Super-Universal Regularized Newton Method

Nikita Doikov

Joint work with Konstantin Mishchenko and Yurii Nesterov

September 27, 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 method
- Newton's method, classical approach

II. Modern techniques

- Cubic regularization and Tensor methods
- Gradient regularization
- Super-universality

III. Experiments and 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 $k \geq 0$:

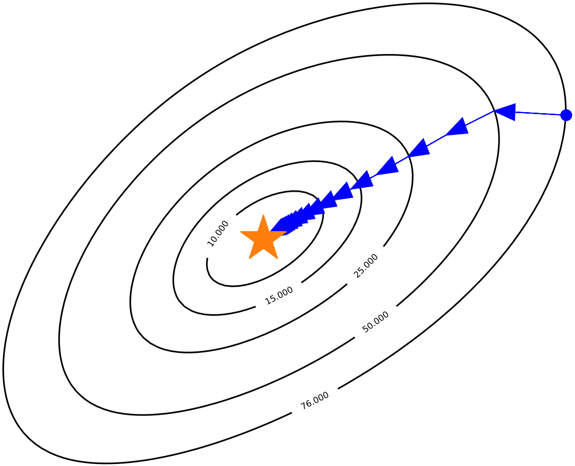
$$\boxed{x_{k+1} \;:=\; x_k - \alpha_k \nabla f(x_k), \quad \text{for some} \quad \alpha_k > 0}$$

[Cauchy, 1847]

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

The Gradient Method: Trajectory

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



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

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

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

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

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

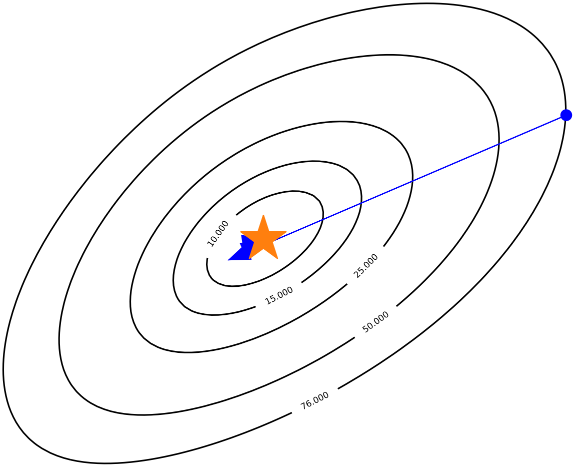
Newton's Method. Iterate, for $k \geq 0$:

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

[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:

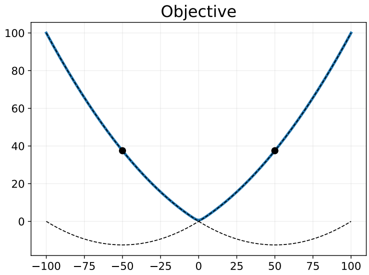
$$\mathcal{O}(\log \log \tfrac{1}{\varepsilon})$$

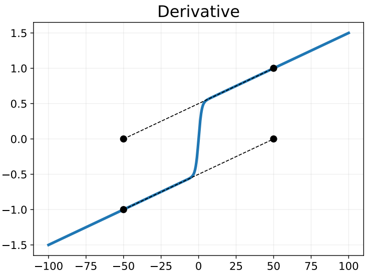
  iterations to find an ε-solution, **when in the neighbourhood of the optimum**
- Global convergence — ?

Newton's Method: Global Behaviour

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

- 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]

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

- **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 method
- Newton's method, classical approach

II. Modern techniques

- Cubic regularization and Tensor methods
- Gradient regularization
- Super-universality

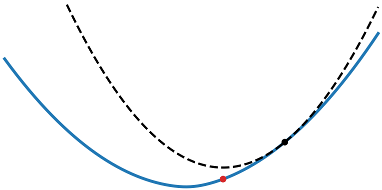
III. Experiments and conclusions

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**:

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

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

New assumption: Hessian is Lipschitz continuous

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

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

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

Newton method with cubic regularization:

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

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

Fix $p \geq 1$. Taylor's approximation:

$$f(y) \approx \Omega_p(x;y) \stackrel{\text{def}}{=} f(x) + \sum_{i=1}^{p} \tfrac{1}{i!} D^i f(x)[y-x]^i$$

Assume that p-th derivative is Lipschitz continuous:

$$\|D^p f(x) - D^p f(y)\| \leq L_p \|x-y\|, \quad \forall x, y$$

- Then we have a global bound for the approximation error:
$$|f(y) - \Omega_p(f,x;y)| \leq \tfrac{L_p}{(p+1)!}\|y-x\|^{p+1}, \quad \forall x, y$$

