Lecture 16

* HW2 due on Gradescope on Wednesday
  → If you're presenting in the 1st week, can get extension.
  → Please mark pages on Gradescope when you submit.

RECAP

GD:

1. Initialize w_1
2. For $t = 1, \dots, T$
3. $\quad w_{t+1} \leftarrow w_t - \eta \nabla f(w_t)$

Thm (convergence rate of GD) Let f be a convex, ρ-Lipschitz function i.e.

$$|f(x) - f(y)| \leq \rho \|x - y\|_2 \quad \forall \; x, y \in \mathbb{R}^d.$$

Let $w^* = \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} f(w)$. Let $\|w^* - w_1\| \leq B$, where w_1 is initialization for GD. Suppose we run GD for T steps with $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$ & let $\bar{w} = \frac{1}{T} \sum_{t=1}^{T} w_t$.

Then $\bar{w}$ satisfies

$$f(\bar{w}) - f(w^*) \leq \frac{B\rho}{\sqrt{T}}.$$

If $T \geq \frac{B^2 \rho^2}{\varepsilon^2}$, $f(\bar{w}) - f(w^*) \leq \varepsilon$.

Stochastic gradient descent (SGD)

1. Initialize w_1
2. For $t = 1, \ldots, T$
3. Choose any v_t s.t. $\mathbb{E}[v_t \mid w_t] = \nabla f(w_t)$.
4. $w_{t+1} \leftarrow w_t - \eta v_t$

Thm (convergence of SGD) Let f be a convex, ρ-Lipschitz function. Let $w^* = \operatorname{argmin}_{w \in \mathbb{R}^d} f(x)$ & $\|w^* - w_1\| \leq B$ (where w_1 is the initialization). Suppose we run SGD for T steps with step size $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$ & $\|v_t\| \leq \rho \; \forall t$. Let $\bar{w} = \frac{1}{T} \sum_{t=1}^{T} w_t$. Then,

$$\mathbb{E}[f(\bar{w})] - f(w^*) \leq \frac{B\rho}{\sqrt{T}}.$$

Proof Let $v_{1:t}$ to denote the sequence $v_1, \ldots, v_t$.

By convexity,

$$\mathbb{E}_{v_{1:T}}\left[f(\bar{w}) - f(w^*)\right] \leq \mathbb{E}_{v_{1:T}}\left[\frac{1}{T}\sum_{t=1}^{T} f(w_t) - f(w^*)\right]$$

Lemma (Iterative update) Let $v_1, \dots, v_T$ be an arbitary sequence of vectors. Consider any algorithm with an update rule

$$w_{t+1} = w_t - \eta v_t .$$

Then

$$\sum_{t=1}^{T} \langle w_t - w^*, v_t \rangle \le \frac{\|w^* - w_1\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|v_t\|^2$$

If $\|w^* - w_1\| \le B$ & $\|v_t\| \le \rho$ and we set $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$,

$$\frac{1}{T} \sum_{t=1}^{T} \langle w_t - w^*, v_t \rangle \le \frac{B\rho}{\sqrt{T}} .$$

By above Lemma,

$$\mathbb{E}_{v_{1:T}} \left[\frac{1}{T} \sum_{t=1}^{T} \langle w_t - w^*, v_t \rangle \right] \le \frac{B\rho}{\sqrt{T}}$$

∴ we only need to show

$$\mathbb{E}_{v_{1:T}} \left[\frac{1}{T} \sum_{t=1}^{T} f(w_t) - f(w^*) \right] \le \mathbb{E}_{v_{1:T}} \left[\frac{1}{T} \sum_{t=1}^{T} \langle w_t - w^*, v_t \rangle \right]$$

$$\mathbb{E}_{v_{1:T}} \left[\frac{1}{T} \sum_{t=1}^{T} f(w_t) - f(w^*) \right] = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{v_{1:T}} \left[f(w_t) - f(w^*) \right]$$

$$= \frac{1}{T} \sum_{t=1}^{T} \mathop{\mathbb{E}}_{v_{1:t-1}} \left[f(w_t) - f(w^*) \right]$$

SGD requires $\mathop{\mathbb{E}}_{v_t} \left[v_t \mid w_t \right] = \nabla f(w_t)$

Since w_t only depends on $v_{1:t-1}$

$$\therefore \quad \mathop{\mathbb{E}}_{v_t} \left[v_t \mid v_{1:t-1} \right] = \nabla f(w_t).$$

$$\mathop{\mathbb{E}}_{v_{1:T}} \left[\frac{1}{T} \sum_{t=1}^{T} f(w_t) - f(w^*) \right] \leq \frac{1}{T} \sum_{t=1}^{T} \mathop{\mathbb{E}}_{v_{1:t-1}} \left[\langle w_t - w^*, \mathop{\mathbb{E}}_{v_t} \left[v_t \mid v_{1:t-1} \right] \rangle \right]$$

$$= \frac{1}{T} \sum_{t=1}^{T} \mathop{\mathbb{E}}_{v_{1:t-1}} \left[\mathop{\mathbb{E}}_{v_t} \left[\langle w_t - w^*, v_t \rangle \mid v_{1:t-1} \right] \right]$$

$$= \frac{1}{T} \sum_{t=1}^{T} \mathop{\mathbb{E}}_{v_{1:t}} \left[\langle w_t - w^*, v_t \rangle \right]$$

(because $\mathop{\mathbb{E}}_{v_{1:t-1}} \left[\mathop{\mathbb{E}}_{v_t} \left[x \mid v_{1:t-1} \right] \right] = \mathop{\mathbb{E}}_{v_{1:t}} \left[x \right]$)

$$= \frac{1}{T} \sum_{t=1}^{T} \mathop{\mathbb{E}}_{v_{1:T}} \left[\langle w_t - w^*, v_t \rangle \right]$$

$$= \mathop{\mathbb{E}}_{v_{1:T}} \left[\frac{1}{T} \sum_{t=1}^{T} \langle w_t - w^*, v_t \rangle \right]$$

and we are done. ∎

Some more points about the convergence rate,

→ Same convergence rate as GD!

