

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

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

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

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

The Gradient Method

Iterate, for $k \geq 0$:

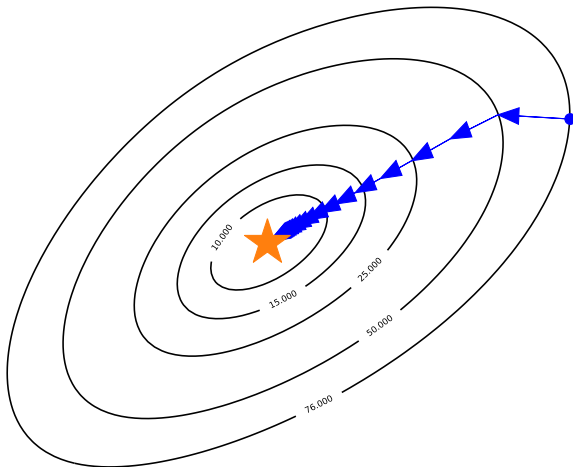
$$x_{k+1} := x_k - \alpha_k \nabla f(x_k), \quad \text{for some } \alpha_k > 0$$

[Cauchy, 1847]

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

The Gradient Method: Trajectory

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



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

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:

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

- ▶ 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

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

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

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

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

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

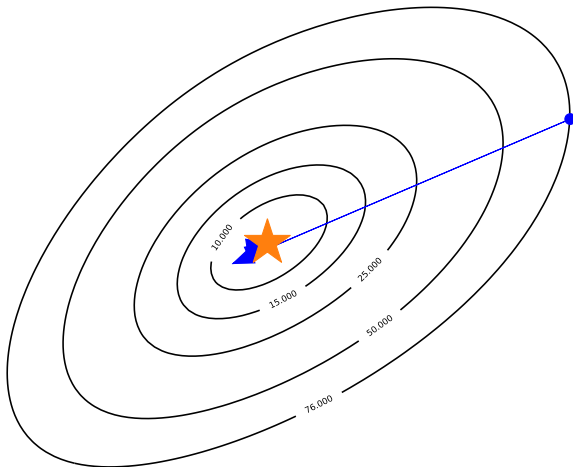
Newton's Method. Iterate, for $k \geq 0$:

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

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

Newton's Method: Trajectory

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



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

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

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

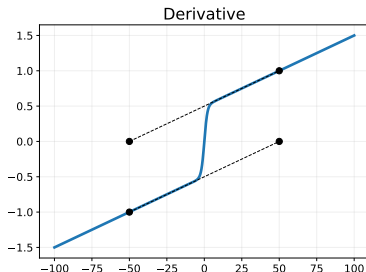
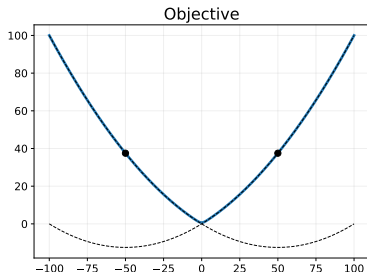
iterations to find an ε -solution, **when in the neighbourhood of the optimum**

- ▶ Global convergence — ?

Newton's Method: Global Behaviour

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

- 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]

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

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

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

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

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

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

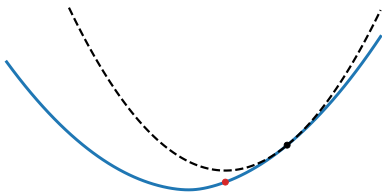
The Gradient Method: a Modern View

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

Assumption: gradient is Lipschitz continuous

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

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



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

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

Cubic Regularization of Newton's Method

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

New assumption: Hessian is Lipschitz continuous

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

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

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

Newton method with **cubic regularization**:

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

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

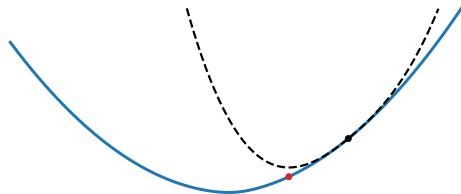
[Nesterov-Polyak, 2006]

Adaptive Cubic Newton

Iterate, for $k \geq 0$:

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

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



$$H = 0.1$$

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

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

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

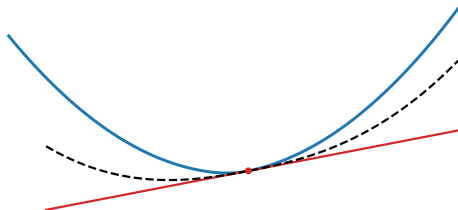
2: Make the Cubic step $x_{k+1} := T_{H_k 2^{i_k}}(x_k)$.

