Ch 14 Stochastic Gradient Descent

Friday, March 20, 2020 3:28 PM

SGD = Stochastic Gradient $\underbrace{\text{Descent}}_{\text{misnomer}}$

$\min_w f(w)$, $f(w) = L_D(w) \equiv \mathbb{E}\, \ell(w, z)$ SA = Stochastic Approximation

or $\hat{L}_S(w) = \frac{1}{m}\sum_{i=1}^{m} \ell(w, z_i)$ SAA = Sample Avg. Approximation = ERM

Algo, same analysis, covers both cases
(ie SAA is a special case of SA w/ $D = \text{Uniform}(\{z_i\}_{i=1}^m)$)

See supplemental notes for basics of gradient descent

$w^{(t+1)} = w^{(t)} - \eta \nabla f(w^{(t)})$ η is learning rate / stepsize

take T steps, output 1) $w^{(T)}$

2) $\underset{w \in \{w^{(1)}, \ldots, w^{(T)}\}}{\text{argmin}} f(w)$ ← can't always evaluate f

or 3) $\bar{w} = \frac{1}{T}\sum_{t=1}^{T} w^{(t)}$

see typed notes

Corollary 14.2 Analysis of GD if f convex, Lipschitz, but not smooth

Let f be convex, ρ-Lipschitz, $w^* \in \text{argmin}_w f(w)$, $\|w^*\| \le B$,

iterate $w^{(t+1)} = w^{(t)} - \eta d_t$, $d_t \in \partial f(w^{(t)})$,

then choosing $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$ gives $f(\bar{w}) - f(w^*) \le \frac{B\rho}{\sqrt{T}}$ ← super slow!

§14.3 SGD

SGD algo: for $t = 1, 2, \ldots, T$

Draw r.v. V_t s.t. $\mathbb{E}(V_t | w^{(t)}) \in \partial f(w^{(t)})$

$w^{(t+1)} = w^{(t)} - \eta V_t$

output $\bar{w} = \frac{1}{T}\sum_{t=1}^{T} w^{(t)}$ other possibilities too (Polyak-Ruppert averaging, various weights)

What might we want to show?

1. First, pick error metric or $\bar{w}$

$$e_t = \begin{cases} f(w^{(t)}) - f(w^*) & \text{Standard choice if } f \text{ convex} \quad (\text{ok}) \\ \|w^{(t)} - w^*\| & \text{choice if } f \text{ strongly convex} \quad (\text{best}) \\ \|\nabla f(w^{(t)})\| & \text{choice if } f \text{ not convex} \quad (\text{weak}) \end{cases}$$

2. V_t, hence $w^{(t)}$, hence e_t, is a random variable

A. $e_t \xrightarrow{\mathbb{P}} 0$ ← or anything convergence in probability (measure)

means $\forall \varepsilon > 0$, $\lim_{t \to \infty} \mathbb{P}(|e_t| > \varepsilon) = 0$ (weak)

B. $e_t \xrightarrow{L^p} 0$ if $\mathbb{E}|e_t|^p = 0$, L^p convergence (p=2 aka quadratic mean) (ok, p=1)

C. $e_t \xrightarrow{a.s.} 0$ if $\mathbb{P}(\lim e_t = 0) = 1$, almost sure convergence aka w/ probability 1 (best)

other types as well (in distribution ...) see prob. textbook

Ex $e_t = \begin{cases} 1 & \text{w.p. } 1/t \\ 0 & \text{w.p. } 1-1/t \end{cases}$ then $\forall r \geq 1$, $\mathbb{E} e_t = 1/t$

so $e_t \xrightarrow{L^r} 0$ ($\forall r$) and $e_t \xrightarrow{\mathbb{P}} 0$

but (I don't think) $e_t \xrightarrow{a.s.} 0$ } Edit: too vague. See end of notes

$e_t = \begin{cases} t^\alpha & \text{wp } 1/t \\ 0 & \text{wp } 1-1/t \end{cases}$ $\quad \alpha = 1/2$ $\mathbb{E} e_t = 1/\sqrt{t} \to 0$ so $e_t \xrightarrow{L^1} 0$

$\mathbb{E} e_t^2 = 1$ so $e_t \not\xrightarrow{L^2} 0$

(Fact: $e_t \xrightarrow{L^r} 0 \Rightarrow e_t \xrightarrow{\mathbb{P}} 0$) $\quad \alpha = 1$ $\mathbb{E} e_t = 1$ So doesn't converge in L^1 even

but $\forall \varepsilon > 0$, $\mathbb{P}(|e_t| > \varepsilon) \leq 1/t \to 0$

so converges in probability

What type to use?

