Analysis of Stochastic Gradient Descent

Review:

apx ERM

$$L_n(f) := \frac{1}{n}\sum_{i=1}^{n} \ell(f, z_i) \quad \to \min \text{ s.t. } f \in \mathcal{F}$$

$$(\xi_t) \overset{iid}{\sim} \text{Unif}([n])$$

$$f_0 \in \mathcal{F} \qquad \text{(init)}$$

$$f_t = \Pi\big(f_{t-1} - \alpha_t \underbrace{\nabla \ell(f_{t-1}, \xi_t)}_{\text{Stoch. grad.}}\big) \qquad t = 1, 2, \ldots$$

$$\mathbb{E}_{\xi_t}[\nabla \ell(f_{t-1}, \xi_t)] = \frac{1}{n}\sum_{i=1}^{n} \nabla \ell(f_{t-1}, z_i)$$

$t = 1, \ldots, T$: $\quad \mathbb{E}[L_n(f_T)] - \min_{f \in \mathcal{F}} L_n(f) \le ?$

$\uparrow$ w.r.t. (ξ_t)

$T \to \infty$

- Stochastic Approximation

— classic paper of Robbins-Monro (1950s)

$\mathcal{F}$: Hilbert space

$\Gamma : \mathcal{F} \to \mathbb{R}$: cont. diff., has a finite inf.:

$$\Gamma^* := \inf_{f \in \mathcal{F}} \Gamma(f) < \infty$$

$f_1 \in \mathcal{F}$: init

$(\xi_t)_{t \geq 1}$: iid random elements of some space $\mathcal{S}$

$g : \mathcal{F} \times \mathcal{S} \to \mathcal{F}$: stoch. gradients

$$f_{t+1} = f_t - \alpha_t \, g(f_t, \xi_t) \qquad \text{(SA update)}$$

$(\alpha_t)_{t \geq 1}$: positive nonincreasing seq. of step sizes

$(f_t)_{t \geq 1}$: random process w/ values in $\mathcal{F}$

$f_1 \in \mathcal{F}$: deterministic init

$f_{t+1} = G_t(f_t, \xi_t)$ $\quad$ ξ_t iid

$\Rightarrow$ $(f_t)_{t \geq 1}$ is a Markov process

$\forall V : \mathcal{F} \to \mathbb{R}$

$\mathbb{E}[V(f_t) | f_1, \dots, f_{t-1}] = \mathbb{E}[V(f_t) | f_{t-1}]$

Goals: i) control expected optimality gap

$$\Delta_t := \mathbb{E}[\Gamma(f_t)] - \Gamma^*$$

ii) control expected sq. norms of gradients:

$$\mathbb{E}\|\nabla \Gamma(f_t)\|^2$$

Examples

$$\Gamma(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f, z_i) \qquad \text{(emp. loss)}$$

opt. gap.: $\Gamma(f_t) - \Gamma^* = \left(\text{loss of } f_t\right) - \left(\text{min emp loss}\right)$

- $\xi_t \overset{iid}{\sim} \text{unif}([n]) \qquad \mathcal{S} = [n]$

$$g(f, \xi) := \nabla \ell(f, z_\xi)$$

$$\mathbb{E}_\xi\, g(f, \xi) = \frac{1}{n}\sum_{i=1}^{n} \nabla\ell(f, z_i) = \nabla\Gamma(f)$$

- $\xi_t \overset{iid}{\sim} \text{unif}\left(\binom{[n]}{k}\right) \qquad 1 \le k < n$

$\mathcal{S} = \binom{[n]}{k}$ = all subsets of $[n]$ of card. k

$$\xi = \{i_1, \dots, i_k\} \subset [n]$$

$$g(f, \xi) := \frac{1}{k}\sum_{j \in \xi} \nabla\ell(f, z_\xi) \qquad \text{[mini-batches]}$$

Both cases: $\mathbb{E}_\xi [g(f, \xi)] = \nabla\Gamma(f)$ (unbiased stoch. gradients)

SGD/SA: $f_{t+1} = f_t - \alpha_t\left(\nabla\Gamma(f_t) + \text{noise}\right)$

Assumptions:

1) $\exists\ 0 < \mu \leq 1$ s.t.

$$\langle \nabla \Gamma(f_t), \mathbb{E}_{\xi_t}[g(f_t, \xi_t)] \rangle \geq \mu \|\nabla \Gamma(f_t)\|^2 \quad \forall t$$

e.g. if $\mathbb{E}_{\xi}[g(f, \xi)] = \nabla \Gamma(f)$, then we have $\mu = 1$ + equality (unbiased stoch. grads)

2) $\exists\ B \geq 0,\ B_G \geq 0$ s.t.

