OPT2022: 14th Annual Workshop on Optimization for Machine Learning

ProxSkip for Stochastic Variational Inequalities: A Federated Learning Algorithm for Provable Communication Acceleration

Siqi Zhang SZHAN207@JHU.EDU
Nicolas Loizou NLOIZOU@JHU.EDU
Department of Applied Mathematics and Statistics, Johns Hopkins University, USA

Abstract

Recently Mishchenko et al. [11] proposed and analyzed ProxSkip, a provably efficient method for minimizing the sum of a smooth (f) and an expensive nonsmooth proximable (R) function (i.e. $\min_{x\in\mathbb{R}^d} f(x)+R(x)$). The main advantage of ProxSkip, is that in the federated learning (FL) setting, offers provably an effective acceleration of communication complexity.

This work extends this approach to the more general regularized variational inequality problems (VIP). In particular, we propose ProxSkip-VIP algorithm, which generalizes the original ProxSkip framework of [11] to VIP, and we provide convergence guarantees for a class of structured non-monotone problems. In the federated learning setting, we explain how our approach achieves acceleration in terms of the communication complexity over existing state-of-the-art FL algorithms.

1. Introduction

Minimax optimization and, more generally, variational inequality problems (VIPs) appear in a wide range of applications in machine learning including Generative Adversarial Networks (GANs) [6], adversarial training of neural networks [9, 15] and distributionally robust learning [17]. Motivated by these applications, in this work, we consider the following regularized variational inequality problem (VIP): find $x^*\in\mathbb{R}^d$, such that

$$\langle F(x^*), x-x^*\rangle + R(x) - R(x^*) \geq 0, \quad \forall x\in\mathbb{R}^d, \tag{1}$$

where $F:\mathbb{R}^d\to\mathbb{R}^d$ and $R:\mathbb{R}^d\to\mathbb{R}$ is the regularizer (a proper lower semicontinuous convex function). This problem is quite general and covers a wide range of possible problem formulations. For example, special cases of (1) are the regularized minimization problems [10] and minimax problems [12]:

$$\min_{x\in\mathbb{R}^d} f(x)+R(x) \quad \text{and} \quad \min_{x_1\in\mathbb{R}^{d_1}}\max_{x_2\in\mathbb{R}^{d_2}} f(x_1,x_2)+R(x_1,x_2). \tag{2}$$

In this work, we are interested in the situations when operator F is accessible through the calls of unbiased stochastic oracle. This is natural when F has an expectation form $F(x)=\mathbb{E}_{\xi\sim\mathcal{D}}[F_\xi(x)]$ or a finite-sum form $F(x)=\frac{1}{n}\sum_{i=1}^n F_i(x)$. In this scenario, one of the most popular algorithms for solving (1) is the stochastic proximal method[13]

$$x_{t+1} = \mathbf{prox}_{\gamma R}(x_t - \gamma g_t),$$

where $\mathbf{prox}_{\gamma R}(x) \triangleq \operatorname{argmin}_{z\in\mathbb{R}^d}\left\{R(z)+\frac{1}{2\gamma}\|z-x\|^2\right\}$, g_t is an unbiased estimator of $F(x_t)$ (i.e. $\mathbb{E}[g_t]=F(x_t)$) and $\gamma>0$ is the step-size of the method.

Typically computing the proximal operator is easy and cheap. However, in our work, following the approach of Mishchenko et al. [11], we are interested in the situation when the evaluation of the proximity operator is expensive. That is, we assume that the computation of $\mathbf{prox}_{\gamma R}$ is costly relative to the evaluation of the unbiased estimator g_t. It is in this scenario that ProxSkip-VIP (Alg. 1) thrives, as it skips the evaluation of the proximity operator and it requires its computation only once every few iterations.

In the federated learning setting (see Sec. 4 for more details) ProxSkip-VIP can be interpreted as a new distributed method performing local steps. In that scenario, ProxSkip-VIP becomes equivalent to the update rule of Algorithm (2) (ProxSkip-VIP-FL) where the computation of the proximity operator becomes equivalent to communications between workers. Thus, skipping proximity operator's computation means that the algorithm performs local updates (it skips communication). See [11] for the full exposition of this connection.

Main Contributions Our main contributions are summarized below:

- We generalize the ProxSkip framework proposed in [11] for minimization problems into the VIP regime, and proposed the ProxSkip-VIP algorithm.
- We prove that ProxSkip-VIP converges linearly to a neighborhood of the optimal set when problem (1), has μ-quasi-strongly monotone and L-star-cocoercive operator F. This is a class of structure non-monotone problems. As a corollary of our results, for the deterministic regime where $g_t = F(x_t)$, ProxSkip-VIP converges linearly to the exact solution.
- We extend the ProxSkip-VIP method into the federated learning setting, and propose ProxSkip-VIP-FL; we show that the algorithm enjoys an improved communication complexity over existing literature of local Stochastic Methods for solving VIPs. Numerical experiment results shows that our proposed algorithm outperforms over existing algorithms.

