Strong Convexity, Stability, and Generalization

PART 1 — Linear Algebra & Strong Convexity Basics

1. Why is $A + 2\lambda m I$ always invertible?

Question: Could A be $-2\lambda m I$?

Core facts:

- A is PSD $\Rightarrow \mu_i(A) \geq 0$ for all eigenvalues.
- $2\lambda m I$ is PD $\Rightarrow$ all eigenvalues are $2\lambda m > 0$.

  Therefore eigenvalues of $A + 2\lambda m I$ are

$$\mu_i(A) + 2\lambda m,$$

and since $\mu_i(A) \geq 0$ and $2\lambda m > 0$,

$$\mu_i(A) + 2\lambda m > 0.$$

Hence $A + 2\lambda m I$ is PD $\Rightarrow$ invertible $\Rightarrow$ the solution is unique.

2. PD $\iff$ Strong Convexity (Hessian view)

Definition:

$$f \text{ is } \alpha\text{-strongly convex} \iff \nabla^2 f(x) \succeq \alpha I$$

(i.e. the minimum eigenvalue of $\nabla^2 f(x)$ is at least α).

Example.

$$f(w_1, w_2) = w_1^2 + w_2^2, \qquad \nabla f = (2w_1, 2w_2), \qquad \nabla^2 f = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}.$$

Eigenvalues are $2, 2$, so $\lambda_{\min} = 2$ and f is 2-strongly convex.

3. Why does strong convexity imply stability?

Key inequality:

$$f(v) - f(w^*) \geq \lambda \|v - w^*\|^2.$$

Interpretation: the minimum is "sharp". If one sample changes, the minimizer cannot move too much. This is the essence of stability.

PART 2 — Stability Definition

Definition 13.3 (On-average-replace-one stability).

$$\mathbb{E}_{S,z',i}\Big[\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)\Big] \le \epsilon(m).$$

Why the probability space D^{m+1}?

- S contains m i.i.d. samples from D.
- z' is one additional i.i.d. sample used for replacement.
- Total $m+1$ i.i.d. draws $\Rightarrow D^{m+1}$.

Why evaluate at z_i? We compare two models (trained on S and $S^{(i)}$) on the same point z_i: this directly measures how much the model changes at the replaced location.

PART 3 — RLM (Regularized ERM)

$$A(S) = \arg\min_w \Big(L_S(w) + \lambda \|w\|^2\Big), \qquad L_S(w) = \frac{1}{m}\sum_{j=1}^{m} \ell(w, z_j).$$

Define

$$f_S(w) = L_S(w) + \lambda \|w\|^2 .$$

PART 4 — (13.7) Strong Convexity Core Inequality

Lemma 13.5(3) (core form). If f is α-strongly convex and $u = \arg\min f$, then

$$f(w) - f(u) \ge \frac{\alpha}{2} \|w - u\|^2 .$$

Why is f_S (2λ)-strongly convex?

- $\lambda \|w\|^2$ is 2λ-strongly convex,
- L_S is convex,
- strong convex + convex $\Rightarrow$ strong convex (same coefficient).

Set $\alpha = 2\lambda$, $f = f_S$, $u = A(S)$, $w = v$:

$$f_S(v) - f_S(A(S)) \ge \frac{2\lambda}{2} \|v - A(S)\|^2 \quad \Rightarrow \quad \boxed{f_S(v) - f_S(A(S)) \ge \lambda \|v - A(S)\|^2} \tag{13.7}$$

PART 5 — (13.8), (13.9), (13.10) Exact Flow

0) Setting/Notation (the book's core symbols)

- Data sample: $S = (z_1, \ldots, z_m)$.
- One-point replaced sample:

$$S^{(i)} = (z_1, \ldots, z_{i-1}, z', z_{i+1}, \ldots, z_m), \quad z' \sim D.$$

- Empirical risk:

$$L_S(w) = \frac{1}{m} \sum_{j=1}^{m} \ell(w, z_j).$$

- RLM algorithm:

$$A(S) = \arg\min_{w} \big(L_S(w) + \lambda \|w\|^2 \big).$$

- Objective:

$$f_S(w) = L_S(w) + \lambda \|w\|^2 .$$

- Difference vector:

$$\Delta := A(S^{(i)}) - A(S).$$

(13.8) Decompose S vs $S^{(i)}$ into two losses

Start:

$$f_S(v) - f_S(u) = \big(L_S(v) + \lambda \|v\|^2 \big) - \big(L_S(u) + \lambda \|u\|^2 \big).$$

