Finite-Sum Coupled Compositional Stochastic Optimization

Theory and Applications

Bokun Wang and Tianbao Yang

ICML
International Conference
On Machine Learning

Finite-Sum Optimization

$$\min_{h \in \mathcal{H}} \hat{R}(h), \quad \hat{R}(h) = \frac{1}{n} \sum_{i=1}^{n} L(h(\mathbf{x}_i), y_i).$$

Finite-Sum Optimization

$$\min_{h \in \mathcal{H}} \hat{R}(h), \quad \hat{R}(h) = \frac{1}{n} \sum_{i=1}^{n} L(h(\mathbf{x}_i), y_i).$$

Hypothesis parameterized by **w**

$$\min_{\mathbf{w} \in \Omega} F(\mathbf{w}), \quad F(\mathbf{w}) := \frac{1}{n} \sum_{\mathbf{z}_i \in \mathcal{D}} \ell(\mathbf{w}; \mathbf{z}_i)$$

Finite-Sum Optimization

$$\min_{h\in\mathcal{H}} \hat{R}(h), \quad \hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} L(h(\mathbf{x}_i), y_i).$$

Hypothesis parameterized by **w**

$$\min_{\mathbf{w}\in\Omega} F(\mathbf{w}), \qquad F(\mathbf{w}) := \frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} \ell(\mathbf{w};\mathbf{z}_i)$$

$$n = |\mathcal{D}|$$

Gradient Descent

$$\mathbf{w} \leftarrow \mathbf{w} - \eta\frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} \nabla\ell(\mathbf{w};\mathbf{z}_i)$$

Finite-Sum Optimization

$$\min_{h \in \mathcal{H}} \hat{R}(h), \quad \hat{R}(h) = \frac{1}{n} \sum_{i=1}^{n} L(h(\mathbf{x}_i), y_i).$$

Hypothesis parameterized by **w**

$$\min_{\mathbf{w} \in \Omega} F(\mathbf{w}), \quad F(\mathbf{w}) := \frac{1}{n} \sum_{\mathbf{z}_i \in \mathcal{D}} \ell(\mathbf{w}; \mathbf{z}_i)$$

Gradient Descent

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla F(\mathbf{w})$$

$$\boxed{\frac{1}{n} \sum_{\mathbf{z}_i \in \mathcal{D}} \nabla \ell(\mathbf{w}; \mathbf{z}_i)}$$

Expensive when n is large!

Finite-Sum <u>Stochastic</u> Optimization

$$\min_{h \in \mathcal{H}} \hat{R}(h), \quad \hat{R}(h) = \frac{1}{n} \sum_{i=1}^{n} L(h(\mathbf{x}_i), y_i).$$

Hypothesis parameterized by **w**

$$\min_{\mathbf{w} \in \Omega} F(\mathbf{w}), \quad F(\mathbf{w}) := \frac{1}{n} \sum_{\mathbf{z}_i \in \mathcal{D}} \ell(\mathbf{w}; \mathbf{z}_i)$$

Stochastic Gradient Descent

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \hat{\nabla} F(\mathbf{w})$$

Unbiased estimator, e.g., $\nabla \ell(\mathbf{w}; \mathbf{z}_i)$

$$\mathbb{E}[\hat{\nabla} F(\mathbf{w})] = \nabla F(\mathbf{w})$$

Finite-Sum Stochastic Optimization

$$\min_{h\in\mathcal{H}} \hat{R}(h), \quad \hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} L(h(\mathbf{x}_i), y_i).$$

Hypothesis parameterized by **w**

$$\min_{\mathbf{w}\in\Omega} F(\mathbf{w}), \qquad F(\mathbf{w}) := \frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} \ell(\mathbf{w};\mathbf{z}_i)$$

Stochastic Gradient Descent

$$\mathbf{w} \leftarrow \mathbf{w} - \eta\hat{\nabla}F(\mathbf{w})$$

Unbiased estimator, e.g., $\nabla\ell(\mathbf{w};\mathbf{z}_i)$

Independent of n. Looks good?

Finite-Sum Stochastic Optimization

$$\min_{h \in \mathcal{H}} \hat{R}(h), \quad \hat{R}(h) = \frac{1}{n} \sum_{i=1}^{n} L(h(\mathbf{x}_i), y_i).$$

Hypothesis parameterized by $\mathbf{w}$

$$\min_{\mathbf{w} \in \Omega} F(\mathbf{w}), \quad F(\mathbf{w}) := \frac{1}{n} \sum_{\mathbf{z}_i \in \mathcal{D}} \ell(\mathbf{w}; \mathbf{z}_i)$$

Stochastic Gradient Descent

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla \ell(\mathbf{w}; \mathbf{z}_i)$$

Finite-Sum Stochastic Optimization

$$\min_{h \in \mathcal{H}} \hat{R}(h), \quad \hat{R}(h) = \frac{1}{n} \sum_{i=1}^{n} L(h(\mathbf{x}_i), y_i).$$

Hypothesis parameterized by **w**

$$\min_{\mathbf{w} \in \Omega} F(\mathbf{w}), \quad F(\mathbf{w}) := \frac{1}{n} \sum_{\mathbf{z}_i \in \mathcal{D}} \ell(\mathbf{w}; \mathbf{z}_i)$$

Stochastic Gradient Descent

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla \ell(\mathbf{w}; \mathbf{z}_i)$$

Could still be expensive (if not infeasible)!

Smooth Average Precision (AP) Maximization

$$F(\mathbf{w}) = -\frac{1}{|\mathcal{S}_+|}\sum_{\mathbf{x}_i\in\mathcal{S}_+}\frac{\sum_{\mathbf{x}\in\mathcal{S}_+}\ell(h_{\mathbf{w}}(\mathbf{x})-h_{\mathbf{w}}(\mathbf{x}_i))}{\sum_{\mathbf{x}\in\mathcal{S}}\ell(h_{\mathbf{w}}(\mathbf{x})-h_{\mathbf{w}}(\mathbf{x}_i))}$$

Smooth Average Precision (AP) Maximization

$$F(\mathbf{w}) = -\frac{1}{|\mathcal{S}_+|} \underbrace{\sum_{\mathbf{x}_i \in \mathcal{S}_+}}_{\text{Positive Sample}} \frac{\sum_{\mathbf{x} \in \mathcal{S}_+} \ell(h_{\mathbf{w}}(\mathbf{x}) - h_{\mathbf{w}}(\mathbf{x}_i))}{\sum_{\mathbf{x} \in \mathcal{S}} \ell(h_{\mathbf{w}}(\mathbf{x}) - h_{\mathbf{w}}(\mathbf{x}_i))}$$

Smooth Average Precision (AP) Maximization

$$F(\mathbf{w}) = -\frac{1}{|\mathcal{S}_+|} \sum_{\mathbf{x}_i \in \mathcal{S}_+} \frac{\sum_{\mathbf{x} \in \mathcal{S}_+} \ell(h_{\mathbf{w}}(\mathbf{x}) - h_{\mathbf{w}}(\mathbf{x}_i))}{\sum_{\mathbf{x} \in \mathcal{S}} \ell(h_{\mathbf{w}}(\mathbf{x}) - h_{\mathbf{w}}(\mathbf{x}_i))}$$

