- Recall the regularized loss minimization problem:

$$\hat{\theta} \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \Big\{ \underbrace{\mathcal{L}(\theta; \{z_i\}_{i=1}^n) + \lambda R(\theta)}_{F} \Big\} \qquad -(*)$$

- So far we have discussed how to bound the statistical error $\hat{\Delta} = \hat{\theta} - \theta^*$, where $\theta^* \in \mathbb{R}^d$ is the true parameter to be estimated. Two key concepts are
  (i) decomposable norm,
  (ii) restricted strong convexity

- Our next task is to discuss the algorithmic aspects of (*). We shall consider the setting where
  (A1) $\mathcal{L}$ is convex and continuously differentiable on $\mathbb{R}^d$
  (A2) R is a norm on $\mathbb{R}^d$ and $\lambda = 1$.

- Under the above setting, (*) is a convex optimization problem. Hence, the first-order condition

$$0 \in \nabla \mathcal{L}(\theta) + \partial R(\theta) \qquad -(**)$$

is necessary and sufficient for optimality. Here, recall that

$$\partial R(\theta) = \{ s \in \mathbb{R}^d : R(\gamma) \geq R(\theta) + s^T(\gamma - \theta) \ \forall \gamma \}$$

is the subdifferential of R at θ, which is guaranteed to be non-empty, convex, and compact because R is convex on $\mathbb{R}^d$. In fact, one can show the following:

Claim (Exercise): $\partial R(\theta) = \{ s \in \mathbb{R}^d : R^*(s) \leq 1, \ s^T\theta = R(\theta) \}$,

where $R^*(s) = \sup_{\theta : R(\theta) = 1} \theta^T s$ is the dual norm of R.

- Various methods for solving (*) are motivated by different representations of the generalized equation (**). One such representation is a fixed point equation. Specifically, define the proximal map associated with R, denoted $\text{prox}_R : \mathbb{R}^d \to \mathbb{R}^d$, by

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

Note that the optimization problem on the RHS of (†) has a unique minimizer (why?). Hence, $\text{prox}_R(\theta)$ is well defined.

Now, observe that $\hat{\theta}$ is optimal for (*) if and only if

$$\hat{\theta} = \text{prox}_R(\hat{\theta} - \nabla \mathcal{L}(\hat{\theta})). \quad -(\Delta)$$

Indeed, by considering the first-order optimality condition of (†), we see that for any $\theta \in \mathbb{R}^d$,

$$0 \in \text{prox}_R(\theta) - \theta + \partial R(\text{prox}_R(\theta))$$

Hence,

$\hat{\theta}$ is optimal for (*)

$\Leftrightarrow$ $0 \in -(\hat{\theta} - \nabla \mathcal{L}(\hat{\theta})) + \hat{\theta} + \partial R(\hat{\theta})$

$\Leftrightarrow$ $\hat{\theta} = \text{prox}_R(\hat{\theta} - \nabla \mathcal{L}(\hat{\theta}))$.

- The fixed point equation (Δ) suggests the following natural iterative procedure for solving $(\star)$:

$$\theta^{k+1} \leftarrow \text{prox}_{\alpha_k R}(\theta^k - \alpha_k \nabla \mathcal{L}(\theta^k)) \qquad \text{— (PGM)}$$

where $\alpha_k > 0$ is the step size. This is known as the proximal gradient method.

- Here is another way of understanding (PGM). By definition, we have

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

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

We see that

$$\gamma \mapsto \mathcal{L}(\theta^k) + \nabla \mathcal{L}(\theta^k)^T (\gamma - \theta^k) + \frac{1}{2\alpha_k} \| \gamma - \theta^k \|_2^2$$

is a quadratic approximation of $\mathcal{L}$ at θ^k, using $\frac{1}{\alpha_k} I$ as the approximate Hessian.

- The PGM is a first-order method, as it uses only the first-order optimality condition of $(\star)$. Now, it is natural to ask whether (PGM) converges to a solution to the fixed point equation (Δ), and if so, whether we can determine the rate of convergence.

Convergence Analysis of the PGM

- Our first goal is to show that the PGM converges to a solution to (*) under the following additional assumption:

(A3) $\nabla \mathcal{L}$ is Lipschitz continuous with parameter $L > 0$.

We shall discuss the generality of (A3) later.