Since S and $S^{(i)}$ differ only at index i,

$$L_S(v) = L_{S^{(i)}}(v) + \frac{1}{m}\big(\ell(v, z_i) - \ell(v, z')\big), \qquad L_S(u) = L_{S^{(i)}}(u) + \frac{1}{m}\big(\ell(u, z_i) - \ell(u, z')\big).$$

Plugging these into $L_S(v) - L_S(u)$ gives:

$$\boxed{\begin{aligned} f_S(v) - f_S(u) = \big(L_{S^{(i)}}(v) + \lambda \|v\|^2 \big) - \big(L_{S^{(i)}}(u) + \lambda \|u\|^2 \big) \\ + \frac{1}{m}\Big(\ell(v, z_i) - \ell(u, z_i) + \ell(u, z') - \ell(v, z') \Big). \end{aligned}} \tag{13.8}$$

(13.9) Use minimizer property to flip sign and upper bound

Choose

$$v = A(S^{(i)}), \qquad u = A(S).$$

The first big bracket in (13.8) becomes

$$f_{S^{(i)}}(v) - f_{S^{(i)}}(u) = f_{S^{(i)}}(A(S^{(i)})) - f_{S^{(i)}}(A(S)) \leq 0,$$

since $A(S^{(i)})$ minimizes $f_{S^{(i)}}$. Thus dropping it yields:

$$\boxed{f_S(A(S^{(i)})) - f_S(A(S)) \leq \frac{\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)}{m} + \frac{\ell(A(S), z') - \ell(A(S^{(i)}), z')}{m}} \tag{13.9}$$

(13.10) Combine (13.7) + (13.9)

Plug $v = A(S^{(i)})$ into (13.7):

$$f_S(A(S^{(i)})) - f_S(A(S)) \geq \lambda \left\| A(S^{(i)}) - A(S) \right\|^2 = \lambda \left\| \Delta \right\|^2 .$$

Together with (13.9), we obtain the sandwich inequality:

$$\boxed{\lambda \left\| \Delta \right\|^2 \leq \frac{\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)}{m} + \frac{\ell(A(S), z') - \ell(A(S^{(i)}), z')}{m}} \tag{13.10}$$

PART 6 — Lipschitz Branch: (13.11) → $\|\Delta\|$ bound → Stability

Assume ρ-Lipschitz:

$$|\ell(w, z) - \ell(u, z)| \leq \rho \left\| w - u \right\| .$$

Apply to RHS of (13.10):

$$\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i) \leq \rho \left\| \Delta \right\| , \qquad \ell(A(S), z') - \ell(A(S^{(i)}), z') \leq \rho \left\| \Delta \right\| .$$

Thus

$$\lambda \left\| \Delta \right\|^2 \leq \frac{2\rho}{m} \left\| \Delta \right\| .$$

If $\|\Delta\| > 0$,

$$\boxed{\left\| \Delta \right\| \leq \frac{2\rho}{\lambda m}.} \tag{13.11}$$

Then again by Lipschitzness,

$$\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i) \leq \rho \left\| \Delta \right\| \leq \boxed{\frac{2\rho^2}{\lambda m}},$$

giving the stability bound.

PART 7 — Smooth Branch: (13.13)(13.14) → $\|\Delta\|$ bound → Stability

(13.13) From β-smoothness definition

Assume β-smoothness:

$$\ell(u, z) \leq \ell(w, z) + \langle \nabla \ell(w, z), u - w \rangle + \frac{\beta}{2} \left\| u - w \right\|^2 .$$

Set $u = A(S^{(i)})$, $w = A(S)$, $u - w = \Delta$:

$$\boxed{\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i) \leq \langle \nabla \ell(A(S), z_i), \Delta \rangle + \frac{\beta}{2} \left\| \Delta \right\|^2} \tag{13.13}$$

(13.14) Cauchy–Schwarz + self-boundedness

Cauchy–Schwarz:

$$\langle \nabla \ell(A(S), z_i), \Delta \rangle \leq \|\nabla \ell(A(S), z_i)\| \, \|\Delta\| \, .$$

Self-boundedness (nonnegative + smooth):

$$\|\nabla \ell(w, z)\|^2 \leq 2\beta \, \ell(w, z) \quad \Rightarrow \quad \|\nabla \ell(w, z)\| \leq \sqrt{2\beta \, \ell(w, z)}.$$

Hence

$$\boxed{\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i) \leq \sqrt{2\beta \, \ell(A(S), z_i)} \, \|\Delta\| + \frac{\beta}{2} \|\Delta\|^2} \tag{13.14}$$

A symmetric bound holds for the z'-term.