Positive Sample

Total Sample

$$\mathcal{S} = \mathcal{S}_+ \cup \mathcal{S}_-$$

Smooth Average Precision (AP) Maximization

$$\min_{\mathbf{w}\in\Omega} F(\mathbf{w}), \quad F(\mathbf{w}) := \frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} \ell(\mathbf{w};\mathbf{z}_i)$$

$$F(\mathbf{w}) = -\frac{1}{|\mathcal{S}_+|}\sum_{\mathbf{x}_i\in\mathcal{S}_+} \underbrace{\frac{\sum_{\mathbf{x}\in\mathcal{S}_+}\ell(h_{\mathbf{w}}(\mathbf{x})-h_{\mathbf{w}}(\mathbf{x}_i))}{\sum_{\mathbf{x}\in\mathcal{S}}\ell(h_{\mathbf{w}}(\mathbf{x})-h_{\mathbf{w}}(\mathbf{x}_i))}}_{\ell(\mathbf{w};\mathbf{z}_i)}$$

Smooth Average Precision (AP) Maximization

$$\min_{\mathbf{w}\in\Omega} F(\mathbf{w}), \quad F(\mathbf{w}) := \frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} \ell(\mathbf{w};\mathbf{z}_i)$$

$$F(\mathbf{w}) = -\frac{1}{|\mathcal{S}_+|}\sum_{\mathbf{x}_i\in\mathcal{S}_+} \frac{\sum_{\mathbf{x}\in\mathcal{S}_+}\ell(h_\mathbf{w}(\mathbf{x})-h_\mathbf{w}(\mathbf{x}_i))}{\sum_{\mathbf{x}\in\mathcal{S}}\ell(h_\mathbf{w}(\mathbf{x})-h_\mathbf{w}(\mathbf{x}_i))}$$

$$\ell(\mathbf{w};\mathbf{z}_i)$$

Stochastic Gradient Descent

$$\mathbf{w} \leftarrow \mathbf{w} - \eta\nabla\ell(\mathbf{w};\mathbf{z}_i)$$

Smooth Average Precision (AP) Maximization

$$\min_{\mathbf{w}\in\Omega} F(\mathbf{w}), \quad F(\mathbf{w}) := \frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} \ell(\mathbf{w};\mathbf{z}_i)$$

$$F(\mathbf{w}) = -\frac{1}{|\mathcal{S}_+|}\sum_{\mathbf{x}_i\in\mathcal{S}_+} \underbrace{\frac{\sum_{\mathbf{x}\in\mathcal{S}_+}\ell(h_{\mathbf{w}}(\mathbf{x})-h_{\mathbf{w}}(\mathbf{x}_i))}{\sum_{\mathbf{x}\in\mathcal{S}}\ell(h_{\mathbf{w}}(\mathbf{x})-h_{\mathbf{w}}(\mathbf{x}_i))}}_{\ell(\mathbf{w};\mathbf{z}_i)}$$

Stochastic Gradient Descent

$$\mathbf{w} \leftarrow \mathbf{w} - \eta\nabla\ell(\mathbf{w};\mathbf{z}_i)$$

Unbiased estimator is still expensive !

Robust Logistic Regression

$$\min_{\mathbf{w}} \frac{1}{n} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}} \left[\log\left(1 + \exp\left(-y_i \mathbb{E}_{\xi|\mathbf{x}_i}\left[\xi^T \mathbf{w}\right]\right)\right) \right]$$

Robust Logistic Regression

$$\min_{\mathbf{w}\in\Omega} F(\mathbf{w}), \quad F(\mathbf{w}) := \frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} \ell(\mathbf{w};\mathbf{z}_i)$$

$$\min_{\mathbf{w}} \frac{1}{n}\sum_{(\mathbf{x}_i,y_i)\in\mathcal{D}} \left[\log\left(1+\exp\left(-y_i\mathbb{E}_{\xi|\mathbf{x}_i}\left[\xi^T\mathbf{w}\right]\right)\right)\right]$$

$$\ell(\mathbf{w};\mathbf{z}_i)$$

Robust Logistic Regression

$$\min_{\mathbf{w}\in\Omega} F(\mathbf{w}), \quad F(\mathbf{w}) := \frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} \ell(\mathbf{w};\mathbf{z}_i)$$

$$\min_{\mathbf{w}} \frac{1}{n}\sum_{(\mathbf{x}_i,y_i)\in\mathcal{D}} \underbrace{\left[\log\left(1+\exp\left(-y_i\mathbb{E}_{\xi|\mathbf{x}_i}\left[\xi^T\mathbf{w}\right]\right)\right)\right]}_{\ell(\mathbf{w};\mathbf{z}_i)}$$

Stochastic Gradient Descent

$$\mathbf{w} \leftarrow \mathbf{w} - \eta\nabla\ell(\mathbf{w};\mathbf{z}_i)$$

Robust Logistic Regression

$$\min_{\mathbf{w}\in\Omega} F(\mathbf{w}), \quad F(\mathbf{w}) := \frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} \ell(\mathbf{w};\mathbf{z}_i)$$

$$\min_{\mathbf{w}} \frac{1}{n} \sum_{(\mathbf{x}_i, y_i)\in\mathcal{D}} \left[\log\left(1+\exp\left(-y_i \mathbb{E}_{\xi|\mathbf{x}_i}\left[\xi^T \mathbf{w}\right]\right)\right)\right]$$

$$\ell(\mathbf{w};\mathbf{z}_i)$$

Stochastic Gradient Descent

$$\mathbf{w} \leftarrow \mathbf{w} - \eta\nabla\ell(\mathbf{w};\mathbf{z}_i)$$

Infeasible !

Finite-Sum Coupled Compositional Stochastic Optimization (FCCO)

$$\min_{\mathbf{w}\in\Omega} F(\mathbf{w}),$$

$$F(\mathbf{w}) := \frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} f_i(g(\mathbf{w};\mathbf{z}_i,\mathcal{S}_i))$$

How is it related to finite-sum stochastic optimization?

