B9145: Reliable Statistical Learning **Spring 2023**

Lecture 2: Stochastic Gradient Descent

Lecturer: Hongseok Namkoong *Scribe: Tianyu Wang*

2.1 Proof of Dudley's entropy integral bound

Theorem 1 (Dudley's entropy integral)**.** *Let $\{V_h : h \in \mathcal{T}\}$ be a sub-Gaussian process w.r.t. d on $\mathcal{T}$. For any $\delta \in [0, D]$,*

$$\mathbb{E} \sup_{h \in \mathcal{T}} V_h \le \mathbb{E}\left[\sup_{h,h' \in \mathcal{T}} V_h - V_{h'}\right] \le 2\mathbb{E}\left[\sup_{d(\gamma,\gamma') \le \delta, \gamma, \gamma' \in \mathcal{T}} (V_\gamma - V_{\gamma'})\right] + 32 \int_{\delta/4}^{D} \sqrt{\log N(\mathcal{T}, d, \varepsilon)} d\varepsilon$$

Remark 1. *Setting $\delta = 0$ gives $\mathbb{E} \sup_{h \in \mathcal{T}} V_h \le 32 \int_0^\infty \sqrt{\log N(\mathcal{T}, d, \varepsilon)} d\varepsilon$. ($N(\mathcal{T}, d, \delta) = 1$ for any $\delta \ge D$)*

Proof. We begin with the inequality established before:

$$\sup_{h,h' \in \mathcal{T}} (V_h - V_{h'}) \le 2 \sup_{d(\gamma,\gamma') \le \delta, \gamma, \gamma' \in \mathcal{T}, d(\gamma,\gamma') \le \delta} (V_\gamma - V_{\gamma'}) + 2 \max_{1 \le j \le N} |V_{h_j} - V_{h_1}|. \tag{2.1}$$

Instead of bounding the last term via the max lemma, we use a chaining argument.

Recall that $U := \{h_j\}_{j=1}^N$ is a δ-cover of $\mathcal{T}$. For each $m = 1, 2, \ldots, L$, define $U_m :=$ minimal $(D2^{-m})$-cover of U_{m-1}, where we allow for any element of $\mathcal{T}$ to be used in forming the cover.

Since U is finite, for $L = \lceil \log_2(D/\delta) \rceil$ such that $2^{-L} \le \frac{\delta}{D}$. We can set $U_L = U$. By definition, $|U_m| \le N(\mathcal{T}, d, D2^{-m})$. For each m, we define $\pi_m : U \to U_m$ such that $\pi_m(h) = \text{argmin}_{\tilde{h} \in U_m} d(h, \tilde{h})$. Using this, we can construct a chaining process for any $h \in U$, where we define $\gamma_L = h$, $\gamma_{m-1} = \pi_{m-1}(\gamma_m)$ recursively for $m = L, L-1, \ldots, 2$.

By construction, we have the *chaining relation*:

$$V_h - V_{\gamma_1} = \sum_{m=2}^{L} (V_{\gamma_m} - V_{\gamma_{m-1}}),$$

and therefore, $|V_h - V_{\gamma_1}| \le \sum_{m=2}^L \sup_{\gamma \in U_m} |V_\gamma - V_{\pi_{m-1}(\gamma)}|$. See for an illustration of this setup in Figure 2.1. Similarly, for any other $h' \in \mathcal{T}$, we have the same bound with γ'_m. Therefore, we arrive at:

$$\begin{aligned}|V_h - V_{h'}| &= |V_{\gamma_1} - V_{\gamma'_1} + V_h - V_{\gamma_1} + V_{\gamma'_1} - V_{h'}| \\ &\le |V_{\gamma_1} - V_{\gamma'_1}| + |V_h - V_{\gamma_1}| + |V_{\gamma'_1} - V_{h'}| \\ &\le \max_{\gamma_1, \gamma'_1 \in U_1} |V_{\gamma_1} - V_{\gamma'_1}| + 2 \sum_{m=2}^{L} \sup_{\gamma \in U_m} |V_\gamma - V_{\pi_{m-1}(\gamma)}|,\end{aligned}$$

where we apply the chaining technique for the second and third term in the second inequality. From previous lemma, we know

$$\mathbb{E}\left[\max_{\gamma_1, \gamma'_1 \in U_1} |V_{\gamma_1} - V_{\gamma'_1}|\right] \le 2D \sqrt{\log N\left(\mathcal{T}, d, \frac{D}{2}\right)}.$$

