- In the last lecture, we introduced the proximal gradient method for solving the regularized loss minimization problem

$$\hat{v} \triangleq \min_{\theta \in \mathbb{R}^d} \{ \mathcal{L}(\theta) + R(\theta) \}. \qquad \text{— } (*)$$

- We have seen that strong convergence rate results for the PGM can be derived when $(*)$ possesses the following error bound (EB) property:

(Local Error Bound) For any $v \geq \hat{v}$, there exist $\mu, \varepsilon > 0$ s.t. $\operatorname{dist}(\theta, \Theta) \leq \mu \cdot \|E(\theta)\|_2$ for any θ satisfying $F(\theta) \leq v$ and $\|E(\theta)\|_2 \leq \varepsilon$.

(Here, Θ is the set of optimal solutions to $(*)$, assumed to be non-empty, and

$$E(\theta) = \operatorname{prox}_R(\theta - \nabla \mathcal{L}(\theta)) - \theta$$

is the first-order residual error associated with θ.

- The local EB can be viewed as a regularity property of $(*)$. As such, it may not hold for an arbitrary instance of $(*)$. An important research direction in optimization is to identify instances of $(*)$ for which the local EB holds. In the sequel, we shall present several classes of instances of $(*)$ for which the local EB holds.

Scenario 1: $\mathcal{L}$ strongly convex and $\nabla\mathcal{L}$ Lipschitz continuous

By definition, there exist $\kappa > 0$ and $L > 0$ s.t.

(1) $$\mathcal{L}(\gamma) \geq \mathcal{L}(\theta) + \nabla\mathcal{L}(\theta)^T(\gamma - \theta) + \frac{\kappa}{2}\|\gamma - \theta\|_2^2 \quad \forall \theta, \gamma$$

and

(2) $$\|\nabla\mathcal{L}(\gamma) - \nabla\mathcal{L}(\theta)\|_2 \leq L \cdot \|\gamma - \theta\|_2 \quad \forall \theta, \gamma.$$

It is a routine exercise to show that (1) is equivalent to

$$(\nabla\mathcal{L}(\gamma) - \nabla\mathcal{L}(\theta))^T(\gamma - \theta) \geq \kappa\|\gamma - \theta\|_2^2 \quad \forall \theta, \gamma$$

Now, let $\hat{\theta}$ be the unique optimal solution to $(*)$. Then, for any $\theta \in \mathbb{R}^d$,

(Δ) $$\kappa \cdot \mathrm{dist}(\theta, \Theta)^2 = \kappa \cdot \|\theta - \hat{\theta}\|_2^2 \leq (\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\hat{\theta}))^T(\theta - \hat{\theta}).$$

Moreover, the first-order optimality conditions imply that

$$-\nabla\mathcal{L}(\hat{\theta}) \in \partial R(\hat{\theta}) \quad \left(\text{since } \hat{\theta} \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \{\mathcal{L}(\theta) + R(\theta)\}\right)$$

$$-[\nabla\mathcal{L}(\theta) + E(\theta)] \in \partial R(\theta + E(\theta))$$

$$\left(\text{since } \operatorname{prox}_R(\theta - \nabla\mathcal{L}(\theta)) = \underset{\gamma \in \mathbb{R}^d}{\operatorname{argmin}} \left\{\frac{1}{2}\|\theta - \nabla\mathcal{L}(\theta) - \gamma\|_2^2 + R(\gamma)\right\}\right)$$

By definition of the subdifferential, we have

$$R(\theta + E(\theta)) \geq R(\hat{\theta}) - \nabla\mathcal{L}(\hat{\theta})^T(\theta + E(\theta) - \hat{\theta})$$

$$R(\hat{\theta}) \geq R(\theta + E(\theta)) - [\nabla\mathcal{L}(\theta) + E(\theta)]^T(\hat{\theta} - \theta - E(\theta))$$

