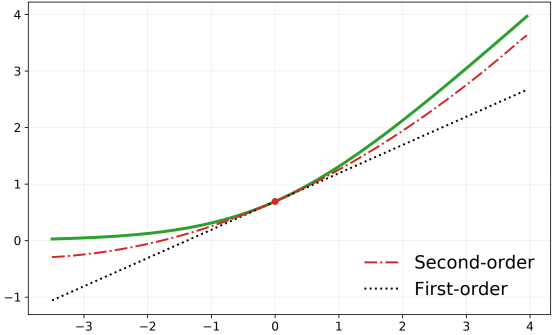
Theorem. Set $\gamma_k = \frac{3}{k+3} \quad \Rightarrow \quad$ **global rate:** $f(x_k) - f^* \leq \mathcal{O}(1/k^2)$

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

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \quad \leq \quad L_2 \|x - y\|$$

Convexity + Smoothness $\Rightarrow$ **tighter lower bound**: $\exists \gamma_{x,y} \in (0, 1]$

$$f(y) \quad \geq \quad f(x) + \langle \nabla f(x), y - x \rangle + \frac{\gamma_{x,y}}{2} \langle \nabla^2 f(x)(y - x), y - x \rangle$$



Cubic Newton vs. Contracting Newton

Cubic Newton, H_k	**Contracting Newton**, γ_k
global *upper* approximation	global *lower* model
fixed Euclidean norm	affine-invariant
-	bounded domain

Complexity

- Convex functions: $\mathcal{O}(\frac{1}{\sqrt{\varepsilon}})$; $\gamma_k = \frac{3}{3+k}$
- Strongly convex functions: $\mathcal{O}(\boldsymbol{\omega} \log \frac{f(x_0)-f^*}{\varepsilon})$; $\gamma_k = \frac{1}{1+\boldsymbol{\omega}}$
- Local quadratic convergence: $\gamma_k = 1$

$$\boxed{x_{k+1} \;\; = \;\; \operatorname*{argmin}_y \Big\{ \, \Omega_2(x_k; y) \; : \; y \in x_k + \gamma_k (Q - x_k) \, \Big\}}$$

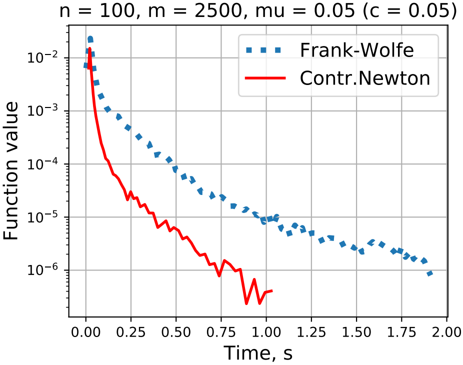
How to compute the iteration?

- At each step we solve the subproblem inexactly by the first-order Frank-Wolfe algorithm
- We have full control over the required accuracy

Theorem. To reach $f(x_K) - f^* \leq \varepsilon$ it needs

- $K = \mathcal{O}\big(\frac{1}{\sqrt{\varepsilon}}\big)$ oracle calls for f
- $\mathcal{O}(\frac{1}{\varepsilon})$ linear minimization oracle calls for Q totally

$$\min_{x \in \mathbb{R}^n_+} \left\{ f(x) \;=\; \mu \log\Big(\sum_{i=1}^m e^{(\langle a_i, x\rangle - b_i)/\mu} \Big) \;:\; \sum_{i=1}^n x^{(i)} = 1 \right\}$$



two times faster

Finite-sum minimization: $f(x) = \frac{1}{M} \sum_{i=1}^{M} f_i(x)$

Big $M \;\Rightarrow\;$ computing $\nabla f(x)$ and $\nabla^2 f(x)$ is **expensive**

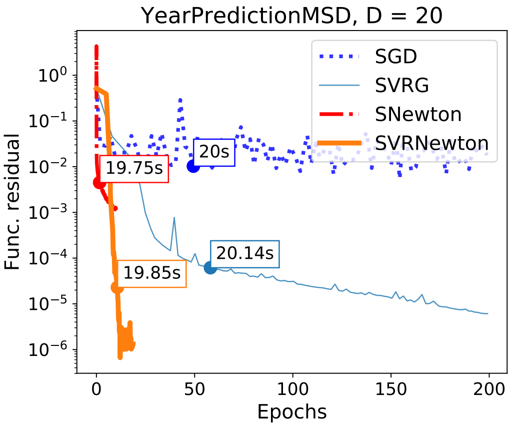
Idea: sample random batch $B \subseteq \{1, \dots, M\}$ and approximate

$$\nabla f(x) \;\approx\; \frac{1}{|B|} \sum_{i \in B} \nabla f_i(x), \qquad \nabla^2 f(x) \;\approx\; \frac{1}{|B|} \sum_{i \in B} \nabla^2 f_i(x)$$