$$\mathbb{E}_{\xi_t} \|g(f_t, \xi_t)\|^2 \leq B + B_G \|\nabla \Gamma(f_t)\|^2 \quad \forall t$$

(can show: $B_G \geq \mu^2$)

– can verify directly for above example

$\xi \sim \mathrm{Unif}([n])$, $\quad g(f, \xi) = \nabla \ell(f, z_\xi)$

$$\mathbb{E}_\xi \|g(f, \xi)\|^2 = \frac{1}{n} \sum_{i=1}^{n} \|\nabla \ell(f, z_\xi)\|^2$$

$$\|\nabla \Gamma(f)\|^2 = \left\|\frac{1}{n} \sum_{i=1}^{n} \nabla \ell(f, z_i)\right\|^2$$

B, B_G would depend on $\max_i \|\nabla \ell(f, z_i)\|^2$

Convex Functions

- M-smooth: $\|\nabla\Gamma(f) - \nabla\Gamma(f')\| \le M\|f - f'\|$
- m-strongly convex:

$$\Gamma(f') - \left(\Gamma(f) + \langle \nabla\Gamma(f), f' - f\rangle\right) \ge \frac{m}{2}\|f - f'\|^2$$

Lemma Γ M-smooth $\Rightarrow$

$$\Gamma(f') - \left(\Gamma(f) + \langle \nabla\Gamma(f), f' - f\rangle\right) \le \frac{M}{2}\|f - f'\|^2$$

$(\Rightarrow \quad m \le M)$

– Fixed step sizes: $\alpha_t = \alpha \quad \forall t$

$\Gamma(f_{t+1}) - \Gamma(f_t)$ $\qquad f_{t+1} - f_t = -\cdot\ g_t$

$= \Gamma(f_t - \alpha g_t) - \Gamma(f_t)$ $\qquad g_t := g(f_t, \xi_t)$

$\le \langle \nabla\Gamma(f_t), f_{t+1} - f_t\rangle + \frac{M}{2}\|f_{t+1} - f_t\|^2$

$= -\alpha \langle \nabla\Gamma(f_t), g_t\rangle + \frac{M}{2}\alpha^2\|g_t\|^2$

$\mathbb{E}_{\xi_t}\{\Gamma(f_{t+1}) - \Gamma(f_t)\}$

$= -\alpha \langle \nabla\Gamma(f_t), \mathbb{E}_{\xi_t} g_t\rangle + \frac{M}{2}\alpha^2 \mathbb{E}_{\xi_t}\|g_t\|^2$

$\le -\alpha\mu \|\nabla\Gamma(f_t)\|^2 + \frac{M\alpha^2}{2}\left(B + B_G\|\nabla\Gamma(f_t)\|^2\right)$

$$= \frac{\alpha^2 MB}{2} - \underbrace{\alpha\left(\mu - \frac{\alpha M B_G}{2}\right)}_{\text{want to be } > 0} \underbrace{\|\nabla\Gamma(f_t)\|^2}_{\text{relate to } \Gamma(f_t) - \Gamma^*}$$

- choose $\alpha M B G \le \mu \iff \alpha \le \frac{\mu}{MBG}$

$$\mu - \frac{\alpha M B G}{2} \geq \frac{\mu}{2}$$

$$\mathbb{E}_{\xi_t} \{\Gamma(f_{t+1}) - \Gamma(f_t)\} \le \frac{\alpha^2 M B}{2} - \frac{\alpha\mu}{2} \|\nabla\Gamma(f_t)\|^2$$

By strong convexity, $\forall f \in \mathcal{F}$

$$\Gamma(f) \geq \Gamma(f_t) + \langle \nabla\Gamma(f_t), f - f_t \rangle + \frac{m}{2}\|f - f_t\|^2$$

$$\geq \Gamma(f_t) + \min_{g \in \mathcal{F}} \{\langle \nabla\Gamma(f_t), g \rangle + \frac{m}{2}\|g\|^2\}$$

$$= \Gamma(f_t) - \frac{1}{2m}\|\nabla\Gamma(f_t)\|^2$$

$$\Rightarrow -\|\nabla\Gamma(f_t)\|^2 \le 2m\left(\Gamma(f) - \Gamma(f_t)\right), \forall f$$

$$-\|\nabla\Gamma(f_t)\|^2 \le 2m(\Gamma^* - \Gamma(f_t))$$

$$\mathbb{E}_{\xi_t} \{\Gamma(f_{t+1}) - \Gamma(f_t)\}$$

$$\le \frac{\alpha^2 M B}{2} - \alpha\mu m(\Gamma(f_t) - \Gamma^*)$$

$$\Delta_t := \mathbb{E}\{\Gamma(f_t)\} - \Gamma^*$$

$$\Delta_{t+1} - \Delta_t \le \frac{\alpha^2 M B}{2} - \alpha\mu m \Delta_t$$

$$\Delta_{t+1} \leq \frac{\alpha^2 MB}{2} + (1-\alpha\mu m)\Delta_t \qquad \Delta_1 = F(f_1) - F^*$$

