MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Cubic Regularization is the Key! The First Accelerated Quasi-Newton Method with a Global Convergence Rate of $O(k^{-2})$ for Convex Functions

Dmitry Kamzolov
Research Associate, MBZUAI, Abu Dhabi, UAE

Our Team

Klea Ziu

Artem Agafonov

Martin Takáč

https://arxiv.org/pdf/2302.04987.pdf

Problem formulation

Convex optimization problem

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

$f(x)$ is convex function with Lipschitz-continuous gradient and Hessian with constants L_1, L_2

Problem formulation

Convex optimization problem

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

$f(x)$ is convex function with Lipschitz-continuous gradient and Hessian with constants L_1, L_2

Lipschitz conditions

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L_1 \|x - y\|$$

$$\|\nabla^2 f(x) - \nabla^2 f(y)\|_{op} \leq L_2 \|x - y\|$$

Model

Taylor approximation

$$\Omega_1(f, x; y) = f(x) + \langle \nabla f(x), y - x \rangle, y \in E$$

$$\Omega_2(f, x; y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \left\langle \nabla^2 f(x)(y - x), y - x \right\rangle, y \in E$$

Model

Taylor approximation

$$\Omega_1(f, x; y) = f(x) + \langle \nabla f(x), y - x \rangle, y \in E$$

$$\Omega_2(f, x; y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \left\langle \nabla^2 f(x)(y - x), y - x \right\rangle, y \in E$$

Upper and lower bound

$$|f(y) - \Omega_p(f, x; y)| \leq \frac{L_p}{(p+1)!} \|y - x\|^{p+1}, \quad p \in \{1, 2\}$$

Model

Taylor approximation

$$\Omega_1(f, x; y) = f(x) + \langle \nabla f(x), y - x \rangle, y \in E$$

$$\Omega_2(f, x; y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \left\langle \nabla^2 f(x)(y - x), y - x \right\rangle, y \in E$$

Upper and lower bound

$$|f(y) - \Omega_p(f, x; y)| \leq \frac{L_p}{(p+1)!} \|y - x\|^{p+1}, \quad p \in \{1, 2\}$$

Corollary

$$f(y) \leq \Omega_p(f, x; y) + \frac{L_p}{(p+1)!} \|y - x\|^{p+1}, \quad p \in \{1, 2\}$$

First-order Method

Gradient Method

$$x_{t+1} = \underset{y \in \mathbb{E}}{\operatorname{argmin}} \left\{ f(x_t) + \langle \nabla f(x_t), y - x_t \rangle + \frac{L_1}{2} \|y - x_t\|^2 \right\}$$

First-order Method

Gradient Method

$$x_{t+1} = \operatorname*{argmin}_{y \in \mathbb{E}} \left\{ f(x_t) + \langle \nabla f(x_t), y - x_t \rangle + \frac{L_1}{2} \|y - x_t\|^2 \right\}$$

Step

$$\nabla f(x_t) + L_1(x_{t+1} - x_t) = 0$$

First-order Method

Gradient Method

$$x_{t+1} = \operatorname*{argmin}_{y \in \mathbb{E}} \left\{ f(x_t) + \langle \nabla f(x_t), y - x_t \rangle + \frac{L_1}{2} \|y - x_t\|^2 \right\}$$

Step

$$\nabla f(x_t) + L_1(x_{t+1} - x_t) = 0$$

Explicit form

$$x_{t+1} = x_t - \tfrac{1}{L_1} \nabla f(x_t)$$

First-order Method

Gradient Method

$$x_{t+1} = \operatorname*{argmin}_{y \in \mathbb{E}} \left\{ f(x_t) + \langle \nabla f(x_t), y - x_t \rangle + \frac{L_1}{2} \|y - x_t\|^2 \right\}$$

Step

$$\nabla f(x_t) + L_1 (x_{t+1} - x_t) = 0$$

Explicit form

$$x_{t+1} = x_t - \tfrac{1}{L_1} \nabla f(x_t)$$

Global Convergence

Basic: $T = O\left(\frac{L_1 R^2}{\varepsilon} \right)^{1}$; Optimal: $T = O\left(\frac{L_1 R^2}{\varepsilon} \right)^{1/2}$

Second-order Method

Second-order method

$$x_{t+1} = \operatorname*{argmin}_{y \in \mathbb{E}} \left\{ f(x_t) + \langle \nabla f(x_t), y - x_t \rangle + \frac{1}{2} \left\langle \nabla^2 f(x_t)(y - x_t), y - x_t \right\rangle \right\}$$

Second-order Method

Second-order method

$$x_{t+1} = \underset{y \in \mathbb{E}}{\operatorname{argmin}} \left\{ f(x_t) + \langle \nabla f(x_t), y - x_t \rangle + \frac{1}{2} \left\langle \nabla^2 f(x_t)(y - x_t), y - x_t \right\rangle \right\}$$