Finite-Sum Coupled Compositional Stochastic Optimization (FCCO)

$$\min_{\mathbf{w}\in\Omega} F(\mathbf{w}),$$

$$F(\mathbf{w}) := \frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} f_i(g(\mathbf{w};\mathbf{z}_i,\mathcal{S}_i))$$

Finite-Sum Stochastic Optimization

$$\min_{\mathbf{w}\in\Omega} F(\mathbf{w}),$$

$$F(\mathbf{w}) := \frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} \ell(\mathbf{w};\mathbf{z}_i)$$

Finite-Sum Coupled Compositional Stochastic Optimization (FCCO)

$$\min_{\mathbf{w}\in\Omega} F(\mathbf{w}),$$

$$F(\mathbf{w}) := \frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} f_i(g(\mathbf{w};\mathbf{z}_i,\mathcal{S}_i))$$

Take into account the cost of S_i

Finite-Sum Stochastic Optimization

$$\min_{\mathbf{w}\in\Omega} F(\mathbf{w}),$$

$$F(\mathbf{w}) := \frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} \ell(\mathbf{w};\mathbf{z}_i)$$

Finite-Sum Coupled Compositional Stochastic Optimization (FCCO)

$$\min_{\mathbf{w}\in\Omega} F(\mathbf{w}),$$

$$F(\mathbf{w}) := \frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} f_i(g(\mathbf{w};\mathbf{z}_i,\mathcal{S}_i))$$

Stochastic Gradient (**Biased);**

Sample both $\mathcal{D}$ and $\mathcal{S}_i$

Finite-Sum Stochastic Optimization

$$\min_{\mathbf{w}\in\Omega} F(\mathbf{w}),$$

$$F(\mathbf{w}) := \frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} \ell(\mathbf{w};\mathbf{z}_i)$$

Stochastic Gradient (Unbiased);

Sample $\mathcal{D}$

$$F(\mathbf{w}) := \frac{1}{n} \sum_{\mathbf{z}_i \in \mathcal{D}} f_i(g(\mathbf{w}; \mathbf{z}_i, \mathcal{S}_i))$$

Finite-Sum Coupled Compositional Stochastic Optimization (FCCO)

- AP Maximization

$$F(\mathbf{w}) = -\frac{1}{|\mathcal{S}_+|} \sum_{\mathbf{x}_i \in \mathcal{S}_+} \frac{\sum_{\mathbf{x} \in \mathcal{S}_+} \ell(h_{\mathbf{w}}(\mathbf{x}) - h_{\mathbf{w}}(\mathbf{x}_i))}{\sum_{\mathbf{x} \in \mathcal{S}} \ell(h_{\mathbf{w}}(\mathbf{x}) - h_{\mathbf{w}}(\mathbf{x}_i))}$$

- Bipartite ranking by p-norm Push

$$F(\mathbf{w}) = \frac{1}{|\mathcal{S}_-|} \sum_{\mathbf{z}_i \in \mathcal{S}_-} \left(\frac{1}{|\mathcal{S}_+|} \sum_{\mathbf{z}_j \in \mathcal{S}_+} \ell(h_{\mathbf{w}}(\mathbf{z}_j) - h_{\mathbf{w}}(\mathbf{z}_i)) \right)^p$$

$$F(\mathbf{w}) := \frac{1}{n} \sum_{\mathbf{z}_i \in \mathcal{D}} f_i(g(\mathbf{w}; \mathbf{z}_i, \mathcal{S}_i))$$

Finite-Sum Coupled Compositional Stochastic Optimization (FCCO)

- AP Maximization

$$F(\mathbf{w}) = -\frac{1}{|\mathcal{S}_+|} \sum_{\mathbf{x}_i \in \mathcal{S}_+} \frac{\sum_{\mathbf{x} \in \mathcal{S}_+} \ell(h_{\mathbf{w}}(\mathbf{x}) - h_{\mathbf{w}}(\mathbf{x}_i))}{\sum_{\mathbf{x} \in \mathcal{S}} \ell(h_{\mathbf{w}}(\mathbf{x}) - h_{\mathbf{w}}(\mathbf{x}_i))}$$

- Bipartite ranking by p-norm Push

$$F(\mathbf{w}) = \frac{1}{|\mathcal{S}_-|} \sum_{\mathbf{z}_i \in \mathcal{S}_-} \left(\frac{1}{|\mathcal{S}_+|} \sum_{\mathbf{z}_j \in \mathcal{S}_+} \ell(h_{\mathbf{w}}(\mathbf{z}_j) - h_{\mathbf{w}}(\mathbf{z}_i)) \right)^p$$

$$F(\mathbf{w}) := \frac{1}{n} \sum_{\mathbf{z}_i \in \mathcal{D}} \ell(\mathbf{w}; \mathbf{z}_i)$$

Finite-Sum Stochastic Optimization

- Logistic regression

$$F(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} \ln\left(1 + e^{-y_i \langle \mathbf{w}, \mathbf{x}_i \rangle}\right)$$

- Ridge regression

$$F(\mathbf{w}) = \frac{1}{2n} \sum_{i=1}^{n} \left\| \mathbf{x}_i^\top \mathbf{w} - y_i \right\|_2^2 + \frac{\lambda}{2} \|\mathbf{w}\|_2^2$$

Finite-Sum Coupled Compositional Stochastic Optimization (FCCO)

$$\min_{\mathbf{w}\in\Omega} F(\mathbf{w}),$$

$$F(\mathbf{w}) := \frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} f_i(g(\mathbf{w};\mathbf{z}_i,\mathcal{S}_i))$$

Wait ! We have already seen something similar ...

Hu et al. "Biased stochastic first-order methods for conditional stochastic optimization and applications in meta learning." NeurIPS 2020.

Finite-Sum Coupled Compositional Stochastic Optimization (FCCO)

$$\min_{\mathbf{w} \in \Omega} F(\mathbf{w}),$$

$$F(\mathbf{w}) := \frac{1}{n} \sum_{\mathbf{z}_i \in \mathcal{D}} f_i(g(\mathbf{w}; \mathbf{z}_i, \mathcal{S}_i))$$

Conditional Stochastic Optimization (CSO)

$$\min_{\mathbf{w} \in \Omega} F(\mathbf{w}),$$

$$F(\mathbf{w}) = \mathbb{E}_\xi f_\xi \left(\mathbb{E}_{\zeta|\xi} [g_\zeta(\mathbf{w}; \xi)] \right)$$

Hu et al. "Biased stochastic first-order methods for conditional stochastic optimization and applications in meta learning." NeurIPS 2020.

Finite-Sum Coupled Compositional Stochastic Optimization (FCCO)

Conditional Stochastic Optimization (CSO)

Special Case:

$$\min_{\mathbf{w}\in\Omega} F(\mathbf{w}),$$

$$F(\mathbf{w}) := \frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} f_i(g(\mathbf{w};\mathbf{z}_i,\mathcal{S}_i))$$

Outer problem has finite support

$$\min_{\mathbf{w}\in\Omega} F(\mathbf{w}),$$

$$F(\mathbf{w}) = \mathbb{E}_\xi f_\xi\left(\mathbb{E}_{\zeta|\xi}[g_\zeta(\mathbf{w};\xi)]\right)$$

This could be exploited for better rates !

Wang et al. "Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions." Math. Program. 161(1-2):419–449, 2017.

Finite-Sum Coupled Compositional Stochastic Optimization (FCCO)

$$\min_{\mathbf{w}\in\Omega} F(\mathbf{w}),$$

$$F(\mathbf{w}) := \frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} f_i(g(\mathbf{w};\mathbf{z}_i,\mathcal{S}_i))$$

Finite-Sum Compositional Stochastic Optimization (FCO)

$$\min_{\mathbf{w}\in\Omega} F(\mathbf{w}),$$

$$F(\mathbf{w}) = \frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} f_i(g(\mathbf{w};\mathcal{S}))$$

Wang et al. "Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions." Math. Program. 161(1-2):419–449, 2017.

