Stochastic Semi-Proximal Mirror-Prox

Niao He
Georgia Institute of Technology
nhe6@gatech.edu

Zaid Harchaoui
NYU, Inria
firstname.lastname@nyu.edu

Abstract

We present a direct extension of the Semi-Proximal Mirror-Prox algorithm [3] for minimizing convex composite problems with objectives consisted of three parts – a convex loss represented by stochastic oracle, a proximal-friendly regularizer and another LMO-friendly regularizer. The algorithm leverages stochastic oracle, proximal operators, and linear minimization over problems' domain and attains optimality in two-fold: i) optimal in the number of calls to the stochastic oracle representing the objective, ii) optimal in the number of calls to linear minimization oracle representing problem's domain in the "smooth" saddle point case.

1 Introduction

Last decade demonstrates significant interests in minimizing composite functions of the form:

$$\min_{x\in X} \ \tfrac{1}{n}\textstyle\sum_{i=1}^{n} f_i(x) + h(x)$$

where f_i are convex continuously differentiable functions and h is a convex but perhaps not differentiable function. Such problems arise ubiquitously in machine learning, where the first term usually refers to empirical risk and h is a regularizer. The mainstream algorithms are devoted to address two challenges, i) large number of summation, ii) nonsmoothness of regularization term. To deal with the finite sum, a variety of stochastic and incremental gradient methods have been developed that use only one or a mini batch of components at a time. To handle the nonsmooth regularization and utilize the underlying structure, two kinds of algorithms have been widely studied – the proximal type and conditional gradient type. Proximal type methods ([10, 1, 12]) require computation of a composite proximal operator at each iteration, i.e. solving problems of the form, $\min_{x\in X}\left\{\frac{1}{2}\|x\|_2^2 + \langle \xi, x\rangle + \alpha h(x)\right\}$, given input vector ξ and positive scalar α. Function h that admits easy-to-compute proximal operators, are called *proximal-friendly*. In contrast, conditional gradient type methods [2, 11] operate with the linear minimization oracle (LMO) at each iteration, i.e., solving problems of the form, $\min_{x\in X}\{\langle \xi, x\rangle + \alpha h(x)\}$, much cheaper than computing proximal operators. For instance, in the case of nuclear-norm, LMO only requires computing the leading pair of singular vectors, which is by orders of magnitude faster than full singular value decomposition required when computing proximal operator. Such h, are called *LMO-friendly*.

The scope of this paper. Despite of the much success in this regime, few work has been devoted to the situation where the regularization is given by a mixture of proximal friendly and LMO-friendly components. Such mixtures are often introduced to promote several desired properties of the solution simultaneously, such as sparsity and low rank. While recent work [3, 7] considers only deterministic case, in this paper, we focus on the stochastic setting and aim to solve problems such as
(i) stochastic composite optimization

$$\min_{x=[x_1,x_2]\in X} \mathbf{E}_{\xi}[f(x,\xi)] + h_1(x_1) + h_2(x_2) \tag{1}$$

(ii) stochastic composite saddle point problem

$$\min_{x_1\in X_1} \max_{x_2\in X_2} \mathbf{E}_{\xi}[\Phi(x_1,x_2,\xi)] + h_1(x_1) - h_2(x_2) \tag{2}$$

where $\mathbf{E}[f(x,\xi)]$ is convex in $x \in X$; $\mathbf{E}[\Phi(x_1, x_2, \xi)]$ is convex in $x_1 \in X_1$ and concave in $x_2 \in X_2$; $h_1(x_1)$ is proximal-friendly and $h_2(x_2)$ is LMO-friendly. Such problems indeed arise from many sparse and low-rank regularized models used in recommendation systems and social network prediction. To address them in a unified manner, we will focus on a broad class of variational inequalities with mixed structures, which further generalize problems (1) and (2).

Our contribution. We develop a stochastic variant of the Semi-Proximal Mirror-Prox proposed in [3] that supports and leverages stochastic oracles, proximal operators and linear minimization oracles and achieves the best of three worlds. The algorithm extends the usual conditional gradient to stochastic setting, and enjoys, in terms of LMO calls, a $O(1/\epsilon^2)$ complexity for "smooth case" (e.g f and Φ in (1) and (2) are smooth) as well as a $O(1/\epsilon^4)$ complexity for general nonsmooth case. In both situation, the algorithm attains the optimal $O(1/\epsilon^2)$ complexity of the number of stochastic oracles. To our best knowledge, the algorithm as well as theoretical results presented seem to be novel.

2 Stochastic Semi-Proximal Mirror-Prox

2.1 The situation

Structured Variational Inequalities. We consider the variational inequality VI(X, F):

$$\text{Find } x_* \in X : \langle F(x), x - x_* \rangle \geq 0, \forall x \in X$$

with domain X and operator F that satisfy the assumptions ($\mathbf{A}$.1)–($\mathbf{A}$.4) below.

