Simple Stochastic Gradient Methods for Non-Smooth Non-Convex Regularized Optimization

Michael R. Metel[1] Akiko Takeda[1,2]

[1]RIKEN Center for Advanced Intelligence Project, Tokyo, Japan

[2]Department of Creative Informatics, Graduate School of Information Science and Technology, the University of Tokyo, Tokyo, Japan

June 12, 2019

Problem setting

Regularized optimization problems of the form

$$\min_{w \in \mathbb{R}^d} h(w) := f(w) + g(w)$$

f(w): loss function

- non-convex
- Lipschitz continuous gradient
- $f(w) := \mathbb{E}_\xi[F(w, \xi)]$ or $f(w) := \frac{1}{n} \sum_{j=1}^{n} f_j(w)$

g(w): sparse regularizer

- non-smooth, non-convex
- Lipschitz continuous
- easily computable proximal operator
  $\text{prox}_{\lambda g}(w) := \operatorname*{argmin}_{x \in \mathbb{R}^d} \left\{ \frac{1}{2\lambda} ||w - x||_2^2 + g(x) \right\}$
- SCAD, MCP, log-sum penalty, capped l_1 norm

Research focus

Non-asymptotic convergence results using simple first-order stochastic methods.

Aim is to find an ϵ-stationary solution $\overline{w}$ in expectation,

$$\mathbb{E}\left[\mathrm{dist}(0, \partial h(\overline{w}))\right] \leq \epsilon.$$

Auxiliary function of $h(w)$

$$\min_{w\in\mathbb{R}^d} h(w) := f(w) + g(w)$$

Considered an auxiliary function

$$\tilde{h}_\lambda(w) := f(w) + e_\lambda g(w),$$

where

$$e_\lambda g(w) := \inf_{x\in\mathbb{R}^d} \left\{ \frac{1}{2\lambda} ||w - x||_2^2 + g(x) \right\} \quad \text{(Moreau envelope)}$$

Using iteration w^k construct a smooth majorizing function $E_\lambda^k(w)$ of $h_\lambda(w)$, with

$$\nabla E_\lambda^k(w) = \nabla f(w) + \frac{1}{\lambda}(w - \zeta^\lambda(w^k)), \quad \zeta^\lambda(w^k) \in \text{prox}_{\lambda g}(w^k).$$

Convergence for $h(w)$

Use a mini-batch stochastic gradient algorithm (MBSGA) to minimize

$$\mathbb{E}||\nabla E_\lambda^R(w^R)||_2.$$

Lipschitz continuity of $g(w)$ used to bound

$$\mathbb{E}\left[\text{dist}(0, \partial h(\text{prox}_{\lambda g}(w^R))) - ||\nabla E_\lambda^R(w^R)||_2\right].$$

- Also considered a variance reduced stochastic gradient algorithm (VRSGA) for finite-sum problems.

Convergence results

$$\min_{w \in \mathbb{R}^d} \; h(w) := f(w) + g(w)$$

For an ϵ-stationary point $\bar{w}$ in expectation,

$$\mathbb{E}\left[\text{dist}(0, \partial(h(\bar{w})))\right] \leq \epsilon.$$

Table: Comparison of convergence complexities obtained in (Xu et al., 2018a,b) and this paper.

Algorithm	Finite-sum Assumption	Gradient Call Complexity	Proximal Operator Complexity
SSDC-SPG[a]	$\times$	$O(\epsilon^{-8})$	$O(\epsilon^{-8})$
SSDC-SVRG[a]	$\surd$	$O(n\epsilon^{-4})$	$O(\epsilon^{-4})$
MBSGA	$\times$	$O(\epsilon^{-5})$	$O(\epsilon^{-4})$
VRSGA	$\surd$	$O(n^{2/3}\epsilon^{-3})$	$O(\epsilon^{-3})$
SSDC-SPG[b]	$\times$	$O(\epsilon^{-5})$	$O(\epsilon^{-5})$
SSDC-SVRG[b]	$\surd$	$\tilde{O}(n\epsilon^{-3})$	$\tilde{O}(\epsilon^{-3})$

Experimental results

Application: Binary classification with smooth non-convex loss function and log-sum penalty as regularizer.

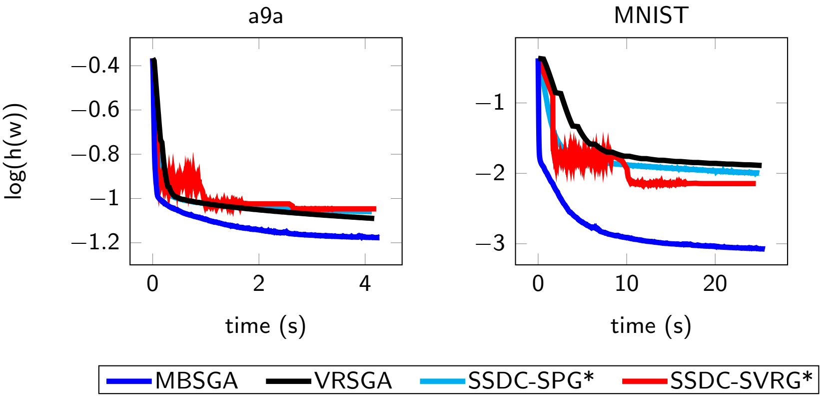

Figure: Comparison of algorithms of this paper and (Xu et al., 2018) (marked with *).

Poster session

Today 06:30 – 09:00 PM @ Pacific Ballroom #104

Bibliography

Xu, Yi, Qi Qi, Qihang Lin, Rong Jin, and Tianbao Yang (2018). "Stochastic Optimization for DC Functions and Non-smooth Non-convex Regularizers with Non-asymptotic Convergence". In: *(a): arXiv preprint arXiv:1811.11829v1, Access date: December 3, 2018, (b): arXiv preprint arXiv:1811.11829v2.*