Basic Tensor Method of order $p \geq 1$:

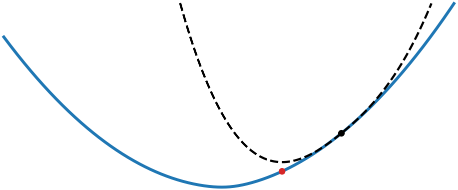
$$x_{k+1} = \operatorname*{argmin}_y \Big\{ \Omega_p(f,x_k;y) + \tfrac{H}{(p+1)!}\|y-x_k\|^{p+1} \Big\}$$

$p = 1$: Gradient Method; $p = 2$: Cubic Newton

$\boxed{\min_{x \in \mathbb{R}^n} f(x)}$ **Basic Tensor Method of order** $p \geq 1$:

$$x_{k+1} \quad = \quad \operatorname*{argmin}_y \Big\{ \Omega_p(x_k; y) + \tfrac{H}{(p+1)!} \|y - x_k\|^{p+1} \Big\}$$

$H \geq 0$ is a regularization parameter



H = 0.1

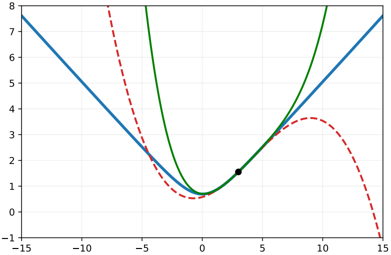
Theorem [Baes, 2009]:

Let $H := L_p \quad \Rightarrow \quad$ global rate $\quad f(x_k) - f^* \leq O\big(1/k^p\big)$.

- How to solve subproblem?

Convex Tensor Model

Note: $\Omega_p(x; y)$ is **nonconvex** for $p \geq 3$.



Theorem [Nesterov, 2018]: Let $f(\cdot)$ be convex and $H \geq pL_p$. Then

$$M(y) \quad := \quad \Omega_p(x; y) + \frac{H}{(p+1)!}\|y - x\|^{p+1}$$

is convex in y.

- **For $p = 3$:** efficient implementation using only second-order oracle. The cost is $\mathcal{O}(n^3)$.

(*Observation:* Third derivative of convex function is weak)

Plan of the Talk

I. Intro

- Gradient method
- Newton's method, classical approach

II. Modern techniques

- Cubic regularization and Tensor methods
- Gradient regularization
- Super-universality

III. Experiments and conclusions

Summary and Issues

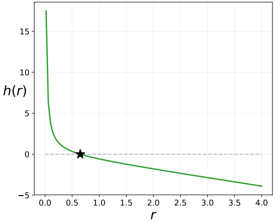
Family of methods

» Which problem class to choose?

» How to solve the subproblem.

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

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 — $\mathcal{O}(n^3)$

Cubic step: $$x^+ \; = \; x \; - \; \Big(\nabla^2 f(x) + \tfrac{Hr^+}{2} I \Big)^{-1} \nabla f(x)$$

- f is convex $\quad \Rightarrow \quad \nabla^2 f(x) \succeq 0$

Then,

$$\begin{aligned} r^+ \; &= \; \|x^+ - x\| \; = \; \Big\| \Big(\nabla^2 f(x) + \tfrac{Hr^+}{2} I \Big)^{-1} \nabla f(x) \Big\| \\ &\leq \; \tfrac{2}{Hr^+} \|\nabla f(x)\| \end{aligned}$$

Idea: Substitute for r^+ in Cubic step its approximation:

$$r^+ \; \approx \; \bar{r} \; \stackrel{\mathrm{def}}{=} \; \sqrt{\tfrac{2}{H} \|\nabla f(x)\|}$$

Gradient regularization

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

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

- One matrix inversion; the rate of Cubic Newton for $H := L_2$

Let $p \in \{2,3\}$. Fix $\nu \in [0,1]$ and define

$$L_{p,\nu} \stackrel{\mathrm{def}}{=} \sup_{x \neq y} \frac{\|D^p f(x) - D^p f(y)\|}{\|x-y\|^{\nu}}$$

$L_{p,\nu}$ is log-convex function of ν: for any $0 \leq \nu_1 \leq \nu_2 \leq 1$ we have

$$L_{p,\nu} \leq \left[L_{p,\nu_1}\right]^{\frac{\nu_2-\nu}{\nu_2-\nu_1}} \left[L_{p,\nu_2}\right]^{\frac{\nu-\nu_1}{\nu_2-\nu_1}} \qquad \forall \nu \in [\nu_1, \nu_2].$$

Define M_q, for $2 \leq q \leq 4$:

$$M_{2+\nu} \stackrel{\mathrm{def}}{=} L_{2,\nu}, \qquad \nu \in [0,1),$$

$$M_{3+\nu} \stackrel{\mathrm{def}}{=} L_{3,\nu}, \qquad \nu \in [0,1].$$

Main Assumption: $\boxed{\inf\limits_{2 \leq q \leq 4} M_q < +\infty}$.