($\mathbf{A}$.1) Set $X \subset E_u \times E_v$ is closed convex and its projection $PX = \{u : x = [u; v] \in X\} \subset U$, where U is convex and closed, E_u, E_v are Euclidean spaces;

($\mathbf{A}$.2) The function $\omega(\cdot) : U \to \mathbf{R}$ is continuously differentiable and also 1-strongly convex w.r.t. some norm $\|\cdot\|$. This defines the Bregman distance

$$V_u(u') = \omega(u') - \omega(u) - \langle \omega'(u), u' - u \rangle \geq \tfrac{1}{2}\|u' - u\|^2;$$

($\mathbf{A}$.3) The operator $F(x = [u, v]) : X \to E_u \times E_v$ is monotone and of form $F(u, v) = [F_u(u); F_v]$ with $F_v \in E_v$ being a constant and $F_u(u) \in E_u$ satisfying the condition

$$\forall u, u' \in U : \|F_u(u) - F_u(u')\|_* \leq L\|u - u'\| + M$$

for some $L < \infty, M < \infty$;

($\mathbf{A}$.4) The linear form $\langle F_v, v \rangle$ of $[u; v] \in E_u \times E_v$ is bounded from below on X and is coercive on X w.r.t. v: whenever $[u^t; v^t] \in X$, $t = 1, 2, ...$ is a sequence such that $\{u^t\}_{t=1}^{\infty}$ is bounded and $\|v^t\|_2 \to \infty$ as $t \to \infty$, we have $\langle F_v, v^t \rangle \to \infty$, $t \to \infty$.

Semi-structured Variational Inequalities. The class of semi-structured variational inequalities allows to go beyond Assumptions ($\mathbf{A}$.1) – ($\mathbf{A}$.4), by assuming more sub-structure. This structure is consistent with what we call a *semi-proximal* setup, which encompasses both the regular *proximal setup* and the regular *linear minimization setup* as special cases. Indeed, we assume further

($\mathbf{S}$.1) *Proximal setup for* X: we assume that $E_u = E_{u_1} \times E_{u_2}$, $E_v = E_{v_1} \times E_{v_2}$, and $U \subset U_1 \times U_2$, $X = X_1 \times X_2$ with $X_i \in E_{u_i} \times E_{v_i}$ and $P_i X = \{u_i : [u_i; v_i] \in X_i\} \subset U_i$ for $i = 1, 2$, where U_1 is convex and closed, U_2 is convex and compact. We also assume that $\omega(u) = \omega_1(u_1) + \omega_2(u_2)$ and $\|u\| = \|u_1\|_{E_{u_1}} + \|u_2\|_{E_{u_2}}$, with $\omega_2(\cdot) : U_2 \to \mathbf{R}$ continuously differentiable such that

$$\omega_2(u_2') \leq \omega_2(u_2) + \langle \nabla \omega_2(u_2), u_2' - u_2 \rangle + \tfrac{L_0}{\kappa}\|u_2' - u_2\|_{E_{u_2}}^{\kappa}, \forall u_2, u_2' \in U_2;$$

for a particular $1 < \kappa \leq 2$ and $L_0 < \infty$. Furthermore, we assume that the $\|\cdot\|_{E_{u_2}}$-diameter of U_2 is bounded by some $D > 0$.

($\mathbf{S}$.2) *Partition of* F: operator F induced by the above partition of X_1 and X_2 can be written as

$$F(x) = [F_u(u); F_v] \text{ with } F_u(u) = [F_{u_1}(u_1, u_2); F_{u_2}(u_1, u_2)], F_v = [F_{v_1}; F_{v_2}].$$

($\mathbf{S}$.3) *Proximal mapping on* X_1: we assume that for any $\eta_1 \in E_{u_1}$ and $\alpha > 0$, we have at our disposal easy-to-compute prox-mappings of the form,

$$\text{Prox}_{\omega_1}(\eta_1, \alpha) := \text{argmin}_{x_1 = [u_1; v_1] \in X_1} \; \{\omega_1(u_1) + \langle \eta_1, u_1 \rangle + \alpha \langle F_{v_1}, v_1 \rangle\}.$$

(**S**.4) *Linear minimization oracle for X_2*: we assume that we we have at our disposal Composite Linear Minimization Oracle (LMO), which given any input $\eta_2 \in E_{u_2}$ and $\alpha > 0$, returns an optimal solution to the minimization problem with linear form, that is,

$$\mathrm{LMO}(\eta_2, \alpha) := \mathrm{argmin}_{x_2=[u_2;v_2]\in X_2} \; \{\langle \eta_2, u_2\rangle + \alpha\langle F_{v_2}, v_2\rangle\}\,.$$

Stochastic Oracles. We are interested in the situation where we only have access to noisy information on $F_u(u)$. More specifically, we assume that the operator is represented by the following stochastic oracle, such that for any $u \in U$, it returns a vector $g(u, \xi)$ satisfying

(**C**.1) *Unbiasedness and bounded variance*: $\mathbf{E}[g(u,\xi)] = F_u(u),\; \mathbf{E}[\|g(u,\xi) - F_u(u)\|_*^2] \le \sigma^2$,

(**C**.2) *Light tail:* $\mathbf{E}\left[\exp\{\|g(u,\xi) - F_u(u)\|_*^2/\sigma^2\}\right] \le \exp\{1\}$ for some $\sigma > 0$.

where $\|\cdot\|_*$ is the dual norm same as in (**A**.3). Note that by Jensen's inequality, (**C**.2) implies (**C**.1).