2. Preliminaries

First, let us introduce the setting of this work.

Assumption 1 (Main Settings) *We assume that problem* (1) *has a unique*[1] *solution* x^* *and that:*

1. *The operator F is μ-quasi-strongly monotone and ℓ-star-cocoercive with $\mu, L > 0$, i.e.,*

$$\langle F(x) - F(x^*), x - x^* \rangle \geq \mu \|x - x^*\|^2, \quad \langle F(x) - F(x^*), x - x^* \rangle \geq \frac{1}{\ell}\|F(x) - F(x^*)\|^2. \tag{3}$$

2. *The function R is a proper lower semicontinuous convex function.*

Note that given that an operator F is L'-Lipschitz continuous and μ'-strongly monotone, it can be shown that the operator F is $(\kappa L')$-star-cocoercive with $\kappa = L'/\mu'$ [8].

The convergence results in this paper will depend on the following operator noise at x^* that is finite for any reasonable sampling: $\sigma^2 \triangleq \mathrm{Var}(g(x^*;\xi)) < +\infty$. Regarding the inherent stochasticity, we further use the following expected cocoercivity assumption [4, 8] to characterize the behavior of the operator estimation

[1] This assumption can be relaxed; but for simplicity of exposition we enforce it.

Assumption 2 (Expected Cocoercivity) *We assume that stochastic operator $g(x;\xi)$ is such that for all $x \in \mathbb{R}^d$ there is $L > 0$:*

$$\mathbb{E}\|g(x;\xi) - g(x^*;\xi)\|^2 \leq L\langle F(x) - F(x^*), x - x^*\rangle.$$

See [4, 8] for more details on this assumption and why is weaker among other bounds on the noise of the stochastic operator.

3. Algorithm: ProxSkip-VIP

In this section, we incorporate the ProxSkip algorithm [11] into our problem (1), and propose the following ProxSkip-VIP algorithm.

Algorithm 1 ProxSkip-VIP

```
Input: Initial point $x_0$, parameters $\gamma_1, \gamma_2, \gamma_3, p$, initial control variate $h_0$, number of iterations $T$
for all $t = 0, 1, ..., T$ do
    $\hat{x}_{t+1} = x_t - \gamma_1(g(x_t;\xi_t) - h_t)$
    Flip a coin $\theta_t$, $\theta_t = 1$ w.p. $p$, otherwise 0
    if $\theta_t = 1$ then
        $x_{t+1} = \mathbf{prox}_{\gamma_2 R}(\hat{x}_{t+1} - \gamma_2 h_t)$
    else
        $x_{t+1} = \hat{x}_{t+1}$
    end if
    $h_{t+1} = h_t + \gamma_3(x_{t+1} - \hat{x}_{t+1})$
end for
Output: $x_T$
```

Here the key step is the randomized prox-skipping and the control variate h_t. The proximal oracle is called very rarely if p is small, which helps to reduce the computational cost if the proximal oracle is expensive; the introducing of h_t helps to stabilize the iterations toward the optimal point.

3.1. Convergence Analysis

The main theorem of this work, on the convergence guarantees of ProxSkip-VIP is presented below.

Theorem 1 (Convergence of ProxSkip-VIP) *Let Assumption 1 and 2 hold, and let $\gamma_1 = \gamma \in \left(0, \min\left\{\frac{1}{\mu}, \frac{1}{2L}\right\}\right)$, $\gamma_1 = \gamma_2 p$, $\gamma_3 = \frac{1}{\gamma_2}$. Then the iterates of ProxSkip-VIP (Alg. 1) satisfy*

$$\mathbb{E}[V_T] \leq \left(1 - \min\left\{\gamma\mu, p^2\right\}\right)^T V_0 + \frac{2\gamma^2\sigma^2}{\min\left\{\gamma\mu, p^2\right\}}. \tag{4}$$

where $V_t \triangleq \|x_t - x_t^*\|^2 + \gamma_2^2\|h_t - F(x_t^*)\|^2$.

We defer the proof of Theorem 1 to Appendix B. As a corollary of the above theorem, we can obtain the following corresponding complexity results (proof is deferred to Appendix C).