Step

$$\nabla f(x_t) + \nabla^2 f(x_t)(y - x_t) = 0$$

Second-order Method

Second-order method

$$x_{t+1} = \operatorname*{argmin}_{y \in \mathbb{E}} \left\{ f(x_t) + \langle \nabla f(x_t), y - x_t \rangle + \frac{1}{2} \left\langle \nabla^2 f(x_t)(y - x_t), y - x_t \right\rangle \right\}$$

Step

$$\nabla f(x_t) + \nabla^2 f(x_t)(y - x_t) = 0$$

Newton Method [1948, L. Kantorovich]

$$x_{t+1} = x_t - h_t \left[\nabla^2 f(x_t) \right]^{-1} \nabla f(x_t)$$

Second-order Method

Second-order method

$$x_{t+1} = \operatorname*{argmin}_{y \in \mathbb{E}} \left\{ f(x_t) + \langle \nabla f(x_t), y - x_t \rangle + \frac{1}{2} \left\langle \nabla^2 f(x_t)(y - x_t), y - x_t \right\rangle \right\}$$

Step

$$\nabla f(x_t) + \nabla^2 f(x_t)(y - x_t) = 0$$

Newton Method [1948, L. Kantorovich]

$$x_{t+1} = x_t - h_t \left[\nabla^2 f(x_t) \right]^{-1} \nabla f(x_t)$$

Global Convergence

$$T = O\left(\frac{L_2 R^3}{\varepsilon} \right)^2$$

Inexact Second-order Methods

Inexact Newton Method

$$x_{t+1} = x_t - h_t B_t^{-1} \nabla f(x_t)$$

Inexact Second-order Methods

Inexact Newton Method

$$x_{t+1} = x_t - h_t {B_t}^{-1} \nabla f(x_t)$$

BFGS [1970, C. Broyden, R. Fletcher, D. Goldfarb, D. Shanno]

$$B_{t+1} = B_t + \frac{y_t y_t^T}{y_t^T s_t} - \frac{B_t s_t s_t^T B_t^T}{s_t^T B_t s_t}$$

BFGS

$$s_t = x_{t+1} - x_t$$
$$y_t = \nabla f(x_{t+1}) - \nabla f(x_t)$$

Global Convergence

$$T = O\left(\frac{L_1 R^2}{\varepsilon}\right)^3$$

Cubic Regularized Newton Method

Cubic Regularization [2006, Yu. Nesterov and Boris T Polyak]

$$x_{t+1} = \underset{y \in \mathbb{E}}{\operatorname{argmin}} \left\{ f(x_t) + \langle \nabla f(x_t), y - x_t \rangle + \frac{1}{2} \left\langle \nabla^2 f(x_t)(y - x_t), y - x_t \right\rangle + \frac{M}{6} \|y - x_t\|^3 \right\}$$

Cubic Regularized Newton Method

Cubic Regularization [2006, Yu. Nesterov and Boris T Polyak]

$$x_{t+1} = \operatorname*{argmin}_{y \in \mathbb{E}} \left\{ f(x_t) + \langle \nabla f(x_t), y - x_t \rangle + \frac{1}{2} \left\langle \nabla^2 f(x_t)(y - x_t), y - x_t \right\rangle + \frac{M}{6} \|y - x_t\|^3 \right\}$$

Global Convergence

Basic: $T = O\left(\frac{L_2 R^3}{\varepsilon}\right)^{1/2}$; Fast: $O\left(\frac{L_2 R^3}{\varepsilon}\right)^{1/3}$; Optimal: $O\left(\frac{L_2 R^3}{\varepsilon}\right)^{2/7}$

Acceleration with 3-rd order smoothness

Basic: $T = O\left(\frac{L_3 R^4}{\varepsilon}\right)^{1/3}$; Fast: $O\left(\frac{L_3 R^4}{\varepsilon}\right)^{1/4}$; Optimal: $O\left(\frac{L_3 R^4}{\varepsilon}\right)^{1/5}$

Damped Newton vs Cubic Newton

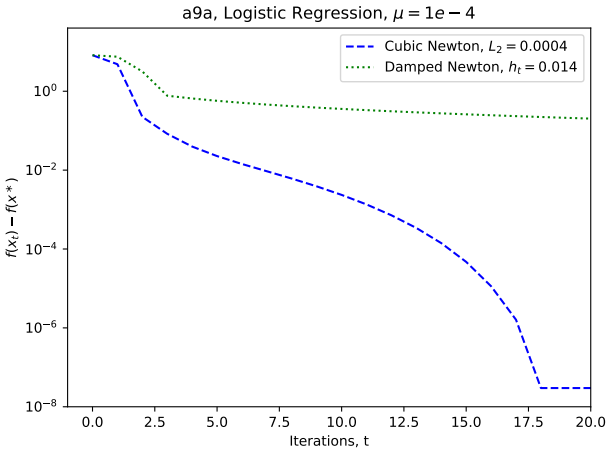

Figure: Damped Newton vs Cubic Newton for best tuned parameters from the starting point $x_0 = 3 * e$, where e is a vector of all ones.

Inexact Cubic Regularization [2017, S. Ghadimi et.al.]

