Randomized Block Cubic Newton Method

Nikita Doikov[1] Peter Richtárik[2, 3, 4]

[1]Higher School of Economics, Russia
[2]King Abdullah University of Science and Technology, Saudi Arabia
[3]The University of Edinburgh, United Kingdom
[4]Moscow Institute of Physics and Technology, Russia

International Conference on Machine Learning, Stockholm

July 12, 2018

Plan of the Talk

1. Review: Gradient Descent and Cubic Newton methods
2. RBCN: Randomized Block Cubic Newton
3. Application: Empirical Risk Minimization

Plan of the Talk

1. Review: Gradient Descent and Cubic Newton methods
2. RBCN: Randomized Block Cubic Newton
3. Application: Empirical Risk Minimization

Review: Classical Gradient Descent

Optimization problem:

$$\min_{x\in\mathbb{R}^N} F(x)$$

- **Main assumption:** gradient of F is Lipschitz-continuous:

$$\|\nabla F(x) - \nabla F(y)\| \;\leq\; L\|x-y\|, \quad \forall x, y \in \mathbb{R}^N.$$

- From which we get the **Global upper bound** for the function:

$$F(y) \;\leq\; F(x) + \langle \nabla F(x), y-x\rangle + \frac{L}{2}\|y-x\|^2, \quad \forall x, y \in \mathbb{R}^N.$$

- The Gradient Descent:

$$x^+ = \operatorname*{argmin}_{y\in\mathbb{R}^N}\left[F(x) + \langle \nabla F(x), y-x\rangle + \frac{L}{2}\|y-x\|^2\right] = x - \tfrac{1}{L}\nabla F(x).$$

Review: Cubic Newton

Optimization problem: $\min_{x \in \mathbb{R}^N} F(x)$.

- **New assumption:** Hessian of F is Lipschitz-continuous:

$$\|\nabla^2 F(x) - \nabla^2 F(y)\| \;\leq\; H\|x - y\|, \quad \forall x, y \in \mathbb{R}^N.$$

- Corresponding **Global upper bound** for the function:

$$Q(x; y) \;\equiv\; F(x) + \langle \nabla F(x), y - x\rangle + \frac{1}{2}\langle \nabla^2 F(x)(y - x), y - x\rangle,$$

then $\quad F(y) \;\leq\; Q(x; y) + \frac{H}{6}\|y - x\|^3, \quad \forall x, y \in \mathbb{R}^N.$

- Newton method with cubic regularization[1]:

$$x^+ \;=\; \operatorname*{argmin}_{y \in \mathbb{R}^N} \Big[Q(x; y) + \frac{H}{6}\|y - x\|^3 \Big]$$
$$\;=\; x - \Big(\nabla^2 F(x) + \frac{H\|x^+ - x\|}{2} I\Big)^{-1} \nabla F(x).$$

[1] Yurii Nesterov and Boris T Polyak. "Cubic regularization of Newton's method and its global performance". In: *Mathematical Programming* 108.1 (2006), pp. 177–205.

Optimization problem: $F^* = \min_{x \in \mathbb{R}^N} F(x)$.

- $F(x^K) - F^* \leq \varepsilon$, What is K – ?
- Let F be **convex**: $F(y) \geq F(x) + \langle \nabla F(x), y - x \rangle$.
- Iteration complexity estimates:

  $\boldsymbol{K = O\left(\frac{1}{\varepsilon}\right)}$ for GD, and $\boldsymbol{K = O\left(\frac{1}{\sqrt{\varepsilon}}\right)}$ for CN (much better).

- But, cost of one iteration: $\boldsymbol{O(N)}$ for GD and $\boldsymbol{O(N^3)}$ for CN.

N is huge for modern applications. Even $\boldsymbol{O(N)}$ is too much!

Our Motivation

Recent advances in **block coordinate** methods.

1. Paul Tseng and Sangwoon Yun. "A coordinate gradient descent method for nonsmooth separable minimization". In: *Mathematical Programming* 117.1-2 (2009), pp. 387–423
2. Peter Richtárik and Martin Takáč. "Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function". In: *Mathematical Programming* 144.1-2 (2014), pp. 1–38
3. Zheng Qu et al. "SDNA: stochastic dual Newton ascent for empirical risk minimization". In: *International Conference on Machine Learning*. 2016, pp. 1823–1832

Computationally effective steps, convergence as for full methods.