Here is a fundamental result:

Proposition 1: Suppose that $\{\alpha_k\}$ satisfy $0 < \underline{\alpha} \le \alpha_k \le \bar{\alpha} < \frac{1}{L}$ for some $\underline{\alpha}, \bar{\alpha}$.

(a) (Sufficient Decrease) There exists $\kappa_1 > 0$ s.t.

$$F(\theta^k) - F(\theta^{k+1}) \ge \kappa_1 \| \theta^k - \theta^{k+1} \|_2^2$$

(b) (Safeguard) There exists $\kappa_2 > 0$ s.t.

$$\| E(\theta^k) \|_2 \le \kappa_2 \| \theta^k - \theta^{k+1} \|_2 ,$$

where

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

can be viewed as the first-order residual error associated with θ. (Clearly, $E(\bar{\theta}) = 0$ if and only if $\bar{\theta}$ is optimal for (*))

- Armed with Proposition 1, we can establish a first convergence result concerning (PGM). Indeed, Proposition 1(a) implies that $\{F(\theta^k)\}_{k \ge 0}$ is monotonically decreasing. Since $F(\theta^k) \ge \hat{v} \triangleq F(\hat{\theta})$ for all k, this implies that $\{F(\theta^k)\}_{k \ge 0}$ converges. Hence, we have $\theta^k - \theta^{k+1} \to 0$.

Now, by Proposition 1(b), we have $\|E(\theta^k)\|_2 \to 0$.

It follows that $F(\theta^k) \to v^*$. Moreover, every accumulation point of $\{\theta^k\}_{k\geq 0}$ is optimal for (*).

Proof of Proposition 1(a)

Since

$$\theta^{k+1} = \text{prox}_{\alpha_k R}(\theta^k - \alpha_k \nabla \mathcal{L}(\theta^k))$$

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

It follows that

$$\frac{1}{2}\| \theta^{k+1} - \theta^k + \alpha_k \nabla \mathcal{L}(\theta^k)\|_2^2 + \alpha_k R(\theta^{k+1})$$

$$\leq \frac{1}{2}\| \alpha_k \nabla \mathcal{L}(\theta^k)\|_2^2 + \alpha_k R(\theta^k).$$

This gives

$$R(\theta^{k+1}) + \nabla \mathcal{L}(\theta^k)^T(\theta^{k+1} - \theta^k) + \frac{1}{2\alpha_k}\|\theta^{k+1} - \theta^k\|_2^2 \leq R(\theta^k).$$

On the other hand, since $\nabla \mathcal{L}$ is Lipschitz continuous, by defining $g(t) = \mathcal{L}(\theta^k + t(\theta^{k+1} - \theta^k))$, we have

$$\mathcal{L}(\theta^{k+1}) - \mathcal{L}(\theta^k) = g(1) - g(0) = \int_0^1 g'(t)\, dt$$

$$= \int_0^1 \nabla \mathcal{L}(\theta^k + t(\theta^{k+1} - \theta^k))^T(\theta^{k+1} - \theta^k)\, dt$$

$$= \nabla \mathcal{L}(\theta^k)^T(\theta^{k+1} - \theta^k) + \int_0^1 \left[\nabla \mathcal{L}(\theta^k + t(\theta^{k+1} - \theta^k)) - \nabla \mathcal{L}(\theta^k) \right]^T (\theta^{k+1} - \theta^k)\, dt$$

$$\leq \nabla \mathcal{L}(\theta^k)^T(\theta^{k+1} - \theta^k) + L\|\theta^{k+1} - \theta^k\|_2^2 \int_0^1 t\, dt$$

$$= \nabla \mathcal{L}(\theta^k)^T(\theta^{k+1} - \theta^k) + \frac{L}{2}\|\theta^{k+1} - \theta^k\|_2^2.$$

Thus,

$$F(\theta^{k+1}) - F(\theta^k) = \mathcal{L}(\theta^{k+1}) - \mathcal{L}(\theta^k) + R(\theta^{k+1}) - R(\theta^k)$$

$$\leq \nabla\mathcal{L}(\theta^k)^T(\theta^{k+1} - \theta^k) + \frac{L}{2}\|\theta^{k+1} - \theta^k\|_2^2 + R(\theta^{k+1}) - R(\theta^k)$$

$$\leq -\frac{1}{2}\left(\frac{1}{\alpha_k} - L\right)\|\theta^{k+1} - \theta^k\|_2^2.$$

It remains to note that the assumption on α_k implies $\frac{1}{\alpha_k} - L > 0$.