Inexactness 1

$$\nabla^2 f(x_t) \preceq B_t \preceq \nabla^2 f(x_t) + \delta I$$

Inexact Cubic Regularization [2017, S. Ghadimi et.al.]

Inexactness 1

$$\nabla^2 f(x_t) \preceq B_t \preceq \nabla^2 f(x_t) + \delta I$$

Inexact Cubic Regularization

$$x_{t+1} = x_t + \underset{h \in \mathbb{E}}{\operatorname{argmin}} \left\{ f(x_t) + \langle \nabla f(x_t), h \rangle + \frac{1}{2} \langle B_t h, h \rangle + \frac{L_2}{6} \|h\|^3 \right\}$$

Inexact Cubic Regularization [2017, S. Ghadimi et.al.]

Inexactness 1

$$\nabla^2 f(x_t) \preceq B_t \preceq \nabla^2 f(x_t) + \delta I$$

Inexact Cubic Regularization

$$x_{t+1} = x_t + \operatorname*{argmin}_{h \in \mathbb{E}} \left\{ f(x_t) + \langle \nabla f(x_t), h \rangle + \frac{1}{2} \langle B_t h, h \rangle + \frac{L_2}{6} \|h\|^3 \right\}$$

Global Convergence

$$T = O(1) \max \left\{ \left(\frac{\delta R^2}{\varepsilon} \right)^1 ; \left(\frac{M R^3}{\varepsilon} \right)^{1/2} \right\}$$

Inexact Cubic Regularization [2020, A. Agafonov et.al]

Inexactness 2

$$\nabla^2 f(x_t) - \delta I \preceq B_t \preceq \nabla^2 f(x_t) + \delta I$$

Inexact Cubic Regularization [2020, A. Agafonov et.al]

Inexactness 2

$$\nabla^2 f(x_t) - \delta I \preceq B_t \preceq \nabla^2 f(x_t) + \delta I$$

Inexact Cubic Regularization

$$x_{t+1} = x_t + \operatorname*{argmin}_{h \in \mathbb{E}} \left\{ f(x_t) + \langle \nabla f(x_t), h \rangle + \frac{1}{2} \langle B_t h, h \rangle + \frac{M}{6} \|h\|^3 + \frac{\delta}{2} \|h\|^2 \right\}$$

Inexact Cubic Regularization [2020, A. Agafonov et.al]

Inexactness 2

$$\nabla^2 f(x_t) - \delta I \preceq B_t \preceq \nabla^2 f(x_t) + \delta I$$

Inexact Cubic Regularization

$$x_{t+1} = x_t + \operatorname*{argmin}_{h \in \mathbb{E}} \left\{ f(x_t) + \langle \nabla f(x_t), h \rangle + \frac{1}{2} \langle B_t h, h \rangle + \frac{M}{6} \|h\|^3 + \frac{\delta}{2} \|h\|^2 \right\}$$

Global Convergence

$$T = O(1) \max \left\{ \left(\frac{\delta R^2}{\varepsilon} \right)^1 ; \left(\frac{M R^3}{\varepsilon} \right)^{1/2} \right\}$$

Inexact Cubic Regularization, NEW

Inexactness 3

$$\|(\nabla^2 f(x_t) - B_{x_t})(x_{t+1} - x_t)\| \leq \delta_{x_t}^{x_{t+1}} \|x_{t+1} - x_t\|$$

Inexact Cubic Regularization, NEW

Inexactness 3

$$\|(\nabla^2 f(x_t) - B_{x_t})(x_{t+1} - x_t)\| \leq \delta_{x_t}^{x_{t+1}} \|x_{t+1} - x_t\|$$

Inexact Cubic Regularization

$$x_{t+1} = x_t + \operatorname*{argmin}_{h \in \mathbb{E}} \left\{ f(x_t) + \langle \nabla f(x_t), h \rangle + \frac{1}{2} \langle B_t h, h \rangle + \frac{M}{6} \|h\|^3 + \frac{\delta_t}{2} \|h\|^2 \right\}$$

Inexact Cubic Regularization, NEW

Inexactness 3

$$\|(\nabla^2 f(x_t) - B_{x_t})(x_{t+1} - x_t)\| \leq \delta_{x_t}^{x_{t+1}} \|x_{t+1} - x_t\|$$

Inexact Cubic Regularization

$$x_{t+1} = x_t + \operatorname*{argmin}_{h \in \mathbb{E}} \left\{ f(x_t) + \langle \nabla f(x_t), h \rangle + \frac{1}{2} \langle B_t h, h \rangle + \frac{M}{6} \|h\|^3 + \frac{\delta_t}{2} \|h\|^2 \right\}$$

Global Convergence

$$T = O(1) \max \left\{ \left(\frac{\delta_T R^2}{\varepsilon} \right)^1 ; \left(\frac{M R^3}{\varepsilon} \right)^{1/2} \right\}$$

