How Does Black-Box Impact the Learning Guarantee of Stochastic Compositional Optimization?

Jun Chen

Huazhong Agricultural University, Wuhan, China

cj850487243@163.com

Oct. 2024

The work is jointed with Hong Chen and Bin Gu.

Stochastic Compositional Optimization[1]

Chen et al.,[1] stated that **stochastic bilevel nested (1)** problem encompasses two popular formulations with the nested structure: **stochastic min-max problem** and **stochastic compositional problem (2)**.

$$\min_{w\in\mathbb{R}^p} \mathbb{E}_{\bar{z}}\left[f_{\bar{z}}\left(w, v^*(w)\right)\right] \quad \text{s.t.} \quad v^*(w) = \arg\min_{v\in\mathbb{R}^d} \mathbb{E}_z\left[h_z(w,v)\right] \tag{1}$$

$$h_z(w,v) = \|v - g_z(w)\|^2 \text{ and } f_{\bar{z}}(w,v) = f_{\bar{z}}(v) \Rightarrow \min_{w\in\mathcal{W}\in\mathbb{R}^p} \mathbb{E}_{\bar{z}}\left[f_{\bar{z}}(\mathbb{E}_z\left[g_z(w)\right])\right] \tag{2}$$

[1] T. Chen, Y. Sun, and W. Yin. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. NeurIPS, 2021.

Background

- **Stochastic Compositional Problems** $$\min_{w\in\mathcal{W}\in\mathbb{R}^p} \mathbb{E}_{\bar{z}}\left[f_{\bar{z}}(\mathbb{E}_z\left[g_z(w)\right])\right] \qquad (2)$$

$$\min_{w\in\mathbb{R}^d} \max_{p\in\triangle_n} F_p(w) = \sum_{i=1}^{n} p_i \ell(w; z_i) - h(p, 1/n)$$

$$\|$$

$$\lambda \sum_{i=1}^{n} p_i \log(np_i)$$

$$p_i^*(w) = \frac{\exp(\ell(w; z_i)/\lambda)}{\sum_{i=1}^n \exp(\ell(w; z_i)/\lambda)}$$

$$\min_{w\in\mathbb{R}^d} F(w) = \lambda \log\left(1/n \sum_{i=1}^{n} \exp(\ell(w; z_i)\lambda)\right)$$

Distributionally Robust Optimization (DRO)[2]

$$\min_{w\in\mathbb{R}^d} \{\mathbb{E}[(h_w(x) - a(w))^2 | y = 1] + \mathbb{E}[(h_w(x') - b(w))^2 | y' = -1] + (1 - a(w) + b(w))^2\}$$

$$a(w) = \mathbb{E}[h_w(x) | y = 1] \qquad b(w) = \mathbb{E}[h_w(x') | y' = -1]$$

AUC maximization[3]

[2] Q. Qi, Z. Guo, Y. Xu, R. Jin, and T. Yang. An online method for a class of distributionally robust optimization with non-convex objectives. NeurIPS, 2021.
[3]T. Yang, and Y. Ying. AUC maximization in the era of big data and AI: A survey. ACM Computing Surveys, 2023.

Background

[4]

Algorithm 1 (Black-box) SCGD / SCSC

Require: v_1, w_1: initial outer model and inner models; β, η_1: initial learning rates

for all $t = 1, ..., T-1$ **do**

Randomly sample $i_t \in [n]$, obtain $g_{z_{i_t}}(w_t)$ and $\nabla g_{z_{i_t}}(w_t)$ (**Inner black-box:** obtain $\tilde{\nabla} g_{z_{i_t}}(w_t)$ similar to Equation ($\boxed{4}$))

SCGD: Update $v_{t+1} = (1-\beta)v_t + \beta g_{z_{i_t}}(w_t)$

SCSC: Update $v_{t+1} = (1-\beta)v_t + \beta g_{z_{i_t}}(w_t) + (1-\beta)(g_{z_{i_t}}(w_t) - g_{z_{i_t}}(w_{t-1}))$