2.2 Stochastic Semi-Proximal Mirror-Prox

We present our Stochastic Semi-Proximal Mirror-Prox in Algoirthm 1. The algorithm blends composite Mirror Prox and composite Conditional Gradient to exploit the best efficiency out of the mixed structure. At each iteration t, we estimate the operator F by taking the average of m_t vectors returned by the stochastic oracle. For sub-domain X_2 given by LMO, instead of computing exactly the prox-mapping, we mimick the prox-mapping via a conditional gradient algorithm (denoted as **CCG**) directly from [3]. For the sub-domain X_1, we compute the prox-mapping as it is.

Algorithm 1 Stochastic Semi-Proximal Mirror-Prox Algorithm for Semi-VI(X, F)

```
Input: stepsizes $\gamma_t > 0$, accuracies $\epsilon_t \ge 0$, $t = 1, 2, \ldots$
[1] Initialize $x^1 = [x_1^1; x_2^1] \in X$, where $x_1^1 = [u_1^1; v_1^1]; x_2^1 = [u_2^1; v_2^1]$.
for $t = 1, 2, \ldots, T$ do
  [2] Set $u^t = [u_1^t; u_2^t]$, compute $[g_1^t; g_2^t] = \frac{1}{m_t}\sum_{j=1}^{m_t} g(u^t, \xi_j^t)$ and $y^t = [y_1^t; y_2^t]$ that
        $y_1^t := [\widehat{u}_1^t; \widehat{v}_1^t] \;=\; \mathrm{Prox}_{\omega_1}(\gamma_t g_1^t - \omega_1'(u_1^t), \gamma_t)$
        $y_2^t := [\widehat{u}_2^t; \widehat{v}_2^t] \;=\; \mathbf{CCG}(X_2, \omega_2(\cdot) + \langle \gamma_t g_2^t - \omega_2'(u_2^t), \cdot\rangle, \gamma_t F_{v_2}; \epsilon_t)$
  [3] Set $\widehat{u}^t = [\widehat{u}_1^t; \widehat{u}_2^t]$, compute $[\widehat{g}_1^t; \widehat{g}_2^t] = \frac{1}{m_t}\sum_{j=m_t+1}^{2m_t} g(\widehat{u}^t, \xi_j^t)$ and $x^{t+1} = [x_1^{t+1}; x_2^{t+1}]$ that
        $x_1^{t+1} := [u_1^{t+1}; v_1^{t+1}] \;=\; \mathrm{Prox}_{\omega_1}(\gamma_t \widehat{g}_1^t - \omega_1'(u_1^t), \gamma_t)$
        $x_2^{t+1} := [u_2^{t+1}; v_2^{t+1}] \;=\; \mathbf{CCG}(X_2, \omega_2(\cdot) + \langle \gamma_t \widehat{g}_2^t - \omega_2'(u_2^t), \cdot\rangle, \gamma_t F_{v_2}; \epsilon_t)$
end for
Output: $\bar{x}_T := [\bar{u}_T; \bar{v}_T] = \left(\sum_{t=1}^T \gamma_t\right)^{-1} \sum_{t=1}^T \gamma_t y^t$
```

The CCG routine [3] is designed to solve smooth semi-linear problems: $\min_{x=[u;v]\in X}\{\phi^+(u,v) = \phi(u) + \langle c, v\rangle\}$. For the input pair $(X, \phi, c; \epsilon)$, the algorithm works as follows:

$$\begin{array}{ll}
(a) & x^1 := [u^1; v^1] \in X; \\
(b) & \widehat{x}^t := [\widehat{u}^t; \widehat{v}^t] = \mathrm{argmin}_{x=[u;v]\in X}\{\langle \nabla\phi(u^t), u\rangle + \langle c, v\rangle\} \\
 & \text{compute } \delta_t = \langle \nabla\phi^+(x^t), x^t - \widehat{x}^t\rangle, \text{return if } \delta_t \le \epsilon \\
(c) & x^{t+1} := [u^{t+1}; v^{t+1}] \text{ s.t. } \phi^+(x^{t+1}) \le \phi^+(x^t + \frac{2}{t+1}(\widehat{x}^t - x^t))
\end{array} \qquad (\mathbf{CCG}(\mathbf{X}, \phi, \mathbf{c}; \epsilon))$$

We provide the convergence analysis in the next section.

3 Convergence Results

3.1 Main Results

Let us denote the resolution of the execution protocal $\mathcal{I}_T = \{y^t, F(y^t)\}_{t=1}^T$ and a collection $\lambda^T = \{\lambda_t\}_{t=1}^T$ with $\lambda_t = \gamma_t / \sum_{t=1}^T \gamma_t$ on any set $X' \subset X$ by

$$\mathrm{Res}(X'|\mathcal{I}_T, \lambda^T) = \sup_{x\in X'} \sum_{t=1}^T \lambda_t \langle F(y^t, y^t - x\rangle.$$

The resolution can be used to certify the accuracy of the approximate solution $\bar{x}_T = \sum_{t=1}^T \lambda_t y^t$ to generic convex problems not limited to variational inequalities (see [8] for details). We arrive at the following