Cubic L-BFGS Method

Cubic L-BFGS Method

$$x_{t+1} = x_t + \operatorname*{argmin}_{h \in \mathbb{E}} \left\{ f(x_t) + \langle \nabla f(x_t), h \rangle + \frac{1}{2} \langle B_t h, h \rangle + \frac{M}{6} \|h\|^3 + \frac{\delta_t}{2} \|h\|^2 \right\}$$

Cubic L-BFGS Method

Cubic L-BFGS Method

$$x_{t+1} = x_t + \underset{h \in \mathbb{E}}{\operatorname{argmin}} \left\{ f(x_t) + \langle \nabla f(x_t), h \rangle + \frac{1}{2} \langle B_t h, h \rangle + \frac{M}{6} \|h\|^3 + \frac{\delta_t}{2} \|h\|^2 \right\}$$

L-BFGS

$$B_t = B_t^m = B_t^{m-1} + \frac{y_m y_m^T}{y_m^T s_m} - \frac{B_k^m s_m (B_k^m s_m)^T}{s_m^T B_k^m s_m}$$

Subproblem solution

Subproblem's gradient

$$\nabla f(x_t) + (B_t + \delta_t I)h^* + \frac{M}{2}\|h^*\|h^* = 0$$

Subproblem solution

Subproblem's gradient

$$\nabla f(x_t) + (B_t + \delta_t I)h^* + \frac{M}{2}\|h^*\|h^* = 0$$

Solution/Line-search intuition

The solution of the subproblem can be formulated as

$$h^* = -\left(B_t + \delta I + \tfrac{M}{2}\|h^*\| I\right)^{-1} \nabla f(x_t)$$

Solution

$$\|h^*\| = \arg\min_{\tau \geq 0} \left\{ \left\langle \left(B_t + \delta_t I + \tfrac{L}{2}\tau I\right)^{-1} \nabla f(x_t), \nabla f(x_t) \right\rangle + \tfrac{M}{6}\tau^2 \right\}$$

Computational complexity

Sherman-Morrison-Woodbury formula for m-rank approximation

$I^d \in \mathbb{R}^{d \times d}$, $A = \lambda I^d \in \mathbb{R}^{d \times d}$, $U \in \mathbb{R}^{d \times m}$, $V \in \mathbb{R}^{m \times d}$, $c = \nabla f(x_t) \in \mathbb{R}^d$

$$c^T (A + UV)^{-1} c = c^T A^{-1} c - c^T A^{-1} U (I^m + VU)^{-1} V A^{-1} c$$

Computational complexity

Sherman-Morrison-Woodbury formula for m-rank approximation

$I^d \in \mathbb{R}^{d \times d}$, $A = \lambda I^d \in \mathbb{R}^{d \times d}$, $U \in \mathbb{R}^{d \times m}$, $V \in \mathbb{R}^{m \times d}$, $c = \nabla f(x_t) \in \mathbb{R}^d$

$$c^T (A + UV)^{-1} c = c^T A^{-1} c - c^T A^{-1} U (I^m + VU)^{-1} V A^{-1} c$$

Complexity

$VU -$ **multiplication:** $O(m^2 d)$,

$(I^m + VU)^{-1} -$ **inversion:** $O(m^3)$,

Total: $O(m^2 d)$.

Cubic complexity

SVD $- O(d^3)$

Hessian approximation

L-BFGS

$$B_t = B_t^m = B_t^{m-1} + \frac{y_m y_m^T}{y_m^T s_m} - \frac{B_k^m s_m (B_k^m s_m)^T}{s_m^T B_k^m s_m}$$

History

$$s_m = x_m - x_{m-1},$$
$$y_m = \nabla f(x_m) - \nabla f(x_{m-1})$$

Complexity

Long-term memory $- O(md)$; **Computations** $- O(1)$ **gradients**

Accuracy

$$\|B_t - \nabla^2 f(x_t)\| \leq mL_1$$

Hessian approximation

Damped L-BFGS

$$B_t = B_t^m = B_t^{m-1} + \frac{1}{m}\frac{y_m y_m^T}{y_m^T s_m} - \frac{B_k^m s_m (B_k^m s_m)^T}{s_m^T B_k^m s_m}$$

History

$$s_m = x_m - x_{m-1},$$
$$y_m = \nabla f(x_m) - \nabla f(x_{m-1})$$

Complexity

Long-term memory $- O(md)$; **Computations** $- O(1)$ **gradients**

Accuracy

$$\|B_t - \nabla^2 f(x_t)\| \leq L_1$$

Hessian approximation

L-BFGS

$$B_t = B_t^m = B_t^{m-1} + \frac{y_m y_m^T}{y_m^T s_m} - \frac{B_k^m s_m (B_k^m s_m)^T}{s_m^T B_k^m s_m}$$