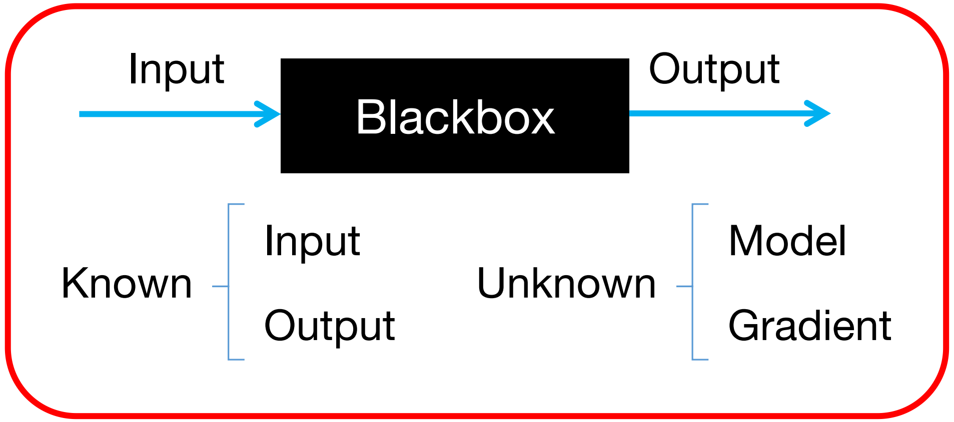
Randomly sample $j_t \in [m]$, obtain $\nabla f_{\bar{z}_{j_t}}(v_{t+1})$ (**Outer black-box:** obtain $\tilde{\nabla} f_{\bar{z}_{j_t}}(v_{t+1})$)

Update $w_{t+1} = w_t - \eta_t \nabla g_{z_{i_t}}(w_t) \nabla f_{\bar{z}_{j_t}}(v_{t+1}) / w_{t+1} = w_t - \eta_t \nabla g_{z_{i_t}}(w_t) \tilde{\nabla} f_{\bar{z}_{j_t}}(v_{t+1})$

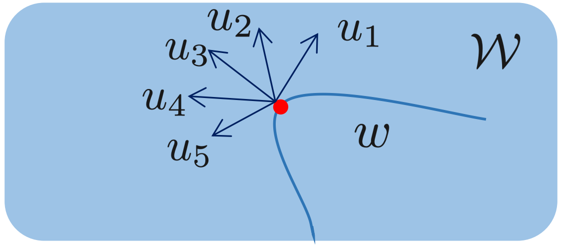
$/ w_{t+1} = w_t - \eta_t \tilde{\nabla} g_{z_{i_t}}(w_t) \nabla f_{\bar{z}_{j_t}}(v_{t+1}) / w_{t+1} = w_t - \eta_t \tilde{\nabla} g_{z_{i_t}}(w_t) \tilde{\nabla} f_{\bar{z}_{j_t}}(v_{t+1})$

end for

Ensure: Final model w_T



$$\tilde{\nabla} g(w) = \mathbb{E}_u \left[\frac{g(w+\mu u) - g(w)}{\mu} u \right]$$