Finite-Sum Coupled Compositional Stochastic Optimization (FCCO)

$$\min_{\mathbf{w}\in\Omega} F(\mathbf{w}),$$

$$F(\mathbf{w}) := \frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} f_i(g(\mathbf{w};\mathbf{z}_i,\mathcal{S}_i))$$

Coupled

Finite-Sum Compositional Stochastic Optimization (FCO)

$$\min_{\mathbf{w}\in\Omega} F(\mathbf{w}),$$

$$F(\mathbf{w}) = \frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} f_i(g(\mathbf{w};\mathcal{S}))$$

Wang et al. "Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions." Math. Program. 161(1-2):419–449, 2017.

Finite-Sum Coupled Compositional Stochastic Optimization (FCCO)

$$F(\mathbf{w}) := \frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} f_i(g(\mathbf{w};\mathbf{z}_i,\mathcal{S}_i))$$

How about reformulating FCCO as FCO?

Finite-Sum Compositional Stochastic Optimization (FCO)

$$F(\mathbf{w}) = \frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} f_i(g(\mathbf{w};\mathcal{S}))$$

Wang et al. "Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions." Math. Program. 161(1-2):419–449, 2017.

Finite-Sum Coupled Compositional Stochastic Optimization (FCCO)

Finite-Sum Compositional Stochastic Optimization (FCO)

$$F(\mathbf{w}) := \frac{1}{n}\sum_{\mathbf{z}_i \in \mathcal{D}} f_i(g(\mathbf{w}; \mathbf{z}_i, \mathcal{S}_i))$$

Doable. But the FCCO formulation leads to more efficient algorithm !

$$F(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_i(\mathbf{g}(\mathbf{w}; \mathcal{S}))$$

$$\mathbf{g}(\mathbf{w}; \mathcal{S}) = \left[g(\mathbf{w}; \mathbf{z}_1, \mathcal{S}_1)^\top, \ldots, g(\mathbf{w}; \mathbf{z}_n, \mathcal{S}_n)^\top\right]^\top$$

$$\mathcal{S} = \mathcal{S}_1 \cup \cdots \mathcal{S}_i \cdots \cup \mathcal{S}_n$$

$$\hat{f}_i(\cdot) = f_i(\mathbb{I}_i \cdot) \quad \mathbb{I}_i := [0_{d\times d}, \ldots, I_{d\times d}, \ldots, 0_{d\times d}]$$

Algorithm and Theory

Ghadimi et al. "A single timescale stochastic approximation method for nested stochastic optimization." SIAM J. Optim., 30:960–979,2020.

The NASA Algorithm for Finite-Sum Compositional Stochastic Optimization (FCO)

$$F(\mathbf{w}) = \frac{1}{n}\sum_{\mathbf{z}_i \in \mathcal{D}} f_i(g(\mathbf{w}; \mathcal{S}))$$

Sample mini-batches $\mathcal{B}_1 \subset \mathcal{D}, \mathcal{B}_2 \subset \mathcal{S}$

$$u \leftarrow (1-\gamma)u + \gamma g(\mathbf{w}; \mathcal{B}_2)$$

$$\mathbf{v} \leftarrow (1-\beta)\mathbf{v} + \beta \frac{1}{|\mathcal{B}_1|}\sum_{\mathbf{z}_i \in \mathcal{B}_1} \nabla g(\mathbf{w}; \mathcal{B}_2)\nabla f_i(u)$$

$$\mathbf{w} \longleftarrow \mathbf{w} - \eta\mathbf{v}$$

Apply NASA to Finite-Sum Coupled Compositional Stochastic Optimization?

$$F(\mathbf{w}) := \frac{1}{n}\sum_{\mathbf{z}_i \in \mathcal{D}} f_i(g(\mathbf{w}; \mathbf{z}_i, \mathcal{S}_i))$$

reformulation

$$F(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_i(\mathbf{g}(\mathbf{w}; \mathcal{S}))$$

$$\mathbf{g}(\mathbf{w}; \mathcal{S}) = \left[g(\mathbf{w}; \mathbf{z}_1, \mathcal{S}_1)^\top, \ldots, g(\mathbf{w}; \mathbf{z}_n, \mathcal{S}_n)^\top\right]^\top$$

$$\mathcal{S} = \mathcal{S}_1 \cup \cdots \mathcal{S}_i \cdots \cup \mathcal{S}_n$$

$$\hat{f}_i(\cdot) = f_i(\mathbb{I}_i \cdot) \quad \mathbb{I}_i := [0_{d\times d}, \ldots, I_{d\times d}, \ldots, 0_{d\times d}]$$

$$u \leftarrow (1-\gamma)u + \gamma g(\mathbf{w}; \mathcal{B}_2)$$

All n components of u are updated in every iteration!

Apply NASA to Finite-Sum Coupled Compositional Stochastic Optimization?

$$F(\mathbf{w}) := \frac{1}{n} \sum_{\mathbf{z}_i \in \mathcal{D}} f_i(g(\mathbf{w}; \mathbf{z}_i, \mathcal{S}_i))$$

reformulation

$$F(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}_i(\mathbf{g}(\mathbf{w}; \mathcal{S}))$$

$$\mathbf{g}(\mathbf{w}; \mathcal{S}) = \left[g(\mathbf{w}; \mathbf{z}_1, \mathcal{S}_1)^\top, \ldots, g(\mathbf{w}; \mathbf{z}_n, \mathcal{S}_n)^\top\right]^\top$$

$$\mathcal{S} = \mathcal{S}_1 \cup \cdots \mathcal{S}_i \cdots \cup \mathcal{S}_n$$

$$\hat{f}_i(\cdot) = f_i(\mathbb{I}_i \cdot) \quad \mathbb{I}_i := [0_{d\times d}, \ldots, I_{d\times d}, \ldots, 0_{d\times d}]$$

$$u \leftarrow (1-\gamma)u + \gamma g(\mathbf{w}; \mathcal{B}_2)$$

All n components of u are updated in every iteration!

Not efficient when n is large.

