Extending the Reach of Big Data Optimization: Randomized Algorithms for Minimizing Relatively Smooth Functions

Filip Hanzely Peter Richtárik

University of Edinburgh and KAUST

Optimization Problem

$$\text{minimize } f(x) \text{ subject to } x \in Q$$

(Q: subset of $\mathbb{R}^n$; f: convex but not necessarily smooth function)

Main Assumptions

L - relative smoothness of $f(x)$ w.r.t. $h(x)$: $Lh(x) - f(x)$ is convex

μ - relative strong convexity of $f(x)$ w.r.t. $h(x)$: $f(x) - \mu h(x)$ is convex

- Standard smoothness and strong convexity arises as a special case for $h(x) = \frac{1}{2}\|x\|^2$
- Key tool – Bregman divergence

$$D_h(x, y) := h(x) - h(y) - \langle \nabla h(y), x - y \rangle$$

- Equivalent formulations (L - relative smoothness):

$$D_f(x, y) \leq L D_h(x, y)$$

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq L \langle \nabla h(x) - \nabla h(y), x - y \rangle$$

$$\nabla^2 f(x) \preccurlyeq L \nabla^2 h(x)$$

Baseline Algorithm – Primal Gradient Scheme [1, 2]

$$x^{t+1} \leftarrow \text{argmin}_{x \in Q} \langle \nabla f(x^t), x \rangle + L D_h(x, x^t)$$

Convergence result

$$f(x^T) - f(x^*) \leq \frac{\mu D_h(x^*, x^0)}{(1 + \frac{\mu}{L-\mu})^T - 1}$$

- Key tool for analysis – Three point property

$$\phi(x) + D_h(x, z) \geq \phi(z^+) + D_h(z^+, z) + D_h(x, z^+)$$

(ϕ: convex; any $x \in Q$; $z^+ = \text{argmin}_{x' \in Q} \{\phi(x') + D_h(x', z)\}$)

Main Contributions

First stochastic algorithms for minimization of relatively smooth functions:

- Generalized version of Stochastic Gradient Descent:
  Relative Stochastic Gradient Descent
- Generalized version of Randomized Coordinate Descent:
  Relative Randomized Coordinate Descent

Relative Stochastic Gradient Descent

- Access to unbiased estimator of gradient
- Useful when computation of gradient estimator is much cheaper than computation of full gradient

$$x^{t+1} \leftarrow \text{argmin}_{x \in Q} \langle g^t, x \rangle + \underbrace{(L + t\mu/2)}_{\text{stepsize}} D_h(x, x^t)$$

(g^t: unbiased estimator of $\nabla f(x^t)$)

Convergence result

$$E\left[\min_{t \leq T} f(x^t) - f(x^*)\right] \leq \frac{(L - \mu/2)^2 D_h(x^*, x^0)}{LT + (T-1)T\mu/4} + \frac{\sigma^2}{L + (T-1)\mu/4}$$

- Analysis provided for a range of stepsize parameters

Relative Randomized Coordinate Descent

- h is separable:

$$h(x) = \sum_{i=1}^n h_i(x_i)$$

w-relative strong convexity (extension for separable h):

$$D_f(x, y) \geq D_h(x, y)_w$$

$$D_h(y, z)_u = \sum_{i=1}^n u_i \left(h_i(y_i) - h_i(z_i) - \langle \nabla h_i(z_i), y_i - z_i \rangle\right)$$

h-ESO (Expected Separable Overapproxmation, defined in [3] in case of h being L2 norm):

$$E\left[f(x + \sum_{i \in \hat{S}} q_i e_i)\right] \leq f(x) + \langle \nabla f(x), q \rangle_p + D_h(x + q, x)_{p \circ v}$$

($\hat{S}$: random subset of $\{1, 2, \ldots, n\}$; $P(i \in \hat{S}) = p_i$; $p = (p_1, \ldots, p_n)^\top$; v: vector of parameters for h–ESO)

$$x_j^{t+1} \leftarrow \text{argmin}_{x \in Q_j^t} \langle \nabla f(x^t), x \rangle + D_h(x, x^t)_v$$

($Q_j^t = Q \cap \{x^t + \gamma e_j | \gamma \in \mathbb{R}\}$; j belong to random subset of $\{1, 2, \ldots, n\}$)

Convergence result

$$E\left[\min_{t \leq T} f(x^t) - f(x^*)\right] \leq \frac{\left(D_h(x, x^0)_v + f(x^0) - f(x^*)\right)\Delta}{\left(\frac{1}{1 - p_1 \Delta}\right)^T - 1 - p_1 \Delta}$$

($\Delta = \min_i \frac{w_i}{v_i}$; assume that $p_1 = p_2 = \cdots = p_n$; Most important term; guaranteeing linear rate)

References

[1] Bauschke, Heinz H., Jérôme Bolte, and Marc Teboulle. "A descent Lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications." *Mathematics of Operations Research* (2016).

[2] Lu, Haihao, Robert M. Freund, and Yurii Nesterov. "Relatively-Smooth Convex Optimization by First-Order Methods, and Applications." arXiv preprint arXiv:1610.05708 (2016).

[3] Qu, Zheng, and Peter Richtárik. "Coordinate descent with arbitrary sampling II: Expected separable overapproximation." Optimization Methods and Software 31.5 (2016): 858-884.