Stability of Stochastic Gradient Descent

- Learning algo $A : \mathcal{Z}^* \to \mathcal{F}$

$$\mathcal{Z}^* = \bigcup_{n \geq 1} \mathcal{Z}^n$$

- Properties:

  - consistency:

$$\lim_{n\to\infty} \sup_{P \in \mathcal{P}} \left\{ \mathbb{E}[L_P(A(Z^n))] - L_P^* \right\} = 0$$

where $L_P(f) = \mathbb{E}[\ell(f, Z)] \qquad Z \sim P$

$$L_P^* = \inf_{f \in \mathcal{F}} L_P(f)$$

  - AERM:

$$\lim_{n\to\infty} \sup_{P \in \mathcal{P}} \mathbb{E}[L_n(A(Z^n)) - L_n^*] = 0$$

where $L_n(f) := \frac{1}{n}\sum_{i=1}^{n} \ell(f, z_i)$

$$L_n^* := \inf_{f \in \mathcal{F}} L_n(f)$$

  - stability:

$$\lim_{n\to\infty} \sup_{P \in \mathcal{P}} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\ell(A(Z^n), Z_i') - \ell(A(Z^n_{(i)}), Z_i')\right] = 0$$

where $Z^n = (Z_1, \ldots, Z_n)$, $Z'^n = (Z_1', \ldots, Z_n')$ iid $\sim P$

$$Z^n_{(i)} = (Z_1, \ldots, Z_{i-1}, Z_i', Z_{i+1}, \ldots, Z_n)$$

$$Z^n \overset{A}{\longmapsto} A(Z^n), \quad Z^n_{(i)} \longmapsto A(Z^n_{(i)})$$

NB: stability $\Leftrightarrow$ generalization (on avg)

$$\lim_{n\to\infty} \sup_{P \in \mathcal{P}} \left| \mathbb{E}[L_n(A(Z^n)) - L(A(Z^n))] \right| = 0$$

$$\boxed{\text{AERM} + \text{stability} \Rightarrow \text{consistency}}$$

Gen.: $\lim_{n\to\infty} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left| L_n(A(Z^n)) - L(A(Z^n)) \right| = 0$

e.g. $0 \le \ell \le 1$

AERM + gen. on avg. $\Rightarrow$ gen.

Stab. on avg. $\Leftarrow$ uniform stability:

$$\sup_{z \in \mathcal{Z}} \left| \mathbb{E}[\ell(A(Z^n), z)] - \mathbb{E}[\ell(A(Z^n_{(i)}), z)] \right| \to 0$$

E.g. $f \mapsto \ell(f, z)$ is Lipschitz, unif. in z

$$\sup_{z \in \mathcal{Z}} |\ell(f, z) - \ell(f', z)| \le L \|f - f'\|$$

so if $A(Z^n) \approx A(Z^n_{(i)})$ $\forall i$

then $\sup_{z \in \mathcal{Z}} |\ell(A(Z^n), z) - \ell(A(Z^n_{(i)}), z)|$

$\le L \|A(Z^n) - A(Z^n_{(i)})\|$.

Stochastic Gradient Descent

Goal: prove uniform stability of SGD (Hardt - Recht - Singer 2016)

ERM:

$$\min_{f \in \mathcal{F}} L_n(f)$$

$\mathcal{F}$: closed, convex subset of H.S. $\mathcal{H}$

$$L_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f, z_i)$$

$z_1, \dots, z_n \overset{iid}{\sim} P$

Approximate ERM: gradient descent

- assume $f \mapsto \ell(f, z)$ is diff. $(\forall z)$

$\nabla \ell(f, z)$: gradient

GD: $f_0 \in \mathcal{F}$ (init.)