Proof of Proposition 1(b)

- The proof uses the following result:

Lemma: Let θ, γ be arbitrary. Then, the function $g_1 : \mathbb{R}_{++} \to \mathbb{R}_+$ given by

$$g_1(\alpha) = \frac{1}{\alpha}\|\text{prox}_{\alpha R}(\theta - \alpha\gamma) - \theta\|_2$$

is decreasing, and the function $g_2 : \mathbb{R}_{++} \to \mathbb{R}_+$ given by

$$g_2(\alpha) = \|\text{prox}_{\alpha R}(\theta - \alpha\gamma) - \theta\|_2$$

is increasing.

- Armed with the lemma, it suffices to observe

$$\|\theta^{k+1} - \theta^k\|_2 = \|\text{prox}_{\alpha_k R}(\theta^k - \alpha_k\nabla\mathcal{L}(\theta^k)) - \theta^k\|_2$$

$$\geq \|\text{prox}_{\underline{\alpha} R}(\theta^k - \underline{\alpha}\nabla\mathcal{L}(\theta^k)) - \theta^k\|_2$$

$$\geq \min\{1, \underline{\alpha}\} \cdot \|\text{prox}_R(\theta^k - \nabla\mathcal{L}(\theta^k)) - \theta^k\|_2$$

$$= \min\{1, \underline{\alpha}\} \cdot \|E(\theta^k)\|_2.$$

- Now, let us sketch the proof of the lemma.

• Consider the function $h : \mathbb{R}_{++} \times \mathbb{R}^d \to \mathbb{R}$ given by

$$h(\alpha, \omega) = \gamma^T(\omega - \theta) + \frac{1}{2\alpha}\|\omega - \theta\|_2^2 + R(\omega).$$

Note that $h(\alpha, \cdot)$ is convex for any given $\alpha > 0$. Now, consider the so-called Moreau envelope

$$H(\alpha) = \inf_{\omega \in \mathbb{R}^d} h(\alpha, \omega). \qquad \text{— } (\Delta\Delta)$$

• We claim that the optimal value of $(\Delta\Delta)$ is attained by

$$\omega^* = \text{prox}_{\alpha R}(\theta - \alpha\gamma). \qquad \text{— } (\Delta\Delta\Delta)$$

Indeed, the first-order optimality condition associated with $(\Delta\Delta)$ is

$$0 \in \gamma + \frac{1}{\alpha}(\omega - \theta) + \partial R(\omega), \qquad (\diamond)$$

while by definition of prox, ω^* in $(\Delta\Delta\Delta)$ satisfies the first-order optimality condition

$$0 \in \omega^* - (\theta - \alpha\gamma) + \alpha \partial R(\omega^*), \qquad (\diamond\diamond)$$

It can be readily seen that $(\diamond)$ and $(\diamond\diamond)$ are identical, which establishes the claim. In fact, by the strict convexity of $h(\alpha, \cdot)$, ω^* is the unique optimal solution to $(\Delta\Delta)$.

• Intuitively, the continuous differentiability of $h(\cdot, \omega)$ for each ω and the uniqueness of ω^* should imply the differentiability of H and

$$H'(\alpha) = \frac{\partial h(\alpha, \omega^*)}{\partial \alpha} = -\frac{1}{2\alpha^2}\|\omega^* - \theta\|_2^2. \qquad (\Omega)$$

$$= -\frac{1}{2} g_1(\alpha)^2$$

The rigorous proof of this relies on the following result, which is known as an envelope theorem in the literature.

Fact: Let X be a metric space and P an open subset of $\mathbb{R}$. Let $h: P \times X \to \mathbb{R}$ be s.t. $\frac{\partial h}{\partial p}$ exists and is continuous on $P \times X$. For each $p \in P$, let $x^* = x^*(p)$ be the unique optimal solution to $\inf_{x \in X} h(p,x)$. Suppose that x^* is continuous. Then, the value function $p \mapsto h(p, x^*(p))$ is continuously differentiable and the gradient is given by $\frac{\partial h(p, x^*(p))}{\partial p} = \frac{\partial h(p,x)}{\partial p}\Big|_{x = x^*(p)}$.

