Stochastic Variance Reduced Primal-Dual Algorithms for Empirical Composition Optimization

UNIVERSITY of FLORIDA — Tencent AI Lab — NEURAL INFORMATION PROCESSING SYSTEMS

Adithya M. Devraj[1] and Jianshu Chen[2]

[1]University of Florida, Gainesville, FL [2]Tencent AI Lab, Bellevue, WA

Empirical Composition Optimization

Objective:
$$\min_{\theta} \frac{1}{n_X}\sum_{i=0}^{n_X-1} \phi_i\Big(\frac{1}{n_{Y_i}}\sum_{j=0}^{n_{Y_i}-1} f_\theta(x_i, y_{ij})\Big) + g(\theta)$$

- $(x_i, y_{ij}) \in \mathbb{R}^{m_x} \times \mathbb{R}^{m_y}$ is the (i,j)-th data sample
- $f_\theta : \mathbb{R}^{m_x} \times \mathbb{R}^{m_y} \to \mathbb{R}^{\ell}$ is parameterized by $\theta \in \mathbb{R}^d$
- $\phi_i : \mathbb{R}^{\ell} \to \mathbb{R}^{+}$ is a convex *merit function*
- $g(\theta)$ is a μ-strongly convex regularizer
- *There is an empirical average both inside and outside the nonlinear merit function*

Motivating Examples

I. Unsupervised sequence classification [1]:

$$\min_\theta \Big\{ -\sum_{i=0}^{n_X-1} p_{\rm LM}(x_i) \log\Big(\frac{1}{n_Y}\sum_{j=0}^{n_Y-1} f_\theta(x_i, y_j)\Big)\Big\}$$

- $p_{\rm LM} : \mathbb{R}^{m_x} \to [0,1]$: known language model
- $f_\theta : \mathbb{R}^{m_x} \times \mathbb{R}^{m_y} \to \mathbb{R}^{\ell}$: prediction of n-gram frequency by a parameterized sequence classifier

II. Risk-averse learning [2, 4, 5, 7]:

$$\min_\theta \Big\{ -\frac{1}{n}\sum_{i=0}^{n-1}\langle x_i, \theta\rangle + \frac{1}{n}\sum_{i=0}^{n-1}\Big(\langle x_i, \theta\rangle - \frac{1}{n}\sum_{j=0}^{n-1}\langle x_j, \theta\rangle\Big)^2\Big\}$$

- $\{x_i \in \mathbb{R}^d : 1 \le i \le n\}$: vector consisting of the rewards from d assets at time i
- $\theta \in \mathbb{R}^d$: weight vector on the d assets

III. MDP policy evaluation [2, 5]:

$$\min_\theta \Big\{ \frac{1}{S}\sum_{i=1}^{S}\Big(\langle \Psi_i, \theta\rangle - \sum_{j=1}^{S} P^\pi_{i,j}\big(r_{i,j} + \gamma\langle\Psi_j, \theta\rangle\big)\Big)^2\Big\}$$

- $P^\pi \in \mathbb{R}^{S\times S}$: state transition probability matrix
- γ: discount factor
- $\{\Psi_i \in \mathbb{R}^d : 1 \le i \le S\}$: feature vectors
- $\theta \in \mathbb{R}^d$: weight vector for the d features

Challenges: Biased Gradients

- Gradient of the objective with respect to θ:

$$\frac{1}{n_X}\sum_{i=0}^{n_X-1}\Big[\frac{1}{n_{Y_i}}\sum_{j=0}^{n_{Y_i}-1}\frac{\partial}{\partial\theta} f_\theta(x_i, y_{ij})\Big]^{\intercal}\Big[\phi_i'\Big(\frac{1}{n_{Y_i}}\sum_{j=0}^{n_{Y_i}-1} f_\theta(x_i, y_{ij})\Big)\Big] + \frac{\partial}{\partial\theta} g(\theta)$$

- The following "sampled stochastic gradient" is biased:

$$\Big[\frac{\partial}{\partial\theta} f_\theta(x_i, y_{ij})\Big]^{\intercal}\Big[\phi_i'\big(f_\theta(x_i, y_{ij})\big)\Big] + \frac{\partial}{\partial\theta} g(\theta)$$

- *Cannot* directly apply SGD-like techniques

Our Approach: Primal-Dual Formulation

For $\psi : \mathbb{R}^{\ell} \to \mathbb{R}$, its convex conjugate $\psi^* : \mathbb{R}^{\ell} \to \mathbb{R}$ is defined:

$$\psi^*(y) = \sup_{x\in\mathbb{R}^\ell}(\langle x, y\rangle - \psi(x)) \overset{\text{Strong Duality}}{\Leftrightarrow} \psi(x) = \sup_{y\in\mathbb{R}^\ell}(\langle x, y\rangle - \psi^*(y))$$