Adding the inequalities yield

$$0 \geq (\nabla\mathcal{L}(\theta) + E(\theta) - \nabla\mathcal{L}(\hat{\theta}))^T(\theta + E(\theta) - \hat{\theta})$$

Which implies that

$$
\begin{aligned}
&(\nabla \mathcal{L}(\theta) - \nabla \mathcal{L}(\hat{\theta}))^T (\theta - \hat{\theta}) + \|E(\theta)\|_2^2 \\
&\leq E(\theta)^T (\hat{\theta} - \theta + \nabla \mathcal{L}(\hat{\theta}) - \nabla \mathcal{L}(\theta)) \\
&\leq \left[\|\nabla \mathcal{L}(\hat{\theta}) - \nabla \mathcal{L}(\theta)\|_2 + \mathrm{dist}(\theta, \Theta) \right] \cdot \|E(\theta)\|_2 \\
&\leq (L+1) \cdot \mathrm{dist}(\theta, \Theta) \cdot \|E(\theta)\|_2 . \qquad \text{— } (**)
\end{aligned}
$$

Putting (Δ) and $(**)$ together yields the inequality

$$\mathrm{dist}(\theta, \Theta) \leq \frac{L+1}{\kappa} \|E(\theta)\|_2 .$$

Scenario 2: $\mathcal{L}$ takes the form $\mathcal{L}(\theta) = h(A\theta)$ for some linear operator $A \in \mathbb{R}^{n \times d}$; h is strongly convex on compact sets, continuously differentiable, and ∇h is Lipschitz continuous; R has polyhedral epigraph; Θ is compact.

- Recall that $h : \mathbb{R}^n \to \mathbb{R}$ is strictly convex if $\forall\, u \neq v \in \mathbb{R}^n$ and $\alpha \in (0,1)$,

$$h(\alpha u + (1-\alpha) v) < \alpha h(u) + (1-\alpha) h(v).$$

- Note that for a strongly convex function h, there exists $\kappa > 0$ s.t.

$$h(\alpha u + (1-\alpha) v) \leq \alpha h(u) + (1-\alpha) h(v) - \alpha(1-\alpha) \frac{\kappa}{2} \|u - v\|_2^2$$

for any $u, v \in \mathbb{R}^n$ and $\alpha \in [0,1]$, which implies that it is strictly convex. However, the converse is not true.

- Recall that $\mathrm{epi}(R) = \{ (\theta, t) \in \mathbb{R}^d \times \mathbb{R} : R(\theta) \leq t \}$. We say that R has polyhedral epigraph if $\mathrm{epi}(R)$ is polyhedral. Examples of norms R having polyhedral epigraphs are $R(\theta) = \|\theta\|_1$ and $R(\theta) = \|\theta\|_\infty$

- This scenario covers many practical applications:
e.g.: $R(\theta) = \|\theta\|_1$ and $\mathcal{L}$ is one of the following:

(a) Least squares regression

$$\mathcal{L}(\theta) = \frac{1}{2n}\|y - A\theta\|_2^2 \Rightarrow h(u) = \frac{1}{2n}\|y-u\|_2^2$$

(b) Logistic regression

$$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\log(1+\exp(-b_i a_i^T\theta))$$

$$\Rightarrow h(u) = \frac{1}{n}\sum_{i=1}^{n}\log(1+\exp(-b_i u_i))$$

It can be easily verified that the function h in (a) is strongly convex but that in (b) is only strongly convex on compact sets.

- Let us characterize the optimal solution set of $(*)$ in this scenario.