- To apply the fact, we need to check $\frac{\partial h}{\partial \alpha}$ exists and is continuous, which is obvious, and $w^* = w^*(\alpha)$ is continuous in α, which follows from Theorem 2.26 of Rockafellar and Wets, Variational Analysis, Springer, 2004. Hence, $(\Diamond)$ is established.

- In fact, the same theorem in Rockafellar and Wets shows that H is convex, which implies H' is increasing. This shows that g_1 is decreasing, as required.

- To prove that g_2 is increasing, define $\tilde{H}(\alpha) = H(1/\alpha)$. Observe that $\tilde{H}$ is differentiable with

$$\tilde{H}'(\alpha) = -\frac{1}{\alpha^2} H'(1/\alpha) = \frac{1}{2} \| \operatorname{prox}_{\frac{1}{\alpha} R}(\theta - \tfrac{1}{\alpha}\gamma) - \theta \|_2^2 .$$

Moreover, $\tilde{H}$, being the pointwise infimum of linear functions in α, is concave. Hence, $\tilde{H}'$ is decreasing. By renaming $1/\alpha$ as α, we see that g_2 is increasing.

- To estimate the convergence rate of (PGM), we need to measure its progress towards optimality. The following provides a way of achieving that:

Proposition 2: Let Θ be the optimal solution set of (*), assumed to be non-empty (in particular, Θ is convex and closed). Under the setting of Proposition 1, the following holds:

(c) (Cost-to-Go Estimate) There exists $K_3 > 0$ s.t.

$$F(\theta^{k+1}) - \hat{v} \le K_3 \left[\operatorname{dist}(\theta^k, \Theta)^2 + \| \theta^k - \theta^{k+1} \|_2^2 \right]$$

where $\operatorname{dist}(\theta^k, \Theta)^2 = \inf_{\theta \in \Theta} \| \theta^k - \theta \|_2$.

Proof: Let $\bar{\theta}^k$ be the projection of θ^k onto Θ.

By definition of θ^{k+1}, we have

$$\frac{1}{2} \| \theta^{k+1} - \theta^k + \alpha_k \nabla \mathcal{L}(\theta^k) \|_2 + \alpha_k R(\theta^{k+1})$$

$$\le \frac{1}{2} \| \bar{\theta}^k - \theta^k + \alpha_k \nabla \mathcal{L}(\theta^k) \|_2 + \alpha_k R(\bar{\theta}^k).$$

Upon expanding both sides, we obtain.

$$R(\theta^{k+1}) - R(\bar{\theta}^k) + \nabla \mathcal{L}(\theta^k)^T (\theta^{k+1} - \bar{\theta}^k)$$

$$\le \frac{1}{2\alpha_k} \| \bar{\theta}^k - \theta^k \|_2^2 \le \frac{1}{2\alpha} \operatorname{dist}(\theta^k, \Theta)^2.$$

On the other hand, by the mean-value theorem, there exists $\hat{\theta}^k \in [\bar{\theta}^k, \theta^{k+1}]$ s.t.

$$\mathcal{L}(\theta^{k+1}) - \mathcal{L}(\bar{\theta}^k) = \nabla\mathcal{L}(\hat{\theta}^k)^T(\theta^{k+1} - \bar{\theta}^k).$$

It follows that

$$F(\theta^{k+1}) - F(\bar{\theta}^k) = F(\theta^{k+1}) - \hat{v}$$

$$= \mathcal{L}(\theta^{k+1}) - \mathcal{L}(\bar{\theta}^k) + R(\theta^{k+1}) - R(\bar{\theta}^k)$$

$$= \nabla\mathcal{L}(\hat{\theta}^k)^T(\theta^{k+1} - \bar{\theta}^k) + R(\theta^{k+1}) - R(\bar{\theta}^k)$$

$$= \nabla\mathcal{L}(\theta^k)^T(\theta^{k+1} - \bar{\theta}^k) + R(\theta^{k+1}) - R(\bar{\theta}^k)$$

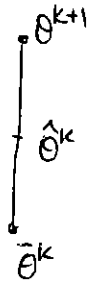
$$+ \left(\nabla\mathcal{L}(\hat{\theta}^k)^T - \nabla\mathcal{L}(\theta^k)\right)^T(\theta^{k+1} - \bar{\theta}^k)$$

$$\leq \frac{1}{2\underline{\alpha}}\,\mathrm{dist}(\theta^k, \Theta)^2 + L \cdot \|\hat{\theta}^k - \theta^k\|_2 \cdot \|\theta^{k+1} - \bar{\theta}^k\|_2.$$