→ the convergence rate does not improve with the β-smoothness assumption.

→ for strongly convex functions, the convergence rate goes from $\frac{1}{\sqrt{T}} \rightarrow \frac{1}{T}$

we can't get the e^{-Tk} rates.

So some tradeoffs in terms of cost per iteration & # iterations.

Learning with SGD

Suppose we are interested in minimizing the risk

$$R(w) = \mathop{\mathbb{E}}_{z \sim D} [\ell(w, z)]$$

Using ERM, we sample n datapoints & minimize training error. Use generalization bounds to ensure small test error.

SGD gives a different way to look at this,

$$\nabla R(w_t) = \nabla \mathop{\mathbb{E}}_{z \sim D} [\ell(w_t, z)] = \mathop{\mathbb{E}}_{z \sim D} [\nabla \ell(w_t, z)]$$

just the gradient at z!

$\therefore$ If we set $v_t = \nabla \ell(w_t, z)$ (for $z \sim D$)

$$\mathbb{E}[v_t] = \nabla R(w_t)$$

SGD for minimizing $R(w)$

1. initialize w_1
2. for $t = 1 \dots T$
3. $z \leftarrow$ Fresh(c, D)
4. get $v_t = \nabla \ell(w_t, z)$
5. update $w_{t+1} = w_t - \eta v_t$
6. Output $\bar{w} = \frac{1}{T} \sum_{t=1}^{T} w_t$

Corollary Consider a convex, ρ-Lipschitz function $\ell(w, z)$. Let $w^* = \operatorname{argmin}_{w \in \mathcal{H}} R(w)$. Let $\|w^* - w_1\| \leq B$.

Then if we run SGD for T iterations with $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$ where $T \geq \frac{B^2 \rho^2}{\epsilon^2}$, then the output $\bar{w}$ satisfies,

$$\mathbb{E}[R(\bar{w})] \leq \min_{w \in \mathcal{H}} R(w) + \epsilon.$$

Note SGD is implementable in the SQ model (we need to extend SQ model to allow real-valued queries).

Online Learning

PAC learning / Statistical learning :

Learn under probabilistic assumptions on data
(train/test from some distribution)
Generalization theory to do well on unseen test points.

Online learning

No probabilistic assumptions, predict well on a datapoint as we see it.

Example : weather forecasting

→ We're interested in predicting rain/no rain.
→ Every night, make prediction about next day.
→ Next day, we see whether or not it rained.

There's no train/test split. Every example is both training & test example.

Formally,

At every time step t

→ learner receives an input $x_t \in X$.
→ makes prediction $p_t \in Y$.
→ see true label $y_t \in Y$. suffer loss $\ell(p_t, y_t)$.

Think $Y = \{0, 1\}$

$\ell(p_t, y_t) = \mathbb{1}\{p_t \neq y_t\}$.

Realizability

Def (Mistake bound model) Let $\mathcal{H}$ be a hypothesis class and A be a online learning algo. Given any sequence $S = (x_1, h^*(x_1)), \ldots, (x_T, h^*(x_T))$ of T labelled data points where $h^* \in \mathcal{H}$, let $M_A(S)$ be the # mistakes A makes on the sequence S. We denote by $M_A(\mathcal{H})$ to the supremum of $M_A(S)$ over all possible S.

If there exists an algorithm A that satisfies a mistake bound of the form $M_A(\mathcal{H}) \leq B < \infty$, we say $\mathcal{H}$ is online learnable in the mistake bound model.

Note : B should be independent of length of sequence T.

$\therefore$ as $T \to \infty$, avg # mistakes $\left(\leq \frac{B}{T}\right) \to 0$.

In PAC learning, ERM choose any consistent hypothesis over training set.

Alg: Consistent

① Initialize $V_1 = \mathcal{H}$

② for $t = 1, \dots, T$

receive x_t

Choose any $h \in V_t$

predict $p_t = h(x_t)$

receive $y_t = h^*(x_t)$, loss $\mathbb{1}(p_t \neq y_t)$

update $V_{t+1} = \{h \in V_t : h(x_t) = y_t\}$.

Proposition: Let $\mathcal{H}$ be a finite hypothesis class. The above algorithm gets a mistake bound

$$M_{\text{consistent}}(\mathcal{H}) \leq |\mathcal{H}| - 1.$$

Alg: Halving

① Initialize $V_1 = \mathcal{H}$

② for $t = 1, \dots, T$

receive x_t

predict $p_t = \underset{r \in \{0,1\}}{\operatorname{argmax}} \left| \{h \in V_t : h(x_t) = r\} \right|$

(if tie, $p_t = 1$)

receive $y_t = h^*(x_t)$, loss $\mathbb{1}(p_t \neq y_t)$

update $V_{t+1} = \{h \in V_t : h(x_t) = y_t\}$.

Proposition Let $\mathcal{H}$ be a finite hypothesis class. Then halving algorithm satisfies the mistake bound $M_{Halving}(\mathcal{H}) \leq \log_2(|\mathcal{H}|)$.

Proof Whenever the algo. errors, $|V_{t+1}| \leq \frac{|V_t|}{2}$.

If M is # mistakes

$$|V_{T+1}| \leq |\mathcal{H}| 2^{-M}$$

as $|V_{T+1}| \geq 1$

$$\Rightarrow \quad M \leq \log_2(|\mathcal{H}|).$$

■