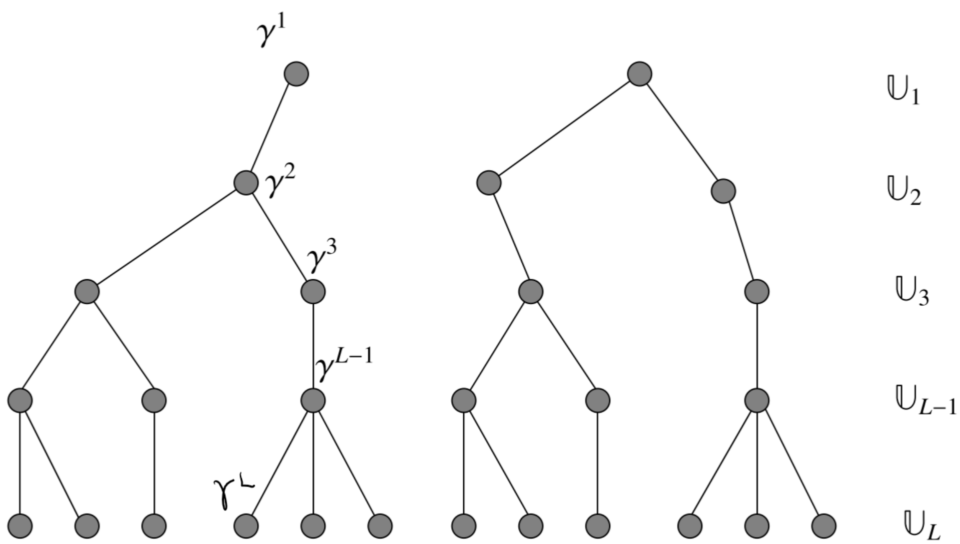

Figure 2.1: Illustration of the chaining relationship (extracted from Figure 5.3 in Wainwright (2019))

And since $\max_{\gamma \in U_m} d(\gamma, \pi_{m-1}(\gamma)) \leq D2^{-(m-1)}$ and $|U_m| \leq N(\mathcal{T}, d, D2^{-m})$, we have:

$$\mathbb{E}\left[\max_{h,h' \in U} |V_h - V_{h'}|\right] \leq 2D2^{-(m-1)}\sqrt{\log N(\mathcal{T}, d, D2^{-m})}.$$

Combining the pieces, we conclude that: $\mathbb{E}\left[\sup_{h,h' \in U} |V_h - V_{h'}|\right] \leq 4\sum_{m=1}^{L} D2^{-(m-1)}\sqrt{\log N\left\{\mathcal{T}, d, D2^{-m}\right\}}$. Since the metric entropy $\log N(\mathcal{T}, d, \delta)$ is decreasing in δ, we have:

$$D2^{-m}\sqrt{\log N(\mathcal{T}, d, D2^{-m}} \leq 2\int_{D2^{-(m+1)}}^{D2^{-m}} \sqrt{\log N(\mathcal{T}, d, \varepsilon)} d\varepsilon.$$

Therefore, $2\mathbb{E}\left[\sup_{h,h' \in U} |V_h - V_{h'}|\right] \leq 32\int_{\delta/4}^{D} \sqrt{\log N(\mathcal{T}, d, \varepsilon)} d\varepsilon$. Combining with Equation (2.1), we get the result. $\square$

2.2 Stochastic Gradient Descent (SGD)

Definition 1. *A function $R : \mathbb{R}^d \to \mathbb{R}$ is convex if $\forall \theta, \theta' \in \mathbb{R}^d$,*

$$R(t\theta + (1-t)\theta') \leq tR(\theta) + (1-t)R(\theta'), \forall t \in [0, 1].$$

Lemma 1. *Let the function $R : \mathbb{R}^d \to \mathbb{R}$ be differentiable on the interior of its domain. Then R is convex iff $R(\theta') \geq R(\theta) + \nabla R(\theta)^\top(\theta' - \theta), \forall \theta, \theta' \in \mathbb{R}^d$.*

This result shows that in convex functions, first order approximation is a global minimization.

Proof. "If" part: $\forall \theta, \theta' \in \mathbb{R}^d$, define $\theta_t = t\theta + (1-t)\theta'$. Combining:

$$R(\theta) \geq R(\theta_t) + \nabla R(\theta_t)^\top(\theta - \theta_t),$$
$$R(\theta') \geq R(\theta_t) + \nabla R(\theta_t)^\top(\theta' - \theta_t),$$

we have: $tR(\theta) + (1-t)R(\theta') \geq R(\theta_t) + \nabla R(\theta_t)^\top (t\theta + (1-t)\theta' - \theta_t) = R(\theta_t), \forall t \in [0,1]$.

