Recall the setting:

$$v^* = \inf_{\theta \in \mathbb{R}^d} \{ \varphi(\theta) \triangleq F(\theta) + r(\theta) \} \quad \text{— (P)}$$

where $F : \mathbb{R}^d \to \mathbb{R}$ is ρ-weakly convex and $r = \mathbb{1}_C$ for some closed convex set C.

We assume that F has a finite-sum structure:

$$F(\theta) = \sum_{i=1}^{n} F_i(\theta).$$

Projected stochastic subgradient method (PSSM)

a) Sample $\{1, \dots, n\}$ uniformly at random to get ξ_t

$$\Pr[\xi_t = i] = \frac{1}{n} \quad ; \quad i = 1, \dots, n.$$

b) $\theta^{t+1} \leftarrow \Pi_C(\theta^t - \alpha_t s(\theta^t, \xi_t))$, $\quad s(\theta^t, \xi_t) \in \partial F_{\xi_t}(\theta^t)$

To analyze PSSM, we need a measure that can monitor its progress. As discussed, the norm of the Moreau envelope of φ is a good candidate.

<u>Definition:</u> Given $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ and $\lambda > 0$, define

$$\text{prox}_{\lambda\varphi}(\theta) = \underset{\gamma \in \mathbb{R}^d}{\text{argmin}} \left\{ \varphi(\gamma) + \frac{1}{2\lambda} \|\gamma - \theta\|_2^2 \right\} \quad \text{(proximal map)}$$

$$\varphi_\lambda(\theta) = \min_{\gamma \in \mathbb{R}^d} \left\{ \varphi(\gamma) + \frac{1}{2\lambda} \|\gamma - \theta\|_2^2 \right\} \quad \text{(Moreau envelope)}$$

We say that $\bar{\theta} \in \mathbb{R}^d$ is an (ε, δ)-approximate stationary point of (P) if $\bar{\theta} \in \{ \gamma : \text{dist}(0, \partial\varphi(\gamma)) \le \varepsilon \} + \delta B(0,1)$.

With the above setup, let us analyze PSSM under the following assumptions:

A) $\mathbb{E}_\xi[s(\theta, \xi)] \in \partial F(\theta)$, $\quad \forall\, \theta \subseteq U$ open, $U \supseteq \text{dom}(r)$

A) $\mathbb{E}_{\xi}[s(\theta, \xi)] \in \partial F(\theta), \quad \forall\, \theta \in U$ open, $U \supseteq \mathrm{dom}(r)$

(unbiasedness)

B) $\mathbb{E}_{\xi}[\| s(\theta, \xi) \|_2^2] \le L^2, \quad \forall\, \theta \in \mathrm{dom}(r)$ (variance control)

C) After T iterations, output θ^{t^*}, where $t^* \in \{0, 1, \dots, T\}$ is sampled according to

$$\Pr[t^* = t] = \frac{\alpha_t}{\sum_{j=0}^{T} \alpha_j}.$$

Theorem: (Iteration Complexity of PSSM)

For any $\bar{\rho} > \rho$,

$$\mathbb{E}\Big[\varphi_{\underbrace{1/\bar{\rho}}_{\lambda}}(\theta^{t+1}) \Big] \le \mathbb{E}\Big[\varphi_{1/\bar{\rho}}(\theta^t) \Big] - \frac{\alpha_t(\bar{\rho} - \rho)}{\bar{\rho}} \mathbb{E}\Big[\| \nabla \varphi_{1/\bar{\rho}}(\theta^t) \|_2^2 \Big]$$

$$+ \frac{\alpha_t^2 \bar{\rho} L^2}{2},$$

("sufficient decrease" in terms of $\varphi_{1/\bar{\rho}}$)

So that

$$\mathbb{E}\Big[\| \nabla \varphi_{1/\bar{\rho}}(\theta^{t^*}) \|_2^2 \Big] \le \frac{\bar{\rho}}{\bar{\rho} - \rho} \, \frac{(\varphi_{1/\bar{\rho}}(\theta^0) - v^*) + \frac{\bar{\rho} L^2}{2} \sum_{t=0}^{T} \alpha_t^2}{\sum_{t=0}^{T} \alpha_t}.$$

In particular, if we take $\bar{\rho} = 2\rho$, $\alpha_t = \sqrt{\frac{\Delta}{\rho L^2 (T+1)}}$ for some $\Delta > \varphi_{1/2\rho}(\theta^0) - v^*$, then

$$\mathbb{E}\Big[\| \nabla \varphi_{1/2\rho}(\theta^{t^*}) \|_2^2 \Big] \le 4 \sqrt{\frac{\rho \Delta L^2}{T+1}}.$$

Proof: Let $\hat{\theta}^t = \mathrm{prox}_{\frac{1}{\bar{\rho}} \varphi}(\theta^t)$

$$\mathbb{E}_t\Big[\varphi_{1/\bar{\rho}}(\theta^{t+1}) \Big] \le \mathbb{E}_t\Big[\varphi(\hat{\theta}^t) + \frac{\bar{\rho}}{2} \| \hat{\theta}^t - \theta^{t+1} \|_2^2 \Big]$$

expectation with info up to time t

$$= \mathbb{E}_t\Big[F(\hat{\theta}^t) + \frac{\bar{\rho}}{2} \| \hat{\theta}^t - \theta^{t+1} \|_2^2 \Big]$$