Theorem 3.1. *Let the stepsizes γ_t satisfy $0 < \gamma_t \leq (\sqrt{3}L)^{-1}$ and $\Theta[X'] = \sup_{[u;v]\in X'} V_{u^1}(u)$ for any $X' \subset X$. For a sequence of inexact prox-mappings with inexactness $\epsilon_t \geq 0$ and batch size $m_t > 0$, we have under assumption $(\boldsymbol{C}.1)$ that*

$$\mathbf{E}[\text{Res}(X'|\mathcal{I}_T, \lambda^T)] \leq \mathcal{M}_0(T) := \left(\textstyle\sum_{t=1}^T \gamma_t\right)^{-1} \left(2\Theta[X'] + \tfrac{7}{2}\textstyle\sum_{t=1}^T \gamma_t^2 (M^2 + \tfrac{2\sigma^2}{m_t}) + 2\textstyle\sum_{t=1}^T \epsilon_t\right).$$

Moreover, if assumption $(\boldsymbol{C}.2)$ holds, then for any $\Lambda > 0$,

$$\text{Prob}\left\{\text{Res}(X'|\mathcal{I}_T, \lambda^T) \geq \mathcal{M}_0(T) + \Lambda\mathcal{M}_1(T)\right\} \leq \exp\{-\Lambda^2/3\} + \exp\{-\Lambda\}$$

where $\mathcal{M}_1(T) = (\sum_{t=1}^T \gamma_t)^{-1} \left(\frac{7}{2} \sum_{t=1}^T \frac{\gamma_t^2\sigma^2}{m_t} + 3\Theta[X]\sqrt{\sum_{t=1}^T \frac{\gamma_t^2\sigma^2}{m_t}}\right).$

As a corollary, we can derive (based on immediate results in [8])

(i) when F is a monotone vector field, the resulting efficiency estimate takes place for the dual gap of variational inequalities, i.e. $\mathbf{E}[\epsilon_{\text{VI}}(\bar{x}_T \big| X, F)] \leq \mathcal{M}_0(T)$;

(ii) when F stems from the (sub)gradient of a convex minimization problem $\min_{x\in X} f(x)$ (for instance, problem (1)) with optimal solution being x_*, the resulting efficiency estimate takes place for the suboptimality, i.e. $\mathbf{E}[f(\bar{x}_T) - f(x_*)] \leq \mathcal{M}_0(T)$;

(iii) when F stems from a convex-concave saddle point problem $\min_{x^1\in X_1} \max_{x^2\in X_2} \Phi(x^1, x^2)$ (for instance, problem (2)), along with two induced convex optimization problems

$$\begin{array}{rcll} \text{Opt}(P) & = & \min_{x^1\in X_1} \left[\overline{\Phi}(x^1) = \sup_{x^2\in X_2} \Phi(x^1, x^2)\right] & (P) \\ \text{Opt}(D) & = & \max_{x^2\in X_2} \left[\underline{\Phi}(x^2) = \inf_{x^1\in X_1} \Phi(x^1, x^2)\right] & (D) \end{array}$$

the resulting efficiency estimate is inherited both by primal and dual suboptimality gaps, i.e. $\mathbf{E}[\overline{\Phi}(\bar{x}_T^1) - \text{Opt}(P)] \leq \mathcal{M}_0(T)$ and $\mathbf{E}[\text{Opt}(D) - \underline{\Phi}(\bar{x}_T^2)] \leq \mathcal{M}_0(T)$.

3.2 When F_u is Lipschitz continuous ($M = 0$)

In the case when F_u is a Lipschitz continuous monotone operator with $L > 0$ and $M = 0$, we have

Proposition 3.1. *Under the assumptions $(\boldsymbol{A}.1) - (\boldsymbol{A}.4)$, $(\boldsymbol{S}.1) - (\boldsymbol{S}.4)$ with $M = 0$ and the proximal setup on U_2 being Euclidean setup. Setting stepsize $\gamma_t = (\sqrt{2}L)^{-1}$ and batch size $m_t = O(\gamma_t^2\sigma^2 T/\Theta[X]), t = 1, \ldots, T$, for the Stochastic Semi-Proximal Mirror-Prox algorithm to return an stochastic ϵ-solution to the variational inequality $VI(X, F)$ represented by stochastic oracle in $(\boldsymbol{C}.1) - (\boldsymbol{C}.2)$, the total number of stochastic oracle calls required does not exceed*

$$\mathcal{N}_{SO} = O(1)\sigma^2\Theta[X]/\epsilon^2$$

and the total number of calls to the Linear Minimization Oracle does not exceed

$$\mathcal{N}_{LMO} = O(1)L^2D^2\Theta[X]/\epsilon^2$$

where $\sigma^2, D, \Theta[X]$ are defined previously.

