Finite-Sum Coupled Compositional Stochastic Optimization: Theory and Applications

Bokun Wang and Tianbao Yang

bokun-wang.github.io github.com/bokun-wang/SOX

ICML International Conference On Machine Learning

IOWA

The Problem

Minimize a sum of compositions: inner-level function of each summand **coupled** with outer index.

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

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

Coupled

Finite-Sum **C**oupled **C**ompositional **O**ptimization (**FCCO**)

Applications

1. Average precision maximization

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

$h_{\mathbf{w}}(\mathbf{z})$: model prediction; $\mathcal{S}_+$: positive data; $\mathcal{S}$: whole data; $\ell(\cdot)$: monotonically decreasing upper bound on the 0-1 loss (e.g. hinge loss).

$$f_i(\cdot) = -\frac{[\cdot]_1}{[\cdot]_2},\quad g(\mathbf{w};\mathbf{z}_i,\mathcal{S}) = \begin{bmatrix}\sum_{\mathbf{z}\in\mathcal{S}_+}\ell(h_{\mathbf{w}}(\mathbf{z})-h_{\mathbf{w}}(\mathbf{z}_i))\\ \sum_{\mathbf{z}\in\mathcal{S}}\ell(h_{\mathbf{w}}(\mathbf{z})-h_{\mathbf{w}}(\mathbf{z}_i))\end{bmatrix}.$$

2. Ranking by p-norm push

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

$p>1$: order; $\mathcal{S}_-$: negative data; others same as AP maximization.

$$f_i(\cdot) = (\cdot)^p,\quad g(\mathbf{w};\mathbf{z}_i,\mathcal{S}_+) = \textstyle\sum_{\mathbf{z}_j\in\mathcal{S}_+}\ell\left(h_{\mathbf{w}}(\mathbf{z}_j)-h_{\mathbf{w}}(\mathbf{z}_i)\right).$$

3. Neighborhood component analysis

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

A: embedding matrix; $\mathcal{D}$: data set; $\mathcal{C}_i$: nearest-neighbour set of $\mathbf{z}_i$; $\mathcal{S}_i$: leave-one-out set of $\mathbf{z}_i$.

$$f_i(\cdot) = -\frac{[\cdot]_1}{[\cdot]_2},\quad g(A;\mathbf{z}_i,\mathcal{S}_i) = \begin{bmatrix}\sum_{\mathbf{z}\in\mathcal{C}_i}\exp(-\|A\mathbf{z}_i-A\mathbf{z}\|^2)\\ \sum_{\mathbf{z}\in\mathcal{S}_i}\exp(-\|A\mathbf{z}_i-A\mathbf{z}\|^2)\end{bmatrix}.$$

4. Listwise Ranking, e.g. ListNet

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

$h_{\mathbf{w}}(\mathbf{x}_i^\mathbf{q};\mathbf{q})$: prediction score of item $\mathbf{x}_\mathbf{q}$; $\mathcal{Q} = \{\mathbf{q}_1,\ldots,\mathbf{q}_n\}$: queries; $\mathcal{S}_\mathbf{q} = \{(\mathbf{x}_1^\mathbf{q},y_1^\mathbf{q}),\ldots,(\mathbf{x}_{n_\mathbf{q}}^\mathbf{q},y_{n_\mathbf{q}}^\mathbf{q})\}$: item-relevance pairs.

$$f_i(\cdot) = \log(\cdot),\quad \mathcal{D} = \{(\mathbf{q},\mathbf{x}_i^\mathbf{q})\mid P(y_i^\mathbf{q})>0\},$$
$$g\left(\mathbf{w};\mathbf{x}_i^\mathbf{q},\mathcal{S}_\mathbf{q}\right) = \textstyle\sum_{\mathbf{x}\in\mathcal{S}_\mathbf{q}}\exp\left(h_{\mathbf{w}}(\mathbf{x};\mathbf{q})-h_{\mathbf{w}}\left(\mathbf{x}_i^\mathbf{q};\mathbf{q}\right)\right).$$

Applications (cont'd)

