Randomized Block Cubic Newton Method

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

nikitad101@gmail.com   peter.richtarik@kaust.edu.sa

[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

Motivation

We consider a convex optimization problem: $\min_{x \in Q} F(x)$, $Q \subset \mathbb{R}^N$.

- Gradient Descent requires $K = O(1/\varepsilon)$ iterations to solve the problem with accuracy ε: $F(x^K) - F^* \leq \varepsilon$. Cubic Newton [1] requires $O(1/\sqrt{\varepsilon})$ iterations, which is much faster.
- But the costs of one iteration are $O(N)$ and $O(N^3)$ respectively. This is too high for huge-scale applications.
- Modern block coordinate methods [3, 2] have *computationally effective* steps (updating only a subset of coordinates per step) and their *convergence rate* is similar to that of the full methods. Thus, they are suitable for huge-scale applications.
- **Our aim** is to develop a fast second-order method with global complexity guarantees and low cost of every iteration. By utilizing the problem structure, we combine the block-coordinate randomization technique and cubic regularization of Newton's method.

Problem Structure

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

$$F(x) \;\equiv\; \underbrace{\phi(x)}_{\substack{\text{twice}\\\text{differentiable}}} \;+\; \underbrace{g(x)}_{\text{differentiable}} \;+\; \underbrace{\psi(x)}_{\substack{\text{nonsmooth}\\\text{but simple}}}$$

For a given space decomposition

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

assume that ϕ and ψ are block separable:

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

There are no assumptions on the block separability of $g : \mathbb{R}^N \to \mathbb{R}$.

Thus, our **Convex Optimization Problem** has the following structure:

$$\min_{x \in Q} \Big[\, F(x) \equiv \sum_{i=1}^n \phi_i(x_{(i)}) + g(x) + \sum_{i=1}^n \psi_i(x_{(i)}) \Big],$$

where $Q \subset \mathbb{R}^N$ is a *simple* convex set.

Assumptions

- Every ϕ_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 is differentiable, and for some fixed symmetric positive semi-definite matrices $A \succeq G \succeq 0$ we have, for all $x, y \in \mathbb{R}^N$:
  - $g(y) \leq g(x) + \langle \nabla g(x), y - x\rangle + \frac{1}{2}\langle A(y-x), y-x\rangle$,
  - $g(y) \geq g(x) + \langle \nabla g(x), y - x\rangle + \frac{1}{2}\langle G(y-x), y-x\rangle$.
- ψ_i are arbitrary convex functions, possibly nondifferentiable but *simple*.

Model of the Objective

- Fix a subset of blocks: $S \subset \{1, \ldots, n\}$.
- For $y \in \mathbb{R}^N$ denote by $y_{[S]} \in \mathbb{R}^N$ the vector with zeroed $i \notin S$.
- $$M_{H,S}(x; y) \equiv F(x) + \langle \nabla \phi(x), y_{[S]}\rangle + \frac{1}{2}\langle \nabla^2 \phi(x) y_{[S]}, y_{[S]}\rangle + \frac{H}{6}\|y_{[S]}\|^3$$
$$+ \langle \nabla g(x), y_{[S]}\rangle + \frac{1}{2}\langle A y_{[S]}, y_{[S]}\rangle + \sum_{i \in S}\Big[\psi_i(x_{(i)} + y_{(i)}) - \psi_i(x_{(i)})\Big].$$
- From smoothness: $F(x + y) \leq M_{H,S}(x; y)$ $\forall x, y \in \mathbb{R}^N$ for $H \geq \sum_{i \in S} H_i$.
- Step of the method: $T_{H,S}(x) \equiv \underset{\substack{y \in \mathbb{R}^N_{[S]}\\ \text{s.t. } x + y \in Q}}{\operatorname{argmin}} M_{H,S}(x; y)$.

Algorithm 1: Randomized Block Cubic Newton (RBCN)

Parameters: starting point $x^0 \in Q$, uniform random distribution $\hat{S}$.

Iteration $k \geq 0$:

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: $x^{k+1} \overset{\text{def}}{=} x^k + T_{H_k,S_k}(x^k)$.

Convergence Results

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

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

Theorem 1. Under our assumptions, it is enough to set
$$K = O\left(\frac{1}{\varepsilon} \cdot \frac{n}{\tau} \cdot \left(1 + \log \frac{1}{\rho}\right)\right), \quad \text{where } \tau \equiv \mathbb{E}[|\hat{S}|].$$

Theorem 2. Let $\sigma \in [0, 1]$ be a special *condition number* (see our paper).

We have a bound for it: $\sigma \geq \lambda_{\min}(G) \,/\, \lambda_{\max}(A)$.