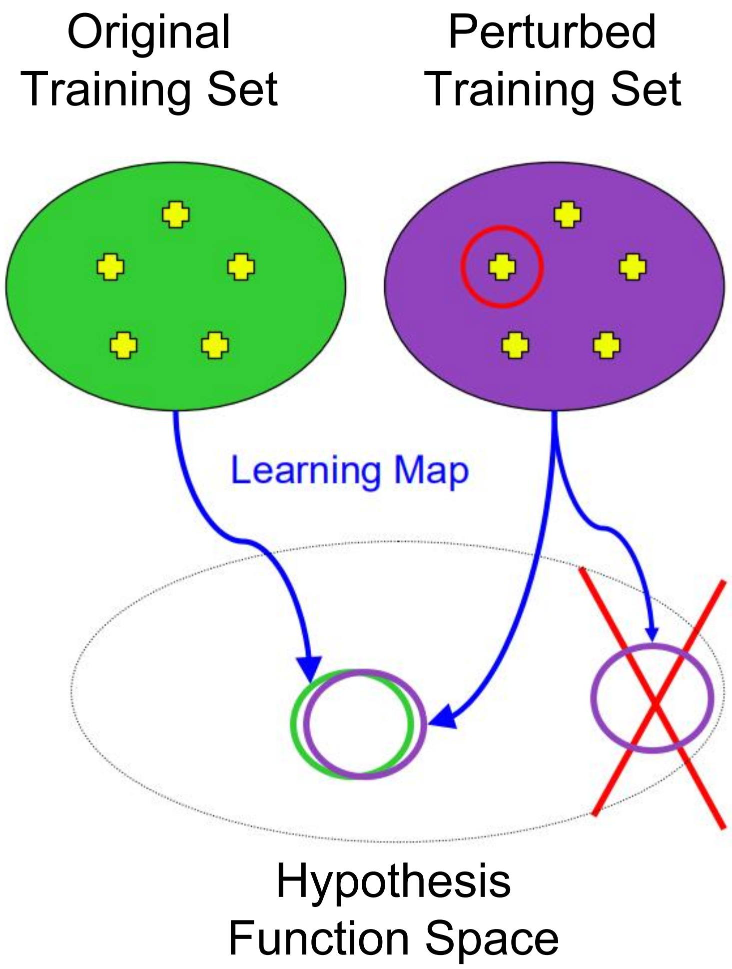


b finite differences

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

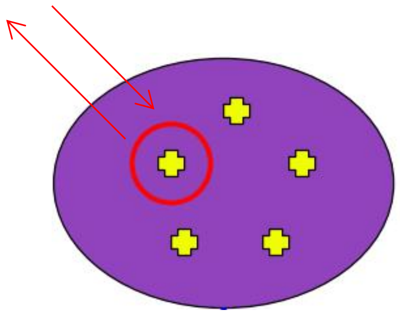
Preliminary

Algorithmic Stability Analysis

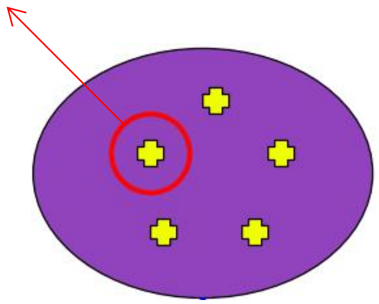


Small perturbation of the training set should not change the trained model parameters much[5].

- Perturbation



Replace a sample



Delete a sample

[5] O. Bousquet, and A. Elisseeff. Stability and generalization. Journal of Machine Learning Research, 2: 499–526, 2002.

Preliminary

- **Assumptions**

➢ Lipschitz continuity $\|g_z(w) - g_z(w')\| \le L_g \|w - w'\|$ $\quad |f_{\bar{z}}(v) - f_{\bar{z}}(v')| \le L_f \|v - v'\|$

➢ Bounded variance $\mathbb{E}_z\left[\|g_z(w) - g(w)\|^2\right] \le V_g$

➢ Smoothness

$$\begin{cases} \|\nabla g_z(w) - \nabla g_z(w')\| \le \alpha_g \|w - w'\|, \quad \|\nabla f_{\bar{z}}(v) - \nabla f_{\bar{z}}(v')\| \le \alpha_f \|v - v'\| \\ \|\nabla f_{\bar{z}}(g_z(w)) - \nabla f_{\bar{z}}(g_z(w'))\| \le \alpha \|w - w'\| \end{cases}$$