5. Survival Analysis

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

$\mathbf{z}_i$: patient feature; $h_{\mathbf{w}}(\mathbf{z}_i)$: predicted risk; $E_i=1$: observable event of interest (e.g., death); T_i: time interval between data collection and event; $\mathcal{S}(t) = \{i: T_i\geq t\}$: patients at risk of failure at time t.

$$f_i(\cdot) = -\log(\cdot),\quad \mathcal{D} = \{\mathbf{z}_i\mid E_i=1\},$$
$$g(\mathbf{w};\mathbf{z}_i,\mathcal{S}(T_i)) = \textstyle\sum_{j\in\mathcal{S}(T_i)}\exp\left(h_{\mathbf{w}}(\mathbf{z}_j)-h_{\mathbf{w}}(\mathbf{z}_i)\right).$$

6. Latent Variable Model

$$F(\mathbf{w}) = -\sum_{(\mathbf{x}_i,y_i)\in\mathcal{D}}\log\sum_{\mathbf{s}\in\mathcal{S}}\Pr\left(y_i\mid\mathbf{s},\mathbf{x}_i\right)\Pr\left(\mathbf{s}\mid\mathbf{x}_i\right).$$

$\mathbf{s}$: discrete latent variable; $\mathcal{S}$: support set of the latent variable $\mathbf{s}$.

$$f_i(\cdot) = -\log(\cdot),\quad \mathcal{D} = \{(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_n,y_n)\}.$$
$$g(\mathbf{w};(\mathbf{s}_i,y_i),\mathcal{S}) = \textstyle\sum_{\mathbf{s}\in\mathcal{S}}\Pr\left(y_i\mid\mathbf{s},\mathbf{x}_i\right)\Pr\left(\mathbf{s}\mid\mathbf{x}_i\right).$$

Why FCCO?

- Finite-sum optimization (FO) problem:

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

Algorithms for FO: SGD, SGDm, Adam...

Why not view applications 1.-6. as FO?

$\Longrightarrow$ Computing an **unbiased** stochastic estimator of $\nabla F(\mathbf{w})$ is expensive when inner batch $\mathcal{S}_i$ is large!

- Finite-sum compositional optimization (FCO):

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

Algorithms for FCO: SCGD, NASA, ALSET... We can reformulate an FCCO problem as FCO problem by re-writing $F(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n \hat{f}_i(\mathbf{g}(\mathbf{w};\mathcal{S}))$, where $\hat{f}_i(\cdot) = f_i(\mathbb{I}_i\cdot)$, $\mathbb{I}_i = [0_{d\times d},\ldots,I_{d\times d},\ldots,0_{d\times d}]$, $\mathbf{g}(\mathbf{w};\mathcal{S}) = \left[g\left(\mathbf{w};\mathbf{z}_1,\mathcal{S}_1\right)^\top,\ldots,g\left(\mathbf{w};\mathbf{z}_n,\mathcal{S}_n\right)^\top\right]^\top$, $\mathcal{S} = \mathcal{S}_1\cup\cdots\mathcal{S}_i\cdots\cup\mathcal{S}_n$.

Why not view applications 1.-6. as FCO?

$\Longrightarrow$ FCO algorithms compute stochastic estimators for **all** n components of $\mathbf{g}(\mathbf{w};\mathcal{S})$ (inefficient)!

Why FCCO? (cont'd)

- Conditional stochastic optimization (CSO):

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

Algorithms for CSO: BSGD, $O(\epsilon^{-2})$ batch size. FCCO is a special case of CSO whose outer problem has finite support).

Why not view applications 1.-6. as CSO?

$\Longrightarrow$ CSO algorithm has worse sample complexity because the finite-sum structure is not used!