Aim: To create a second-order method with global complexity guarantees and low cost of every iteration.

Plan of the Talk

1. Review: Gradient Descent and Cubic Newton methods
2. RBCN: Randomized Block Cubic Newton
3. Application: Empirical Risk Minimization

Problem Structure

- Consider the following decomposition of $F : \mathbb{R}^N \to \mathbb{R}$:

$$F(x) \quad \equiv \quad \underbrace{\phi(x)}_{\substack{\text{twice} \\ \text{differentiable}}} \quad + \quad \underbrace{g(x)}_{\text{differentiable}}$$

- For a given space decomposition

$$\mathbb{R}^N \equiv \mathbb{R}^{N_1} \times \cdots \times \mathbb{R}^{N_n}, \quad x \equiv \left(x_{(1)}, \ldots, x_{(n)}\right), \quad x_{(i)} \in \mathbb{R}^{N_i},$$

  assume block-separable structure of ϕ:

$$\phi(x) \equiv \sum_{i=1}^{n} \phi_i(x_{(i)}).$$

- Block separability for $g : \mathbb{R}^N \to \mathbb{R}$ is not fixed.

Main Assumptions

Optimization problem: $\min_{x \in Q} F(x)$, where

$$F(x) \;\equiv\; \sum_{i=1}^{n} \phi_i(x_{(i)}) \;+\; g(x).$$

- Every $\phi_i(x_{(i)})$, $i \in \{1, \ldots, n\}$ is twice-differentiable and convex, with Lipschitz-continuous Hessian:

$$\|\nabla^2 \phi_i(x) - \nabla^2 \phi_i(y)\| \;\leq\; H_i \|x - y\|, \quad \forall x, y \in \mathbb{R}^{N_i}.$$

- $g(x)$ is differentiable, and for some fixed positive-semidefinite matrices $A \succeq G \succeq 0$ we have bounds, for all $x, y \in \mathbb{R}^N$:
  - $g(y) \leq g(x) + \langle \nabla g(x), y - x \rangle \,+\, \frac{1}{2}\langle A(y-x), y - x\rangle$,
  - $g(y) \geq g(x) + \langle \nabla g(x), y - x \rangle \,+\, \frac{1}{2}\langle G(y-x), y - x\rangle$.
- $Q \subset \mathbb{R}^N$ is a <u>simple</u> convex set.

Model of the Objective

$$\text{Objective:} \quad F(x) \;\equiv\; \sum_{i=1}^{n} \phi_i(x_{(i)}) \;+\; g(x)$$

We want to build a **model** of F.

- Fix subset of blocks: $S \subset \{1, \ldots, n\}$.
- For $y \in \mathbb{R}^N$ denote by $y_{[S]} \in \mathbb{R}^N$ a vector with zeroed $i \notin S$.
- $$M_{H,S}(x; y) \;\equiv\; F(x) \; + \langle \nabla \phi(x), y_{[S]} \rangle + \frac{1}{2} \langle \nabla^2 \phi(x) y_{[S]}, y_{[S]} \rangle$$ $$+ \; \tfrac{H}{6} \| y_{[S]} \|^3 \; + \langle \nabla g(x), y_{[S]} \rangle + \frac{1}{2} \langle A y_{[S]}, y_{[S]} \rangle.$$
- From smoothness: $F(x + y) \;\leq\; M_{H,S}(x; y), \;\; \forall x, y \in \mathbb{R}^N$ for $H \geq \sum_{i \in S} H_i$.

- Method step: $T_{H,S}(x) \equiv \mathop{\mathrm{argmin}}_{\substack{y \in \mathbb{R}^N_{[S]} \\ \text{s.t. } x+y \in Q}} M_{H,S}(x;y)$.
- <u>Algorithm:</u>

Initialization: choose $x^0 \in \mathbb{R}^N$, uniform random distribution $\hat{S}$.