Corollary 2 *Let all assumptions of Theorem 1 be satisfied. If we further set $\gamma \leq \frac{\mu\epsilon}{2\sigma^2}$ and $p = \sqrt{\gamma\mu}$, we have $\mathbb{E}[V_T] \leq \epsilon$ with iteration complexity and the number of calls of the proximal oracle $\mathbf{prox}(\cdot)$ as*

$$\mathcal{O}\left(\max\left\{\frac{L}{\mu}, \frac{\sigma^2}{\mu^2\epsilon}\right\}\ln\left(\frac{1}{\epsilon}\right)\right) \quad \textit{and} \quad \mathcal{O}\left(\sqrt{\max\left\{\frac{L}{\mu}, \frac{\sigma^2}{\mu^2\epsilon}\right\}}\ln\left(\frac{1}{\epsilon}\right)\right). \tag{5}$$

Deterministic ProxSkip: As a corollary of our results, for the deterministic regime where $g_t = F(x_t)$, ProxSkip-VIP converges linearly to the exact solution since $\sigma^2 = 0$. In this scenario, under the same assumptions with Theorem 1, the iterates of ProxSkip (Alg. 1) satisfy:

$$\mathbb{E}[V_T] \leq \left(1 - \min\left\{\gamma\mu, p^2\right\}\right)^T V_0.$$

In this setting, we get $\mathbb{E}[V_T] \leq \epsilon$ with iteration complexity and number of calls of the proximal oracle $\mathbf{prox}(\cdot)$ as

$$\mathcal{O}\left(\frac{L}{\mu}\ln\left(\frac{1}{\epsilon}\right)\right) \quad \text{and} \quad \mathcal{O}\left(\sqrt{\frac{L}{\mu}}\ln\left(\frac{1}{\epsilon}\right)\right) \tag{6}$$

4. Connection with Federated Learning

Let us now explain how ProxSkip-VIP works in the federated learning setting, i.e., find $x^* \in \mathbb{R}^d$ such that

$$\langle F(x^*), x - x^*\rangle \geq 0, \quad \forall x \in \mathbb{R}^d, \tag{7}$$

where $F(x) \triangleq \frac{1}{n}\sum_{i=1}^n f_i(x)$ and $f_i(x) \triangleq \mathbb{E}_{\xi_i\sim\mathcal{D}_i}[f_i(x;\xi_i)]$, the data ξ_i follows an unknown distribution $\mathcal{D}_i$ $(i = 1, 2, \cdots, n)$. We highlight that following a similar approach as in section 1, the federated learning minimization and federated minimax problems [11, 14] can easily obtained as special cases of (7). As mentioned in [13], the problem (7) can be recast into the problem (1) while

$$F(x) \triangleq \frac{1}{n}\sum_{i=1}^n f_i(x_i), \quad f_i(x) \triangleq \mathbb{E}_{\xi_i\sim\mathcal{D}_i}[f_i(x;\xi_i)] \tag{8}$$

where $x_i \in \mathbb{R}^d$, $x = (x_1, x_2, \cdots, x_n) \in \mathbb{R}^{nd}$, and

$$R(x) = R((x_1, x_2, \cdots, x_n)) \triangleq \begin{cases} 0 & \text{if } x_1 = x_2 = \cdots = x_n \\ +\infty & \text{otherwise.} \end{cases} \tag{9}$$

Note that $\mathbf{prox}_{\gamma R}(x) = (\bar{x}, \bar{x}, \cdots, \bar{x})$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$, which is easy to compute [13].

4.1. Algorithm: ProxSkip-VIP-FL

With the above reformulation (8), we propose the following ProxSkip-VIP-FL algorithm based on Algorithm 1 for the federated learning problem 7, which is presented below.

Different from the centralized setting we discussed in Section 3, the federated learning framework (8) often characterized by a heterogeneous environment, i.e., the distributions $\{\mathcal{D}_i\}_i$ are not identical.

4.2. Convergence Analysis

Let us now present the convergence guarantees of ProxSkip-VIP-FL (Alg. 2).

Theorem 3 (Complexity of ProxSkip-VIP-FL) *Lets assume the same setting as in Corollary 2. Then ProxSkip-VIP-FL achieves $\mathbb{E}[V_T] \leq \epsilon$ (where V_T is defined in Theorem 1),*

- *with iteration complexity*

$$\mathcal{O}\left(\max\left\{\frac{L}{\mu}, \frac{\sigma^2}{\mu^2\epsilon}\right\}\ln\left(\frac{1}{\epsilon}\right)\right)$$