Most ML results show L^1 convergence,

$\mathbb{E}(|e_t|^1) \to 0$

(or, since usually $e_t \geq 0$, $\mathbb{E} e_t \to 0$)

However, if convergence is fast enough,

or for simplest cases (original Robbins-Monro '50's)

prove $\underline{L^2}$ or $\underline{\text{almost sure}}$

Thm 14.8 [SSS] L^1 convergence of SGD assuming ...

Let f be convex, w^* a minimizer, $\|w^*\| \leq B$,

$\|v_t\| \leq \rho$ $\forall t \in [T]$ (w.p. 1) (like ρ-Lipschitz), then

$$0 \leq \mathbb{E} f(\bar{w}) - f(w^*) \leq \frac{B\rho}{\sqrt{T}}, \quad \text{i.e. for } \varepsilon \text{ error, } T \geq \frac{B^2\rho^2}{\varepsilon^2}$$

proof:

(v_t) is a stochastic process

$\mathcal{F}_T := \sigma(v_t : t \leq T)$, then $(\mathcal{F}_T)_{T \in \mathbb{N}}$ is a "filtration"

background

used to help w/ conditional probabilities

write $\mathbb{E}(v_t \mid \{v_{t-1}, v_{t-2}, \ldots, v_0\})$ as $\mathbb{E}(v_t \mid \mathcal{F}_{t-1})$

and use "law of total expectation" aka "tower property"

$\mathbb{E}(\mathbb{E}(X \mid \mathcal{F})) = \mathbb{E}(X)$

i.e. Simpler notation, $\mathbb{E}_\alpha g(\alpha) = \mathbb{E}_\beta(\mathbb{E}_\alpha[g(\alpha) \mid \beta])$

See, e.g.,

https://ocw.mit.edu/courses/sloan-school-of-management/15-070j-advanced-stochastic-processes-fall-2013/lecture-notes/MIT15_070JF13_Lec9.pdf

~~By default, write $\mathbb{E}$ to mean $\mathbb{E}[\cdot \mid \mathcal{F}_T]$~~

Define $\bar{w} = \frac{1}{T}\sum_t w^{(t)}$, $f(\bar{w}) \leq \frac{1}{T}\sum f(w^{(t)})$ by Jensen,

eg

$$f(\bar{w}) - f^* \le \frac{1}{T}\sum_{t=1}^{T} f(w^{(t)}) - f^*$$

$$\mathbb{E}[\text{"\_\_\_\_"}] \le \mathbb{E}[\text{"\_\_\_\_\_\_"}]$$

then via deterministic bounds (lemma 14.1)

if $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$, $\|w^*\| \le B$, $\|v_t\| \le \rho$,

$$\mathbb{E}\left[\frac{1}{T}\sum_t \langle w^{(t)} - w^*, v_t\rangle\right] \le \frac{B\rho}{\sqrt{T}}$$

Now claim

$$\mathbb{E}\left[\frac{1}{T}\sum_t f(w^{(t)}) - f^*\right] \le \underbrace{\qquad\qquad}$$ to conclude proof.

$$\mathbb{E}\left[\frac{1}{T}\sum\langle w^{(t)} - w^*, v_t\rangle\right] = \frac{1}{T}\sum_t \mathbb{E}\langle \underbrace{w^{(t)} - w^*}, v_t\rangle$$

recall $w^{(t)} = w^{(t-1)} - \eta v_{t-1}$

$$= \frac{1}{T}\sum_t \mathbb{E}\left(\mathbb{E}\left[\langle w^{(t)} - w^*, v_t\rangle \mid \mathcal{F}_{t-1}\right]\right)$$

not random for now

and $\underbrace{\mathbb{E}[v_t \mid \mathcal{F}_{t-1}]}_{g_t} \in \partial f(w^{(t)})$ so by convexity $f(w^{(t)}) - f^* \le \langle w^{(t)} - w^*, g_t\rangle$

$$\ge \frac{1}{T}\sum_t \mathbb{E}\, f(w^{(t)}) - f^*$$

$$= \mathbb{E}\left(\frac{1}{T}\sum f(w^{(t)}) - f^*\right) \quad \square$$

§14.5 Learning w/ SGD

i.e. let $f(w) = L_{\mathcal{D}}(w) := \mathbb{E}_{z\sim\mathcal{D}}[\ell(w,z)]$

can't compute $f(w)$ or $\nabla f(w)$ since we don't know $L_{\mathcal{D}}$!