Fix $q \in [2, 4]$. Choose $M_q > 0$.

Iteration, $k \geq 0$**:**

$$x_{k+1} \;\; = \;\; x_k - \Big(\nabla^2 f(x_k) + \lambda_k I \Big)^{-1} \nabla f(x_k),$$

with $\lambda_k := (6 M_q \|\nabla f(x_k)\|^{q-2})^{\frac{1}{q-1}}$.

Theorem. Global convergence rate:

$$f(x_k) - f^* \;\; \leq \;\; 6 M_q D^q \Big(\tfrac{32(q-1)}{k} \Big)^{q-1} \; + \; \|\nabla f(x_0)\| D \exp\big(-\tfrac{k}{4}\big)$$

where D is the diameter of the initial sublevel set.

Note: $\|\nabla f(x_0)\| D \exp\big(-\frac{k}{4}\big) \leq \varepsilon$ for $k \geq 4 \ln \frac{\|\nabla f(x_0)\| D}{\varepsilon}$.

Plan of the Talk

I. Intro
  - Gradient method
  - Newton's method, classical approach

II. Modern techniques
  - Cubic regularization and Tensor methods
  - Gradient regularization
  - Super-universality

III. Experiments and conclusions

Which problem class to choose?

Global rate: $f(x_k) - f^* \leq O\left(\frac{M_q D^q}{k^{q-1}}\right), \quad 2 \leq q \leq 4.$

$q = 2$: **Bounded variation of the Hessian**

$$\Rightarrow \qquad f(y) \;\leq\; \Omega_2(x; y) + \frac{M_2}{2}\|y - x\|^2, \quad \forall x, y$$

$q = 3$: **Lipschitz continuity of the Hessian**

$$\Rightarrow \qquad f(y) \;\leq\; \Omega_2(x; y) + \frac{M_3}{6}\|y - x\|^3, \quad \forall x, y$$

$q = 4$: **Lipschitz continuity of third derivative**

$$\Rightarrow \qquad f(y) \;\leq\; \Omega_3(x; y) + \frac{M_4}{24}\|y - x\|^4, \quad \forall x, y$$

Our objective can belong to several problem classes simultaneously!

Main Lemma

Consider the step $x^+ = x - \left(\nabla^2 f(x) + \lambda I\right)^{-1} \nabla f(x)$
with

$$\lambda \;\; := \;\; H \|\nabla f(x)\|^{\alpha}, \quad 0 \leq \alpha \leq 1$$

Lemma. Let $\frac{q-2}{q-1} \leq \alpha \leq 1$, and $H \geq \left(6M_q\right)^{\frac{1}{q-1}} \left(\frac{1}{\|\nabla f(x)\|}\right)^{\alpha - \frac{q-2}{q-1}}$.

Then

$$\langle \nabla f(x^+), x - x^+ \rangle \;\; \geq \;\; \tfrac{1}{4\lambda} \|\nabla f(x^+)\|^2.$$

Initialization. Choose $x_0 \in \mathbb{R}^n$. Fix arbitrary $\alpha \in \left[\frac{2}{3}, 1\right]$, $H_0 > 0$.

Iteration $k \geq 0$:

Find smallest $j_k \geq 0$ s.t. for $\lambda_k := 4^{j_k} H_k \|\nabla f(x_k)\|^\alpha$ and

$$x^+ \;=\; x_k - \left(\nabla^2 f(x_k) + \lambda_k I\right)^{-1} \nabla f(x_k)$$

it holds

$$\langle \nabla f(x^+), x_k - x^+ \rangle \;\geq\; \tfrac{1}{4\lambda_k} \|\nabla f(x^+)\|^2.$$

Set $x_{k+1} = x^+$ and $H_{k+1} = \frac{4^{j_k} H_k}{4}$.

Theorem. The method is well defined. We have

$$f(x_k) - f^* \;\; \leq \;\; 6 M_q D^q \left(\tfrac{32(q-1)}{k} \right)^{q-1} \; + \; \|\nabla f(x_0)\| D \exp\left(-\tfrac{k}{4}\right)$$

- The average number of adaptive steps per iterations is two.