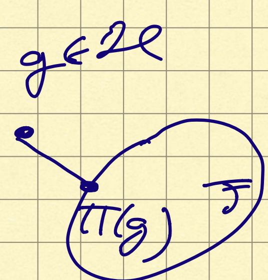
$$f_t = \Pi\left(f_{t-1} - \alpha_t \nabla L_n(f_{t-1})\right) \qquad t = 1, 2, \dots$$

where $\nabla L_n(f) := \frac{1}{n}\sum_{i=1}^{n} \nabla \ell(f, z_i)$

$(\alpha_t)_{t \geq 1}^{\infty}$: positive step sizes

$\Pi : \mathcal{H} \to \mathcal{F}$: projection onto $\mathcal{F}$

$$\Pi(g) := \underset{f \in \mathcal{F}}{\operatorname{argmin}} \| g - f \|^2$$



Opt. theory: $\lim_{t \to \infty} L_n(f_t) \to L_n^*$

under various regularity assumptions (including carefully tuned step sizes)

<u>Drawback:</u> computation of f_t takes $O(n)$ steps (compute $\nabla L_n(f_{t-1})$)

<u>Compromise:</u> stochastic approx. (SA)

$$\nabla L_n(f) = \frac{1}{n}\sum_{i=1}^{n} \nabla \ell(f, z_i)$$

$$I \sim \mathrm{Unif}([n])$$

$$\nabla L_n(f) = \mathbb{E}_I[\ell(f, z_I)]$$

$\Rightarrow$ $\ell(f, z_I)$ is an unbiased est. of $\nabla L_n(f)$

(only I is random!)

<u>GD:</u> $f_{t-1} - \alpha_t \nabla L_n(f_{t-1})$ vs. <u>SGD:</u> $f_{t-1} - \alpha_t \nabla \ell(f_{t-1}, z_I)$

$$f_{t-1} - \alpha_t \nabla \ell(f_{t-1}, z_I)$$
$$= f_{t-1} - \alpha_t [\nabla L_n(f_{t-1}) + \xi_t]$$

where $\xi_t := \nabla \ell(f_{t-1}, z_I) - \nabla L_n(f_{t-1})$

$$\mathbb{E}_I[\xi_t] = 0$$

Details matter; this is just heuristic.

<u>Gain:</u> each iteration now takes $O(1)$ steps.

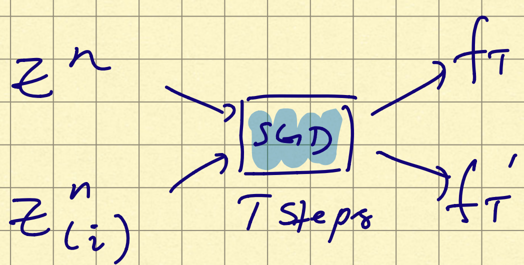
Let $(I_t)_{t\geq 1}^{\infty}$ be iid from Unif$([n])$
(other choices are possible)

<u>SGD:</u> $f_0 \in \mathcal{F}$ (init, indep of Z^n)

$t = 1, 2, \dots, T$:

$$f_t = \Pi\left(f_{t-1} - \alpha_t \nabla \ell(f_{t-1}, Z_{I_t})\right)$$

$A(Z^n) = f_T$ — stability?



<u>Notation:</u> $G_{\varphi, \alpha}(f) := \Pi(f - \alpha \nabla \varphi(f))$

$\varphi : \mathcal{F} \to \mathbb{R}$ differentiable

$\alpha > 0$

Π: proj. onto $\mathcal{F}$

$(f_t)_{t=0}^{T}$ (on Z^n) $\qquad (f_t')_{t=0}^{T}$ (on $Z^n_{(i)}$)

$f_0 = f_0'$ (same init)

$f_t = G_t(f_{t-1}) \qquad f_t' = G_t'(f_{t-1}')$

where $G_t = G_{\ell(\cdot, Z_{I_t}), \alpha_t}$, G_t' analogous

$$Z^n = (z_1, \dots, z_{i-1}, z_i, z_{i+1}, \dots, z_n)$$
$$Z^n_{(i)} = (z_1, \dots z_{i-1}, z_i', z_{i+1}, \dots, z_n)$$

Let $\delta_t := \|f_t - f_t'\|$ $\quad\quad \delta_0 = 0$

$$\delta_t = \|f_t - f_t'\|$$
$$= \|G_t(f_{t-1}) - G_t'(f_{t-1}')\|$$

two cases: $\begin{cases} I_t \neq i & (G_t = G_t') \quad ① \\ I_t = i & (G_t \neq G_t') \quad ② \end{cases}$