<u>Proposition 1</u>: There exist a $\bar{y} \in \mathbb{R}^n$ s.t. for all $\hat{\theta} \in \Theta$,

$$A\hat{\theta} = \bar{y} \quad \text{and} \quad \nabla\mathcal{L}(\hat{\theta}) = A^T\nabla h(\bar{y}) \triangleq \bar{g}.$$

In particular,

$$\Theta = \{\theta \in \mathbb{R}^d : A\theta = \bar{y},\ -\bar{g} \in \partial R(\theta)\}.$$

<u>Proof</u>: Let $\theta_1, \theta_2 \in \Theta$. Set $\bar{y}_1 = A\theta_1$, $\bar{y}_2 = A\theta_2$. By the strict convexity of h, if $\bar{y}_1 \neq \bar{y}_2$, then

$$h\left(\frac{\bar{y}_1+\bar{y}_2}{2}\right) < \frac{1}{2}h(\bar{y}_1) + \frac{1}{2}h(\bar{y}_2),$$

or equivalently,

$$\mathcal{L}\left(\frac{\theta_1+\theta_2}{2}\right) < \frac{1}{2}\mathcal{L}(\theta_1) + \frac{1}{2}\mathcal{L}(\theta_2).$$

Moreover, by the convexity of R,

$$R\left(\frac{\theta_1+\theta_2}{2}\right) \leq \frac{1}{2}R(\theta_1)+\frac{1}{2}R(\theta_2).$$

Adding the above inequalities gives

$$F\left(\frac{\theta_1+\theta_2}{2}\right) < \frac{1}{2}\hat{v}+\frac{1}{2}\hat{v} = \hat{v},$$

which is impossible. Thus, $\bar{y}_1 = \bar{y}_2$.

Now, it remains to compute

$$\nabla \mathcal{L}(\theta) = \nabla h(A\theta) = A^T \nabla h(A\theta).$$

This completes the proof.

(Remark: The proof above does not require ∇h to be Lipschitz continuous and R has polyhedral epigraph.)

- Proposition 1 allows us to write

$$\Theta = \Theta_{\mathcal{L}} \cap \Theta_R, \quad \text{where}$$

$$\Theta_{\mathcal{L}} = \{\theta \in \mathbb{R}^d : A\theta = \bar{y}\} \text{ and } \Theta_R = \{\theta \in \mathbb{R}^d : -\bar{g} \in \partial R(\theta)\}.$$

Clearly, $\Theta_{\mathcal{L}}$ is polyhedral. It turns out that Θ_R is also polyhedral. This relies on the following classic convex analysis facts (see the corresponding results in Rockafellar: Convex Analysis, Princeton University Press, 1970):

(I) (Theorem 19.2) If R has polyhedral epigraph, so does its conjugate $\tilde{R}$ (recall that $\tilde{R}(\gamma) = \sup_\theta \{\theta^T\gamma - R(\theta)\}$).

(II) (Corollary 23.5.1) $\partial\tilde{R} = (\partial R)^{-1}$; i.e.,

$$\partial\tilde{R}(\gamma) = (\partial R)^{-1}(\gamma) = \{\theta : \gamma \in \partial R(\theta)\}.$$

(III) (Theorem 23.10) If R has polyhedral epigraph and $R(\theta)$ is finite, then $\partial R(\theta)$ is a polyhedron.

Now, note that $\Theta_R = (\partial R)^{-1}(-\bar{g})$. By (II), we have $\Theta_R = \partial \hat{R}(-\bar{g})$. Hence, by (I) and (III), Since R has polyhedral epigraph, the same holds for $\hat{R}$, and hence $\partial \tilde{R}(-\bar{g})$ is a polyhedron as long as $\hat{R}(-g)$ is finite. To prove the latter, we first recall that R is a norm and note the following fact:

<u>Fact</u>: Given a norm R, its conjugate $\tilde{R}$ is given by

$$\tilde{R}(\gamma) = \begin{cases} 0 & \text{if } R^*(\gamma) \leq 1, \\ +\infty & \text{otherwise;} \end{cases}$$

i.e., $\hat{R}$ is the indicator of the Unit ball of the dual norm of R.

Now, since $\Theta \neq \phi$, there exists an $\hat{\theta} \in \Theta_R = \partial \tilde{R}(-\bar{g})$. By definition of the subdifferential, we have

$$\tilde{R}(g) \geq \hat{R}(-\bar{g}) + \hat{\theta}^T(g+\bar{g}) \quad \forall g \in \mathbb{R}^d.$$

Taking $g=0$, we have $\quad 0 \geq \tilde{R}(-\bar{g}) + \hat{\theta}^T \bar{g}$, which implies that $\tilde{R}(-\bar{g})$ is finite.