Algorithm 2 ProxSkip-VIP-FL

```
Input: Initial point $x_{1,0} = x_{2,0} = \cdots = x_{n,0} \in \mathbb{R}^d$, parameters $\gamma_1, \gamma_2, \gamma_3, p \in \mathbb{R}$, initial control
    variate for each client $h_{1,0}, h_{2,0}, \cdots, h_{n,0} \in \mathbb{R}^d$ , number of iterations $T$
for all $t = 0, 1, ..., T$ do
    Server: Flip a coin $\theta_t$, $\theta_t = 1$ w.p. $p$, otherwise $0$. Send $\theta_t$ to all workers
    for each workers $i \in [n]$ in parallel do
        $\hat{x}_{i,t+1} = x_{i,t} - \gamma_1(g_i(x_{i,t}; \xi_{i,t}) - h_{i,t})$        // Local update with control variate
        if $\theta_t = 1$ then
            Worker: $x'_{i,t+1} = \hat{x}_{i,t+1} - \gamma_2 h_{i,t}$, sends $x'_{i,t+1}$ to the server
            Server: computes $x_{i,t+1} = \frac{1}{n}\sum_{i=1}^{n} x'_{i,t+1}$ and send to workers        // Communication
        else
            $x_{i,t+1} = \hat{x}_{i,t+1}$        // Otherwise skip the communication step
        end if
        $h_{i,t+1} = h_{i,t} + \gamma_3(x_{i,t+1} - \hat{x}_{i,t+1})$
    end for
end for
Output: $x_T$
```

- *and communication complexity as*

$$\mathcal{O}\left(\sqrt{\max\left\{\frac{L}{\mu}, \frac{\sigma^2}{\mu^2\epsilon}\right\}}\ln\left(\frac{1}{\epsilon}\right)\right).$$

Comparison with Existing Literature The above theorem provides the complexity results of Algorithm 2 in the federated learning setting. We note that our approach is quite general as we do not make strong assumptions on the choice of the unbiased estimator $g_i(x_{i,t}; \xi_{i,t})$. For example, in federated minimax problems, a recently proposed method is the Local SGDA algorithm [5]. With our framework, one is able to use the same (mini-batch) gradient estimator $g_i(x_{i,t}; \xi_{i,t})$ from [5] in our Algorithm 2. The benefit of this is that our result will avoid the dependence on the condition number κ (when ϵ is small enough), and we can attain an improved communication and iteration complexity for solving (7) (see comparison in Table 1).

In Table 1 we provide a more detailed comparison of our theoretical convergence guarantees of Algorithm 2 (Theorem 3) with existing literature in federated learning. It is clear that the proposed approach outperforms the other algorithms (Local SGDA, Local SEG, FedAvg-S) in terms of iteration and communication complexities (when ϵ is small enough). Finally let us highlight that in our analysis of ProxSkip algorithm we do not require an assumption on bounded heterogeneity.

5. Numerical Experiment

In this section we conduct a numerical experiment on a toy example to test the efficiency of our proposed algorithm. Following the setting in [16], we consider the problem (7) with $x = (x_1, x_2) \in \mathbb{R}^{d_1 \times d_2}$ and

$$f_i(x) = -\left[\frac{1}{2}\|x_2\|^2 - b_i^\top x_2 + x_2^\top A_i x_1\right] + \frac{\lambda}{2}\|x_1\|^2, \tag{10}$$

Algorithm	Setting[1]	# Communication[2]	# Iteration
Local SEG [2, 3]	SM, LS	$\mathcal{O}\left(\max\left(\kappa \ln \frac{1}{\epsilon}, \frac{p\sigma^2}{\mu^2 n\epsilon}, \frac{\kappa\xi}{\mu\sqrt{\epsilon}}, \frac{\sqrt{p}\kappa\sigma}{\mu\sqrt{\epsilon}}\right)\right)$	$\mathcal{O}\left(\max\left(\frac{\kappa}{p} \ln \frac{1}{\epsilon}, \frac{\sigma^2}{\mu^2 n\epsilon}, \frac{\kappa\xi}{p\mu\sqrt{\epsilon}}, \frac{\kappa\sigma}{\mu\sqrt{p\epsilon}}\right)\right)$
Local SGDA [5]	SM, LS	$\mathcal{O}\left(\sqrt{\frac{\kappa^2(\xi^2+\sigma^2)}{\mu\epsilon}}\right)$	$\mathcal{O}\left(\frac{\kappa^2(\xi^2+\sigma^2)}{\mu\epsilon}\right)$
FedAvg-S [7]	SM, LS	$\tilde{\mathcal{O}}\left(\frac{p\sigma^2}{n\mu^2\epsilon} + \frac{\sqrt{p}\kappa\sigma}{\mu\sqrt{\epsilon}} + \frac{\kappa\xi}{\mu\sqrt{\epsilon}}\right)$	$\tilde{\mathcal{O}}\left(\frac{\sigma^2}{n\mu^2\epsilon} + \frac{\kappa\sigma}{\mu\sqrt{p\epsilon}} + \frac{\kappa\xi}{p\mu\sqrt{\epsilon}}\right)$
Ours (Theorem 3)	SM, LS[3]	$\tilde{\mathcal{O}}\left(\sqrt{\max\left\{\kappa^2, \frac{\sigma^2}{\mu^2\epsilon}\right\}}\right)$	$\tilde{\mathcal{O}}\left(\max\left\{\kappa^2, \frac{\sigma^2}{\mu^2\epsilon}\right\}\right)$