Derive the $\|\Delta\|$ bound from (13.10)

Let

$$\ell_1 := \ell(A(S), z_i), \qquad \ell_2 := \ell(A(S^{(i)}), z').$$

Plugging (13.14) (and the symmetric one) into (13.10):

$$\lambda \|\Delta\|^2 \leq \frac{1}{m} \Big(\sqrt{2\beta \ell_1} \, \|\Delta\| + \frac{\beta}{2} \|\Delta\|^2 + \sqrt{2\beta \ell_2} \, \|\Delta\| + \frac{\beta}{2} \|\Delta\|^2 \Big).$$

So

$$\lambda \|\Delta\|^2 \leq \frac{\sqrt{2\beta}}{m} (\sqrt{\ell_1} + \sqrt{\ell_2}) \, \|\Delta\| + \frac{\beta}{m} \|\Delta\|^2 \, .$$

Multiply by m and rearrange:

$$(\lambda m - \beta) \, \|\Delta\|^2 \leq \sqrt{2\beta} (\sqrt{\ell_1} + \sqrt{\ell_2}) \, \|\Delta\| \, .$$

If $\lambda m > \beta$ and $\|\Delta\| > 0$:

$$\boxed{\|\Delta\| \leq \frac{\sqrt{2\beta}}{\lambda m - \beta} (\sqrt{\ell_1} + \sqrt{\ell_2})}.$$

If $\beta \leq \lambda m/2$, then $\lambda m - \beta \geq \lambda m/2$ and

$$\boxed{\|\Delta\| \leq \frac{\sqrt{8\beta}}{\lambda m} (\sqrt{\ell_1} + \sqrt{\ell_2})}.$$

This leads to the smooth-case stability bound (loss difference bound) by plugging the Δ bound back into (13.14) and using standard inequalities (e.g. $(a + b)^2 \leq 3(a^2 + b^2)$).

PART 8 — Fitting–Stability Tradeoff: (13.15)(13.16) + Corollaries

(13.15) Decomposition identity

$$\boxed{\mathbb{E}[L_D(A(S))] = \mathbb{E}[L_S(A(S))] + \mathbb{E}[L_D(A(S)) - L_S(A(S))]} \tag{13.15}$$

- First term: *fitting*.
- Second term: *generalization gap* (controlled by stability).

Tradeoff. Increasing λ improves stability (smaller gap) but worsens fitting.

(13.16) Bound on fitting via minimizer property

From optimality of $A(S)$:

$$L_S(A(S)) + \lambda \|A(S)\|^2 \le L_S(w^*) + \lambda \|w^*\|^2 .$$

Drop $\lambda \|A(S)\|^2 \ge 0$:

$$L_S(A(S)) \le L_S(w^*) + \lambda \|w^*\|^2 .$$

Taking expectation and using $\mathbb{E}_S[L_S(w^*)] = L_D(w^*)$:

$$\boxed{\mathbb{E}[L_S(A(S))] \le L_D(w^*) + \lambda \|w^*\|^2} \tag{13.16}$$

Corollary 13.8 (Lipschitz oracle inequality)

Combine (13.15) + (13.16) + Lipschitz stability gap:

$$\boxed{\mathbb{E}[L_D(A(S))] \le L_D(w^*) + \lambda \|w^*\|^2 + \frac{2\rho^2}{\lambda m}}$$

Corollary 13.9 (PAC-like bound)

Choosing $\lambda \asymp 1/\sqrt{m}$ (e.g. optimizing $\lambda B^2 + \frac{2\rho^2}{\lambda m}$ under $\|w^*\| \le B$) yields

$$\mathbb{E}[L_D(A(S))] \le \min_w L_D(w) + O\left(\frac{1}{\sqrt{m}}\right).$$

Corollary 13.10/13.11 (Smooth case final result)

Using the smooth-case stability bound inside (13.15) and combining gives a multiplicative bound:

$$\mathbb{E}[L_D(A(S))] \le \left(1 + \frac{48\beta}{\lambda m}\right) \mathbb{E}[L_S(A(S))].$$

Setting $\lambda = \frac{48\beta}{m}$ gives

$$\text{true risk} \le 2 \times \text{empirical risk}.$$