Finite-Sum Coupled Compositional Stochastic Optimization (FCCO)

(NEW) The SOX Algorithm

$$F(\mathbf{w}) := \frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} f_i(g(\mathbf{w};\mathbf{z}_i,\mathcal{S}_i))$$

Sample mini-batches $\mathcal{B}_1^t \subset \mathcal{D}, \mathcal{B}_{i,2}^t \subset \mathcal{S}_i$

$$u_i^t = \begin{cases} (1-\gamma)u_i^{t-1} + \gamma g\left(\mathbf{w}^t;\mathbf{z}_i,\mathcal{B}_{i,2}^t\right), & \mathbf{z}_i \in \mathcal{B}_1^t \\ u_i^{t-1}, & \mathbf{z}_i \notin \mathcal{B}_1^t \end{cases}$$

$$\mathbf{v}^t = (1-\beta)\mathbf{v}^{t-1} + \beta\frac{1}{B_1}\sum_{\mathbf{z}_i\in\mathcal{B}_1^t}\nabla g\left(\mathbf{w}^t;\mathbf{z}_i,\mathcal{B}_{i,2}^t\right)\nabla f_i\left(u_i^{t-1}\right)$$

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \eta_t\mathbf{v}^t$$

Finite-Sum Coupled Compositional Stochastic Optimization (FCCO)

(NEW) The SOX Algorithm

$$F(\mathbf{w}) := \frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} f_i(g(\mathbf{w};\mathbf{z}_i,\mathcal{S}_i))$$

Sample mini-batches $\mathcal{B}_1^t \subset \mathcal{D}, \mathcal{B}_{i,2}^t \subset \mathcal{S}_i$

$$u_i^t = \begin{cases}(1-\gamma)u_i^{t-1} + \gamma g\left(\mathbf{w}^t;\mathbf{z}_i,\mathcal{B}_{i,2}^t\right), & \mathbf{z}_i\in\mathcal{B}_1^t\\ u_i^{t-1}, & \mathbf{z}_i\notin\mathcal{B}_1^t\end{cases}$$

Only update those sampled components in the outer loop !

$$\mathbf{v}^t = (1-\beta)\mathbf{v}^{t-1} + \beta\frac{1}{B_1}\sum_{\mathbf{z}_i\in\mathcal{B}_1^t}\nabla g\left(\mathbf{w}^t;\mathbf{z}_i,\mathcal{B}_{i,2}^t\right)\nabla f_i\left(u_i^{t-1}\right)$$

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \eta_t\mathbf{v}^t$$

Finite-Sum Coupled Compositional Stochastic Optimization (FCCO)

(NEW) The SOX Algorithm

$$F(\mathbf{w}) := \frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} f_i(g(\mathbf{w};\mathbf{z}_i,\mathcal{S}_i))$$

Sample mini-batches $\mathcal{B}_1^t \subset \mathcal{D}, \mathcal{B}_{i,2}^t \subset \mathcal{S}_i$

$$u_i^t = \begin{cases}(1-\gamma)u_i^{t-1} + \gamma g\big(\mathbf{w}^t;\mathbf{z}_i,\mathcal{B}_{i,2}^t\big), & \mathbf{z}_i \in \mathcal{B}_1^t\\ u_i^{t-1}, & \mathbf{z}_i \notin \mathcal{B}_1^t\end{cases}$$

$$\mathbf{v}^t = (1-\beta)\mathbf{v}^{t-1} + \beta\frac{1}{B_1}\sum_{\mathbf{z}_i\in\mathcal{B}_1^t}\nabla g\big(\mathbf{w}^t;\mathbf{z}_i,\mathcal{B}_{i,2}^t\big)\nabla f_i\big(u_i^{t-1}\big)$$

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \eta_t\mathbf{v}^t$$

Only update those sampled components in the outer loop !

Data sampling is independent of n (unlike NASA)

Finite-Sum Coupled Compositional Stochastic Optimization (FCCO)

$$F(\mathbf{w}) := \frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} f_i(g(\mathbf{w};\mathbf{z}_i,\mathcal{S}_i))$$

(NEW) The SOX Algorithm

$$u_i^t = \begin{cases} (1-\gamma)u_i^{t-1} + \gamma g(\mathbf{w}^t;\mathbf{z}_i,\mathcal{B}_{i,2}^t), & \mathbf{z}_i \in \mathcal{B}_1^t \\ u_i^{t-1}, & \mathbf{z}_i \notin \mathcal{B}_1^t \end{cases}$$

$$\Updownarrow$$

$$u_i^t = \begin{cases} u_i^{t-1} - \gamma\left(u_i^{t-1} - g(\mathbf{w}^t;\mathbf{z}_i,\mathcal{B}_{i,2}^t)\right), & \mathbf{z}_i \in \mathcal{B}_1^t \\ u_i^t, & \mathbf{z}_i \notin \mathcal{B}_1^t \end{cases}$$

$$\min_{\mathbf{u}=[u_1,\ldots,u_n]^\top} \frac{1}{2}\sum_{\mathbf{z}_i\in\mathcal{D}} \left\| u_i - g(\mathbf{w}^t;\mathbf{z}_i,\mathcal{S}_i)\right\|^2$$

Stochastic block coordinate descent

Finite-Sum Coupled Compositional Stochastic Optimization (FCCO)

(NEW) The SOX Algorithm

$$F(\mathbf{w}) := \frac{1}{n}\sum_{\mathbf{z}_i\in\mathcal{D}} f_i(g(\mathbf{w};\mathbf{z}_i,\mathcal{S}_i))$$

Sample mini-batches $\mathcal{B}_1^t \subset \mathcal{D}, \mathcal{B}_{i,2}^t \subset \mathcal{S}_i$

$$u_i^t = \begin{cases} (1-\gamma)u_i^{t-1} + \gamma g\left(\mathbf{w}^t;\mathbf{z}_i,\mathcal{B}_{i,2}^t\right), & \mathbf{z}_i \in \mathcal{B}_1^t \\ u_i^{t-1}, & \mathbf{z}_i \notin \mathcal{B}_1^t \end{cases}$$

$$\mathbf{v}^t = (1-\beta)\mathbf{v}^{t-1} + \beta\frac{1}{B_1}\sum_{\mathbf{z}_i\in\mathcal{B}_1^t}\nabla g(\mathbf{w}^t;\mathbf{z}_i,\mathcal{B}_{i,2}^t)\nabla f_i(u_i^{t-1})$$

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \eta_t\mathbf{v}^t$$

u_i^t is more intuitive (and also works in practice).

Finite-Sum Coupled Compositional Stochastic Optimization (FCCO)

(NEW) The SOX Algorithm

$$F(\mathbf{w}) := \frac{1}{n}\sum_{\mathbf{z}_i \in \mathcal{D}} f_i(g(\mathbf{w}; \mathbf{z}_i, \mathcal{S}_i))$$

Sample mini-batches $\mathcal{B}_1^t \subset \mathcal{D}, \mathcal{B}_{i,2}^t \subset \mathcal{S}_i$

$$u_i^t = \begin{cases} (1-\gamma)u_i^{t-1} + \gamma g\left(\mathbf{w}^t; \mathbf{z}_i, \mathcal{B}_{i,2}^t\right), & \mathbf{z}_i \in \mathcal{B}_1^t \\ u_i^{t-1}, & \mathbf{z}_i \notin \mathcal{B}_1^t \end{cases}$$

$$\mathbf{v}^t = (1-\beta)\mathbf{v}^{t-1} + \beta \frac{1}{B_1} \sum_{\mathbf{z}_i \in \mathcal{B}_1^t} \nabla g\left(\mathbf{w}^t; \mathbf{z}_i, \mathcal{B}_{i,2}^t\right) \nabla f_i\left(u_i^{t-1}\right)$$

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \eta_t \mathbf{v}^t$$

u_i^t is more intuitive (and also works in practice).