[1] SM: strongly monotone, LS: (Lipschitz) smooth. $\kappa \triangleq L/\mu$, L and μ are the modulus of SM and LS. $\sigma^2 \triangleq \text{Var}(g(x^*;\xi)) < +\infty$ (or uniform bound on $\text{Var}(g(\cdot;\xi))$). ξ^2 represents the bounded heterogeneity, i.e., $g_i(x;\xi_i)$ is an unbiased estimator of $f_i(x)$ for any $i \in \{1, 2, \cdots, n\}$, and $\xi_i^2(x) \triangleq \sup_{x\in\mathbb{R}^d} \|f_i(x) - F(x)\|^2 \leq \xi^2 \leq +\infty$.

[2] p is the probability of synchronization, by setting ϵ is small enough, we can take $p = \mathcal{O}(\sqrt{\epsilon})$, which recovers $\mathcal{O}(1/\sqrt{\epsilon})$ communication complexity dependence on ϵ in our result. $\tilde{\mathcal{O}}(\cdot)$ hides the logarithmic terms.

[3] Our algorithm works for quasi-strongly monotone and star-cocoercive, which is more general than the SM and LS setting, note that an L-LS and μ-SM operator can be shown to be (κL)-star-cocoercive [8].

Table 1: Comparison of federated learning algorithms for solving VIPs with strongly monotone and Lipscitz operator. Comparison is in terms of both iteration and communication complexities.

here we set the number of clients $n = 100$, and $d_1 = d_2 = 20$, $\lambda = 0.1$, $b_i \sim \mathcal{N}(0, s_i^2 I_{d_2})$ where $s_i \sim \text{Unif}(0, 20)$, $A_i = t_i I_{d_1 \times d_2}$ and $t_i \sim \text{Unif}(0, 1)$. It is easy to show that the quadratic objective function satisfies Assumption 1. Our goal is to have a fair comparison between the Local SGDA [5] and the proposed ProxSkip-VIP-FL (Alg. 2). We fine-tuned the stepsizes for both algorithms using grid-search in $[0.1, 0.5]$, and plot the comparison in terms of communication rounds of the two methods in Fig. 1. As it shows in Fig. 1, ProxSkip has better performance in terms of communication rounds compared to Local SGDA [5].

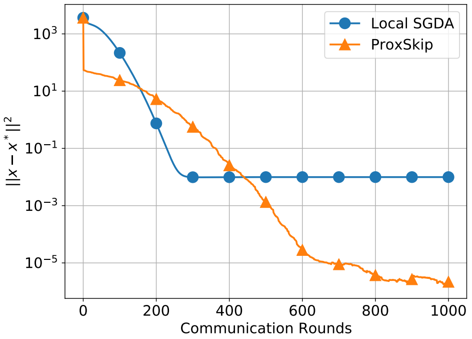

Figure 1: Comparison between Local SGDA [5] and ProxSkip-VIP-FL. The parameters are fine-tuned with grid-search for both algorithms.

References

[1] Amir Beck. *First-order methods in optimization*. SIAM, 2017.

[2] Aleksandr Beznosikov, Valentin Samokhin, and Alexander Gasnikov. Distributed saddle-point problems: Lower bounds, optimal and robust algorithms. *arXiv preprint arXiv:2010.13112*, 2020.

[3] Aleksandr Beznosikov, Pavel Dvurechensky, Anastasia Koloskova, Valentin Samokhin, Sebastian U Stich, and Alexander Gasnikov. Decentralized local stochastic extra-gradient for variational inequalities. *arXiv preprint arXiv:2106.08315*, 2021.

[4] Aleksandr Beznosikov, Eduard Gorbunov, Hugo Berard, and Nicolas Loizou. Stochastic gradient descent-ascent: Unified theory and new efficient methods. *arXiv preprint arXiv:2202.07262*, 2022.

[5] Yuyang Deng and Mehrdad Mahdavi. Local stochastic gradient descent ascent: Convergence analysis and communication efficiency. In *International Conference on Artificial Intelligence and Statistics*, pages 1387–1395. PMLR, 2021.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

[7] Charlie Hou, Kiran K Thekumparampil, Giulia Fanti, and Sewoong Oh. Efficient algorithms for federated saddle point optimization. *arXiv preprint arXiv:2102.06333*, 2021.

[8] Nicolas Loizou, Hugo Berard, Gauthier Gidel, Ioannis Mitliagkas, and Simon Lacoste-Julien. Stochastic gradient descent-ascent and consensus optimization for smooth games: Convergence analysis under expected co-coercivity. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[9] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[10] Yu Malitsky. Proximal extrapolated gradient methods for variational inequalities. *Optimization Methods and Software*, 33(1):140–164, 2018.