***Transformed* min-max objective:**

$$\min_\theta \max_w \Big\{ \underbrace{\frac{1}{n_X}\sum_{i=0}^{n_X-1}\Big[\Big\langle \overbrace{\frac{1}{n_{Y_i}}\sum_{j=0}^{n_{Y_i}-1} f_\theta(x_i, y_{ij})}^{\overline{f}_i(\theta)}, w_i\Big\rangle - \phi_i^*(w_i)\Big]}_{L(\theta, w)} + g(\theta)\Big\}, \quad w := \{w_0, \ldots, w_{n_X-1}\}$$

- No more non-linear compositions of empirical averages
- Can sample unbiased gradients with respect to w_i's and θ
- Maximization is *decoupled* over w_i's

SVRPDA - I: Main Ideas

I. Dual step: Stochastic variance reduced coordinate ascent

- Batch gradient ascent for dual variables: For each $1 \le i \le n_X$,

$$w_i^{(k)} = \arg\min_{w_i}\Big\{ -\Big\langle \frac{1}{n_{Y_i}}\sum_{j=0}^{n_{Y_i}-1} f_{\theta^{(k-1)}}(x_i, y_{ij}), w_i\Big\rangle + \phi_i^*(w_i) + \frac{1}{2\alpha_w}\|w_i - w_i^{(k-1)}\|^2\Big\}$$

  - Evaluating the full batch gradient is expensive; So is updating all n_X dual variables

- **Key Idea 1:** *Exploit decoupled dual maximization over w_i's:*
  - At each iteration k, randomly sample index i and update w_i; Keep $\{w_j, j \neq i\}$ unchanged
- **Key Idea 2:** Replace the full gradient with respect w_i, with *low variance stochastic gradient*, using the SVRG technique [3]: $\delta_k^w = f_{\theta^{(k-1)}}(x_{i_k}, y_{i_k j_k}) - f_{\tilde\theta}(x_{i_k}, y_{i_k j_k}) + \overline{f}_{i_k}(\tilde\theta)$

II. Primal step: Stochastic variance reduced gradient descent

- Batch gradient descent update for primal variable:

$$\theta^{(k)} = \arg\min_\theta\Big\{\Big\langle \sum_{i=0}^{n_X-1}\sum_{j=0}^{n_{Y_i}-1}\frac{1}{n_X n_{Y_i}} f'_{\theta^{(k-1)}}(x_i, y_{ij}) w_i^{(k)}, \theta\Big\rangle + \frac{1}{2\alpha_\theta}\|\theta - \theta^{(k-1)}\|^2\Big\}$$

  - Once again, computational cost can be very large
- As before, we can use the SVRG technique to replace full gradient with a low variance stochastic gradient: $\delta_k^\theta = f'_{\theta^{(k-1)}}(x_{i_k'}, y_{i_k' j_k'})\tilde w_{i_k'} - f'_{\tilde\theta}(x_{i_k'}, y_{i_k' j_k'})\tilde w_{i_k'} + L'_\theta(\tilde\theta, \tilde w)$

III. Low complexity stochastic variance reduced estimator

- *Faster the reference variables $\tilde\theta$ and $\tilde w$ are updated, lower the variance of stochastic gradient, and faster the convergence;* But also requires more complexity
- **Key Trick:** "Free" full batch gradient update to obtain $L'_\theta(\tilde\theta, w^{(k)})$
  - Use the fact that a single w_i is updated in each iteration
  - Exploit the linearity of objective in dual variables

$$L'(\tilde\theta, w^{(k)}) = L'(\tilde\theta, w^{(k-1)}) + \frac{1}{n_X}\overline{f}'_{i_k}(\tilde\theta)\big(w_{i_k}^{(k)} - w_{i_k}^{(k-1)}\big)$$

- Replace $L'_\theta(\tilde\theta, \tilde w)$ with $L'_\theta(\tilde\theta, w^{(k)})$ in naive SVRG gradient estimator

SVRPDA - II: Main Ideas

- Updating $L'(\tilde\theta, w^{(k)})$ requires storing $\overline{f}'_i(\tilde\theta)$, for each $1 \le i \le n_X$; Storage complexity can be very high
- Heuristic: Replace $\overline{f}'_i(\tilde\theta)$ with sampled $f'_{\tilde\theta}(x_i, y_{ij})$ in update rule, resulting in low storage complexity SVRPDA - II

Main Results

Assumptions:

- g is μ-strongly convex, and each ϕ_i is $1/\gamma$-smooth and B_w-Lipschitz
- $f_\theta(x_i, y_{ij})$ is B_θ-smooth in θ, and its gradients are uniformly bounded by B_f
- For each given w in its domain, the $L(\theta, w)$ is convex in θ