- To summarize, we now know that both

$$\Theta_{\mathcal{L}} = \{\theta \in \mathbb{R}^d : A\theta = \bar{y}\}$$

$$\text{and} \quad \Theta_R = \{\theta \in \mathbb{R}^d : -\bar{g} \in \partial R(\theta)\}$$

are polyhedral.

- The above observations motivate us to develop estimates of point-polyhedron distances. The following result, which is known as the Hoffman error bound, is fundamental

Theorem 1: Let $P = \{ z \in \mathbb{R}^n : Az \leq b \}$ be a non-empty polyhedron. Then, there exists a constant $c > 0$, which depends only on A, s.t.

$$\text{dist}(x, P) \leq c \cdot \| (Ax - b)^+ \|_2 \quad \text{for all } x \in \mathbb{R}^n.$$

Corollary 1: Let $\{P_1, \dots, P_M\}$ be a finite collection of polyhedra. Suppose that $P = \bigcap_{i=1}^{M} P_i \neq \emptyset$. Then, there exists a constant $\alpha > 0$ s.t.

$$\text{dist}(x, P)^2 \leq \alpha \cdot \sum_{i=1}^{M} \text{dist}(x, P_i)^2 \quad \text{for all } x \in \mathbb{R}^n;$$

i.e., $\{P_1, \dots, P_M\}$ is linearly regular.

Proof of Corollary 1: Let $H_j = \{ z \in \mathbb{R}^n : a_j^T z \leq b_j \}$ for $j = 1, \dots, L$ and $\{K_1, \dots, K_M\}$ be a partition of $\{1, \dots, L\}$ s.t. $P_i = \bigcap_{j \in K_i} H_j$. Consider

$$\text{dist}(x, H_j)^2 = \min \{ \| x - z \|_2^2 : a_j^T z \leq b_j \}.$$

The optimal solution z^* satisfies the KKT conditions

$$z - x + \mu a_j = 0$$
$$\mu \geq 0$$
$$a_j^T z \leq b_j$$
$$\mu (a_j^T z - b_j) = 0$$

A routine calculation shows that $\mu^* = \frac{(a_j^T x - b_j)^+}{\|a_j\|_2^2}$

and hence $\text{dist}(x, H_j) = \frac{(a_j^T x - b_j)^+}{\|a_j\|_2}$. It follows from

Theorem 1 that

$$\begin{aligned} \text{dist}(x, P)^2 &\leq c' \sum_{j=1}^{L} \text{dist}(x, H_j)^2 \\ &= c' \sum_{i=1}^{M} \sum_{j \in K_i} \text{dist}(x, H_j)^2 \\ &\leq c'' \sum_{i=1}^{M} \text{dist}(x, P_i)^2 \end{aligned}$$

Since $\text{dist}(x, H_j) \leq \text{dist}(x, P_i)$ for all $j \in K_i$.

- Armed with the above results, we immediately have

$$\text{dist}(\theta, \bar{\Theta}) \leq c \cdot \left[\text{dist}(\theta, \Theta_{\mathcal{L}}) + \text{dist}(\theta, \Theta_R) \right] \qquad (+)$$

for some constant $c > 0$. By Theorem 1, we have

$$\text{dist}(\theta, \Theta_{\mathcal{L}}) \leq c' \cdot \|A\theta - \bar{y}\|_2 . \qquad (++)$$

It is natural to apply a similar argument to $\text{dist}(\theta, \Theta_R)$. However, we do not have an explicit description of the polyhedral set Θ_R. Fortunately, we have the following result, which, roughly speaking, states that $(\partial R)^{-1}$ is Lipschitz continuous:

Proposition 2: (Outer Lipschitz Continuity of $(\partial R)^{-1}$).