[11] Konstantin Mishchenko, Grigory Malinovsky, Sebastian Stich, and Peter Richtárik. Proxskip: Yes! local gradient steps provably lead to communication acceleration! finally! In *International Conference on Machine Learning*. PMLR, 2022.

[12] Balamurugan Palaniappan and Francis Bach. Stochastic variance reduction methods for saddle-point problems. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[13] Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and trends® in Optimization*, 1(3):127–239, 2014.

[14] Pranay Sharma, Rohan Panda, Gauri Joshi, and Pramod Varshney. Federated minimax optimization: Improved convergence analyses and algorithms. In *International Conference on Machine Learning*, pages 19683–19730. PMLR, 2022.

[15] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

[16] Davoud Ataee Tarzanagh, Mingchen Li, Christos Thrampoulidis, and Samet Oymak. FedNest: Federated bilevel, minimax, and compositional optimization. In *International Conference on Machine Learning*, pages 21146–21179. PMLR, 2022.

[17] Yaodong Yu, Tianyi Lin, Eric V Mazumdar, and Michael Jordan. Fast distributionally robust learning with variance-reduced min-max optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 1219–1250. PMLR, 2022.

Supplementary Material

Appendix A. Useful Lemmas

Lemma 4 *For any optimal solution $x^* \in X^*$ of (1), we have*

$$x^* = \mathbf{prox}_{\gamma_2 R}(x^* - \gamma_2 F(x^*)). \tag{11}$$

Proof Note that

$$\langle F(x^*), x - x^* \rangle + R(x) - R(x^*) \geq 0, \tag{12}$$

next for any $x \in \mathbb{R}^d$,

$$\begin{aligned}
& R(x^*) + \frac{1}{2\gamma_2}\|x^* - x^* + \gamma_2 F(x^*)\|^2 \leq R(x) + \frac{1}{2\gamma_2}\|x - x^* + \gamma_2 F(x^*)\|^2 \\
\Longleftrightarrow \quad & R(x^*) + \frac{1}{2\gamma_2}\|\gamma_2 F(x^*)\|^2 \leq R(x) + \frac{1}{2\gamma_2}\|x - x^*\|^2 + \frac{1}{2\gamma_2}\|\gamma_2 F(x^*)\|^2 + \langle F(x^*), x - x^* \rangle \\
\Longleftrightarrow \quad & R(x^*) \leq R(x) + \frac{1}{2\gamma_2}\|x - x^*\|^2 + \langle F(x^*), x - x^* \rangle \\
\Longleftarrow \quad & R(x^*) \leq R(x) + \langle F(x^*), x - x^* \rangle,
\end{aligned}$$

which concludes the proof, note that this conclusion indicates that the two parameters in the RHS above should be identical. ■

Lemma 5 (Firm Nonexpansivity of the Proximal Operator [1]) *Let f be a proper closed and convex function, then for any $x, y \in \mathbb{R}^d$ we have*

$$\big\langle x - y, \mathbf{prox}_f(x) - \mathbf{prox}_f(y) \big\rangle \geq \big\|\mathbf{prox}_f(x) - \mathbf{prox}_f(y)\big\|^2, \tag{13}$$

or equivalently,

$$\big\|\big(x - \mathbf{prox}_f(x)\big) - \big(y - \mathbf{prox}_f(y)\big)\big\|^2 + \big\|\mathbf{prox}_f(x) - \mathbf{prox}_f(y)\big\|^2 \leq \|x - y\|^2. \tag{14}$$

The following result is helpful in the proof.

Lemma 6 *With Assumption 1 and 2, we have*

$$\mathbb{E}\|g(x;\xi) - F(x^*)\|^2 \leq 2L\langle F(x) - F(x^*), x - x^* \rangle + 2\sigma^2. \tag{15}$$

Proof

$$\begin{aligned}
\mathbb{E}\|g(x;\xi) - F(x^*)\|^2 &\leq 2\mathbb{E}\Big[\|g(x;\xi) - g(x^*;\xi)\|^2 + \|g(x^*;\xi) - F(x^*)\|^2\Big] \\
&\leq 2L\langle F(x) - F(x^*), x - x^* \rangle + 2\sigma^2,
\end{aligned} \tag{16}$$

which concludes the proof. ■

Appendix B. Proof of Theorem 1

Proof Note that

$$x_{t+1} = \begin{cases} \mathbf{prox}_{\gamma_2 R}(\hat{x}_{t+1} - \gamma_2 h_t) & \text{with probability } p \\ \hat{x}_{t+1} & \text{with probability } 1-p, \end{cases} \tag{17}$$

and

$$h_{t+1} = h_t + \gamma_3(x_{t+1} - \hat{x}_{t+1}) = \begin{cases} h_t + \gamma_3\big(\mathbf{prox}_{\gamma_2 R}(\hat{x}_{t+1} - \gamma_2 h_t) - \hat{x}_{t+1}\big) & \text{with probability } p \\ h_t & \text{with probability } 1-p. \end{cases} \tag{18}$$