3.3 When F_u is bounded ($L = 0$)

In the case when F_u is a uniformly bounded monotone operator with $M > 0$ and $L = 0$, we have

Proposition 3.2. *Under same assumptions with $L = 0$. Setting stepsize $\gamma_t = O(1)\frac{\sqrt{\Theta[X]}}{M\sqrt{T}}$ and batch size $m_t = O(1)\frac{\sigma^2}{M^2}, t = 1, \ldots, T$, for the Stochastic Semi-Proximal Mirror-Prox algorithm to return an stochastic ϵ-solution to the variational inequality $VI(X, F)$, the total number of stochastic oracle calls required does not exceed*

$$\mathcal{N}_{SO} = O(1)\sigma^2\Theta[X]/\epsilon^2$$

and the total number of calls to the Linear Minimization Oracle does not exceed

$$\mathcal{N}_{LMO} = O(1)M^4D^2\Theta[X]/\epsilon^4.$$

Discussion. The stochastic variant of Semi-Proximal Mirror-Prox algorithm is designed to solve a broad class of variational inequalities allowing to cover the stochastic composite minimization and saddle point problems in (1) and (2). The algorithm enjoys the *optimal complexity bounds*, i.e. $O(1/\epsilon^2)$, both in terms of the number of calls to stochastic oracle (see [9]) and the number of calls to linear minimization oracle (see [6]) when the underlying problem is in the saddle point form and associated monotone operator is Lipschitz continuous. In the general nonsmooth situation, the algorithm enjoys still optimal complexity bound $O(1/\epsilon^2)$ in terms of the number of calls to stochastic oracles, but that of the linear minimization oracle becomes $O(1/\epsilon^4)$.

References

[1] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[2] Zaid Harchaoui, Anatoli Juditsky, and Arkadi Nemirovski. Conditional gradient algorithms for norm-regularized smooth convex optimization. *Mathematical Programming*, pages 1–38, 2013.

[3] Niao He and Zaid Harchaoui. Semi-proximal mirror-prox for nonsmooth composite minimization. *Neural Information Processing Systems (NIPS)*, 2015. `http://arxiv.org/pdf/1507.01476.pdf`.

[4] Niao He, Anatoli Juditsky, and Arkadi Nemirovski. Mirror prox algorithm for multi-term composite minimization and semi-separable problems. *Computational Optimization and Applications*, 61(2):275–319, 2015.

[5] Anatoli Juditsky, Arkadi Nemirovski, Claire Tauvel, et al. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

[6] Guanghui Lan. The complexity of large-scale convex programming under a linear optimization oracle. *arXiv*, 2013.

[7] Cun Mu, Yuqian Zhang, John Wright, and Donald Goldfarb. Scalable robust matrix recovery: Frank-wolfe meets proximal methods. *arXiv preprint arXiv:1403.7588*, 2014.

[8] Arkadi Nemirovski, Shmuel Onn, and Uriel G Rothblum. Accuracy certificates for computational problems with convex structure. *Mathematics of Operations Research*, 35(1):52–78, 2010.

[9] A. S. Nemirovsky and D. B. Yudin. Problem complexity and method efficiency in optimization. *Wiley-Interscience Series in Discrete Mathematics*, 1983.

[10] Yu Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

[11] Yurii Nesterov. Complexity bounds for primal-dual methods minimizing the model of objective function. Technical report, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2015.

[12] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):123–231, 2013.

4 Technical Details.

4.1 Problems (1) and (2) in the Form of Semi-Structured Variational Inequality

The stochastic composite optimization in the form of (1)

$$\min_{u=[u_1,u_2]\in U=U_1\times U_2} \mathbf{E}_\xi[f(u,\xi)] + h_1(u_1) + h_2(u_2)$$

can be written as

$$\min_{u\in X, v_1\geq h_1(u_1), v_2\geq h_2(u_2)} \mathbf{E}_\xi[f(u,\xi)] + v_1 + v_2.$$

The variational inequality VI(X,F) associated with the above problem is given by

$$X = \{x = [u = [u_1,u_2]; v = [v_1,v_2]] : u_1 \in U_1, u_2 \in U_2, v_1 \geq h_1(u_1), v_2 \geq h_2(u_2)\},$$
$$F(x = [u = [u_1,u_2]; v = [v_1,v_2]]) = [[\partial_{u_1}\mathbf{E}_\xi[f(u,\xi)], \partial_{u_2}\mathbf{E}_\xi[f(u,\xi)]]; [1,1]].$$

Clearly, the above VI(X,F) satisfies the assumptions $(\mathbf{A}.1)-(\mathbf{A}.4)$, $(\mathbf{S}.1)-(\mathbf{S}.4)$.

Similarly, for the stochastic composite saddle point problem in the form of (2)

$$\min_{u_1\in U_1}\max_{u_2\in U_2} \mathbf{E}_\xi[\Phi(u_1,u_2,\xi)] + h_1(u_1) - h_2(u_2)$$

and its reformulation

$$\min_{u_1\in U_1, v_1\geq h_1(v_1)}\max_{u_2\in U_2, v_2\leq h_2(v_2)} \mathbf{E}_\xi[\Phi(u_1,u_2,\xi)] + v_1 - v_2$$

the variational inequality VI(X,F) associated with the above problem is given by

$$X = \{x = [u = [u_1,u_2]; v = [v_1,v_2]] : u_1 \in U_1, u_2 \in U_2, v_1 \geq h_1(u_1), v_2 \geq h_2(u_2)\},$$
$$F(x = [u = [u_1,u_2]; v = [v_1,v_2]]) = [[\partial_{u_1}\mathbf{E}_\xi[\Phi(u_1,u_2,\xi)], -\partial_{u_2}\mathbf{E}_\xi[\Phi(u_1,u_2,\xi)]]; [1,1]].$$

Clearly, the above VI(X,F) also satisfies the assumptions $(\mathbf{A}.1)-(\mathbf{A}.4)$, $(\mathbf{S}.1)-(\mathbf{S}.4)$.