Sampling

$$s_m = \textit{random vector},$$
$$y_m = \nabla^2 f(x_t) s_m.$$

Complexity

Long-term memory $- 0;$ **Computations** $- O(m)$ **gradients**

Accuracy

$$\|B_t - \nabla^2 f(x_t)\| \leq L_1$$

Cubic L-BFGS Method Convergence Theorem

Theorem

Let $f(x)$ be a convex function $f(x)$ has L_1-Lipschitz-continuous gradient and L_2-Lipschitz-continuous Hessian, B_t is an m-memory Damped L-BFGS approximation, then the convergence rate is

$$T = O(1) \max \left\{ \left(\frac{L_1 R^2}{\varepsilon} \right)^1 ; \left(\frac{M R^3}{\varepsilon} \right)^{1/2} \right\}.$$

Adaptive Accelerated Inexact Cubic Regularized Newton

Linear model

$$l(x, y) = f(y) + \langle \nabla f(y), x - y \rangle$$

Nesterov's estimating sequence

$$\psi_t(x) = \sum_{i=2}^{3} \frac{\kappa_i^t}{i} \|x - x_0\|^i + \sum_{j=0}^{t-1} \frac{\alpha_j}{A_j} l(x, x_{j+1})$$

Inexact Cubic Newton operator

$$S_{M,\delta}(x) = x + \operatorname{argmin}_{h \in \mathbb{R}^d} \left\{ f(x) + \langle \nabla f(x), h \rangle + \frac{1}{2} \langle B_x h, h \rangle + \frac{M}{6} \|h\|^3 + \frac{\delta}{2} \|h\|^2 \right\}$$

Adaptive Accelerated Inexact Cubic Regularized Newton

Algorithm 2 Adaptive Accelerated Inexact Cubic Regularized Newton

Input: $y_0 = x_0$ is starting point; constants $M \geq 2L_2$; increase multiplier γ_{inc}; starting inexactness $\delta_0 \geq 0$; non-negative non-decreasing sequences $\{\kappa_2^t\}_{t\geq 0}$, $\{\kappa_3^t\}_{t\geq 0}$, $\{\alpha_t\}_{t\geq 0}$, and $\{A_t\}_{t\geq 0}$.
for $t \geq 0$ **do**
  $v_t = (1-\alpha_t)x_t + \alpha_t y_t$,
  $x_{t+1} = S_{M,\delta_t}(v_t)$
  while $\langle \nabla f(x_{t+1}), v_t - x_{t+1}\rangle \leq \min\left\{ \frac{\|\nabla f(x_{t+1})\|_*^2}{4\delta_t}, \frac{\|\nabla f(x_{t+1})\|_*^{\frac{3}{2}}}{(3M)^{\frac{1}{2}}} \right\}$ **do**
    $\delta_t = \delta_t \gamma_{inc}$
    $x_{t+1} = S_{M,\delta_t}(v_t)$
  Compute $y_{t+1} = \arg\min_{x\in\mathbb{R}^d} \psi_{t+1}(x)$

Adaptive Accelerated Cubic L-BFGS Method Convergence Theorem

Theorem

Let $f(x)$ be a convex function $f(x)$ has L_1-Lipschitz-continuous gradient and L_2-Lipschitz-continuous Hessian, B_t is an m-memory Damped L-BFGS approximation, then the convergence rate is

$$T = O(1) \max \left\{ \left(\frac{L_1 R^2}{\varepsilon} \right)^{1/2}; \left(\frac{M R^3}{\varepsilon} \right)^{1/3} \right\}.$$

Experiments: strongly convex case

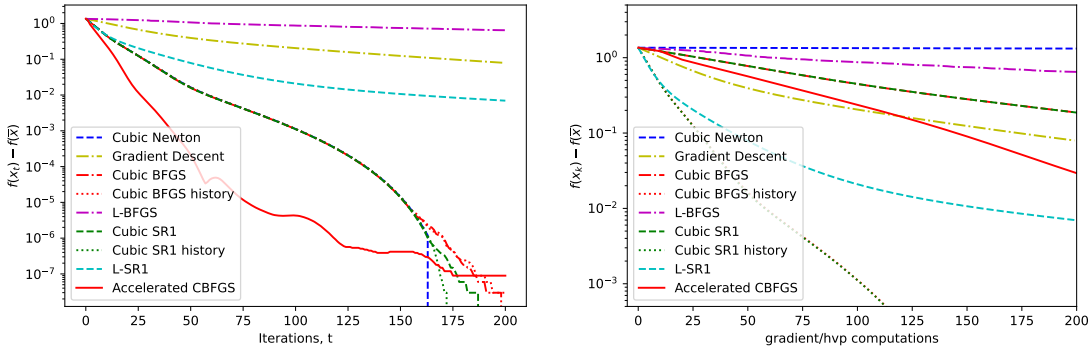

Figure: Comparison of QN methods and CRN methods for `MNIST` dataset using theoretical parameters in strongly convex case $\mu = 10^{-4}$.

