Consider the problem

$$V^* = \inf_{\theta \in \mathbb{R}^d} \left\{ \varphi(\theta) \triangleq F(\theta) + r(\theta) \right\} \qquad \text{— (P)}$$

where

$F: \mathbb{R}^d \to \mathbb{R}$ is ρ-weakly convex; i.e., for $\rho > 0$

$$F(\gamma) \geq F(\theta) + s^T(\gamma - \theta) - \frac{\rho}{2}\|\gamma - \theta\|_2^2, \quad \forall\, \theta, \gamma \in \mathbb{R}^d, \; s \in \partial F(\theta)$$

(e.g., $F(\theta) = \mathcal{L}(\theta) + R_\lambda(\theta)$; note that F need not be smooth)

and

$r: \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a closed convex function

(r being closed means $\text{epi}(r) = \{(\theta, t) \in \mathbb{R}^d \times \mathbb{R} : r(\theta) \leq t\}$ is a closed set)

(e.g., $r(\theta) = \mathbb{I}_{\{\|\theta\|_1 \leq R\}}(\theta) = \begin{cases} 0 & \text{if } \|\theta\|_1 \leq R, \\ +\infty & \text{otherwise.} \end{cases}$; note that r need not be smooth)

Note that (P) is a non-smooth, non-convex optimization problem with good structure.

Remark: For simplicity, we assume that $r(\theta) = \mathbb{I}_C(\theta)$, where $C \subseteq \mathbb{R}^d$ is a closed convex set.

Q: What kind of methods can be used to solve (P)?

Idea: Projected subgradient method

$$\theta^{t+1} \leftarrow \Pi_C\left(\theta^t - \alpha_t\, s(\theta^t)\right), \quad s(\theta^t) \in \partial F(\theta^t)$$

Taking this idea further, in many applications, F has a finite-sum structure; $F(\theta) = \sum_{i=1}^{n} F_i(\theta)$. This can be tackled by stochastic methods:

Projected stochastic subgradient method (PSSM)

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

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

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

e.g. Linear model with additively corrupted covariates

$$\mathcal{L}(\theta) = \frac{1}{2} \theta^T \hat{\Gamma} \theta - \hat{\gamma}^T \theta , \quad \hat{\Gamma} = \frac{1}{n} Z^T Z - \Sigma_w , \quad \hat{\gamma} = \frac{1}{n} Z^T y$$

where $Z \in \mathbb{R}^{n \times d}$, $\Sigma_w \in S^d_+$. Then,

$$\mathcal{L}(\theta) = \frac{1}{2}\left(\frac{1}{n} \|Z\theta\|_2^2 - \|\Sigma_w^{1/2}\theta\|_2^2 \right) - \frac{1}{n} y^T Z \theta$$

$$= \sum_{i=1}^{n} \left[\frac{1}{2n} (z_i^T \theta)^2 - (u_i^T \theta)^2 - \frac{1}{n} y_i (z_i^T \theta) \right],$$

where z_i = i^{th} row of Z , u_i = i^{th} row of $\Sigma_w^{1/2}$.

Q: How do we analyze PSSM?

First-order optimality condition of (P):

$$0 \in \partial \varphi(\theta)$$

A solution θ satisfying the above is called stationary point.

To measure the progress of an iterative method, some ideas include

(i) Function value gap: $\varphi(\theta^t) - v^*$

but (P) is non-convex, so this gap need not go to 0.

(ii) Stationarity measure: $\mathrm{dist}(0, \partial\varphi(\theta^t))$

(in the smooth case, $\mathrm{dist}(0, \underbrace{\partial\varphi(\theta^t)}_{\text{singleton: } \{\nabla\varphi(\theta^t)\}}) = \|\nabla\varphi(\theta^t)\|_2$ under

reasonable definitions of the subdifferential) but (P) is non-smooth, so $\mathrm{dist}(0, \partial\varphi(\theta^t))$ need not go to 0.

(e.g., $\varphi(\theta) = |\theta|$, $\theta^t = \frac{1}{t}$, $t \geq 1$. Then,

$$\mathrm{dist}(0, \underbrace{\partial\varphi(\theta^t)}_{=\{1\}}) = 1 \quad \forall t \geq 1 \,)$$

As it turns out, one can use the proximal map and the associated Moreau envelope to measure the progress.

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

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

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

<u>Fact:</u> If φ is ρ-weakly convex and $\lambda < 1/\rho$, then

φ_λ is smooth with $\nabla \varphi_\lambda(\theta) = \frac{1}{\lambda}(\theta - \mathrm{prox}_{\lambda\varphi}(\theta))$.

<u>Idea:</u> How about using $\|\nabla \varphi_\lambda(\theta)\|_2$ as a stationarity measure?

<u>Properties of the Proximal Map and Moreau Envelope</u>

<u>Claim:</u> Let $\hat{\theta} = \mathrm{prox}_{\lambda\varphi}(\theta)$. Then,

(i) $\|\hat{\theta} - \theta\|_2 = \lambda \|\nabla \varphi_\lambda(\theta)\|_2$ (by the fact)

(ii) $\varphi(\hat{\theta}) \leq \varphi(\theta)$

<u>Proof:</u>

$$\varphi(\hat{\theta}) \leq \varphi(\hat{\theta}) + \frac{1}{2\lambda}\|\hat{\theta} - \theta\|_2^2 \leq \varphi(\theta) + \cancelto{0}{\frac{1}{2\lambda}\|\theta - \theta\|_2^2}$$

(iii) $\mathrm{dist}(0, \partial\varphi(\hat{\theta})) \leq \|\nabla \varphi_\lambda(\theta)\|_2$

<u>Proof:</u> By the optimality condition,

$$0 \in \partial\varphi(\hat{\theta}) + \frac{1}{\lambda}(\hat{\theta} - \theta) = \partial\varphi(\hat{\theta}) - \nabla\varphi_\lambda(\theta)$$

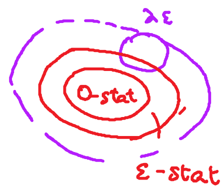
$$\Rightarrow \exists\, s \in \partial\varphi(\hat{\theta}) \text{ s.t. } s - \nabla\varphi_\lambda(\theta) = 0 \;\Rightarrow\; \mathrm{dist}(0, \partial\varphi(\hat{\theta})) \le \|s\|_2 = \|\nabla\varphi_\lambda(\theta)\|_2 .$$

Implication of Claim:

By (iii), if $\|\nabla\varphi_\lambda(\theta)\|_2 \le \varepsilon$, then $\hat{\theta}$ is called an ε-stationary point.

By (i), $\|\hat{\theta} - \theta\|_2 \le \lambda\varepsilon$. Hence,

$$\theta \in \underbrace{\{\, \gamma : \mathrm{dist}(0, \partial\varphi(\gamma)) \le \varepsilon \,\}}_{\hat{\theta} \text{ is in this set}} + \lambda\varepsilon \cdot B(0,1)$$



This motivates us to call θ an $(\varepsilon, \lambda\varepsilon)$-approximate stationary point.