Idea #2: at some iterations recompute the full gradient and Hessian — *variance reduction* [Schmidt-Roux-Bach, 2011]

New algorithm: **Stochastic Contracting Newton Method** with global complexities: $\mathcal{O}(\frac{1}{\varepsilon^{1/2}})$ iterations and $\mathcal{O}(\frac{1}{\varepsilon^{3/2}})$ total random samples among all batches

Experiments: Stochastic Methods for Logistic Regression



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

Plan of the Talk

I. Intro

- Gradient methods
- Newton's method, classical approach

II. Modern techniques

- Cubic regularization
- Contracting-point methods
- Acceleration

III. Conclusions

Problem class: convex functions with Lipschitz Hessian

Basic Cubic Newton: $f(x_k) - f^* \leq \mathcal{O}(1/k^2)$

Accelerated Cubic Newton: $\mathcal{O}(1/k^3)$ [Nesterov, 2008]

Accelerated second-order prox: $\mathcal{O}(1/k^{3.5})$ [Monteiro-Svaiter, 2013]

(extra one-dimensional search each iteration)

Optimal rate, matching the lower bound [Arjevani-Shamir-Shiff, 2019]

Problem class: convex functions with bounded <u>second</u> and <u>fourth</u> derivatives

Superfast second-order schemes: $\mathcal{O}(1/k^5)$ [Nesterov, 2020; Kamzolov-Gasnikov, 2020]

- Approximation of third derivative by finite-differences

- **Contracting-Point Method:**

$$\boxed{\begin{array}{rcl} v_{k+1} & \approx & \operatorname*{argmin}_{x \in Q} f(\gamma_k x + (1-\gamma_k) x_k), \\ x_{k+1} & = & \gamma_k v_{k+1} + (1-\gamma_k) x_k. \end{array}}$$

One step of 2-order Taylor's approximation $\Rightarrow$ $\mathcal{O}(1/k^2)$-rate.

- **Contracting Proximal Method** [D-Nesterov, 2019]:

$$\boxed{\begin{array}{rcl} v_{k+1} & \approx & \operatorname*{argmin}_{x \in Q} \Big\{ f(\gamma_k x + (1-\gamma_k) x_k) + \frac{1}{A_{k+1}} \beta_d(v_k; x) \Big\}, \\ x_{k+1} & = & \gamma_k v_{k+1} + (1-\gamma_k) x_k. \end{array}}$$

$\beta_d(v_k; x) = d(x) - d(v_k) - \langle \nabla d(v_k), x - v_k \rangle$ is Bregman divergence

Theorem. Set $d(x) = \frac{1}{3}\|x\|^3$, $A_k = \frac{k^3}{L_2}$, $\gamma_k = 1 - \frac{A_k}{A_{k+1}}$, then

$\tilde{\mathcal{O}}(1)$ steps of the basic method $\Rightarrow$ $\mathcal{O}(1/k^3)$-rate.

Plan of the Talk

I. Intro

- Gradient methods
- Newton's method, classical approach

II. Modern techniques

- Cubic regularization
- Contracting-point methods
- Acceleration

III. Conclusions

Conclusions

1. To globalize the Newton's method we need to do regularization
   - Cubic Newton — explicit regularizer, $\|\cdot\|^3$
   - Contracting Newton — implicit regularization by contraction
   - Acceleration: prox-point and contracting-point together

2. We can solve the composite problems

$$\min_x \Big\{ F(x) \;\; := \;\; f(x) + \psi(x) \Big\}$$

   and

$$\min_x \Big\{ F(x) \;\; := \;\; \phi(f(x)) \Big\}$$

   where f is a *smooth* component

3. In practice: we can use stochastic approximations and inexact methods with first-order subsolvers, preserving the global rates

Open Questions

- Efficient implementation: parallel and distributed systems

  Note: computation of $\nabla^2 f(x)h$ for any $h \in \mathbb{R}^n$ has the same cost as for $\nabla f(x)$

- Worst-case complexities can be too pessimistic

  $\Rightarrow$ benefits of using two-level schemes

- Theory of the Damped Newton: $x^+ = x - \alpha \nabla^2 f(x)^{-1} \nabla f(x)$

  Hint: different problem classes
  - **Self-Concordant Functions** [Nesterov-Nemirovski, 1994; Dvurechensky-Nesterov, 2018]
  - **Generalized S.C.** [Bach, 2010; Sun-Tran–Dinh, 2019]
  - **Hessian Stability** [Karimireddy-Stich-Jaggi, 2018]

- Quasi-Newton methods

- Nonconvex problems