➢ Weak bounded function and gradient

$$\begin{cases} \|g_z(w + \mu u) - g_z(w)\| \le M_g, \quad |f_{\bar{z}}(v + \mu u) - f_{\bar{z}}(v)| \le M_f \\ \|\nabla g_z(w + \mu u) - \nabla g_z(w)\| \le M'_g, \quad \|\nabla f_{\bar{z}}(v + \mu u) - \nabla f_{\bar{z}}(v)\| \le M'_f \end{cases}$$

➢ PL condition $\mathbb{E}\left[\|\nabla F_S(w)\|^2\right] \ge 2\gamma \mathbb{E}\left[F_S(w) - F_S(w(S))\right]$

Theoretical Analysis

- ## Generalization Bound of General SCO

Co-coercivity property of convex and smooth function

Assume the function f is convex and α-smooth. Then, for any w, w', we have

$$\langle \nabla f(w) - \nabla f(w'), w - w' \rangle \geq \frac{1}{\alpha} \|\nabla f(w) - \nabla f(w')\|^2.$$

Theorem 1 (Convex)

Let Lipschitz continuity and smoothness assumptions hold and the function $f(g(w))$ is convex. Assume that the randomized algorithm *A* for SCO problem brings the model sequences $\{w_t\}_{t=1}^T$ and $\left\{w_t^{i,z}\right\}_{t=1}^T \left(\left\{w_t^{j,\bar{z}}\right\}_{t=1}^T\right)$ on datasets S and $S^{i,z}$ $\left(S^{j,\bar{z}}\right)$ with the step size sequence $\{\eta_t\}_{t=1}^T$. For SCGD with $\eta_t \leq \frac{2\beta}{\alpha t}$ and SCSC with $\eta_t \leq \frac{2}{\alpha t}$, the final output $A(S) = w_T$ has the following generalization bound

$$\mathbb{E}[|F(w_T) - F_S(w_T)|] \leq \mathcal{O}\left(\left(n^{-1} + m^{-1}\right) \log T + n^{-\frac{1}{2}}\right).$$

Theoretical Analysis

- **Generalization Bound of General SCO**

	Yang et al.,[6]		Theorem 1	
➢ Order	$\mathcal{O}\left(\max\left\{n^{-1}, m^{-1}\right\} \cdot \max\left\{n^{\frac{1}{2}}, m^{\frac{1}{2}}\right\}\right)$	×	$\mathcal{O}\left(\left(n^{-1}+m^{-1}\right)\log T + n^{-\frac{1}{2}}\right)$	√
➢ T	$T = \mathcal{O}\left(\max\left\{n^{\frac{7}{2}}, m^{\frac{7}{2}}\right\}\right)$	×	$T = \mathcal{O}(\max\{n, m\})$ [7]	√
➢ η_t	$\eta = \mathcal{O}\left(T^{-\frac{6}{7}}\right)$	×	$\eta_t = \mathcal{O}(t^{-1})$ [8]	√
➢ β	$\beta = \mathcal{O}\left(T^{-\frac{4}{7}}\right)$	×	Without special restriction [9]	√

[6] M. Yang, X. Wei, T. Yang, and Y. Ying. Stability and generalization of stochastic compositional gradient descent algorithms. ICML, 2024.
[7] M. Wang, E. Fang, and H. Liu. Stochastic compositional gradient descent: Algorithms for minimizing compositions of expected-value functions. Mathematical Programming, 161(1-2):419–449, 2017.
[8] J. Zhang and L. Xiao. A stochastic composite gradient method with incremental variance reduction. NeurIPS, 2019.
[9] T. Chen, Y. Sun, and W. Yin. Solving stochastic compositional optimization is nearly as easy as solving stochastic optimization. IEEE Transactions on Signal Processing, 69:4937–4948, 2021.