For simplicity, we denote $P(x_t) \triangleq \mathbf{prox}_{\gamma_2 R}(\hat{x}_{t+1} - \gamma_2 h_t)$, so we have

$$\begin{aligned} &\mathbb{E}_{\xi_t}[V_{t+1}] \\ &= p\Big(\big\|P(x_t) - x_{t+1}^*\big\|^2 + \gamma_2^2\big\|h_t + \gamma_3(P(x_t) - \hat{x}_{t+1}) - F(x_{t+1}^*)\big\|^2\Big) \\ &\qquad + (1-p)\Big(\big\|\hat{x}_{t+1} - x_{t+1}^*\big\|^2 + \gamma_2^2\big\|h_t - F(x_{t+1}^*)\big\|^2\Big) \\ &= p\Big(\big\|P(x_t) - x_{t+1}^*\big\|^2 + \big\|P(x_t) - (\hat{x}_{t+1} - \gamma_2 h_t) - \gamma_2 F(x_{t+1}^*)\big\|^2\Big) \\ &\qquad + (1-p)\Big(\big\|\hat{x}_{t+1} - x_{t+1}^*\big\|^2 + \gamma_2^2\big\|h_t - F(x_{t+1}^*)\big\|^2\Big) \end{aligned} \tag{19}$$

next note that $x_t^* = \mathbf{prox}_{\gamma_2 R}(x_t^* - \gamma_2 F(x_t^*))$, we have

$$\begin{aligned} &\big\|P(x_t) - (\hat{x}_{t+1} - \gamma_2 h_t) - \gamma_2 F(x_{t+1}^*)\big\|^2 \\ &= \big\|P(x_t) - (\hat{x}_{t+1} - \gamma_2 h_t) - \big(\mathbf{prox}_{\gamma_2 R}\big(x_{t+1}^* - \gamma_2 F(x_{t+1}^*)\big) - (x_{t+1}^* - \gamma_2 F(x_{t+1}^*))\big)\big\|^2 \end{aligned} \tag{20}$$

so by Lemma 5, we have

$$\begin{aligned} &\mathbb{E}_{\xi_t}[V_{t+1}] \\ &\leq p\big\|\hat{x}_{t+1} - \gamma_2 h_t - x_{t+1}^* + \gamma_2 F(x_{t+1}^*)\big\|^2 + (1-p)\Big(\big\|\hat{x}_{t+1} - x_{t+1}^*\big\|^2 + \gamma_2^2\big\|h_t - F(x_{t+1}^*)\big\|^2\Big) \\ &= \big\|\hat{x}_{t+1} - x_{t+1}^*\big\|^2 + \gamma_2^2\big\|h_t - F(x_{t+1}^*)\big\|^2 - 2\gamma_2 p\big\langle \hat{x}_{t+1} - x_{t+1}^*, h_t - F(x_{t+1}^*)\big\rangle, \end{aligned} \tag{21}$$

let

$$w_t \triangleq x_t - \gamma_1 g(x_t; \xi_t), \quad w_t^* \triangleq x_t^* - \gamma_1 F(x_t^*), \tag{22}$$

recall that $\gamma = \gamma_1 = \gamma_2 p$, so we have

$$\begin{aligned} &\big\|\hat{x}_{t+1} - x_{t+1}^*\big\|^2 - 2\gamma_2 p\big\langle \hat{x}_{t+1} - x_{t+1}^*, h_t - F(x_{t+1}^*)\big\rangle \\ &= \big\|w_t - w_{t+1}^* + \gamma\big(h_t - F(x_{t+1}^*)\big)\big\|^2 - 2\gamma\big\langle w_t - w_{t+1}^* + \gamma\big(h_t - F(x_{t+1}^*)\big), h_t - F(x_{t+1}^*)\big\rangle \\ &= \big\|w_t - w_{t+1}^*\big\|^2 - \gamma^2\big\|h_t - F(x_{t+1}^*)\big\|^2, \end{aligned} \tag{23}$$

so we have

$$\mathbb{E}_{\xi_t}[V_{t+1}] \leq \big\|w_t - w_{t+1}^*\big\|^2 + \big(1-p^2\big)\gamma_2^2\big\|h_t - F(x_{t+1}^*)\big\|^2, \tag{24}$$

so we also have $w_t^* \equiv x^* - \gamma F(x^*) \triangleq w^*$.