SOX Algorithm for FCCO

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

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

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

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

Convergence Results

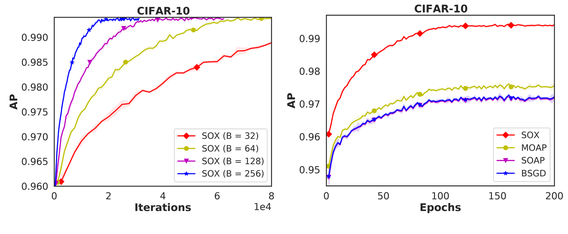
Iteration Complexity

- Strongly convex: $O\left(\frac{n\epsilon^{-1}}{\mu^2B_1B_2}\right)$.
- Convex: $O\left(\frac{n\epsilon^{-3}}{B_1B_2}\right)$.
- Convex and monotone: $O\left(\frac{n\epsilon^{-2}}{B_1}\right)$.
- Non-convex: $O\left(\frac{n\epsilon^{-4}}{B_1B_2}\right)$.

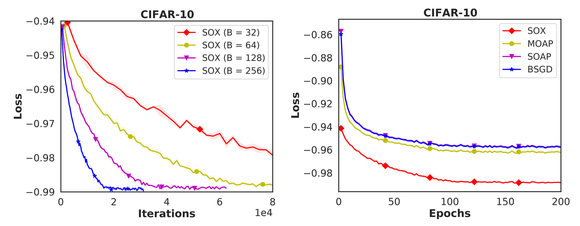
Numerical Experiments

BSGD, SOAP, MOAP are some representative baselines. Total batch size $B = B_1 + B_2$.

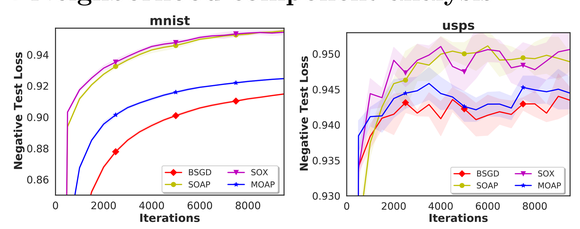
- **AP maximization**



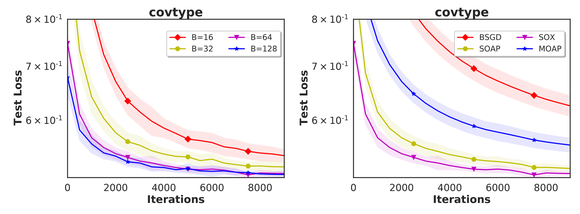
Numerical Experiments (cont'd)



- **Neighborhood component analysis**



- **Ranking by p-norm push**



BS-PnP: boosting-style baseline.

covtype			
Algorithms	BS-PnP	BSGD	SOX
Test Loss (↓)	0.778	0.625 ± 0.018	$\mathbf{0.516\pm0.003}$
Time (s) (↓)	6043.90	$\mathbf{4.20\pm0.08}$	4.62 ± 0.10
ijcnn1			
Algorithms	BS-PnP	BSGD	SOX
Test Loss (↓)	0.268	0.202 ± 0.001	$\mathbf{0.128\pm0.002}$
Time (s) (↓)	648.06	$\mathbf{4.02\pm0.04}$	4.15 ± 0.06

References

[SCGD] Wang, M., Fang, E. X., and Liu, H. *Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions.* Mathematical Programming, 161(1), 419-449.

[NASA] Ghadimi, S., Ruszczynski, A., and Wang, M. *A single timescale stochastic approximation method for nested stochastic optimization.* SIAM Journal on Optimization, 30(1), 960-979.

[BSGD] Hu, Y., Zhang, S., Chen, X., and He, N. *Biased Stochastic First-Order Methods for Conditional Stochastic Optimization and Applications in Meta Learning* Advances in Neural Information Processing Systems, 33, 2759-2770.

[SOAP] Qi, Q., Luo, Y., Xu, Z., Ji, S., and Yang, T. *Stochastic optimization of areas under precision-recall curves with provable convergence.* Advances in Neural Information Processing Systems, 34, 1752-1765.

[MOAP] Wang, G., Yang, M., Zhang, L., and Yang, T. *Momentum accelerates the convergence of stochastic auprc maximization.* International Conference on Artificial Intelligence and Statistics, 25, 3753-3771.