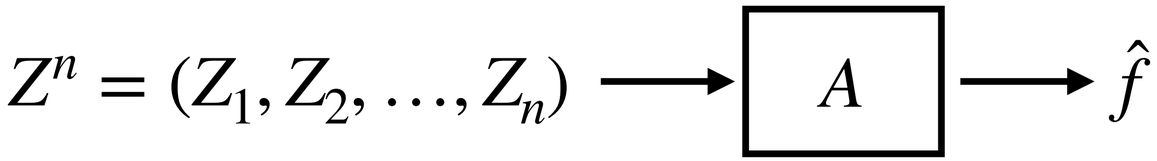
24. Stability of SGD

Recap

- **Last week.** Generalization bounds via algorithmic properties



  - Stability. Generalize, if the algorithmic output (predictor) is stable w.r.t. dataset change

$$\text{dist}(Z^n, \tilde{Z}^n) \leq \delta \quad \rightarrow \quad \text{dist}(A(Z^n), A(\tilde{Z}^n)) \leq \epsilon$$

  - Information. Generalize, if the algorithm output has small shared information with dataset

$$I(Z^n; \hat{f}) \leq \epsilon$$

- **Today.** Apply stability bounds for SGD

Recap: Stochastic Gradient Descent

- **GD.** A procedure which iterates:

$$f_{t+1} = f_t - \eta_t \nabla \hat{R}(f_t)$$

  - Π denotes some projection operation

- **SGD.** A procedure which iterates

$$f_{t+1} = f_t - \eta_t \nabla \ell(\hat{f}_t, Z_{I_t}), \qquad I_t \sim \text{Unif}(\{1, \ldots, n\})$$

  - Single-batch GD with reshuffling

Iteration map

- For simplicity, we write

$$G_{\varphi,\eta} = f - \eta \nabla \varphi(f)$$

  - Using this notation, the SGD can be written as:

$$f_{t+1} = G_{\ell(\cdot,Z_{I_t}),\eta_t}(f_t) \;=: G_t(f_t)$$

- Then, under some conditions, this mapping is contractive

Lemma (Contractive property of GD).

Let φ be a β-smooth and convex function. Then, whenever $\eta \in [0,2/\beta]$, we have:

$$\|G_{\varphi,\eta}(f) - G_{\varphi,\eta}(\tilde{f})\| \leq \|f - \tilde{f}\|$$

If φ is also λ-strongly convex, we have

$$\|G_{\varphi,\eta}(f) - G_{\varphi,\eta}(\tilde{f})\| \leq \sqrt{1 - \eta\lambda(2 - \eta\beta)}\,\|f - \tilde{f}\|$$

Proof sketch

- Observe that

$$
\begin{aligned}
\|G_{\varphi,\eta}(f) - G_{\varphi,\eta}(\tilde{f})\|^2 &= \|f - \tilde{f} - \eta(\nabla\varphi(f) - \nabla\varphi(\tilde{f})\|^2 \\
&= \|f - \tilde{f}\|^2 - 2\eta\langle \nabla\varphi(f) - \nabla\varphi(\tilde{f}), f - \tilde{f}\rangle + \eta^2\|\nabla\varphi(f) - \nabla\varphi(\tilde{f})\|^2 \\
&\leq \|f - \tilde{f}\|^2 - \eta(2 - \eta\beta)\langle \nabla\varphi(f) - \nabla\varphi(\tilde{f}), f - \tilde{f}\rangle \\
&\leq \|f - \tilde{f}\|^2 - \lambda\eta(2 - \eta\beta)\|f - \tilde{f}\|^2 \\
&= (1 - \lambda\eta(2 - \eta\beta))\|f - \tilde{f}\|^2
\end{aligned}
$$

- Convex case can be achieved by letting $\lambda = 0$

Ghost samples

- This contraction property will let us control the stability of SGD

- Fix an index i^*, and generate a "ghost sample" for this index i^*

$$\tilde{Z}^n = (Z_1, \ldots, Z_{i^*-1}, \tilde{Z}_{i^*}, Z_{i^*+1}, \ldots, Z_n)$$

  - Write $\tilde{G}_t(\,\cdot\,)$ to denote updates on $\tilde{Z}^n$

- We'll examine two SGD procedures:

$$f_{t+1} = G_t(f_t), \qquad \tilde{f}_{t+1} = \tilde{G}_t(\tilde{f}_t)$$

  - We share the randomness:
    - **SGD ordering.** Use the same indices $I_1, I_2, \ldots$
    - **Initialization.** Use the same init $f_0 = \tilde{f}_0$

Stability recursion

- We are interested in the stability of SGD, i.e., how much $\tilde{f}_t$ deviate from f_t

$$\delta_t := \|f_t - \tilde{f}_t\|$$

- Easiest way to upper-bound δ_t is by recursion
  - Relate δ_t with δ_{t-1}

Case $I_t \neq i^*$

- Then, we have

$$\begin{aligned}\|f_{t+1} - \tilde{f}_{t+1}\| &= \|G_t(f_t) - \tilde{G}_t(\tilde{f}_t)\| \\ &= \|G_t(f_t) - G_t(\tilde{f}_t)\| \\ &\leq \mathrm{Lip}(G_t) \cdot \|f_t - \tilde{f}_t\|\end{aligned}$$