There exists $\beta > 0$ s.t. for any $g' \in \mathbb{R}^d$, there is a neighborhood $V_{g'}$ of g' s.t. $(\partial R)^{-1}(g'') \subseteq (\partial R)^{-1}(g') + \beta \cdot \|g' - g''\|_2 \cdot B(0,1) \quad \forall g'' \in V_{g'}$

To understand the notion of outer Lipschitz continuity, let us consider the following example:

Example: Consider $R(\theta) = |\theta|$. Then, we have

$$\partial R(\theta) = \begin{cases} 1 & \text{if } \theta > 0, \\ [-1, 1] & \text{if } \theta = 0, \\ -1 & \text{if } \theta < 0. \end{cases}$$

It follows that

$$(\partial R)^{-1}(g) = \begin{cases} \phi & \text{if } |g| > 1, \\ \mathbb{R}_+ & \text{if } g = 1, \\ \mathbb{R}_- & \text{if } g = 1, \\ \{0\} & \text{if } |g| < 1. \end{cases}$$

In particular, it is easy to verify that $(\partial R)^{-1}$ is outer Lipschitz continuous with $\beta = 1$.

- Noting that $\Theta_R = (\partial R)^{-1}(-\bar{g})$, Proposition 2 implies that

$$\text{dist}(\theta, \Theta_R) \leq \beta \cdot \|g - \bar{g}\|_2 \quad \text{for } \theta \in (\partial R)^{-1}(-g), -g \in V_{(-\bar{g})}.$$

Combining this with (+) and (++), we get

$$\text{dist}(\theta, \Theta) \leq \beta' \left(\|A\theta - \bar{y}\|_2 + \|g - \bar{g}\|_2 \right) \text{ for } \theta \in (\partial R)^{-1}(-g), -g \in V_{(-\bar{g})}.$$

- To proceed, recall that $-(\nabla \mathcal{L}(\theta)+E(\theta)) \in \partial R(\theta+E(\theta))$.

Hence, provided $-(\nabla\mathcal{L}(\theta)+E(\theta)) \in V_{(-\bar{g})}$ (which holds if $\mathrm{dist}(\theta,\Theta)$ is small),

$$\mathrm{dist}(\theta+E(\theta),\Theta) \le \beta'\left[\|A(\theta+E(\theta))-\bar{y}\|_2+\|\nabla\mathcal{L}(\theta)+E(\theta)-\bar{g}\|_2\right]$$

$$\le \beta'\cdot\left[\|A\theta-\bar{y}\|_2+\|\nabla\mathcal{L}(\theta)-\bar{g}\|_2+(\|A\|+1)\|E(\theta)\|_2\right]$$

$$\le \beta'\cdot\left[(L\cdot\|A\|+1)\|A\theta-\bar{y}\|_2+(\|A\|+1)\|E(\theta)\|_2\right],$$

where the last inequality follows from

$$\|\nabla\mathcal{L}(\theta)-\bar{g}\|_2=\|A^T\nabla h(A\theta)-A^T\nabla h(\bar{y})\|_2$$

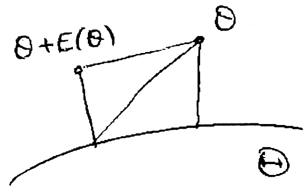
$$\le\|A\|\cdot\|\nabla h(A\theta)-\nabla h(\bar{y})\|_2 \le L\cdot\|A\|\cdot\|A\theta-\bar{y}\|_2$$

with L being the Lipschitz parameter of ∇h.