Theorem: After s outer-loops, to achieve error $\mathsf{E}\|\tilde\theta_s - \theta^*\|^2 < \epsilon$ using SVRPDA-I, total complexity in terms of "number of oracle calls" required is

$$O\big((n_X n_Y + n_X\kappa + n_X)\ln(1/\epsilon)\big)$$

Storage Complexity of SVRPDA

Methods	U_0	$\{\overline{f}_i\}$	$\{\overline{f}'_i\}$	$\theta^{(k)}$	$\tilde\theta$	$\{w_i^{(k)}\}$	δ_k^θ	δ_k^w	Total
SVRPDA-I	$O(d)$	$O(n_X\ell)$	$O(n_X d\ell)$	$O(d)$	$O(d)$	$O(n_X\ell)$	$O(d)$	$O(\ell)$	$O(n_X d\ell)$
SVRPDA-II	$O(d)$	$O(n_X\ell)$	—	$O(d)$	$O(d)$	$O(n_X\ell)$	$O(d)$	$O(\ell)$	$O(d+n_X\ell)$

Comparison with Related Works

Objective in related work:

$$\min_\theta \frac{1}{n_X}\sum_{i=0}^{n_X-1}\phi_i\Big(\frac{1}{n_Y}\sum_{j=0}^{n_Y-1} f_\theta(y_j)\Big)$$

Table: Total complexities of various stochastic composition optimization algorithms. For C-SAGA, $\alpha = 2/3$ in the minibatch setting, and $\alpha = 1$ when batch-size=1.

Methods	**SVRPDA-I**	Comp-SVRG [4]	C-SAGA [5]	MSPBE-SVRG & MSPBE-SAGA [7]	ASCVRG [6]
General problem	$(n_X n_Y + n_X\kappa)\ln\frac{1}{\epsilon}$	—	—	—	—
Special problem	$(n_X + n_Y + n_X\kappa)\ln\frac{1}{\epsilon}$	$(n_X + n_Y + \kappa^3)\ln\frac{1}{\epsilon}$	$(n_X + n_Y + (n_X + n_Y)^\alpha\kappa)\ln\frac{1}{\epsilon}$	—	$(n_X + n_Y)\ln\frac{1}{\epsilon} + \frac{1}{\epsilon^3}$
Special problem with $n_X = 1$	$(n_Y + \kappa)\ln\frac{1}{\epsilon}$	$(n_Y + \kappa^3)\ln\frac{1}{\epsilon}$	$(n_Y + n_Y^\alpha\kappa)\ln\frac{1}{\epsilon}$	$(n_Y + \kappa^2)\ln\frac{1}{\epsilon}$	$n_Y\ln\frac{1}{\epsilon} + \frac{1}{\epsilon^3}$