Stability recursion

Case $I_t = i^*$

- Then, we have

$$\begin{aligned}\|f_{t+1} - \tilde{f}_{t+1}\| &= \|G_t(f_t) - \tilde{G}_t(\tilde{f}_t)\| \\ &\leq \|G_t(f_t) - G_t(\tilde{f}_t)\| + \|G_t(\tilde{f}_t) - \tilde{G}_t(\tilde{f}_t)\| \\ &\leq \mathrm{Lip}(G_t) \cdot \|f_t - \tilde{f}_t\| + \|G_t(\tilde{f}_t) - \tilde{f}_t\| + \|\tilde{f}_t - \tilde{G}_t(\tilde{f}_t)\| \\ &\leq \mathrm{Lip}(G_t) \cdot \|f_t - \tilde{f}_t\| + 2c_t\end{aligned}$$

  - Here, we use the shorthand:

$$c_t = \sup_{f \in \mathcal{F}} \left(\max\{\|G_t(f) - f\|, \|\tilde{G}_t(f) - f\|\} \right)$$

Combining cases

- We have:

$$\delta_{t+1} \leq \mathrm{Lip}(G_t) \cdot \delta_t + 2c_t \mathbf{1}\{I_t = i^*\}$$

Stability: Convex case

- Using this result, we can show the following theorem for the convex case

Theorem (Convex case).

Suppose that, for all z, $\ell(\cdot, z)$ is convex, β-smooth, and L-Lipschitz. Suppose also that we run SGD with $\eta_t \leq 2/\beta$ for T steps.

Then, for any two datasets that differ only in one sample, we have:

$$\sup_z \mathbb{E} \left| \ell(f_T, z) - \ell(\tilde{f}_T, z) \right| \leq \frac{2L^2}{n} \sum_{t=0}^{T-1} \eta_t$$

- Here, the expectation is with respect to the internal randomness of SGD
  - i.e., the selection process I_t
- Longer training = poorer generalization

Proof sketch

- First, note that:

$$\|f - G_{\varphi,\eta}(f)\| = \|\eta \nabla \varphi(f)\| \leq \eta \cdot \mathrm{Lip}(\varphi)$$

- Thus, we have:

$$c_t = \sup_{f \in \mathcal{F}} \left(\max\{ \|G_t(f) - f\|, \|\tilde{G}_t(f) - f\| \} \right) \leq \eta_t L$$

- Combining this with the previous argument and the contraction lemma, we have

$$\delta_{t+1} \leq \delta_t + 2\eta_t L \mathbf{1}\{I_t = i^*\}$$

- Taking expectation on both sides, we get:

$$\mathbb{E}[\delta_{t+1}] \leq \mathbb{E}[\delta_t] + \frac{2\eta_t L}{n}$$

Proof sketch

$$\mathbb{E}[\delta_{t+1}] \leq \mathbb{E}[\delta_t] + \frac{2\eta_t L}{n}$$

- Telescoping, with $\delta_0 = 0$, we get:

$$\mathbb{E}[\|f_T - \tilde{f}_T]\|] = \mathbb{E}[\delta_T] \leq \frac{2L}{n} \sum_{t=0}^{T-1} \eta_t$$

- Using the Lipschitz property of the loss, we get the claim.

Example bound

- If we apply the theorem with:
  - $T = n$ (can't avoid this dependency)
  - $\eta_t = 2/\beta\sqrt{n}$

- Then we get:

$$\mathbb{E}[R(\hat{f}) - \hat{R}(\hat{f})] \leq \frac{4L^2}{\beta\sqrt{n}}$$

Stability: Strongly convex case

Theorem (Strongly convex case).

Suppose that, for all z, $\ell(\cdot, z)$ is λ-strongly convex, β-smooth, and L-Lipschitz. Suppose also that we run SGD with a constant learning rate $\eta \leq 1/\beta$ for T steps.

Then, for any two datasets that differ only in one sample, we have:

$$\sup_z \mathbb{E} \left| \ell(f_T, z) - \ell(\tilde{f}_T, z) \right| \leq \frac{4L^2}{\lambda n}$$

- Note that we do not have an explicit T-dependency here
  - Generalize well, regardless of how long we train
  - Can get $1/n$ dependency

Proof sketch

- Similarly to the previous case, we can get

$$\begin{aligned}\delta_{t+1} &\leq \sqrt{1-\eta\lambda(2-\eta\beta)}\delta_t + 2\eta_t L\mathbf{1}\{I_t = i^*\} \\ &\leq \sqrt{1-\eta\lambda}\delta_t + 2\eta L\mathbf{1}\{I_t = i^*\} \\ &\leq \left(1-\frac{\eta\lambda}{2}\right)\delta_t + 2\eta L\mathbf{1}\{I_t = i^*\}\end{aligned}$$

- Taking expectations on both sides, we get

$$\mathbb{E}[\delta_{t+1}] \leq \left(1-\frac{\eta\lambda}{2}\right)\mathbb{E}[\delta_t] + \frac{2\eta L}{n}$$

- Unwinding the recursion, we get:

$$\begin{aligned}\mathbb{E}[\delta_{t+1}] &\leq \frac{2\eta L}{n}\sum_{t=0}^{T-1}\left(1-\frac{\eta\lambda}{2}\right)^t \\ &\leq \frac{2\eta L}{n}\cdot\frac{2}{\eta\lambda} \qquad = \frac{4L}{n\lambda}\end{aligned}$$

Remarks

- One can show this, even for non-convex cases
    - still requiring convexity and Lipschitz continuity

- A neat way to connect the number of SGD steps with generalization
    - without requiring norms…