Experiments: convex case

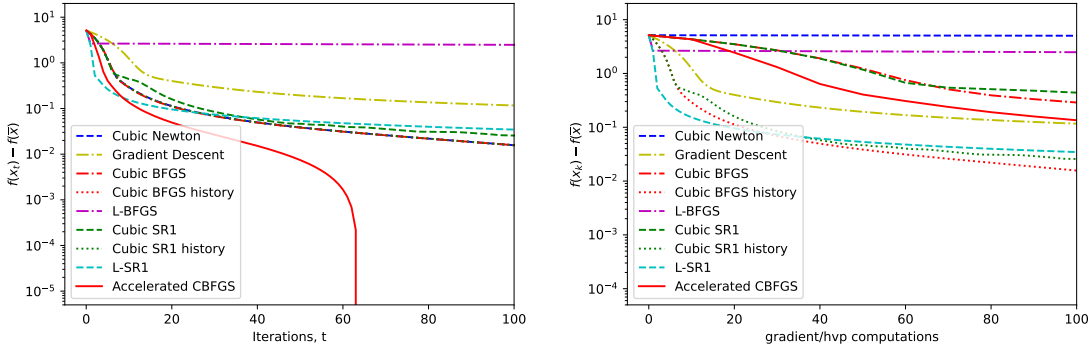

Figure: Comparison of QN methods and CRN methods for `MNIST` dataset using theoretical parameters in the convex case.

`https://arxiv.org/pdf/2302.04987.pdf`

MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

Advancing the lower bounds: An accelerated, stochastic, second-order method with optimal adaptation to inexactness

Our Team

Artem Agafonov
(MZUAI)

Alexander Gasnikov
(MIPT,IITP, HSE)

Ali Kavis
(EPFL)

Kimon Antonakopoulos
(EPFL)

Volkan Cevher
(EPFL)

Martin Takáč
(MBZUAI)

Problem formulation

Convex optimization problem

$$\min_{x \in \mathbb{R}^d} f(x) = \mathbf{E}_\xi \left[f(x, \xi) \right],$$

$f(x)$ is convex function with Lipschitz-continuous gradient and Hessian with constants L_1, L_2

Problem formulation

Convex optimization problem

$$\min_{x \in \mathbb{R}^d} f(x) = \mathbf{E}_\xi \left[f(x, \xi) \right],$$

$f(x)$ is convex function with Lipschitz-continuous gradient and Hessian with constants L_1, L_2

Unbiased stochastic gradient

For all $x \in \mathbb{R}^d$, we assume that stochastic gradients $g(x, \xi) \in \mathbb{R}^d$ satisfy

$$\mathbb{E}[g(x, \xi) \mid x] = \nabla f(x), \quad \mathbb{E}\left[\|g(x, \xi) - \nabla f(x)\|^2 \mid x \right] \leq \sigma_1^2.$$

Inexact Hessian

Unbiased stochastic gradient with bounded variance and inexact Hessian

For all $x \in \mathbb{R}^d$ stochastic gradient $g(x, \xi)$ satisfies (25). For given $x,\ y \in \mathbb{R}^d$ inexact Hessian $H(x)$ satisfies

$$\|(H(x) - \nabla^2 f(x))[y - x]\| \leq \delta_2^{x,y} \|y - x\|.$$

Inexact Hessian

Unbiased stochastic gradient with bounded variance and inexact Hessian

For all $x \in \mathbb{R}^d$ stochastic gradient $g(x, \xi)$ satisfies (25). For given $x,\ y \in \mathbb{R}^d$ inexact Hessian $H(x)$ satisfies

$$\|(H(x) - \nabla^2 f(x))[y - x]\| \leq \delta_2^{x,y} \|y - x\|.$$

Unbiased stochastic gradient with bounded variance and bounded variance stochastic Hessian

For all $x \in \mathbb{R}^d$, stochastic gradient $g(x, \xi)$ satisfies (25) and stochastic Hessian $H(x, \xi)$ satisfies

$$\mathbb{E}\left[\|H(x, \xi) - \nabla^2 f(x)\|_2^2 \mid x\right] \leq \sigma_2^2.$$

Comparison with Existing Methods

Algorithm	Inexactness	Gradient convergence	Hessian convergence	Exact convergence
Accelerated Inexact Cubic Newton (Ghadimi et al., 2017)	exact gradient δ_2-inexact Hessian [1]	✗	$O\left(\frac{\delta_2 R^2}{T^2}\right)$	$O\left(\frac{L_2 R^3}{T^3}\right)$
Accelerated Inexact Tensor Method [2] Agafonov et al. (2020)	δ_1-inexact gradient [3] δ_2-inexact Hessian [4]	$O(\delta_1 R)$	$O\left(\frac{\delta_2 R^2}{T^2}\right)$	$O\left(\frac{L_2 R^3}{T^3}\right)$
Extra-Newton (Antonakopoulos et al., 2022)	stochastic gradient (2) unbiased stochastic Hessian [5]	$O\left(\frac{\sigma_1 R}{\sqrt{T}}\right)$	$O\left(\frac{\sigma_2 R^2}{T^{3/2}}\right)$	$O\left(\frac{L_2 R^3}{T^3}\right)$
Accelerated Stochastic Second-order method [This Paper]	stochastic gradient (2) stochastic Hessian (4)	$O\left(\frac{\sigma_1 R}{\sqrt{T}}\right)$	$O\left(\frac{\sigma_2 R^2}{T^2}\right)$ [6]	$O\left(\frac{L_2 R^3}{T^3}\right)$
Lower bound [This Paper]	stochastic gradient (2) stochastic Hessian (4)	$\Omega\left(\frac{\sigma_1 R}{\sqrt{T}}\right)$	$\Omega\left(\frac{\sigma_2 R^2}{T^2}\right)$	$\Omega\left(\frac{L_2 R^3}{T^{7/2}}\right)$