4.2 Proof of Theorem 3.1

Proof. The proof builds upon [5] and [4]. First of all, we show that

Lemma 4.1. *For any $\epsilon \geq 0$, $x = [u;v] \in X$ and $\xi = [\eta;\zeta] \in E$, let $[u';v'] = P_x^\epsilon(\xi)$, where*

$$P_x^\epsilon(\xi) = \{\widehat{x} = [\widehat{u};\widehat{v}] \in X : \langle \eta + \omega'(\widehat{u}) - \omega'(u), \widehat{u} - s\rangle + \langle \zeta, \widehat{v} - w\rangle \leq \epsilon \ \forall [s;w] \in X\},$$

we have for all $[s;w] \in X$,

$$\langle \eta, u' - s\rangle + \langle \zeta, v' - w\rangle \leq V_u(s) - V_{u'}(s) - V_u(u') + \epsilon. \tag{3}$$

1^0. When applying Lemma 4.1 with $[u;v] = x^t$, $\xi = [\gamma_\tau g^t; \gamma_\tau F_v]$, $[u';v'] = y^t$, and $[s;w] = [u^{t+1};v^{t+1}] = x^{t+1}$ we obtain:

$$\gamma_\tau[\langle g^t, \widehat{u}^t - u^{t+1}\rangle + \langle F_v, \widehat{v}^t - v^{t+1}\rangle] \leq V_{u^t}(u^{t+1}) - V_{\widehat{u}^t}(u^{t+1}) - V_{u^t}(u'_\tau) + \epsilon_\tau \tag{4}$$

and applying Lemma 4.1 with $[u;v] = x_\tau$, $\xi = \gamma_\tau[\widehat{g}^t, F_v]$, $[u';v'] = x^{t+1}$, and $[s;w] = z \in X$ we get:

$$\gamma_\tau[\langle \widehat{g}^t, u^{t+1} - s\rangle + \langle F_v, v^{t+1} - w\rangle] \leq V_{u^t}(s) - V_{u^{t+1}}(s) - V_{u^t}(u^{t+1}) + \epsilon_\tau. \tag{5}$$

Adding (5) to (4) we obtain for every $z = [s;w] \in X$

$$\gamma_t[\langle \widehat{g}^t, \widehat{u}^t - s\rangle + \langle F_v, \widehat{v}^t - w\rangle] \leq V_{u^t}(s) - V_{u^{t+1}}(s) + \sigma_t + 2\epsilon_t\ , \tag{6}$$

for any $[s,w] \in X$, where

$$\sigma_t := \gamma_t\langle \widehat{g}^t - g^t, \widehat{u}^t - u^{t+1}\rangle - V_{\widehat{u}^t}(u^{t+1}) - V_{u^t}(\widehat{u}_t)\ .$$

Let $\Delta_t = F_u(\widehat{u}^t) - \widehat{g}^t$, then for any $z = [s, w] \in X$, we have

$$\sum_{t=1}^{T} \gamma_t \langle F(y^t), y^t - z \rangle \le \Theta[X] + \sum_{t=1}^{T} \sigma_t + \sum_{t=1}^{T} 2\epsilon_t + \sum_{t=1}^{T} \gamma_t \langle \Delta_t, \widehat{u}^t - s \rangle \tag{7}$$

Let $e_t = \|g^t - g^t\|_*$ and $\widehat{e}_t = \|\widehat{g}^t - F_u(\widehat{u}^t)\|_* = \|\Delta_t\|_*$, Then we have

$$\begin{aligned}\|\widehat{g}^t - g^t\|_*^2 &= \|(\widehat{g}^t - F_u(\widehat{u}^t) + (F_u(\widehat{u}^t) - g^t) + (g^t - g^t)\|_*^2 \\ &\le (\widehat{e}_t + L\|\widehat{u}^t - u^t\| + M + e_t)^2 \\ &\le 3L^2\|\widehat{u}^t - u^t\|^2 + 3M^2 + 3(e_t + \widehat{e}_t)^2\end{aligned}$$

Hence,

$$\sigma_t \le \frac{\gamma_t^2}{2}\|\widehat{g}^t - g^t\|_*^2 + \frac{1}{2}\|\widehat{u}^t - u^{t+1}\|_2 - V_{\widehat{u}^t}(u^{t+1}) - V_{u^t}(\widehat{u}_t) \le \frac{\gamma_t^2}{2}\|\widehat{g}^t - g^t\|_*^2 - \frac{1}{2}\|\widehat{u}^t - u^t\|^2.$$

Since the stepsize γ_t satisfy that $3\gamma_t^2 L \le 1$, we further have

$$\sigma_t \le \frac{3\gamma_t^2}{2}[M^2 + (e_t + \widehat{e}_t)^2]. \tag{8}$$

Define a special sequence $\widetilde{u}^t$ such that

$$\widetilde{u}^1 = u^1; \quad \widetilde{u}^{t+1} = \operatorname*{argmin}_{u \in P_u X}\{\langle \gamma_t \Delta_t, u \rangle + V_{\widetilde{u}^t}(u)\}$$