Iterations: $k \geq 0$.

1: Sample $S_k \sim \hat{S}$

2: Find $H_k > 0$ such that

$$F(x^k + T_{H_k,S_k}(x^k)) \leq M_{H_k,S_k}(x^k; x^k + T_{H_k,S_k}(x^k)).$$

3: Make the step: $x^{k+1} \stackrel{\text{def}}{=} x^k + T_{H_k,S_k}(x^k)$.

We want to get: $\mathbb{P}\Big(F(x^K) - F^* \leq \varepsilon\Big) \;\geq\; 1-\rho$

$\varepsilon > 0$ is required **accuracy level**, $\rho \in (0,1)$ is **confidence level**.

Theorem 1. General conditions.

$$K \;=\; O\left(\frac{1}{\varepsilon}\cdot\frac{n}{\boldsymbol{\tau}}\cdot\left(1+\log\frac{1}{\rho}\right)\right), \quad \boldsymbol{\tau} \equiv \mathbb{E}[|\hat{S}|].$$

Theorem 2. $\boldsymbol{\sigma} \in [0,1]$ is a *condition number.* $\boldsymbol{\sigma} \geq \dfrac{\lambda_{\min}(G)}{\lambda_{\max}(A)} > 0.$

$$K \;=\; O\left(\frac{1}{\sqrt{\varepsilon}}\cdot\frac{n}{\boldsymbol{\tau}}\cdot\frac{1}{\boldsymbol{\sigma}}\cdot\left(1+\log\frac{1}{\rho}\right)\right)$$

Theorem 3. Strongly convex case: $\mu \equiv \lambda_{\min}(G) > 0.$

$$K \;=\; O\left(\log\left(\frac{1}{\varepsilon\rho}\right)\cdot\frac{n}{\boldsymbol{\tau}}\cdot\frac{1}{\boldsymbol{\sigma}}\cdot\sqrt{\max\left\{\frac{HD}{\mu}, 1\right\}}\right), \quad D \geq \|x^0 - x^*\|.$$

Plan of the Talk

1. Review: Gradient Descent and Cubic Newton methods

2. RBCN: Randomized Block Cubic Newton

3. Application: Empirical Risk Minimization

ERM problem:

$$\min_{w \in \mathbb{R}^d} \left[\, P(w) \; \equiv \; \sum_{i=1}^{n} \underbrace{\phi_i(b_i^T w)}_{\text{loss}} \; + \; \underbrace{g(w)}_{\text{regularizer}} \, \right]$$

- SVM: $\phi_i(a) = \max\{0, 1 - y_i a\}$,
- Logistic regression: $\phi_i(a) = \log(1 + \exp(-y_i a))$,
- Regression: $\phi_i(a) = (a - y_i)^2$ or $\phi_i(a) = |a - y_i|$,
- Support vector regression: $\phi_i(a) = \max\{0, |a - y_i| - \nu\}$,
- Generalized linear models.

Constrained Problem Reformulation

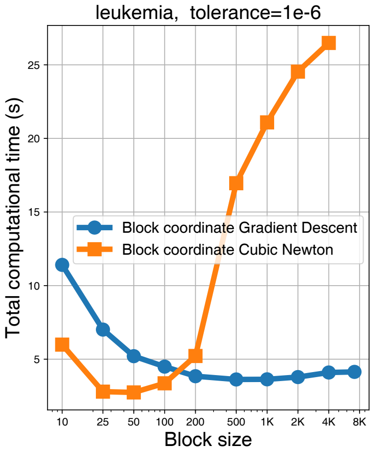
$$\min_{w\in\mathbb{R}^d} P(w) \quad = \quad \min_{w\in\mathbb{R}^d}\Big[\sum_{i=1}^n \phi_i(\underbrace{b_i^T w}_{\equiv \mu_i}) + g(w)\Big]$$