But u_i^{t-1} can help us circumvent some difficulty in theory and derive improved rates.

Convergence Rates

Method	NC	C	SC (PL)	Outer Batch Size $\lvert\mathcal{B}_1\rvert$	Inner Batch Size $\lvert\mathcal{B}_{i,2}\rvert$	Parallel Speed-up
BSGD (Hu et al., 2020)	$O(\epsilon^{-4})$	$O\left(\epsilon^{-2}\right)$	$O\left(\mu^{-1}\epsilon^{-1}\right)^{\dagger}$	1	$O(\epsilon^{-2})$ (NC) $O(\epsilon^{-1})$ (C/SC)	N/A
SOAP (Qi et al., 2021)	$O(n\epsilon^{-5})$	-	-	1	1	N/A
MOAP (Wang et al., 2021)	$O\left(\frac{n\epsilon^{-4}}{B_1}\right)$	-	-	B_1	1	Partial
SOX/SOX-boost (this work)	$O\left(\frac{n\epsilon^{-4}}{B_1 B_2}\right)$	$O\left(\frac{n\epsilon^{-3}}{B_1 B_2}\right)$	$O\left(\frac{n\mu^{-2}\epsilon^{-1}}{B_1 B_2}\right)$	B_1	B_2	Yes
SOX ($\beta = 1$) (this work)	-	$O\left(\frac{n\epsilon^{-2}}{B_1}\right)^{*}$	-	B_1	B_2	Partial

Convergence Rates

Nonconvex

Method	NC	C	SC (PL)	Outer Batch Size $\|\mathcal{B}_1\|$	Inner Batch Size $\|\mathcal{B}_{i,2}\|$	Parallel Speed-up
BSGD (Hu et al., 2020)	$O(\epsilon^{-4})$	$O(\epsilon^{-2})$	$O(\mu^{-1}\epsilon^{-1})^{\dagger}$	1	$O(\epsilon^{-2})$ (NC) $O(\epsilon^{-1})$ (C/SC)	N/A
SOAP (Qi et al., 2021)	$O(n\epsilon^{-5})$	-	-	1	1	N/A
MOAP (Wang et al., 2021)	$O\left(\frac{n\epsilon^{-4}}{B_1}\right)$	-	-	B_1	1	Partial
SOX/SOX-boost (this work)	$O\left(\frac{n\epsilon^{-4}}{B_1B_2}\right)$	$O\left(\frac{n\epsilon^{-3}}{B_1B_2}\right)$	$O\left(\frac{n\mu^{-2}\epsilon^{-1}}{B_1B_2}\right)$	B_1	B_2	Yes
SOX ($\beta=1$) (this work)	-	$O\left(\frac{n\epsilon^{-2}}{B_1}\right)^{*}$	-	B_1	B_2	Partial

Convergence Rates

Nonconvex Convex

Method	NC	C	SC (PL)	Outer Batch Size $\lvert\mathcal{B}_1\rvert$	Inner Batch Size $\lvert\mathcal{B}_{i,2}\rvert$	Parallel Speed-up
BSGD (Hu et al., 2020)	$O(\epsilon^{-4})$	$O\left(\epsilon^{-2}\right)$	$O\left(\mu^{-1}\epsilon^{-1}\right)^{\dagger}$	1	$O(\epsilon^{-2})$ (NC) $O(\epsilon^{-1})$ (C/SC)	N/A
SOAP (Qi et al., 2021)	$O(n\epsilon^{-5})$	-	-	1	1	N/A
MOAP (Wang et al., 2021)	$O\left(\frac{n\epsilon^{-4}}{B_1}\right)$	-	-	B_1	1	Partial
SOX/SOX-boost (this work)	$O\left(\frac{n\epsilon^{-4}}{B_1 B_2}\right)$	$O\left(\frac{n\epsilon^{-3}}{B_1 B_2}\right)$	$O\left(\frac{n\mu^{-2}\epsilon^{-1}}{B_1 B_2}\right)$	B_1	B_2	Yes
SOX ($\beta = 1$) (this work)	-	$O\left(\frac{n\epsilon^{-2}}{B_1}\right)^{*}$	-	B_1	B_2	Partial

Convergence Rates

Nonconvex Convex Strongly Convex

Method	NC	C	SC (PL)	Outer Batch Size $\|\mathcal{B}_1\|$	Inner Batch Size $\|\mathcal{B}_{i,2}\|$	Parallel Speed-up
BSGD (Hu et al., 2020)	$O(\epsilon^{-4})$	$O\left(\epsilon^{-2}\right)$	$O\left(\mu^{-1}\epsilon^{-1}\right)^{\dagger}$	1	$O(\epsilon^{-2})$ (NC) $O(\epsilon^{-1})$ (C/SC)	N/A
SOAP (Qi et al., 2021)	$O(n\epsilon^{-5})$	-	-	1	1	N/A
MOAP (Wang et al., 2021)	$O\left(\frac{n\epsilon^{-4}}{B_1}\right)$	-	-	B_1	1	Partial
SOX/SOX-boost (this work)	$O\left(\frac{n\epsilon^{-4}}{B_1 B_2}\right)$	$O\left(\frac{n\epsilon^{-3}}{B_1 B_2}\right)$	$O\left(\frac{n\mu^{-2}\epsilon^{-1}}{B_1 B_2}\right)$	B_1	B_2	Yes
SOX ($\beta = 1$) (this work)	-	$O\left(\frac{n\epsilon^{-2}}{B_1}\right)^{*}$	-	B_1	B_2	Partial

“Twice batch size, half #iterations”

Convergence Rates

Nonconvex

Convex

Strongly Convex

Method	NC	C	SC (PL)	Outer Batch Size $\lvert\mathcal{B}_1\rvert$	Inner Batch Size $\lvert\mathcal{B}_{i,2}\rvert$	Parallel Speed-up
BSGD (Hu et al., 2020)	$O(\epsilon^{-4})$	$O\left(\epsilon^{-2}\right)$	$O\left(\mu^{-1}\epsilon^{-1}\right)^{\dagger}$	1	$O(\epsilon^{-2})$ (NC) $O(\epsilon^{-1})$ (C/SC)	N/A
SOAP (Qi et al., 2021)	$O(n\epsilon^{-5})$	-	-	1	1	N/A
MOAP (Wang et al., 2021)	$O\left(\frac{n\epsilon^{-4}}{B_1}\right)$	-	-	B_1	1	Partial
SOX/SOX-boost (this work)	$O\left(\frac{n\epsilon^{-4}}{B_1 B_2}\right)$	$O\left(\frac{n\epsilon^{-3}}{B_1 B_2}\right)$	$O\left(\frac{n\mu^{-2}\epsilon^{-1}}{B_1 B_2}\right)$	B_1	B_2	Yes
SOX ($\beta = 1$) (this work)	-	$O\left(\frac{n\epsilon^{-2}}{B_1}\right)^{*}$	-	B_1	B_2	Partial