Generalization Bound of General SCO

Almost co-coercivity property of smooth function

Consider the gradient-based optimization method $w_{t+1} = w_t - \eta_t \nabla \hat{f}(w_t)$. For two iteration sequences $\{w_t\}_{t\in[T]}$ and $\{w'_t\}_{t\in[T]}$, if the function $\hat{f}(w_t)$ is ρ-smooth, $\eta_t \le 1/(2\rho)$, and the minimum eigenvalue $\lambda_{\min}\left(\nabla^2 \hat{f}(w_t)\right) \ge -\epsilon$, then

$$\left\langle w_t - w'_t, \nabla \hat{f}(w_t) - \nabla \hat{f}(w'_t) \right\rangle \ge 2\eta_t \left(1 - \frac{\eta_t \rho}{2}\right) \|\nabla \hat{f}(w_t) - \nabla \hat{f}(w'_t)\|^2 - \epsilon \|w_t - w'_t - \eta_t \nabla \hat{f}(w_t) + \eta_t \nabla \hat{f}(w'_t)\|^2.$$

Theorem 2 (Non-convex)

Under the condition of Theorem 1 without convex assumption, for SCGD with $\eta_t \le \frac{1}{2\rho t}$, $\rho = \alpha_g L_f + \beta L_g^2 \alpha_f$ and $\eta_t \le \frac{1}{2\rho t}$, SCSC with $\rho = \alpha_g L_f + L_g^2 \alpha_f$, the final output $A(S) = w_T$ has the following generalization bound

$$\mathbb{E}[|F(w_T) - F_S(w_T)|] \le \mathcal{O}\left(\left(n^{-1} + m^{-1}\right) \boxed{T^{\frac{1}{2}}} \log T + n^{-\frac{1}{2}}\right).$$

Theoretical Analysis

- **Learning Guarantees of Black-Box SCO**

Theorem 3 (Outer Black-Box)

Let the bounded first order gradient, bounded second order gradient, bounded variance and PL assumptions hold. Assume that the randomized algorithm *A* for SCO problem brings the model sequences $\{w_t\}_{t=1}^T$ and $\left\{w_t^{i,z}\right\}_{t=1}^T \left(\left\{w_t^{j,\bar{z}}\right\}_{t=1}^T\right)$ on $\mathcal{D}$ and $\mathcal{D}^{i,z}$ $\left(\mathcal{D}^{j,\bar{z}}\right)$ with the step size sequence $\{\eta_t\}_{t=1}^T$. For the outer black-box SCGD with $\eta_t = \frac{1}{p\gamma t}, p \geq \max\left\{\sqrt{\frac{2\alpha}{\gamma}}, \frac{2\left(\alpha_g M_f + \beta L_g^2 M_f'\right)}{\mu\gamma}\right\}$ and the outer black-box SCSC with $\eta_t = \frac{1}{p\gamma t}, p \geq \max\left\{\sqrt{\frac{2\alpha}{\gamma}}, \frac{2\left(\alpha_g M_f + L_g^2 M_f'\right)}{\mu\gamma}\right\}$, the final output $A(S) = w_T$ has the learning guarantee

$$\mathbb{E}\left[F(w_T) - F(w^*)\right] \leq \mathcal{O}\left(\left(n^{-1} + m^{-1}\right) T^{\frac{1}{2}} \log T + n^{-\frac{1}{2}} + \boxed{\mu^2} + \boxed{b^{-1} d_2}\right),$$

where $d_2 = d - (2p+1)\beta + \left(p + \frac{1}{2}\right)^2 \beta^2$ for SCGD and $d_2 = d - 2p - 1 + \left(p + \frac{1}{2}\right)^2$ for SCSC.