"Only if" part: From the definition of convexity, we have:

$$R(\theta + t(\theta')) \leq R(\theta) + t(R(\theta') - R(\theta))$$

which is equivalent to saying:

$$R(\theta') - R(\theta) \geq \frac{1}{t}(R(\theta + t(\theta' - \theta)) - R(\theta)), \forall t \in (0,1]$$

Then letting $t \to 0$ in the right hand side above yields $\nabla R(\theta)^\top (\theta' - \theta)$. □

This result shows that in convex functions, first order approximation is a global minimization.

First, we consider $\min_{\theta \in \Theta} R(\theta)$ for $R : \mathbb{R}^d \to \mathbb{R}$ differentiable and convex.

Lemma 2 (Optimality Condition)**.** $\theta^* = \operatorname{argmin}_{\theta \in \Theta} R(\theta)$ *iff* $\nabla R(\theta^*)^\top (\theta - \theta^*) \geq 0, \forall \theta \in \Theta$.

Proof. "If" part: From Lemma 1, $R(\theta) - R(\theta^*) \geq \nabla R(\theta^*)^\top (\theta - \theta^*) \geq 0, \forall \theta \in \Theta$.

"Only if" part: $\nabla R(\theta^*)^\top (\theta - \theta^*) = \lim_{t \to 0} \frac{1}{t}(R(\theta^* + t(\theta - \theta^*)) - R(\theta^*)) \geq 0, \forall \theta \in \Theta$. □

Corollary 1. *Let* Θ *be a closed convex set in* $\mathbb{R}^d$*. Define the projection operator* $\Pi_\Theta(\theta) = \operatorname{argmin}_{\theta' \in \Theta} \|\theta - \theta'\|_2$*. Then* $\|\Pi_\Theta(\theta) - \theta'\|_2 \leq \|\theta - \theta'\|_2, \forall \theta' \in \Theta, \forall \theta \in \mathbb{R}^d$.

Proof. We apply Lemma 2 to $R(\theta') := \|\theta - \theta'\|_2$. Then $\forall \theta \in \Theta$, we have:

$$\begin{aligned}
0 &\leq (\Pi_\Theta(\theta) - \theta)^\top (\theta' - \Pi_\Theta(\theta)) \\
&= (\Pi_\Theta(\theta) - \theta' + \theta' - \theta)^\top (\theta' - \Pi_\Theta(\theta)) \\
&= -\|\theta' - \Pi_\Theta(\theta)\|_2^2 + (\theta' - \theta)^\top (\theta' - \Pi_\Theta(\theta)) \\
&\leq -\|\theta' - \Pi_\Theta(\theta)\|_2^2 + \|\theta' - \theta\|_2 \|\theta' - \Pi_\Theta(\theta))\|_2,
\end{aligned}$$

where the last inequality follows by Cauchy-Schwarz inequality. □

Definition 2 (Stochastic Gradient)**.** *A stochastic gradient* $G(\theta)$ *is a random variable s.t.* $\mathbb{E}[G(\theta)] = \nabla R(\theta)$.

We study the first-order optimization method based on the stochastic gradient, where the canonical problem is:

$$\min_{\theta \in \Theta} \{\mathbb{E}\ell(\theta; Z) =: R(\theta)\}.$$

The idea of SGD is to go in the direction of the stochastic gradient, then project to Θ.

The **algorithm** is: let $G_k(\theta)$ be a stochastic gradient of $R(\theta)$. At each iteration k, we set:

$$\theta_{k+1} = \Pi_\Theta(\theta_k - \alpha_k G_k(\theta_k)), \text{ for some stepsize } \alpha_k > 0.$$

Note that we are completely assuming that projections are efficient to compute.

We would like to study the convergence of SGD. Assume $\theta^* \in \operatorname{argmin}_{\theta \in \Theta} R(\theta) > -\infty$ exists.