The sequence defined above satisfies the following relation (see Corollary 2 in [5] for details): for any $z = [s, w] \in X$,

$$\sum_{t=1}^{T} \gamma_t \langle \Delta_t, \widetilde{u}^t - s \rangle \le \Theta[X] + \sum_{t=1}^{t} \frac{\gamma_t^2}{2}\|\Delta_t\|_*^2 = \Theta[X] + \sum_{t=1}^{t} \frac{\gamma_t^2}{2}\widehat{e}_t \tag{9}$$

Combining (7), (8), (9), we end up with

$$\text{Res}(X'|\mathcal{I}_T, \lambda^T) \le (\sum_{t=1}^{T} \gamma_t)^{-1} \left(2\Theta[X] + \sum_{t=1}^{T} \frac{7\gamma_t^2}{2}[M^2 + (e_t^2 + \widehat{e}_t^2)] + \sum_{t=1}^{T} 2\epsilon_t + \sum_{t=1}^{T} \gamma_t \langle \Delta_t, \widehat{u}^t - \widetilde{u}^t \rangle \right) \tag{10}$$

2^0. Under Assumption (**C**.1), we have

$$\mathbf{E}[\Delta_t|\mathcal{F}_t] = 0,\ \mathbf{E}[e_t^2|\mathcal{G}_{t-1}] \le \frac{\sigma^2}{m_t},\ \text{and } \mathbf{E}[\widehat{e}_t^2|\mathcal{F}_t] \le \frac{\sigma^2}{m_t}.$$

where $\mathcal{F}_t = \sigma(\xi_1^1, \dots, \xi_{2m_t}^1, \dots, \xi_1^t, \dots, \xi_{m_t}^t)$ and $\mathcal{G}_t = \sigma(\xi_1^1, \dots, \xi_{2m_t}^1, \dots, \xi_1^t, \dots, \xi_{2m_t}^t)$.

One can further show that $\mathbf{E}[\langle \Delta_t, \widehat{u}^t - \widetilde{u}^t \rangle] = 0$. It follows from (10) that

$$\mathbf{E}[\text{Res}(X'|\mathcal{I}_T, \lambda^T)] \le (\sum_{t=1}^{T} \gamma_t)^{-1} \left(2\Theta[X] + \sum_{t=1}^{T} \frac{7\gamma_t^2}{2}[M^2 + \frac{2\sigma^2}{m_t}] + \sum_{t=1}^{T} 2\epsilon_t \right) \tag{11}$$

which proves the first part of the theorem.

3^0. Under Assumption (**C**.2), we have

$$\mathbf{E}[\exp\{e_t^2/(\sigma/\sqrt{m_t})^2\}] \le \exp\{1\} \text{ and } \mathbf{E}[\exp\{\widehat{e}_t^2/(\sigma/\sqrt{m_t})^2\}] \le \exp\{1\}.$$

Let $C_1 = \sum_{t=1}^{T} \frac{\gamma_t^2 \sigma^2}{m_t}$, it follows from convexity and the above equation that

$$\mathbf{E}\left[\exp\left\{\frac{1}{C_1}\sum_{t=1}^{T} \gamma_t^2 (e_t^2 + \widehat{e}_t^2)\right\}\right] \le \mathbf{E}\left[\frac{1}{C_1}\sum_{t=1}^{T} \frac{\gamma_t^2 \sigma^2}{m_t} \exp\left\{(e_t^2 + \widehat{e}_t^2)/(\sigma/m_t)^2\right\}\right] \le \exp\{2\}.$$

Applying Markov's inequality, we obtain:

$$\forall \Lambda > 0: \text{ Prob}\left(\sum_{t=1}^{T} \gamma_t^2 (e_t^2 + \widehat{e}_t^2) \geq (2+\Lambda) C_1\right) \leq \exp\{-\Lambda\}. \tag{12}$$

Let $\zeta_t = \langle \Delta_t, \widehat{u}^t - \widetilde{u}^t \rangle$. We showed earlier that $\mathbf{E}[\zeta_t] = 0$. since $\|\widehat{u}^t - \widetilde{u}^t\| \leq 2\sqrt{2}\Theta[X]$, then we also have

$$\mathbf{E}[\exp\{\zeta_t^2/(2\sqrt{2}\Theta[X]\sigma/\sqrt{m_t})^2\}] \leq \exp\{1\}$$

Applying the relation $\exp\{x\} \leq x + \exp\{9x^2/16\}$, one has for any $s \geq 0$,

$$\mathbf{E}\left[\exp\left\{s\sum_{t=1}^{T} \gamma_t \zeta_t\right\}\right] \leq \mathbf{E}\left[\frac{9s^2}{16}\exp\left\{\sum_{t=1}^{T}\gamma_t^2 \zeta_t^2\right\}\right] \leq \exp\left\{\frac{9s^2}{16}\sum_{t=1}^{T}\frac{8\sigma^2\Theta[X]^2\gamma_t^2}{m_t}\right\}$$