Since $\|\hat{\theta}^k - \theta^k\|_2 \leq \|\theta^{k+1} - \theta^k\|_2 + \|\theta^{k+1} - \hat{\theta}^k\|_2$

$$\leq \|\theta^{k+1} - \theta^k\|_2 + \|\theta^{k+1} - \bar{\theta}^k\|_2$$

and $\|\theta^{k+1} - \bar{\theta}^k\|_2 \leq \|\theta^{k+1} - \theta^k\|_2 + \mathrm{dist}(\theta^k, \Theta)$.



we have

$$\|\hat{\theta}^k - \theta^k\|_2 \cdot \|\theta^{k+1} - \bar{\theta}^k\|_2$$

$$\leq 2\left(\|\theta^{k+1} - \theta^k\|_2 + \mathrm{dist}(\theta^k, \Theta)\right)^2$$

$$\leq 4\left(\|\theta^{k+1} - \theta^k\|_2^2 + \mathrm{dist}(\theta^k, \Theta)^2\right) \quad \left[\because (a+b)^2 \leq 2(a^2+b^2)\ \forall a, b \in \mathbb{R}\right]$$

Putting the pieces together completes the proof.

- Note that the cost-to-go estimate, $\Lambda^{(c)}$, involves the quantity $\text{dist}(\theta^k, \Theta)$, which is generally difficult to compute. In view of the sufficient decrease (a) and safeguard (b) conditions, it is natural to connect $\text{dist}(\theta^k, \Theta)$ to $\|\theta^k - \theta^{k+1}\|_2$. However, it is not clear how $\|\theta^k - \theta^{k+1}\|_2$ is related to optimality. Instead, we make use of (b) and make the following assumption:

(A4) For any $v \geq \hat{v}$, there exist $\mu, \varepsilon > 0$ s.t.

$\text{dist}(\theta, \Theta) \leq \mu \cdot \|E(\theta)\|_2$ for any θ satisfying $F(\theta) \leq v$ and $\|E(\theta)\|_2 \leq \varepsilon$.

((Local) Error Bound Condition)

- The error bound condition (A4) says that when θ is close to the optimal set Θ (as determined by the conditions $F(\theta) \leq v$ and $\|E(\theta)\|_2 \leq \varepsilon$), the measure $\|E(\theta)\|_2$ is a good surrogate of $\text{dist}(\theta, \Theta)$. If μ is independent of v and $\varepsilon = +\infty$, we say that the error bound is global.

- Note that (A4) is a problem-specific property. In other words, it is independent of the algorithm used.

- (A4) is a major assumption and it is not clear yet whether it can be satisfied. However, if it can be satisfied, we will have strong convergence result.

Indeed, using (a)–(c) and (A4), for θ^k close to Θ, we have

$$F(\theta^{k+1}) - \hat{v} \overset{\substack{(A4)\\+\\(c)}}{\leq} K_4 \left(\| E(\theta^k) \|_2^2 + \| \theta^k - \theta^{k+1} \|_2^2 \right)$$

$$\overset{(b)}{\leq} K_5 \cdot \| \theta^k - \theta^{k+1} \|_2^2$$

$$\overset{(a)}{\leq} K_6 \left[(F(\theta^k) - \hat{v}) - (F(\theta^{k+1}) - \hat{v}) \right].$$

Hence,

$$F(\theta^{k+1}) - \hat{v} \leq \frac{K_6}{1+K_6} (F(\theta^k) - \hat{v}).$$

This implies that for k large enough (so that θ^k is close to Θ), $\{ F(\theta^{k+1}) - \hat{v} \}$ tends to 0 at a geometric rate.

Similarly, by (a),

$$\| \theta^k - \theta^{k+1} \|_2^2 \leq \frac{1}{K_1} (F(\theta^k) - \hat{v}) \quad (\because F(\theta^{k+1}) \geq \hat{v})$$

and hence by (A4) and (b),

$$\operatorname{dist}(\theta^k, \Theta) \leq \mu \cdot \| E(\theta^k) \|_2 \leq \mu' \cdot \| \theta^k - \theta^{k+1} \|_2$$

$$\leq \mu'' \sqrt{F(\theta^k) - \hat{v}}.$$

This shows that $\{\theta^k\}$ approaches Θ at a geometric rate as well.