...but we can draw from $\mathcal{D}$ and use SGD.

i.e., sample $z_t \sim \mathcal{D}$, let $v_t \in \partial \ell(w^{(t)}, v_t)$
↑ wrt w

So immediate corollary

Corollary 14.2 ℓ is ρ-Lipschitz, $\|w^*\| \le B$, then "SA" (no ERM)

$\forall \varepsilon > 0$, running SGD w/ $T \ge \frac{B^2\rho^2}{\varepsilon^2}$ iterations,

w/ $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$ then

$$\mathbb{E}\, L_{\mathcal{D}}(\bar{w}) \le \min_{w\in\mathcal{H}} L_{\mathcal{D}}(w) + \varepsilon$$

expected risk, like we discussed in Stability chapter

↑ we didn't discuss constraints, but many simple ones (and regularizers) easily fit into SGD / GD

T is like m, = # iid samples

(if someone says "epochs", they are in the SAA/ERM setting, and $T = 4m$ is "4 epochs". In the true SA setting, like above corollary, we never reach a single epoch, i.e., $m = \infty$)

Results also hold if f is β-smooth, and for RLM

For SAA work, a good review is

Optimization methods for large-scale machine learning
L Bottou, FE Curtis, J Nocedal
SIAM Review 60 (2), 223-311, 2018 (https://arxiv.org/abs/1606.04838)

For SA in optimization context,
See Nemirovski "Robust SA"
or Nesterov "Primal-dual Avg"

Addendum on almost sure convergence

Def $X_n \xrightarrow{a.s.} X$ if $\mathbb{P}[\{\omega \in \Omega : \lim X_n(\omega) = X(\omega)\}] = 1$

i.e. think of random variable as a function

... so as functions, this is convergence a.e.

Let $X_n = \begin{cases} 1 & \text{w.p. } \frac{n+1}{2n} \\ 0 & \text{else} \end{cases}$, $X = \begin{cases} 1 & \text{w.p. } 1/2 \\ 0 & \text{else} \end{cases}$

Does $X_n \xrightarrow{a.s.} X$? Unclear because our setup was too vague.
Instead, define $\Omega = [0,1]$ w/ uniform measure, and define

$X_n(\omega) = \begin{cases} 1 & 0 \le \omega < \frac{n+1}{2n} \\ 0 & \text{else} \end{cases}$ $X(\omega) = \begin{cases} 1 & 0 \le \omega < 1/2 \\ 0 & \text{else} \end{cases}$

Now it's clear $X_n \xrightarrow{a.s.} X$ since $\{\omega : X_n(\omega) \to X(\omega)\} = [0,1] \setminus \{1/2\}$
which has full measure.

... thus a.s. convergence depends on setup

Ex. $X_n = \begin{cases} 0 & \text{w.p. } 1/2 \\ 1 & \text{w.p. } 1/2 \end{cases}$ $X = \begin{cases} 0 & \text{w.p. } 1/2 \\ 1 & \text{w.p. } 1/2 \end{cases}$

Does $X_n \xrightarrow{a.s.} X$? (you'd think so, since appear the same!)

Depends on setup.

Setup 1: $\Omega = \{H, T\}$ w/ uniform measure

$X_n(\omega) = \begin{cases} 0 & \omega = H \\ 1 & \omega = T \end{cases}$ $X(\omega) = \begin{cases} 0 & \omega = T \\ 1 & \omega = H \end{cases}$

Then X_n doesn't converge a.s. to X

Setup 2: Same as above but $X(\omega) = \begin{cases} 0 & \omega = H \\ 1 & \omega = T \end{cases}$

Then X_n does converge a.s. to X

Setup 3 Same as setup 1 but

$X_n(\omega) = \begin{cases} 0 & n \text{ even, } \omega = H \text{ or } n \text{ odd, } \omega = T \\ 1 & n \text{ even, } \omega = T \text{ or } n \text{ odd, } \omega = H \end{cases}$

Then X_n doesn't converge to anything!

Main tool for a.s.: Borel-Cantelli

Thm: If $\forall \varepsilon > 0$, $\sum_{n=1}^{\infty} \mathbb{P}[\,|X_n - X| > \varepsilon\,] < \infty$

then $X_n \xrightarrow{a.s.} X$

Ex: $X_n = \begin{cases} -1/n & \text{wp } 1/2 \\ 1/n & \text{wp } 1/2 \end{cases}$ $\quad X = 0$

Then $\forall \varepsilon > 0$, $\sum^{\infty} \mathbb{P}[\,|X_n - 0| > \varepsilon\,]$ is finite

since only finite # of nonzero terms,

hence $X_n \to X$ a.s.