Optimal Stochastic and Online Learning with Individual Iterates

Yunwen Lei[1,2], Peng Yang[1], Ke Tang[1] and Ding-Xuan Zhou[3]

[1]Southern University of Science and Technology
[2]Technical University of Kaiserslautern
[3]City University of Hong Kong

{leiyw, yangp, tangk3}@sustech.edu.cn mazhou@cityu.edu.hk

December 11, 2019

Background

Problem: Want to solve optimization problem of composite structure:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \phi(\mathbf{w}) = \mathbb{E}_z[f(\mathbf{w}, z)] + r(\mathbf{w}), \tag{1}$$

where $f : \mathbb{R}^d \times \mathcal{Z} \mapsto \mathbb{R}_+$(loss), $r : \mathbb{R}^d \mapsto \mathbb{R}_+$ (regularizer) are convex.

Data: $\mathbf{z} = \{z_t\}$ drawn i.i.d. from a measure defined over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$

Instantiations: SVMs, Logistic Regression, Lasso, Ridge Regression, etc.

Optimal model: $\mathbf{w}^* = \operatorname{argmin}_{\mathbf{w}\in\mathbb{R}^d} \phi(\mathbf{w})$

Stochastic Composite Mirror Descent

A strongly convex mirror map $\Psi : \mathbb{R}^d \mapsto \mathbb{R}$ to induce a Bregman distance

$$D_\Psi(\mathbf{w}, \tilde{\mathbf{w}}) := \Psi(\mathbf{w}) - [\Psi(\tilde{\mathbf{w}}) + \langle \mathbf{w} - \tilde{\mathbf{w}}, \nabla\Psi(\tilde{\mathbf{w}})\rangle] \geq \frac{\sigma}{2}\|\mathbf{w} - \tilde{\mathbf{w}}\|^2$$

Idea: separate data-fitting term and regularizer

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \underbrace{\langle \mathbf{w} - \mathbf{w}_t, f'(\mathbf{w}_t, z_t)\rangle}_{\text{first-order approximation of } f(\mathbf{w}, z_t) \text{ at } \mathbf{w}_t} + r(\mathbf{w}) + \underbrace{\eta_t^{-1} D_\Psi(\mathbf{w}, \mathbf{w}_t)}_{\text{stabilizer}} \qquad (2)$$

A framework covering many algorithms: (Nemirovsky and Yudin, 1983; Beck and Teboulle, 2003; Zinkevich, 2003; Zhang, 2004; Bach and Moulines, 2013; Bottou et al., 2018; Duchi et al., 2010; Shalev-Shwartz et al., 2011; Hazan and Kale, 2014)

- SGD
- Stochastic Proximal Gradient Descent
- Stochastic Mirror Descent

keep r intact and approximate f by first-order approximation

Existing Work

Problem: How to identify a model from sequence $\{\mathbf{w}_t\}_{t=1}^T$

- LAST: output the last single iterate (Shamir and Zhang, 2013)
- UNI-AVE: average all iterates with uniform weights
- WEI-AVE: weighted average with weight $t+1$ for $\mathbf{w}_t$ (Lacoste-Julien et al., 2012)
- SUFFIX: uniform average of the last half of SGD iterates (Rakhlin et al., 2012)
- RAND: a random iterate drawn from $\{\mathbf{w}_t\}_{t=1}^T$

Problems:

- either suboptimal in the sense of logarithmic factors
- or requires averaging of iterates (sparsity destroyed)

Algorithm with optimal rate, sparsity and good practical behavior?

Motivation and Idea

Key inequality measuring one-step progress:

$$\mathbb{E}[\phi(\mathbf{w}_t)-\phi(\mathbf{w})] \le \eta_t^{-1}\mathbb{E}[D_\Psi(\mathbf{w},\mathbf{w}_t) - D_\Psi(\mathbf{w},\mathbf{w}_{t+1})] + \eta_t C. \tag{3}$$

- If set $\mathbf{w} = \mathbf{w}^*$ and show $\mathbb{E}[D_\Psi(\mathbf{w}^*,\mathbf{w}_t) - D_\Psi(\mathbf{w}^*,\mathbf{w}_{t+1})] = O(\eta_t^2)$, then

  $$\text{optimal convergence} \quad \mathbb{E}[\phi(\mathbf{w}_t)] - \phi(\mathbf{w}^*) = O(\eta_t)$$

  since $\eta_t = 1/\sqrt{t}$ for convex and $\eta_t = 1/t$ for strongly-convex setting.
- By non-negativity of Bregman distance, we find $T^* \in \{T, \ldots, 2T-1\}$ with

  $$D_\Psi(\mathbf{w}^*,\mathbf{w}_{T^*}) - D_\Psi(\mathbf{w}^*,\mathbf{w}_{T^*+1}) \le T^{-1} \underbrace{D_\Psi(\mathbf{w}^*,\mathbf{w}_T)}_{=O(T\eta_T^2)}. \tag{4}$$