① $\|G_t(f_{t-1}) - G_t'(f_{t-1}')\|$
$$= \|G_t(f_{t-1}) - G_t(f_{t-1}')\|$$
$$\leq \eta_t \|f_{t-1} - f_{t-1}'\|$$
$$= \eta_t \delta_{t-1}$$

where $\eta_t := \sup_{f, f' \in \mathcal{F}} \dfrac{\|G_t(f) - G_t(f')\|}{\|f - f'\|}$

(want $\eta_t \leq 1$)

② $\|G_t(f_{t-1}) - G_t'(f_{t-1}')\|$ $\quad\quad (G_t \neq G_t')$

$$\leq \underbrace{\|G_t(f_{t-1}) - G_t(f_{t-1}')\|}_{\leq \eta_t \delta_{t-1}} + \underbrace{\|G_t(f_{t-1}') - G_t'(f_{t-1}')\|}_{(D)}$$

$$(D) \le \| G_t(f_{t-1}') - f_{t-1}' \| + \| G_t'(f_{t-1}') - f_{t-1}' \|$$

$$\le 2c_t$$

where $c_t := \sup_{f \in \mathcal{F}} \max \{ \| G_t(f) - f \|, \| G_t'(f) - f \| \}$

$$\therefore \quad \boxed{\delta_t \le \eta_t \delta_{t-1} + 2c_t \mathbb{1}_{\{I_t = i\}}}$$

Assume $\eta_t \le 1 \quad \forall t$ (guaranteed by choice of α_t)

<u>Assumptions:</u> on $f \mapsto \ell(f, z)$

1) convex
2) L-Lipschitz: $\sup_{z \in \mathcal{Z}} | \ell(f, z) - \ell(f', z) | \le L \| f - f' \|$
3) M-smooth: $\sup_{z \in \mathcal{Z}} \| \nabla \ell(f, z) - \nabla \ell(f', z) \| \le M \| f - f' \|$
4) m-strongly cvx: $(m \geq 0)$

$$f \mapsto \ell(f, z) - \frac{m}{2} \| f \|^2 \quad \text{cvx}$$

<u>Thm</u> Assume 1) – 3); then, for $\alpha_t \le 2/M$,

$$\mathbb{E}[\delta_T] \le \frac{2L^2}{n} \sum_{t=1}^{T} \alpha_t \quad (\text{expect. w.r.t. } (I_t))$$

Proof (sketch)

1) – 3) : $\eta_t \leq 1$

$c_t \leq \alpha_t L$

$$\| G_{\varphi,\alpha}(f) - f \| = \| \Pi(f - \alpha \nabla \varphi(f)) - \Pi(f) \|$$
$$\leq \alpha \| \nabla \varphi(f) \|$$

$$\delta_t \leq \delta_{t-1} + 2\alpha_t L \mathbb{1}_{\{I_t = i\}} \qquad \delta_0 = 0$$

$$\delta_T \leq 2L \sum_{t=1}^{T} \alpha_t \mathbb{1}_{\{I_t = i\}}$$

$$\mathbb{E}[\delta_T] \leq 2L \sum_{t=1}^{T} \alpha_t \, \mathbb{P}[I_t = i]$$
$$= \frac{2L}{n} \sum_{t=1}^{T} \alpha_t . \qquad \blacksquare$$

Corollary $\sup_{z \in Z} \mathbb{E}[\ell(f_T, z) - \ell(f_T'; z)] \leq \frac{2L^2}{n} \sum_{t=1}^{T} \alpha_t .$

E.g. $T = n$

$\alpha_t = \frac{2}{M\sqrt{n}} \qquad t = 1, \dots, T$

$$\sum_{t=1}^{T} \alpha_t = \frac{2\sqrt{n}}{M}$$

$$\Rightarrow \mathbb{E}[\ell(f_T, z) - \ell(f_T'; z)] \leq \frac{4L^2}{M\sqrt{n}} .$$

Thm Assume 1)–4). If $\alpha_t \le \frac{1}{M}$, then

$$\mathbb{E}[\delta_T] \le \frac{4L}{mn} \quad \forall T$$

Key idea: $\alpha_t = 1/M \Rightarrow \eta_t \le 1 - \frac{\alpha m}{2}$

$$= 1 - \frac{m}{2M}$$

$$\delta_t \le \underbrace{\left(1 - \frac{m}{2M}\right)}_{<1} \delta_{t-1} + \frac{2}{M} \mathbb{1}\{I_t = i\}$$