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

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



- ▶ **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*, $\forall x, y$:

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

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*:

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

Cubic Newton, with $H := L_2$:

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

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

Proof:

$$\begin{aligned} & \gamma f^* + (1 - \gamma)f(x_k) \\ & \stackrel{\text{lower}}{\geq} f(x_k) + \gamma \langle \nabla f(x_k), x^* - x_k \rangle \\ & \quad + \frac{\gamma^2}{2} \langle \nabla^2 f(x_k)(x^* - x_k), x^* - x_k \rangle + \frac{\gamma^3 L_2}{6} \|x^* - x_k\|^3 \end{aligned}$$

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

Proof: continue

We have

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

Hence,

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



Universal Complexity

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

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

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

Condition number of degree ν :

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

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

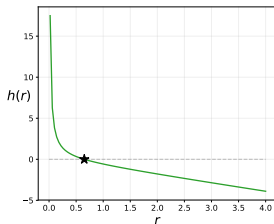
$$\mathcal{O}\left(\inf_{\nu \in [0,1]} [\gamma_\nu^{-1}] \log \frac{f(x_0) - f^*}{\varepsilon}\right)$$

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

How to Compute Iteration?

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

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]:

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

- ▶ One matrix inversion; fast global rates

Stochastic Subspace Cubic Newton

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

where n is Huge

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

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

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

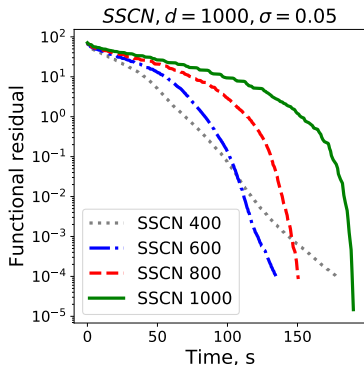
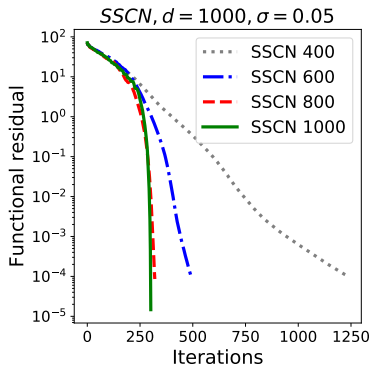
[D-Richtárik, 2018; Hanzely-D-Richtárik-Nesterov, 2020]

Theorem. The method converges globally, after k iterations:

$$\mathbb{E}f(x_k) - f^* \leq \mathcal{O}\left(\frac{n-\tau}{\tau} \cdot \frac{1}{k} + \left(\frac{n}{\tau}\right)^2 \cdot \frac{1}{k^2}\right)$$

Experiment

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



Plan of the Talk

I. Intro

- Gradient methods
- Newton's method, classical approach

II. Modern techniques

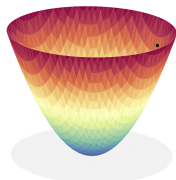
- Cubic regularization
- Contracting-point methods
- Acceleration

III. Conclusions

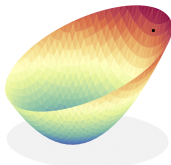
Contraction Technique

Let us consider **contraction** of the objective:

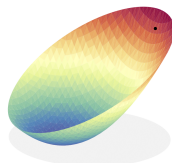
$$g(x) := f(\gamma x + (1 - \gamma)\bar{x}), \quad \gamma \in [0, 1].$$



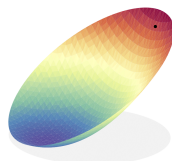
$\gamma = 1$



$\gamma = 0.8$



$\gamma = 0.7$



$\gamma = 0.6$

Note:

$$\nabla g(x) = \gamma \nabla f(\dots),$$

$$\nabla^2 g(x) = \gamma^2 \nabla^2 f(\dots),$$

...