Theoretical Analysis

- **Learning Guarantees of Black-Box SCO**

Corollary 2 (Full Black-Box)

Let the conditions of Theorem 3 hold. For the full black-box SCGD with $\eta_t = \frac{1}{p\gamma t}, p \geq \max\left\{\sqrt{\frac{2\alpha}{\gamma}}, (2\left(\beta L_g M_f' M_g + M_f M_g'\right))\left(\mu^2\gamma\right)\right\}$ and the outer black-box SCSC with $\eta_t = \frac{1}{p\gamma t}, p \geq \max\left\{\sqrt{\frac{2\alpha}{\gamma}}, \frac{2\left(L_g M_f' M_g + M_f M_g'\right)}{\mu^2\gamma}\right\}$, the final output $A(S) = w_T$ has the learning guarantee

$$\mathbb{E}\left[F(w_T) - F(w^*)\right] \leq \mathcal{O}\left(\left(n^{-1} + m^{-1}\right) T^{\frac{1}{2}} \log T + n^{-\frac{1}{2}} + \boxed{\mu^4} + \boxed{b^{-2} d_2^2 + b^{-1} d_2}\right),$$

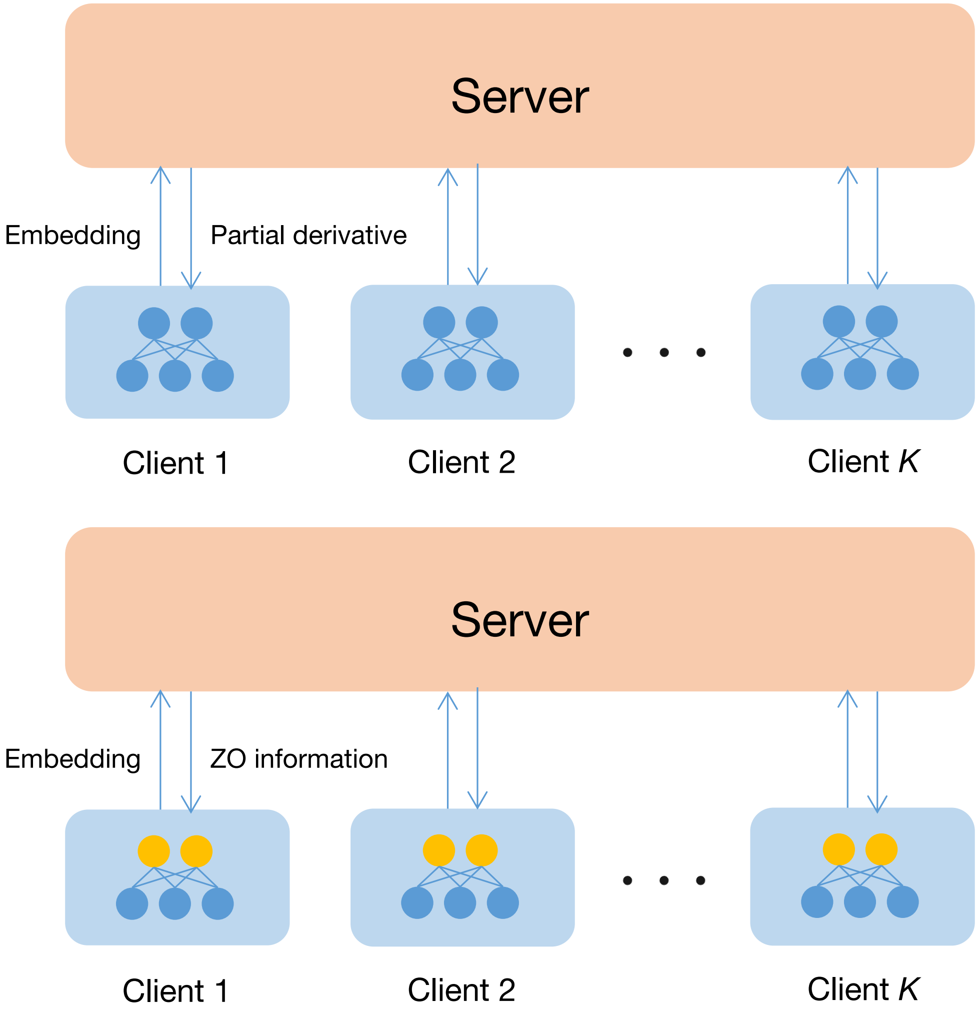
where $d_2 = d - 2\sqrt{\left(p + \frac{1}{2}\right)\beta} + \left(p + \frac{1}{2}\right)\beta$ for SCGD and $d_2 = d - 2\sqrt{\left(p + \frac{1}{2}\right)} + p + \frac{1}{2}$ for SCSC.