Inexact Cubic Regularization [2020, A. Agafonov et.al]

Gradient Inexactness

$$\|g_t - \nabla f(x_t)\| \leq \delta_1$$
$$\|B_t - \nabla^2 f(x_t)\| \leq \delta_2$$

Inexact Cubic Regularization [2020, A. Agafonov et.al]

Gradient Inexactness

$$\|g_t - \nabla f(x_t)\| \leq \delta_1$$
$$\|B_t - \nabla^2 f(x_t)\| \leq \delta_2$$

Inexact Cubic Regularization

$$x_{t+1} = x_t + \operatorname*{argmin}_{h \in \mathbb{E}} \left\{ f(x_t) + \langle g_t, h \rangle + \frac{1}{2} \langle B_t h, h \rangle + \frac{M}{6} \|h\|^3 + \frac{\delta_2 + \delta_1 \gamma_1}{2} \|h\|^2 \right\}$$

Inexact Cubic Regularization [2020, A. Agafonov et.al]

Gradient Inexactness

$$\|g_t - \nabla f(x_t)\| \leq \delta_1$$
$$\|B_t - \nabla^2 f(x_t)\| \leq \delta_2$$

Inexact Cubic Regularization

$$x_{t+1} = x_t + \operatorname*{argmin}_{h \in \mathbb{E}} \left\{ f(x_t) + \langle g_t, h \rangle + \frac{1}{2} \langle B_t h, h \rangle + \frac{M}{6} \|h\|^3 + \frac{\delta_2 + \delta_1 \gamma_1}{2} \|h\|^2 \right\}$$

Global Convergence

$$T = O(1) \max \left\{ \delta_1 R; \left(\frac{\delta_2 R^2}{\varepsilon} \right)^1 ; \left(\frac{M R^3}{\varepsilon} \right)^{1/2} \right\}$$

Accelerated Stochastic Cubic Newton

Algorithm 1 Accelerated Stochastic Cubic Newton

1: **Input:** $y_0 = x_0$ is starting point; constants $M \geq 2L_2$; non-negative non-decreasing sequences $\{\bar{\delta}_t\}_{t\geq 0}, \{\lambda_t\}_{t\geq 0}, \{\kappa_2^t\}_{t\geq 0}, \{\kappa_3^t\}_{t\geq 0}$, and

$$\alpha_t = \frac{3}{t+3}, \quad A_t = \prod_{j=1}^{t}(1-\alpha_j), \quad A_0 = 1,$$
$$\psi_0(x) := \frac{\bar{\kappa}_2^0 + \lambda_0}{2}\|x - x_0\|^2 + \frac{\bar{\kappa}_3^0}{3}\|x - x_0\|^3.$$

2: **for** $t \geq 0$ **do**

3:
$$v_t = (1-\alpha_t)x_t + \alpha_t y_t, \quad x_{t+1} = s^{M,\bar{\delta}_t,\tau}(v_t)$$

4: Compute

$$y_{t+1} = \arg\min_{x\in\mathbb{R}^n}\left\{\psi_{t+1}(x) := \psi_t(x) + \frac{\lambda_{t+1}-\lambda_t}{2}\|x - x_0\|^2 + \sum_{i=2}^{3}\frac{\bar{\kappa}_i^{t+1} - \bar{\kappa}_i^t}{i}\|x - x_0\|^i + \frac{\alpha_t}{A_t}l(x, x_{t+1})\right\}.$$

Stochastic Inexact Cubic Operator, NEW

Stochastic Inexact Cubic Operator

$$x_{t+1} = x_t + \underset{h \in \mathbb{E}}{\operatorname{argmin}} \left\{ f(x_t) + \langle g_t, h \rangle + \frac{1}{2} \langle B_t h, h \rangle + \frac{M}{6} \|h\|^3 + \frac{\bar{\delta}_t}{2} \|h\|^2 \right\}$$

Stochastic Inexact Cubic Operator, NEW

Stochastic Inexact Cubic Operator

$$x_{t+1} = x_t + \operatorname*{argmin}_{h \in \mathbb{E}} \left\{ f(x_t) + \langle g_t, h \rangle + \frac{1}{2} \langle B_t h, h \rangle + \frac{M}{6} \|h\|^3 + \frac{\bar{\delta}_t}{2} \|h\|^2 \right\}$$

Increasing Regularizer

$$\bar{\delta}_t = 2\sigma_2 + \tfrac{\sigma_1 + \tau}{R}(t+3)^{\frac{3}{2}} \quad \bar{\kappa}_2^{t+1} = \tfrac{2\bar{\delta}_t \alpha_t^2}{A_t}, \quad \lambda_t = \tfrac{\sigma_1}{R}(t+3)^{\frac{5}{2}}$$