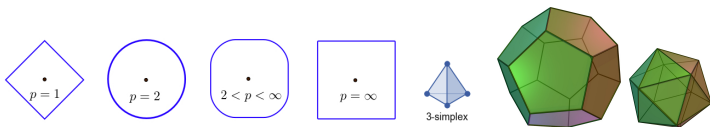
Smoothness properties of $g(\cdot)$ are better than that of $f(\cdot)$

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

Contracting-Point Methods

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

where $Q \subset \mathbb{R}^n$ is a **bounded** convex set, e.g.:



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

$$v_{k+1} \approx \operatorname{argmin}_{v \in Q} f(\gamma_k v + (1 - \gamma_k)x_k),$$

$$x_{k+1} = \gamma_k v_{k+1} + (1 - \gamma_k)x_k$$

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

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

Iterate, for $k \geq 0$:

$$x_{k+1} := \operatorname{argmin}_y \left\{ \Omega_2(x_k; y) : y \in x_k + \gamma_k(Q - x_k) \right\}$$

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 asymmetric **trust region**

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

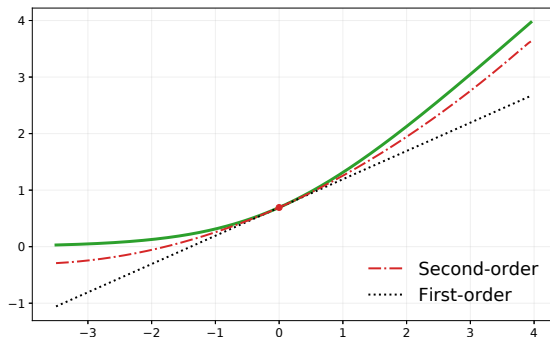
Contracting Newton: Interpretation

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

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

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

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



Cubic Newton vs. Contracting Newton

Cubic Newton, H_k	Contracting Newton, γ_k
global <i>upper</i> approximation	global <i>lower</i> 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: $\mathcal{O}(\omega \log \frac{f(x_0) - f^*}{\varepsilon})$; $\gamma_k = \frac{1}{1+\omega}$
- ▶ Local quadratic convergence: $\gamma_k = 1$

$$x_{k+1} = \operatorname{argmin}_y \left\{ \Omega_2(x_k; y) : y \in x_k + \gamma_k(Q - x_k) \right\}$$

How to compute the iteration?

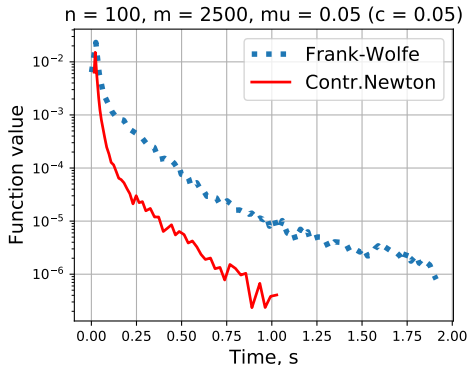
- ▶ At each step we solve the subproblem inexactly by the first-order Frank-Wolfe algorithm
- ▶ We have full control over the required accuracy

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

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

Experiment: Log-sum-exp over the Simplex

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



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**

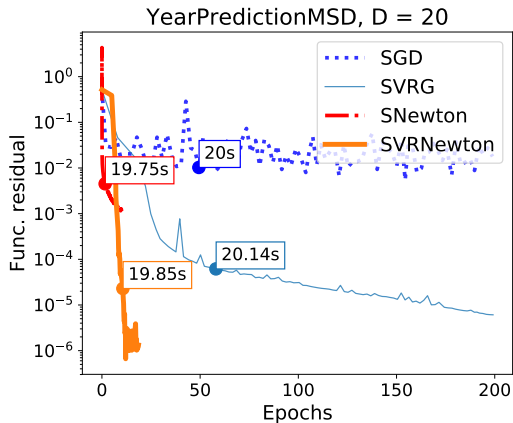
Idea: sample random batch $B \subseteq \{1, \dots, M\}$ and approximate

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

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: $\mathcal{O}(\frac{1}{\epsilon^{1/2}})$ iterations and $\mathcal{O}(\frac{1}{\epsilon^{3/2}})$ 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 second and fourth derivatives

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

- ▶ Approximation of third derivative by finite-differences

General Accelerating Scheme

► Contracting-Point Method:

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

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

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

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

$\beta_d(v_k; x) = d(x) - d(v_k) - \langle \nabla d(v_k), x - v_k \rangle$ 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**

$$\min_x \left\{ F(x) \quad := \quad f(x) + \psi(x) \right\}$$

and

$$\min_x \left\{ F(x) \quad := \quad \phi(f(x)) \right\}$$

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

⇒ 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