Then if $\sigma > 0$, it is enough to set
$$K = O\left(\frac{1}{\sqrt{\varepsilon}} \cdot \frac{n}{\tau} \cdot \frac{1}{\sigma} \cdot \left(1 + \log \frac{1}{\rho}\right)\right).$$

Theorem 3. Strongly convex case. Let $\mu \equiv \lambda_{\min}(G)$ be greater than zero. Then it is enough to set
$$K = O\left(\log\left(\frac{1}{\varepsilon\rho}\right) \cdot \frac{n}{\tau} \cdot \frac{1}{\sigma} \cdot \sqrt{\max\left\{\frac{HD}{\mu}, 1\right\}}\right), \text{ where } D \geq \|x^0 - x^*\|.$$

Empirical Risk Minimization

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

Examples. Logistic regression: $\phi_i(a) = \log(1 + \exp(-y_i a))$, Poisson regression: $\phi_i(a) = \exp(a) - y_i a$, Generalized linear models.

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

- Option 1: Solve this using the primal method (Algorithm 1).

Dual Problem:
$$\min_{w \in \mathbb{R}^d} P(w) \geq \max_{\alpha \in \mathbb{R}^n} D(\alpha) \equiv \underbrace{\sum_{i=1}^n -\phi_i^*(\alpha_i)}_{\substack{\text{separable, twice}\\\text{differentiable}}} - \underbrace{g^*\Big(-B^T\alpha\Big)}_{\text{differentiable}}.$$

- Option 2: Primal-Dual algorithm: updates of the cubic model $\widetilde{M}_{H,S}(\alpha; \cdot)$ for Dual problem.

Algorithm 2: Stochastic Dual Cubic Newton Ascent (SDCNA)

Parameters: starting dual variable $\alpha^0 \in Q$, distribution $\hat{S}$, $H_0 > 0$.

Iteration $k \geq 0$:

1: Make the primal update: $w^k \overset{\text{def}}{=} \nabla g^*\Big(B^T \alpha^k\Big)$

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

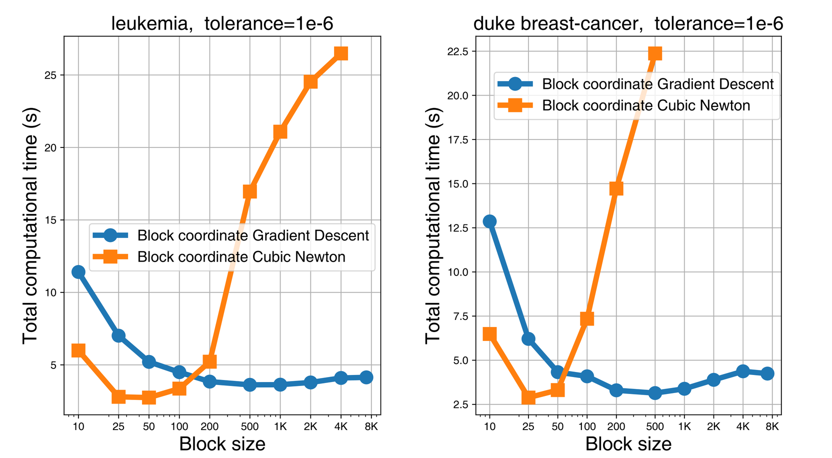
3: While $\widetilde{M}_{H_k,S_k}(\alpha^k; \alpha^k + \widetilde{T}_{H_k,S_k}(\alpha^k)) > -D(\alpha^k + \widetilde{T}_{H_k,S_k}(\alpha^k))$ do
$H_k := 0.5 \cdot H_k$

4: Make the dual update: $\alpha^{k+1} \overset{\text{def}}{=} \alpha^k + \widetilde{T}_{H_k,S_k}(\alpha^k)$

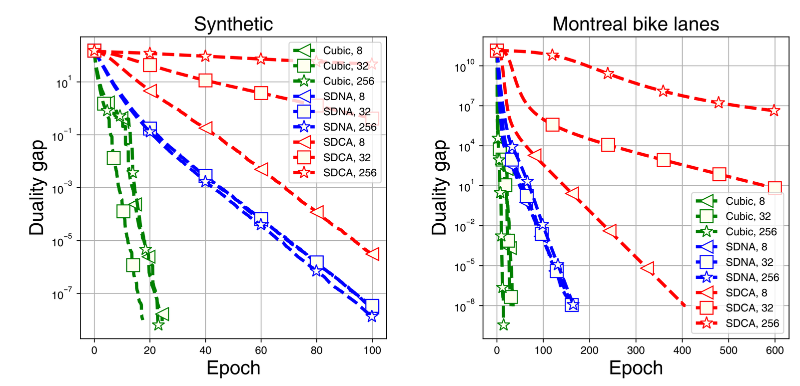
5: Set $H_{k+1} := 2 \cdot H_k$

Experiments

Training Logistic Regression (Algorithm 1)



Training Poisson Regression (Algorithm 2)



SDNA: [2] and **SDCA**: [4] methods.

References

[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

[3] 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

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