Convergence

$$\mathbf{E}\left[f(x_T) - f(x^*)\right] \leq O\left(\tfrac{\tau R}{\sqrt{T}} + \tfrac{\sigma_1 R}{\sqrt{T}} + \tfrac{\sigma_2 R^2}{T^2} + \tfrac{M R^3}{T^3}\right).$$

Experiments: strongly convex case

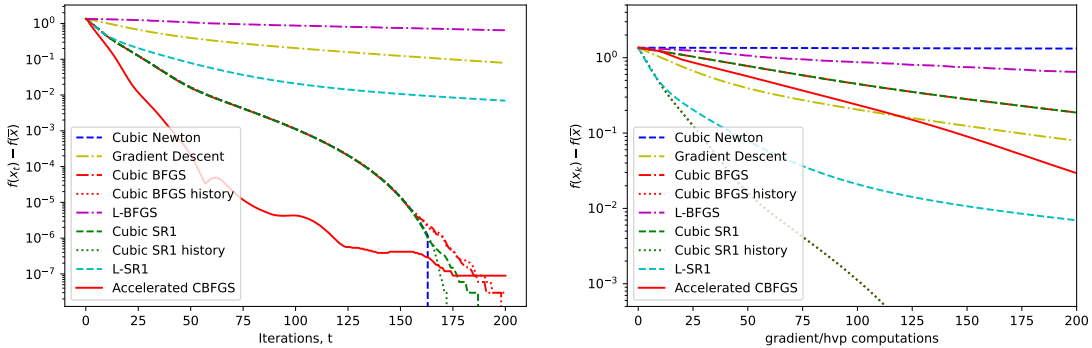

Figure: Comparison of QN methods and CRN methods for `MNIST` dataset using theoretical parameters in strongly convex case $\mu = 10^{-4}$.

Experiments: Same batch-size

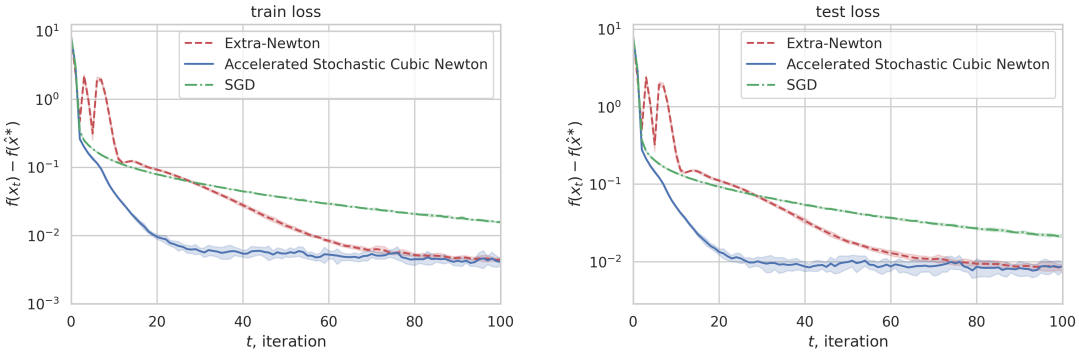

Figure: Train and Test loss where Gradient and Hessian batch sizes are 1500

Experiments: Different batch-size

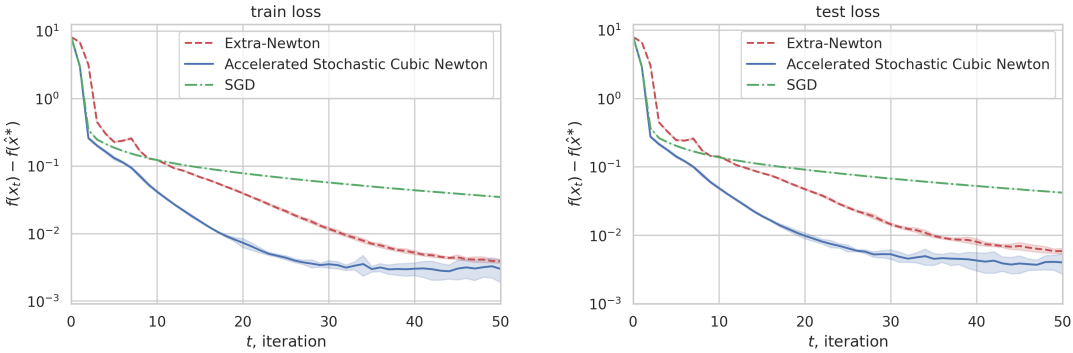

Figure: Train and Test loss where Gradient batch size is 10000, Hessian batch size is 150

MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

Contacts:

kamzolov.dmitry@mbzuai.ac.ae
kamzolov.opt@gmail.com

Mohamed bin Zayed
University of Artificial Intelligence
Masdar City
Abu Dhabi
United Arab Emirates