Applications

- **Vertical Federated Learning (VFL)[11,12]**

Algorithm 2 FOO-based VFL / VFL-CZOFO

Require: v_1, w_1^k: initial global model and K local models; η_0, η_1: initial learning rates
  for all $t = 1, ..., T-1$ **do**
    for all $k \in [K]$ **in parallel do**
      Randomly select a sample $i_t \in [n]$, obtain $g_{z_{i_t}}(w_t^k)$ and $\nabla g_{z_{i_t}}(w_t^k)$
      Send $g_{z_{i_t}}(w_t^k)$ to server
      FOO-based VFL: Receive $\nabla f(g_{z_{i_t}}(w_t^k))$
        Update $w_{t+1} = w_t - \eta_t \nabla g_{z_{i_t}}(w_t^k) \nabla f(g_{z_{i_t}}(w_t^k))$
      VFL-CZOFO: Receive $\tilde{\nabla} f(g_{z_{i_t}}(w_t^k))$
        Update $w_{t+1} = w_t - \eta_t \nabla g_{z_{i_t}}(w_t^k) \tilde{\nabla} f(g_{z_{i_t}}(w_t^k))$
    end for
    Server receives $g_{z_{i_t}}(w_t^k)$ from K clients
    FOO-based VFL: Obtain and send $\nabla f(g_{z_{i_t}}(w_t^k))$ to the k-th client
    VFL-CZOFO: Compute $\tilde{\nabla} f(g_{z_{i_t}}(w_t^k))$ and send it to the k-th client
    Obtain $\nabla f(v_t)$ and update $v_{t+1} = v_t - \eta_0 \nabla f(v_t)$
  end for
Ensure: K final client models $w_T^1, ..., w_T^K$



[11] T. Chen, X. Jin, Y. Sun, and W. Yin. VAFL: a method of vertical asynchronous federated learning, arXiv, 2007.06081, 2020.
[12] G. Wang, B. Gu, Q. Zhang, X. Li, B. Wang, and C. Ling. A unified solution for privacy and communication efficiency in vertical federated learning. NeurIPS, 2023.

Applications

- **Vertical Federated Learning (VFL)**

Corollary 4 (VFL-CZOFO)

Let the bounded first order gradient, bounded second order gradient, and PL assumptions hold. For the k-th client ($k \in [K]$), assume that the randomized VFL-CZOFO algorithm brings the model sequences $\{w_t^k\}_{t=1}^T$ and $\left\{w_t^{i,z,k}\right\}_{t=1}^T$ on S and $S^{i,z}$ with the step size sequence $\{\eta_t\}_{t=1}^T, \eta_t = \frac{1}{p\gamma t}, p \geq \max\left\{\sqrt{\frac{2\alpha}{\gamma}}, \frac{2\left(\alpha_g M_f + L_g^2 M_f'\right)}{\mu\gamma}\right\}$.

Then, the final output $A(S) = w_T^k$ of the k-th client has the generalization guarantee

$$\mathbb{E}\left[F(w_T^k) - F(w^{k*})\right] \leq \mathcal{O}\left(n^{-1}T^{\frac{1}{2}}\log T + \mu^2 + b^{-1}d_2\right).$$

Thanks

Jun Chen

Huazhong Agricultural University, Wuhan, China

cj850487243@163.com

Feb. 2023