Then by the standard analysis on GDA, we have

$$\begin{aligned} \left\|w_t - w_{t+1}^*\right\|^2 &= \|w_t - w^*\|^2 = \|x_t - x^* - \gamma(g(x_t;\xi_t) - F(x^*))\|^2 \\ &= \|x_t - x^*\|^2 - 2\gamma\langle g(x_t;\xi_t) - F(x^*), x_t - x^*\rangle + \gamma^2\|g(x_t;\xi_t) - F(x^*)\|^2, \end{aligned} \tag{25}$$

take the expectation, we have

$$\begin{aligned} \mathbb{E}_{\xi_t}\left[\|w_t - w^*\|^2\right] &= \|x_t - x^*\|^2 - 2\gamma\langle F(x_t) - F(x^*), x_t - x^*\rangle + \gamma^2\mathbb{E}_{\xi_t}\left[\|g(x_t;\xi_t) - F(x^*)\|^2\right] \\ &\le \|x_t - x^*\|^2 - 2\gamma(1-\gamma L)\langle F(x_t) - F(x^*), x_t - x^*\rangle + 2\gamma^2\sigma^2, \end{aligned} \tag{26}$$

so we have

$$\begin{aligned} &\mathbb{E}_{\xi_t}[V_{t+1}] \\ &\le \mathbb{E}_{\xi_t}\left[\|w_t - w^*\|^2 + \left(1 - p^2\right)\gamma_2^2\|h_t - F(x^*)\|^2\right] \\ &\le \mathbb{E}_{\xi_t}\left[\|x_t - x^*\|^2 - 2\gamma(1-\gamma L)\langle F(x_t) - F(x^*), x_t - x^*\rangle + 2\gamma^2\sigma^2 + \left(1 - p^2\right)\gamma_2^2\|h_t - F(x^*)\|^2\right] \\ &\le \mathbb{E}_{\xi_t}\left[(1 - 2\gamma\mu(1-\gamma L))\|x_t - x^*\|^2 + \left(1 - p^2\right)\gamma_2^2\|h_t - F(x^*)\|^2 + 2\gamma^2\sigma^2\right], \end{aligned} \tag{27}$$

recall that $\gamma \le \frac{1}{2L}$, we have

$$\begin{aligned} \mathbb{E}_{\xi_t}[V_{t+1}] &\le \mathbb{E}_{\xi_t}\left[(1-\gamma\mu)\|x_t - x^*\|^2 + \left(1 - p^2\right)\gamma_2^2\|h_t - F(x^*)\|^2 + 2\gamma^2\sigma^2\right] \\ &\le \mathbb{E}_{\xi_t}\left[\left(1 - \min\left\{\gamma\mu, p^2\right\}\right)V_t\right] + 2\gamma^2\sigma^2, \end{aligned} \tag{28}$$

by taking the full expectation, we have

$$\begin{aligned} \mathbb{E}[V_T] &\le \left(1 - \min\left\{\gamma\mu, p^2\right\}\right)\mathbb{E}[V_{T-1}] + 2\gamma^2\sigma^2 \\ &\le \left(1 - \min\left\{\gamma\mu, p^2\right\}\right)^T V_0 + 2\gamma^2\sigma^2\sum_{i=0}^{T-1}\left(1 - \min\left\{\gamma\mu, p^2\right\}\right)^i \\ &\le \left(1 - \min\left\{\gamma\mu, p^2\right\}\right)^T V_0 + \frac{2\gamma^2\sigma^2}{\min\left\{\gamma\mu, p^2\right\}}, \end{aligned} \tag{29}$$

which concludes the proof. ■

Appendix C. Proof of Corollary 2

Proof With the above setting, we know that

$$\min\left\{\gamma\mu, p^2\right\} = \gamma\mu, \tag{30}$$

and

$$\mathbb{E}[V_T] \leq (1-\gamma\mu)^T V_0 + \frac{2\gamma\sigma^2}{\mu}, \tag{31}$$

so it is easy to see that by setting

$$T \geq \frac{1}{\gamma\mu} \ln\left(\frac{2V_0}{\epsilon}\right), \quad \gamma \leq \frac{\mu\epsilon}{4\sigma^2}, \tag{32}$$

we have

$$\mathbb{E}[V_T] \leq \epsilon, \tag{33}$$

which induces the iteration complexity to be

$$T \geq \max\left\{\frac{2L}{\mu}, \frac{4\sigma^2}{\mu^2\epsilon}\right\} \ln\left(\frac{2V_0}{\epsilon}\right) \tag{34}$$

and the corresponding number of calls to the proximal oracle is

$$pT \geq \sqrt{\max\left\{\frac{2L}{\mu}, \frac{4\sigma^2}{\mu^2\epsilon}\right\}} \ln\left(\frac{2V_0}{\epsilon}\right) \tag{35}$$

which concludes the proof. ■