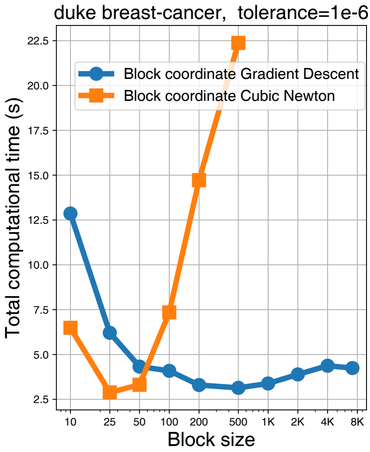
$$= \underbrace{\min_{\substack{w\in\mathbb{R}^d\\ \mu\in\mathbb{R}^n\\ b_i^T w=\mu_i}}}_{\equiv Q} \Big[\underbrace{\sum_{i=1}^n \phi_i(\mu_i)}_{\text{separable, twice differentiable}} \quad + \quad \underbrace{g(w)}_{\text{differentiable}}\Big]$$

- Approximate ϕ_i by second-order models with cubic regularization;
- Treat g as quadratic function;
- Project onto simple constraints

$$Q \equiv \left\{w \in \mathbb{R}^d, \mu \in \mathbb{R}^n \mid b_i^T w = \mu_i\right\}.$$

Proof of Concept: Does second-order information help?





- Training Logistic Regression, $d = 7129$.
- Cubic Newton beats Gradient Descent for $10 \le |S| \le 50$.
- Second-order information improves convergence.

Maximization of the Dual Problem

Initial objective: $P(w) \;\equiv\; \sum\limits_{i=1}^{n} \phi_i(b_i^T w) \;+\; g(w).$

We have **Primal** and **Dual** problems:

$$\min_{w \in \mathbb{R}^d} P(w) \;\geq\; \max_{\alpha \in \mathbb{R}^n} D(\alpha),$$

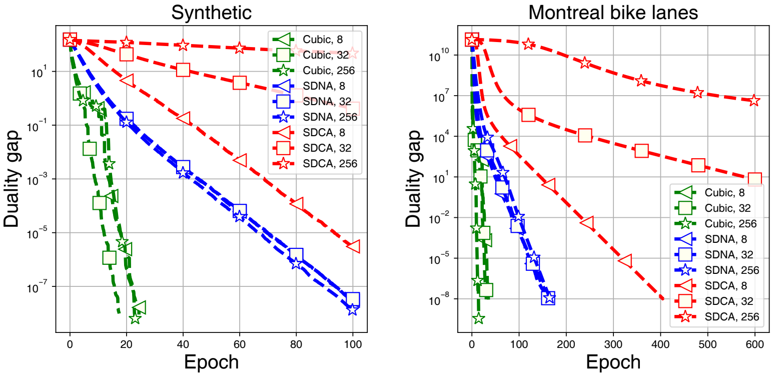
introducing Fenchel Conjugate: $f^*(s) \;\equiv\; \sup\limits_{x} \big[s^T x - f(x) \big]$, we have

$$D(\alpha) \;\equiv\; \underbrace{\sum_{i=1}^{n} -\phi_i^*(\alpha_i)}_{\text{separable, twice differentiable}} \;-\; \underbrace{g^*\Big(-B^T \alpha \Big)}_{\text{differentiable}}.$$

Solve **<u>Dual</u>** problem by our framework:

- Approximate ϕ_i^* by second-order cubic models;
- Treat g^* as quadratic function;
- Project onto $Q \equiv \bigcap_{i=1}^{n} dom\, \phi_i^*$.

- Solving the dual of Poisson regression.



SDNA: Zheng Qu et al. "SDNA: stochastic dual Newton ascent for empirical risk minimization". In: *International Conference on Machine Learning*. 2016, pp. 1823–1832

SDCA: Shai Shalev-Shwartz and Tong Zhang. "Stochastic dual coordinate ascent methods for regularized loss minimization". In: *Journal of Machine Learning Research* 14.Feb (2013), pp. 567–599

Conclusion

New second-order algorithm for convex optimization.

- Based on cubic regularization.
- Utilizes problem structure.
- Does randomized block updates (computationally cheap).
- Has global complexity guarantees.

New Primal-Dual method for Empirical Risk Minimization.

- Outperforms state-of-the-art in terms of number of data accesses.

Thank you for your attention!

See you at Poster #156.