Need $\quad 0 < 1-\alpha\mu m < 1$

$$\alpha \leq \frac{\mu}{MB_G}$$

$$\alpha\mu m \leq \frac{m}{M}\underbrace{\left(\frac{\mu^2}{B_G}\right)}_{\leq 1} \leq \frac{m}{M} < 1$$

$$\Delta_{t+1} \leq (1-\alpha\mu m)\Delta_t + \frac{\alpha^2 MB}{2}$$

$$\Rightarrow \quad \Delta_t \leq \underbrace{(1-\alpha\mu m)^{t-1}\Delta_1}_{\to 0 \text{ as } t\to\infty} + \frac{\alpha^2 MB}{2}$$

$$\frac{\alpha^2 MB}{2} = \frac{\alpha MB}{2}\cdot\alpha \leq \frac{\alpha MB}{2}\cdot\frac{1}{\mu m}$$

$$= \frac{\alpha B}{2\mu}\left(\frac{M}{m}\right)$$

if $F(\cdot)$ is C^2, $\quad \frac{M}{m} \sim$ cond. # of $\nabla^2 F$

<u>Goal:</u> $\quad \mathbb{E}\Delta_t \leq \varepsilon$

1.) choose α s.t. $\quad \frac{\alpha B}{2\mu}\cdot\frac{M}{m} \leq \frac{\varepsilon}{2}$

$$\alpha \leq \frac{\mu m \varepsilon}{MB}$$

2) choose t s.t. $\quad (1-\alpha m\mu)^{t-1} \leq \frac{\varepsilon}{2}$

- Diminishing stepsiz $\quad \alpha_t \to 0$ as $t \to \infty$

Robbins Monro: $\sum_{t=1}^{\infty} \alpha_t = \infty$, $\sum_{t=1}^{\infty} \alpha_t^2 < \infty$ $\Bigg\}$ e.g. $\alpha_t = \frac{c}{\gamma + t}$

$c, \gamma > 0$

Take $\alpha_t = \frac{c}{\gamma + t}$

$\alpha_1 > \alpha_2 > \alpha_3 > \cdots$

choose c, γ carefully:

1) $\alpha_1 \le \frac{\mu}{MBG} \Rightarrow \alpha_t \le \frac{\mu}{MBG}$

$$\Delta_{t+1} \le (1 - \alpha_t m\mu)\Delta_t + \frac{\alpha_t^2 MB}{2}$$

$$= \left(1 - m\mu \frac{c}{\gamma + t}\right)\Delta_t + \frac{c^2 MB}{2(\gamma + t)^2}$$

want: $\Delta_t \le \frac{\nu}{\gamma + t}$ (ν to be tuned)

$$\Delta_{t+1} \le \left(1 - \frac{m\mu c}{\gamma + t}\right)\frac{\nu}{\gamma + t} + \frac{c^2 MB}{2(\gamma + t)^2}$$

$$= \left(\frac{\gamma + t - m\mu c}{(\gamma + t)^2}\right)\nu + \frac{c^2 MB}{2(\gamma + t)^2}$$

$$= \left(\frac{\bar{t} - m\mu c}{\bar{t}^2}\right)\nu + \frac{c^2 MB}{2\bar{t}^2}$$

$\bar{t} := \gamma + t$

$$= \frac{\nu(\bar{t} - 1)}{\bar{t}^2} + \frac{1}{\bar{t}^2}\left[\nu(1 - m\mu c) + \frac{c^2 MB}{2}\right]$$

Choose c, ν, γ s.t. $\longrightarrow \leq 0$

$$\left.\begin{aligned} \frac{t-1}{t^2} &\leq \frac{1}{t+1} \\ \frac{(t-1)(t+1)}{t^2} &\leq 1 \end{aligned}\right\}$$

c, γ, ν tuned

Const. step.: $\Delta_t \leq (1-\alpha m \mu)^{t-1} \Delta_1 + K\alpha$

$\log(\frac{1}{\varepsilon})$ comp. to give ε-opt.

Diminishing step: $\Delta_t \leq \frac{\nu}{\gamma + t}$

$\frac{1}{\varepsilon}$ - comp. to give ε-opt.

$$\frac{\nu}{\gamma + t} \leq \varepsilon \qquad \gamma + t \geq \frac{\nu}{\varepsilon}$$

$$t \geq \frac{\nu}{\varepsilon} - \gamma$$