By Markov's inequality, one has

$$\forall \Lambda > 0: \text{ Prob}\left(\sum_{t=1}^{T}\gamma_t\zeta_t \geq 3\Lambda\Theta[X]\sqrt{\sum_{t=1}^{T}\frac{\sigma^2\gamma_t^2}{m_t}}\right) \leq \exp\{-\Lambda^2/2\} \tag{13}$$

Combing equation (10), (12), and (13), we arrive at

$$\forall \Lambda > 0 \text{ Prob}\big(\text{Res}(X'|\mathcal{I}_T, \lambda^T) \geq \mathcal{M}_0(T) + \Lambda\mathcal{M}_1(T)\big) \leq \exp\{-\Lambda\} + \exp\{-\Lambda^2/2\}$$

where

$$\begin{aligned} \mathcal{M}_0(T) &= (\textstyle\sum_{t=1}^{T}\gamma_t)^{-1}\left(2\Theta[X] + \sum_{t=1}^{T}\frac{7\gamma_t^2}{2}[M^2 + \frac{2\sigma^2}{m_t}] + \sum_{t=1}^{T} 2\epsilon_t\right), \\ \mathcal{M}_1(T) &= (\textstyle\sum_{t=1}^{T}\gamma_t)^{-1}\left(\frac{7}{2}\sum_{t=1}^{T}\frac{\gamma_t^2\sigma^2}{m_t} + 3\Theta[X]\sqrt{\sum_{t=1}^{T}\frac{\gamma_t^2\sigma^2}{m_t}}\right). \end{aligned}$$

□

4.3 Proof of Proposition 3.1

Proof. Let us fix T as the number of Mirror Prox steps, and since $M = 0$, from Theorem 3.1, the efficiency estimate of the variational inequality implies that

$$\mathbf{E}[\epsilon_{\text{VI}}(\bar{x}_T|X,F)] \leq \frac{2\Theta[X] + 7\sum_{t=1}^{T}\gamma_t^2\frac{\sigma^2}{m_t} + 2\sum_{t=1}^{T}\epsilon_t}{\sum_{t=1}^{T}\gamma_t}.$$

Let us fix $\epsilon_t = \frac{\Theta[X]}{T}$ for each $t = 1, \ldots, T$, then from Proposition 3.1 in [3], it takes at most $s = O(1)(\frac{L_0 D^\kappa T}{\Theta[X]})^{1/(\kappa-1)}$ calls to the LMO oracles to generate a point such that $\Delta_s \leq \epsilon_t$. Moreover, we have

$$\mathbf{E}[\epsilon_{\text{VI}}(\bar{x}_T|X,F)] \leq O(1)\frac{L\Theta[X]}{T}.$$

Therefore, to ensure $\mathbf{E}[\epsilon_{\text{VI}}(\bar{x}_T|X,F) \leq \epsilon]$ for a given accuracy $\epsilon > 0$, the number of Mirror Prox steps T is at most $O(\frac{L\Theta[X]}{\epsilon})$. Therefore, the number of stochastic oracle calls used is at most $\mathcal{N}_{\text{SO}} = \sum_{t=1}^{T} m_t = O(\gamma_t^2 T^2\sigma^2/\Theta[X]) = O(1)\frac{\sigma^2\Theta[X]}{\epsilon^2}$. Moreover, the number of linear minimization oracle calls on X_2 needed is at most $\mathcal{N}_{\text{LMO}} = sT = O(1)\left(\frac{L_0 L^\kappa D^\kappa}{\epsilon^\kappa}\right)^{1/(\kappa-1)}\Theta[X]$. In particular, if $\kappa = 2$ and $L_0 = 1$, this becomes $\mathcal{N}_{\text{LMO}} = O(1)\frac{L^2 D^2\Theta[X]}{\epsilon^2}$. □

4.4 Proof of Proposition 3.2

Proof. Let us fix T as the number of Mirror Prox steps, and let , from $\epsilon_t = \frac{\Theta[X]}{T}$ for each $t = 1, \ldots, T$, since $m_t = O(1)\sigma^2/M^2$, from Theorem 3.1, the efficiency estimate of the variational inequality implies that

$$\mathbf{E}[\epsilon_{\text{VI}}(\bar{x}_T|X,F)] \leq O(1)\frac{\sqrt{\Theta[X]}M}{\sqrt{T}}.$$

Therefore, to ensure $\mathbf{E}[\epsilon_{\text{VI}}(\bar{x}_T|X, F) \leq \epsilon]$ for a given accuracy $\epsilon > 0$, the number of Mirror Prox steps T is at most $O(\frac{M^2\Theta[X]}{\epsilon^2})$. Hence, the number of stochastic oracle calls used is at most $\mathcal{N}_{\text{SO}} = \sum_{t=1}^{T} m_t = O(1)\sigma^2 T/M^2 = O(1)\frac{\sigma^2\Theta[X]}{\epsilon^2}$. From Proposition 3.1 in [3], it takes at most $s = O(1)\frac{D^2 T}{\Theta[X]}$ calls to the LMO oracles to generate a point such that $\Delta_s \leq \epsilon_t$. Hence, the number of linear minimization oracle calls on X_2 needed is at most $\mathcal{N}_{\text{LMO}} = sT = O(1)\frac{D^2 T^2}{\Theta[X]} = O(1)\frac{D^2 M^4 \Theta[X]}{\epsilon^4}$. □