It follows that

$$\mathrm{dist}(\theta,\Theta) \le \mathrm{dist}(\theta+E(\theta),\Theta)+\|E(\theta)\|_2$$

$$\le \beta''(\|A\theta-\bar{y}\|_2+\|E(\theta)\|_2).$$



which implies

$$\mathrm{dist}(\theta,\Theta)^2 \le 2(\beta'')^2(\|A\theta-\bar{y}\|_2^2+\|E(\theta)\|_2^2). \qquad -(\diamond)$$

On the other hand, for any θ with $\mathrm{dist}(\theta, \Theta)$ sufficiently small so that $-(\nabla\mathcal{L}(\theta) + E(\theta)) \in V_{(-\bar{g})}$ holds, we have

$$\kappa \| A\theta - \bar{y} \|_2^2 \leq (\nabla h(A\theta) - \nabla h(\bar{y}))^T (A\theta - \bar{y})$$

$$= (\nabla\mathcal{L}(\theta) - \bar{g})^T (\theta - \hat{\theta}), \qquad -(\diamond\diamond)$$

where $\hat{\theta}$ is the projection of θ onto Θ.

Moreover, similar to the derivation of $(**)$, we have

$$(\nabla\mathcal{L}(\theta) - \bar{g})^T (\theta - \hat{\theta}) + \| E(\theta) \|_2^2 \leq \beta''' \cdot \mathrm{dist}(\theta, \Theta) \cdot \| E(\theta) \|_2 \qquad -(\diamond\diamond\diamond)$$

Putting $(\diamond)$, $(\diamond\diamond)$, $(\diamond\diamond\diamond)$ together, we get

$$\mathrm{dist}(\theta, \Theta)^2 \leq \tilde{\beta} \left[\mathrm{dist}(\theta, \Theta) \cdot \| E(\theta) \|_2 + \| E(\theta) \|_2^2 \right]$$

Solving the quadratic inequality yields

$$\mathrm{dist}(\theta, \Theta) \leq \mu \cdot \| E(\theta) \|_2 ,$$

as desired.

- Now, let us show that $g = \nabla\mathcal{L}(\theta) + E(\theta)$ and $\bar{g}$ is close when $\mathrm{dist}(\theta, \Theta)$ is small; so that we can guarantee $-(\nabla\mathcal{L}(\theta) + E(\theta)) \in V_{(-\bar{g})}$. First, we compute

$$\| \nabla\mathcal{L}(\theta) + E(\theta) - \bar{g} \|_2 = \| \nabla\mathcal{L}(\theta) + E(\theta) - \nabla\mathcal{L}(\hat{\theta}) \|_2$$

$\leq L' \cdot \mathrm{dist}(\theta, \Theta) + \| E(\theta) \|_2$, where $\hat{\theta}$ is the projection of θ onto Θ, and L' is the Lipschitz parameter of $\nabla\mathcal{L}$. (note that the Lipschitz continuity of ∇h implies that of $\nabla\mathcal{L}$).

Next, we compute

$$\|E(\theta)\|_2 = \|\operatorname{prox}_R(\theta - \nabla\mathcal{L}(\theta)) - \theta\|_2$$

$$= \|\operatorname{prox}_R(\theta - \nabla\mathcal{L}(\theta)) - \operatorname{prox}_R(\hat{\theta} - \nabla\mathcal{L}(\hat{\theta})) + \hat{\theta} - \theta\|_2$$

(since by optimality of $\hat{\theta}$, we have $\hat{\theta} = \operatorname{prox}_R(\hat{\theta} - \nabla\mathcal{L}(\hat{\theta}))$)

$$\leq \operatorname{dist}(\theta, \Theta) + \|\theta - \hat{\theta} + \nabla\mathcal{L}(\hat{\theta}) - \nabla\mathcal{L}(\theta)\|_2$$

(since $\|\operatorname{prox}_R(\theta) - \operatorname{prox}_R(\gamma)\|_2 \leq \|\theta - \gamma\|_2$ for any $\theta, \gamma \in \mathbb{R}^d$;

See Combettes and Wajs 2005)

$$\leq (L'+2)\operatorname{dist}(\theta, \Theta).$$

It follows that $\|\nabla\mathcal{L}(\theta) + E(\theta) - \bar{g}\|_2 \leq 2(L'+1)\cdot\operatorname{dist}(\theta, \Theta)$.