(Recall: $\varphi(\theta) = F(\theta) + \mathbb{1}_C(\theta)$. Note that $\hat{\theta}^t \in C$, so $\mathbb{1}_C(\hat{\theta}^t) = 0$)

$$= F(\hat{\theta}^t) + \frac{\bar{\rho}}{2} \mathbb{E}_t\Big[\| \underbrace{\Pi_C(\theta^t - \alpha_t s(\theta^t, \xi_t))} - \underbrace{\Pi_C(\hat{\theta}^t)} \|_2^2 \Big]$$

$$= F(\hat{\theta}^t) + \frac{\bar{\rho}}{2}\mathbb{E}_t\Big[\| \underbrace{\Pi_C(\theta^t - \alpha_t s(\theta^t, \xi_t))}_{\theta^{t+1}} - \underbrace{\Pi_C(\hat{\theta}^t)}_{\hat{\theta}^t \ (\because\ \hat{\theta}^t \in C)} \|_2^2\Big]$$

$$\le F(\hat{\theta}^t) + \frac{\bar{\rho}}{2}\mathbb{E}_t\Big[\| (\theta^t - \hat{\theta}^t) - \alpha_t s(\theta^t, \xi_t)\|_2^2\Big]$$

(by the non-expansiveness of Π_C. for any $\theta, \gamma \in \mathbb{R}^d$,
$\|\Pi_C(\gamma) - \Pi_C(\theta)\|_2 \le \|\gamma - \theta\|_2$)

$$= \underline{F(\hat{\theta}^t) + \frac{\bar{\rho}}{2}\|\theta^t - \hat{\theta}^t\|_2^2} + \bar{\rho}\alpha_t \mathbb{E}_t\Big[(\hat{\theta}^t - \theta^t)^T s(\theta^t, \xi_t)\Big]$$
$$+ \frac{\bar{\rho}\alpha_t^2}{2}\mathbb{E}_t\Big[\|s(\theta^t, \xi_t)\|_2^2\Big]$$

$$= \underbrace{\varphi_{1/\bar{\rho}}(\theta^t)}_{\hookrightarrow\ \hat{\theta}^t = \mathrm{prox}_{1/\bar{\rho}\,\varphi}(\theta^t)} + \bar{\rho}\alpha_t(\hat{\theta}^t - \theta^t)^T \underbrace{\mathbb{E}_t\Big[s(\theta^t, \xi_t)\Big]}_{\triangleq s^t \in \partial F(\theta^t) \text{ by assumption (A)}} + \frac{\bar{\rho}\alpha_t^2}{2}\underbrace{\mathbb{E}_t\Big[\|s(\theta^t, \xi_t)\|_2^2\Big]}_{\le L^2 \text{ by assumption (B)}}$$

$$\le \varphi_{1/\bar{\rho}}(\theta^t) + \bar{\rho}\alpha_t\Big[F(\hat{\theta}^t) - F(\theta^t) + \frac{\rho}{2}\|\theta^t - \hat{\theta}^t\|_2^2\Big] + \frac{\bar{\rho}\alpha_t^2 L^2}{2}.$$

(by weak convexity of F)

Since $\theta \mapsto F(\theta) + \frac{\bar{\rho}}{2}\|\theta - \theta^t\|_2^2$ is strongly convex with parameter $\bar{\rho} - \rho$, we have

$$F(\theta^t) - F(\hat{\theta}^t) - \frac{\rho}{2}\|\theta^t - \hat{\theta}^t\|_2^2 = \Big(\underline{F(\theta^t) + \underbrace{\frac{\bar{\rho}}{2}\|\theta^t - \theta^t\|_2^2}_{=0}}\Big) - \Big(\underline{F(\hat{\theta}^t) + \frac{\bar{\rho}}{2}\|\theta^t - \hat{\theta}^t\|_2^2}\Big)$$
$$+ \frac{\bar{\rho} - \rho}{2}\|\theta^t - \hat{\theta}^t\|_2^2 - \underline{0^T(\theta^t - \hat{\theta}^t)}$$

note that $0 \in \partial\big(F(\gamma) + \frac{\bar{\rho}}{2}\|\gamma - \theta^t\|_2^2\big)\big|_{\gamma = \hat{\theta}^t}$ since by definition, $\hat{\theta}^t = \mathrm{prox}_{\frac{1}{\bar{\rho}}\varphi}(\theta^t)$.

$$\ge (\bar{\rho} - \rho)\|\theta^t - \hat{\theta}^t\|_2^2 = \frac{\bar{\rho} - \rho}{\bar{\rho}^2}\|\nabla\varphi_{1/\bar{\rho}}(\theta^t)\|_2^2.$$

use strong convexity of $\theta \mapsto F(\theta) + \frac{\bar{\rho}}{2}\|\theta - \theta^t\|_2^2$

Putting the pieces together yield the first inequality in the theorem