"Twice batch size, half #iterations"

Convergence Rates

Nonconvex

Convex

Strongly Convex

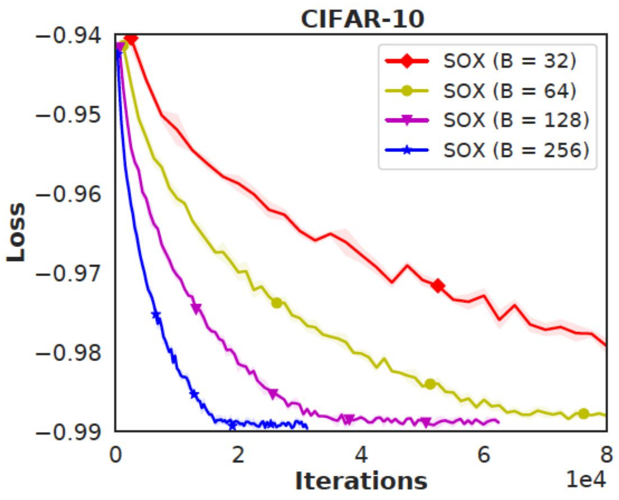
Method	NC	C	SC (PL)	Outer Batch Size $\|\mathcal{B}_1\|$	Inner Batch Size $\|\mathcal{B}_{i,2}\|$	Parallel Speed-up
BSGD (Hu et al., 2020)	$O(\epsilon^{-4})$	$O\left(\epsilon^{-2}\right)$	$O\left(\mu^{-1}\epsilon^{-1}\right)^{\dagger}$	1	$O(\epsilon^{-2})$ (NC) $O(\epsilon^{-1})$ (C/SC)	N/A
SOAP (Qi et al., 2021)	$O(n\epsilon^{-5})$	-	-	1	1	N/A
MOAP (Wang et al., 2021)	$O\left(\frac{n\epsilon^{-4}}{B_1}\right)$	-	-	B_1	1	Partial
SOX/SOX-boost (this work)	$O\left(\frac{n\epsilon^{-4}}{B_1 B_2}\right)$	$O\left(\frac{n\epsilon^{-3}}{B_1 B_2}\right)$	$O\left(\frac{n\mu^{-2}\epsilon^{-1}}{B_1 B_2}\right)$	B_1	B_2	Yes
SOX ($\beta = 1$) (this work)	-	$O\left(\frac{n\epsilon^{-2}}{B_1}\right)^{*}$	-	B_1	B_2	Partial

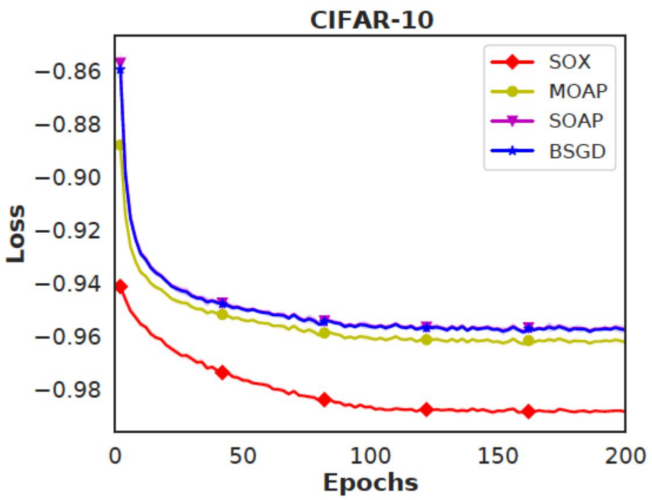
Originally proposed for AP maximization

Numerical Results

AP Maximization

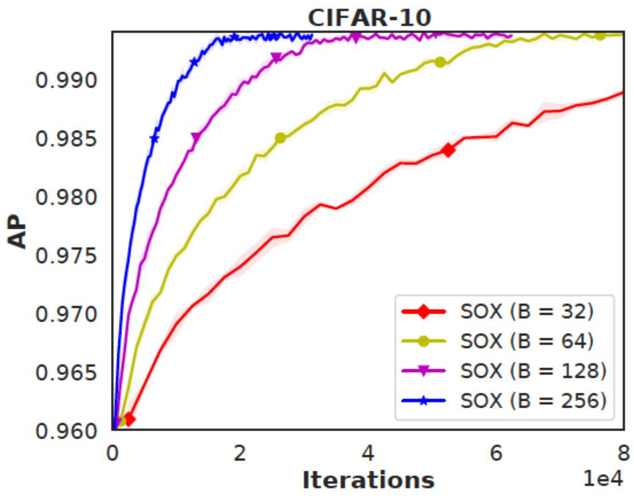
$$F(\mathbf{w}) = -\frac{1}{|\mathcal{S}_+|}\sum_{\mathbf{x}_i \in \mathcal{S}_+} \frac{\sum_{\mathbf{x}\in\mathcal{S}_+} \ell(h_{\mathbf{w}}(\mathbf{x}) - h_{\mathbf{w}}(\mathbf{x}_i))}{\sum_{\mathbf{x}\in\mathcal{S}} \ell(h_{\mathbf{w}}(\mathbf{x}) - h_{\mathbf{w}}(\mathbf{x}_i))}$$

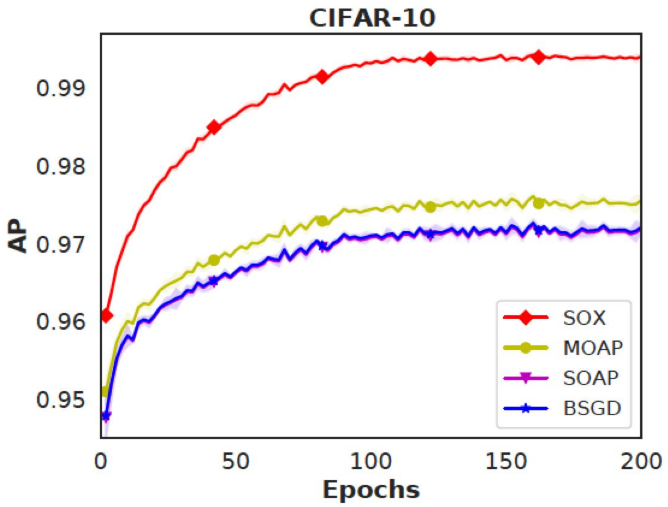




AP Maximization

$$F(\mathbf{w}) = -\frac{1}{|\mathcal{S}_+|}\sum_{\mathbf{x}_i\in\mathcal{S}_+}\frac{\sum_{\mathbf{x}\in\mathcal{S}_+}\ell(h_{\mathbf{w}}(\mathbf{x}) - h_{\mathbf{w}}(\mathbf{x}_i))}{\sum_{\mathbf{x}\in\mathcal{S}}\ell(h_{\mathbf{w}}(\mathbf{x}) - h_{\mathbf{w}}(\mathbf{x}_i))}$$