- $\mathbf{w}^*$ replaced by a surrogate $\bar{\mathbf{w}}_T$ with $\mathbb{E}[\phi(\bar{\mathbf{w}}_T)] - \phi(\mathbf{w}^*) = O(\eta_T)$

Algorithm

- SCMDI: Stochastic Composite Mirror Descent with Individual Iterates

Algorithm 1: SCMDI

Input: $\{\eta_t\}_t, \mathbf{w}_1$ and T.

for $t = 1, 2$ **to** $T - 1$ **do**
  calculate $\mathbf{w}_{t+1}$ by (2)
set $\bar{\mathbf{w}}_T$ as an average of iterates
for $t = T, T+1$ **to** $2T - 1$ **do**
  calculate $\mathbf{w}_{t+1}$ by (2)
  $\triangle \leftarrow D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_t) - D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_{t+1})$
  if $\triangle \leq T^{-1} D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_T)$ **then**
    $T^* \leftarrow t$, $\mathbf{w}_{T^*} \leftarrow \mathbf{w}_t$

- OCMDI: Online Composite Mirror Descent with Individual Iterates
  - update average at 2^t-th iteration, $t = 1, 2, \ldots$
  - no information of T required

Theory

Assumptions 1: the existence of A and $B > 0$ such that

$$\|f'(\mathbf{w}, z)\|_*^2 \leq Af(\mathbf{w}, z) + B \quad \text{and} \quad \|r'(\mathbf{w})\|_*^2 \leq Ar(\mathbf{w}) + B.$$

Convex case: If Assumption 1 and $\eta_t \asymp 1/\sqrt{t}$, then

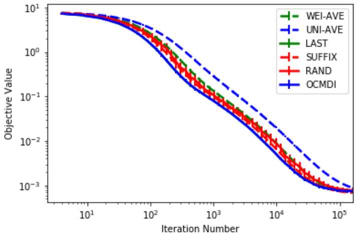
$$\mathbb{E}[\phi(\mathbf{w}_{T^*})] - \phi(\mathbf{w}^*) = O(T^{-\frac{1}{2}}).$$

Strongly convex case: If Assumption 1 and $\eta_t \asymp 1/t$, then

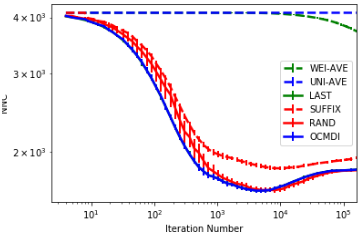
$$\mathbb{E}[\phi(\mathbf{w}_{T^*})] - \phi(\mathbf{w}^*) = O(T^{-1}).$$

Tomography Reconstruction

- Objective function: $\phi(\mathbf{w}) = \frac{1}{n}\|A\mathbf{w} - \mathbf{y}\|_2^2$
  - $A \in \mathbb{R}^{n\times d}$ is a CT-measurement matrix
  - $\mathbf{y} \in \mathbb{R}^n$ is a noisy measurement vector
- $\mathbf{w}^*$ is a sparse image.
- SCMD with (randomized sparse Kaczmarz algorithm)
  - $\Psi(\mathbf{w}) = \lambda\|\mathbf{w}\|_1 + \frac{1}{2}\|\mathbf{w}\|_2^2$
  - $f(\mathbf{w}, z) = \frac{1}{2}(\langle \mathbf{w}, x\rangle - y)^2$
  - $r(\mathbf{w}) = 0$



(a): objective function value



(b): number of non-zero components

Welcome to East Exhibition Hall B + C #164 for more details

Thank You!

References I

F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n). In *Advances in Neural Information Processing Systems*, pages 773–781, 2013.

A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

J. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *Conference on Learning Theory*, pages 14–26, 2010.

E. Hazan and S. Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, 15(1):2489–2512, 2014.

S. Lacoste-Julien, M. Schmidt, and F. Bach. A simpler approach to obtaining an O(1/t) convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*, 2012.

A.-S. Nemirovsky and D.-B. Yudin. *Problem Complexity and Method Efficiency in Optimization.* John Wiley & Sons, 1983.

A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *International Conference on Machine Learning*, pages 449–456, 2012.

S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical programming*, 127(1):3–30, 2011.

O. Shamir and T. Zhang. Stochastic gradient descent for non-smooth optimization convergence results and optimal averaging schemes. In *International Conference on Machine Learning*, pages 71–79, 2013.

T. Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *International Conference on Machine Learning*, pages 919–926, 2004.

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *International Conference on Machine Learning*, pages 928–936, 2003.