Numerical Results: Risk-Averse Learning

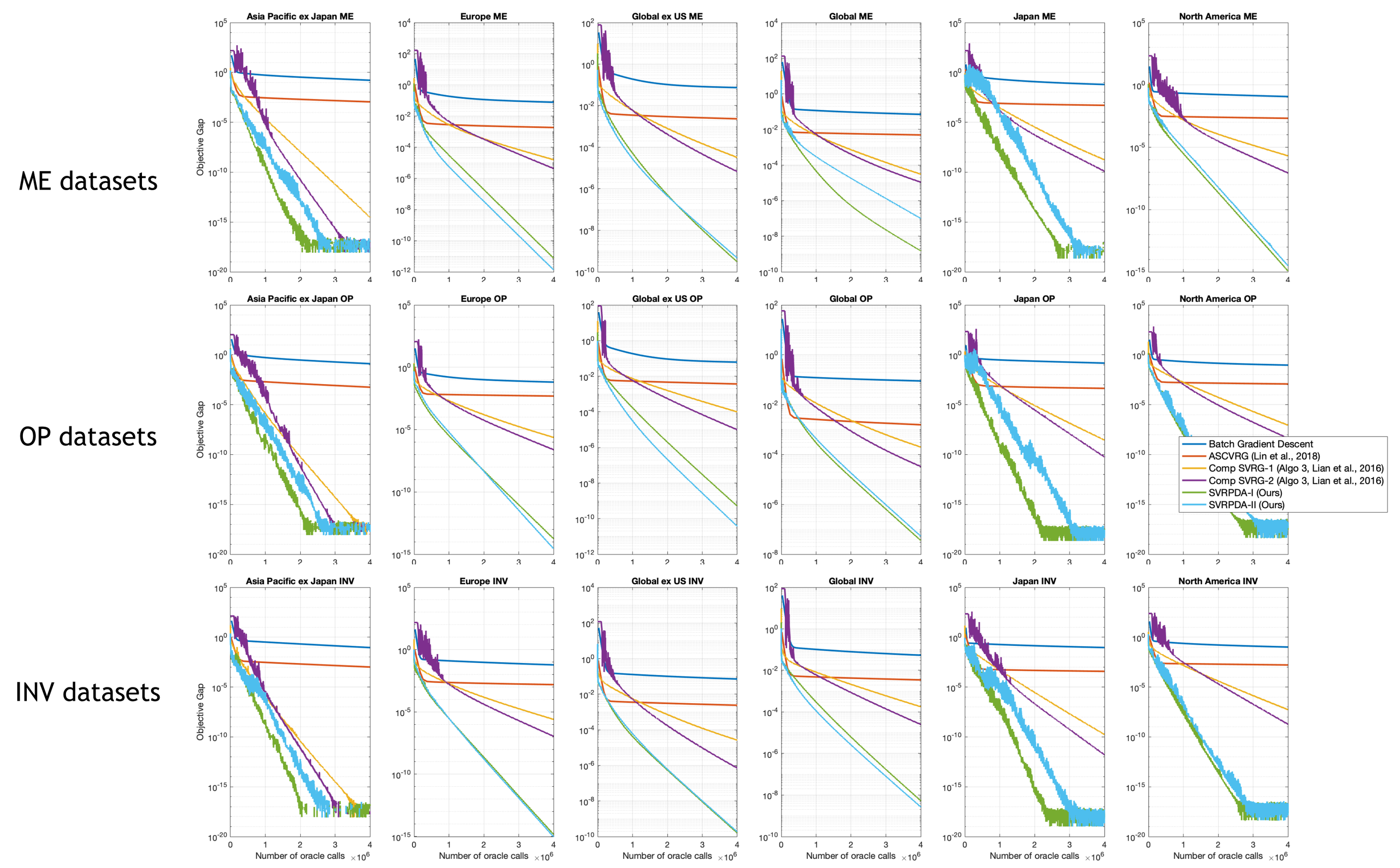

Figure: Risk-averse learning for portfolio optimization, with $n = 7240$ and $d = 25$. Algorithms are evaluated on 18 real-world US Research Returns data-set obtained from Center for Research in Security Prices (CRSP) website.

Numerical Results: MDP Policy Evaluation

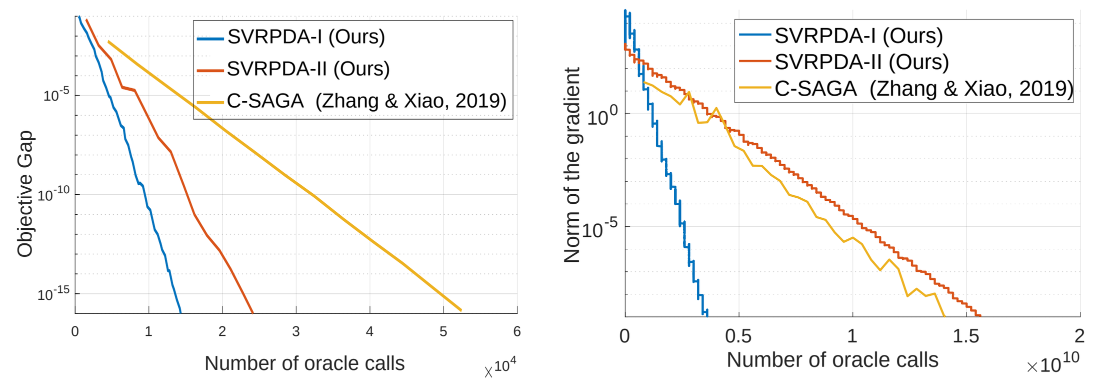

Figure: MDP for policy evaluation, with $S = 10$, $d = 5$ for the left figure, and $S = 10^4$, $d = 10$ for the right figure. Algorithms are evaluated on artificially generated data-set, similar to (Zhang & Xiao, 2019).

References

[1] Y. Liu, J. Chen, & L. Deng, *Unsupervised sequence classification using sequential output statistics*, NeurIPS, 2017.

[2] M. Wang, E. Fang, & H. Liu, *Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions*, Mathematical Programming, 2017.

[3] R. Johnson, & T. Zhang, *Accelerating stochastic gradient descent using predictive variance reduction*, NeurIPS, 2013.

[4] X. Lian, M. Wang, and J. Liu, *Finite-sum composition optimization via variance reduced gradient descent*, AISTATS, 2017.

[5] J. Zhang, & L. Xiao, *A composite randomized incremental gradient method*, ICML, 2019.

[6] T. Lin, C. Fan, M. Wang, & M. I. Jordan, *Improved oracle complexity for stochastic compositional variance reduced gradient*, ArXiv e-prints, 2018.

[7] S. Du, J. Chen, L. Li, L. Xiao, & D. Zhou, *Stochastic variance reduction methods for policy evaluation*, NeurIPS, 2017.