Bipartite Ranking by p-norm Push

$$F(\mathbf{w}) = \frac{1}{|\mathcal{S}_-|} \sum_{\mathbf{z}_i \in \mathcal{S}_-} \left(\frac{1}{|\mathcal{S}_+|} \sum_{\mathbf{z}_j \in \mathcal{S}_+} \ell(h_{\mathbf{w}}(\mathbf{z}_j) - h_{\mathbf{w}}(\mathbf{z}_i)) \right)^p$$

covtype					
Algorithms	BS-PnP	BSGD	SOAP	MOAP	SOX
Test Loss (↓)	0.778	0.625 ± 0.018	0.523 ± 0.004	0.559 ± 0.011	$\mathbf{0.516 \pm 0.003}$
Time (s) (↓)	6043.90	$\mathbf{4.20 \pm 0.08}$	4.32 ± 0.15	4.89 ± 0.06	4.62 ± 0.10
ijcnn1					
Algorithms	BS-PnP	BSGD	SOAP	MOAP	SOX
Test Loss (↓)	0.268	0.202 ± 0.001	$\mathbf{0.128 \pm 0.002}$	0.147 ± 0.001	$\mathbf{0.128 \pm 0.002}$
Time (s) (↓)	648.06	$\mathbf{4.02 \pm 0.04}$	4.04 ± 0.11	4.42 ± 0.05	4.15 ± 0.06

Bipartite Ranking by p-norm Push

$$F(\mathbf{w}) = \frac{1}{|\mathcal{S}_-|} \sum_{\mathbf{z}_i \in \mathcal{S}_-} \left(\frac{1}{|\mathcal{S}_+|} \sum_{\mathbf{z}_j \in \mathcal{S}_+} \ell(h_{\mathbf{w}}(\mathbf{z}_j) - h_{\mathbf{w}}(\mathbf{z}_i)) \right)^p$$

A boosting-style deterministic algorithm

covtype					
Algorithms	BS-PnP	BSGD	SOAP	MOAP	SOX
Test Loss (↓)	0.778	0.625 ± 0.018	0.523 ± 0.004	0.559 ± 0.011	$\mathbf{0.516 \pm 0.003}$
Time (s) (↓)	6043.90	$\mathbf{4.20 \pm 0.08}$	4.32 ± 0.15	4.89 ± 0.06	4.62 ± 0.10

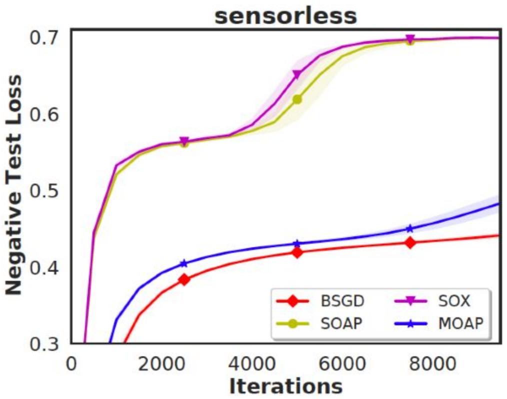
ijcnn1					
Algorithms	BS-PnP	BSGD	SOAP	MOAP	SOX
Test Loss (↓)	0.268	0.202 ± 0.001	$\mathbf{0.128 \pm 0.002}$	0.147 ± 0.001	$\mathbf{0.128 \pm 0.002}$
Time (s) (↓)	648.06	$\mathbf{4.02 \pm 0.04}$	4.04 ± 0.11	4.42 ± 0.05	4.15 ± 0.06

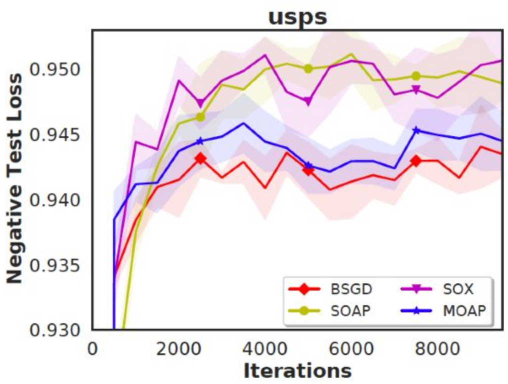
Neighborhood Component Analysis

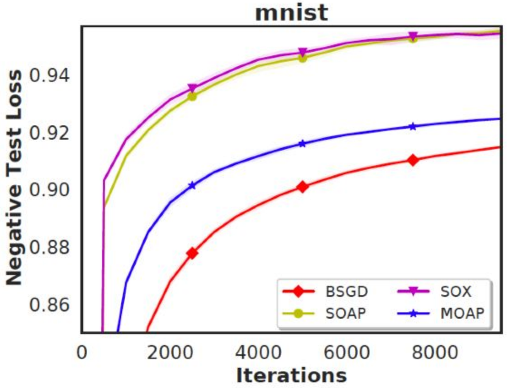
$$F(A) = -\sum_{\mathbf{x}_i \in \mathcal{D}} \frac{\sum_{\mathbf{x} \in \mathcal{C}_i} \exp(-\|A\mathbf{x}_i - A\mathbf{x}\|^2)}{\sum_{\mathbf{x} \in \mathcal{S}_i} \exp(-\|A\mathbf{x}_i - A\mathbf{x}\|^2)}$$

$$\mathcal{C}_i = \{\mathbf{x}_j \in \mathcal{D} : y_j = y_i\}$$

$$\mathcal{S}_i = \mathcal{D} \setminus \{\mathbf{x}_i\}$$







More Potential Applications

Listwise Ranking

$$F(\mathbf{w}) = -\sum_{q} \sum_{\mathbf{x}_i^q \in \mathcal{S}_q} P(y_i^q) \log \frac{\exp(h_{\mathbf{w}}(\mathbf{x}_i^q; \mathbf{q}))}{\sum_{\mathbf{x} \in \mathcal{S}_q} \exp(h_{\mathbf{w}}(\mathbf{x}; \mathbf{q}))}$$

$$\text{queries } \mathcal{Q} = \{\mathbf{q}_1, \ldots, \mathbf{q}_n\}$$

$$\text{items with relavance scores} \quad \mathcal{S}_q = \left\{(\mathbf{x}_1^q, y_1^q), \ldots, \left(\mathbf{x}_{n_q}^q, y_{n_q}^q\right)\right\}$$

Survival Analysis

$$F(\mathbf{w}) = \frac{1}{n} \sum_{i:E_i=1} \log \left(\sum_{j \in \mathcal{S}(T_i)} \exp(h_{\mathbf{w}}(\mathbf{x}_j) - h_{\mathbf{w}}(\mathbf{x}_i)) \right)$$

$\mathbf{x}_i$: patient feature $\quad$ $h_{\mathbf{w}}(\mathbf{x}_i)$: risk predicted by the model

$E_i = 1$: observable event of interest (e.g., death)

T_i : time interval between data collection and the event

$\mathcal{S}(t) = \{i : T_i \geq t\}$ denotes the set of patients still at risk at time t

Thank you !