Initial sublevel set:

$$\mathcal{F}_0 \stackrel{\text{def}}{=} \Big\{x \; : \; f(x) \leq f(x_0)\Big\}.$$

Diameter

$$D \stackrel{\text{def}}{=} \sup_{x,y \in \mathcal{F}_0} \|x - y\|.$$

Symmetrized Bregman Divergence:

$$\beta_f(x,y) \stackrel{\text{def}}{=} \langle \nabla f(x) - \nabla f(y), x - y \rangle > 0.$$

and normalization:

$$V_f \stackrel{\text{def}}{=} \sup_{x,y \in \mathcal{F}_0} \beta_f(x,y), \qquad \xi_f(x,y) \stackrel{\text{def}}{=} \tfrac{1}{V_F} \beta_f(x,y).$$

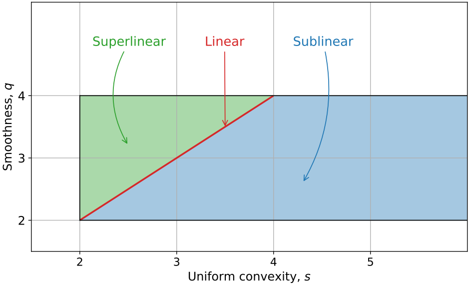
Relative s-size:

$$D_s \stackrel{\text{def}}{=} \sup_{x \neq y} \|x - y\| \cdot \xi_f(x,y)^{-1/s}, \qquad s \geq 2.$$

Table with Complexities

- Level of smoothness $2 \leq q \leq 4$ is fixed.

$2 \leq s < q$	$s = q$	$q < s < \infty$	$s = \infty$
$\left(M_q \frac{D_s^s D^{q-s}}{V_F}\right)^{\frac{1}{q-1}} + \ln\ln\frac{1}{\varepsilon}$	$\left(M_q \frac{D_q^q}{V_F}\right)^{\frac{1}{q-1}} \ln\frac{1}{\varepsilon}$	$\left(M_q \frac{D_s^q}{(V_F^q \varepsilon^{s-q})^{1/s}}\right)^{\frac{1}{q-1}}$	$\left(M_q \frac{D^q}{\varepsilon}\right)^{\frac{1}{q-1}}$



Plan of the Talk

I. Intro

- Gradient method
- Newton's method, classical approach

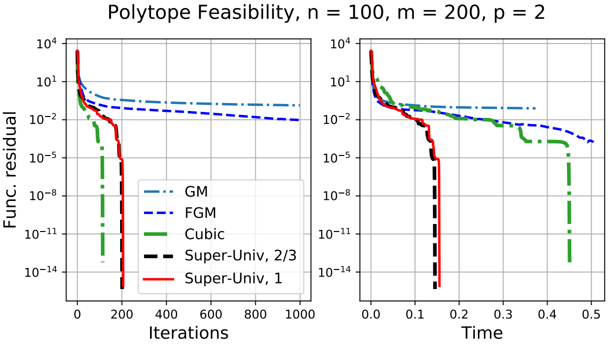
II. Modern techniques

- Cubic regularization and Tensor methods
- Gradient regularization
- Super-universality

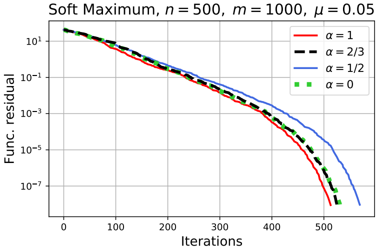
III. Experiments and conclusions

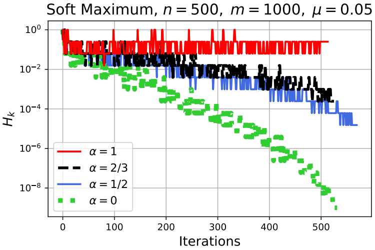
$$\min_{x \in \mathbb{R}^n} \Big[\; f(x) \quad := \quad \sum_{i=1}^{m} \big(\langle a_i, x \rangle - b_i \big)_+^p \; \Big],$$

where $(t)_+ \stackrel{\rm def}{=} \max\{0, t\}$



$$\min_x f(x) \quad := \quad \mu \ln\Big(\sum_{i=1}^{m} \exp\big(\tfrac{\langle a_i, x\rangle - b_i}{\mu}\big) \Big) \quad \approx \quad \max_{1\le i\le m} \big[\langle a_i, x\rangle - b_i\big].$$





Conclusions

1. To globalize the Newton's method we need to do regularization
   - Cubic Newton — explicit regularizer, $\| \cdot \|^3$
   - We can reduce the power to $\| \cdot \|^2$ by **Gradient Regularization**

2. Method $\leftrightarrow$ Problem class

3. **Super-universal** methods: adjust automatically to the best problem class
   - Achieved by using an adaptive search

4. We can solve Composite Problems

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

where ψ is a nonsmooth part (e.g. ℓ_1-regularizer)

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 $\nabla f(x)$

- 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