Theorem 2. *Let Θ be compact. Assume $\exists D > 0$ s.t. $\sup_{\theta\in\Theta} \|\theta - \theta^*\|_2 \leq D$. $\exists M > 0$, s.t. $\mathbb{E}\|G(\theta)\|_2^2 \leq M^2, \forall\theta \in \Theta$.*

Let α_k be the sequence of (decreasing and positive) step sizes, and $\bar{\theta}_K = \frac{1}{K}\sum_{k=1}^K \theta_k$. Then:

$$\mathbb{E}[R(\bar{\theta}_K) - R(\theta^*)] \leq \frac{D^2}{2K\alpha_K} + \frac{M^2}{2K}\sum_{k=1}^K \alpha_k.$$

Proof. We expand on the error $\|\theta_{k+1} - \theta^*\|_2^2$.

$$\begin{aligned}
\frac{1}{2}\|\theta_{k+1} - \theta^*\|_2^2 &= \frac{1}{2}\|\Pi_\Theta(\theta_k - \alpha_k G(\theta_k)) - \theta^*\|_2^2 \\
&\leq \frac{1}{2}\|\theta_k - \alpha_k G(\theta_k) - \theta^*\|_2^2 \\
&= \frac{1}{2}\|\theta_k - \theta^*\|_2^2 - \alpha_k\langle G(\theta_k), \theta_k - \theta^*\rangle + \frac{\alpha_k^2}{2}\|G(\theta_k)\|_2^2 \\
&= \frac{1}{2}\|\theta_k - \theta^*\|_2^2 - \alpha_k\langle \nabla R(\theta_k), \theta_k - \theta^*\rangle + \frac{\alpha_k^2}{2}\|G(\theta_k)\|_2^2 - \alpha_k\langle G(\theta_k) - \nabla R(\theta_k), \theta_k - \theta^*\rangle \\
&\leq \frac{1}{2}\|\theta_k - \theta^*\|_2^2 - \alpha_k(R(\theta_k) - R(\theta^*)) + \frac{\alpha_k^2}{2}\|G(\theta_k)\|_2^2 - \alpha_k\langle G(\theta_k) - \nabla R(\theta_k), \theta_k - \theta^*\rangle,
\end{aligned}$$

where the first inequality follows by the non-expansiveness of Π_Θ in Corollary 1, and the second inequality follows by convexity of $R(\cdot)$.

Then we divide each side by α_k and rearrange:

$$R(\theta_k) - R(\theta^*) \leq \frac{1}{2\alpha_k}(\|\theta_k - \theta^*\|_2^2 - \|\theta_{k+1} - \theta^*\|_2^2) + \frac{\alpha_k}{2}\|G(\theta_k)\|_2^2 - \langle G(\theta_k) - \nabla R(\theta_k), \theta_k - \theta^*\rangle. \tag{2.2}$$

Now, note that:

$$\begin{aligned}
\sum_{k=1}^K \frac{1}{2\alpha_k}(\|\theta_k - \theta^*\|_2^2 - \|\theta_{k+1} - \theta^*\|_2^2) &= \frac{1}{2\alpha_1}\|\theta_1 - \theta^*\|_2^2 - \frac{1}{2\alpha_K}\|\theta_K - \theta^*\|_2^2 + \sum_{k=2}^K(\frac{1}{2\alpha_k} - \frac{1}{2\alpha_{k-1}})\|\theta_k - \theta^*\|_2 \\
&\leq \frac{D^2}{2\alpha_1} + \frac{D^2}{2}\sum_{k=2}^K(\frac{1}{\alpha_k} - \frac{1}{\alpha_{k-1}}) = \frac{D^2}{2\alpha_K}.
\end{aligned}$$

So summing both sides of Equation (2.2) and taking expectation, we have:

$$\mathbb{E}[\sum_{k=1}^K R(\theta_k) - R(\theta^*)] \leq \frac{D^2}{2\alpha_K} + \frac{M}{2}\sum_{k=1}^K \alpha_k - \sum_{k=1}^K \mathbb{E}\langle G(\theta_k - \nabla R(\theta_k)), \theta_k - \theta^*\rangle$$

And notice:

$$\begin{aligned}
\mathbb{E}[\langle G(\theta_k - \nabla R(\theta_k)), \theta_k - \theta^*\rangle] &= \mathbb{E}[\mathbb{E}[\langle G(\theta_k - \nabla R(\theta_k)), \theta_k - \theta^*\rangle | \theta_k]] \\
&= \mathbb{E}[\langle \mathbb{E}[G(\theta_k)|\theta_k] - \nabla R(\theta_k), \theta_k - \theta^*\rangle] = 0.
\end{aligned}$$

Then we get $\mathbb{E}[\sum_{k=1}^K R(\theta_k) - R(\theta^*)] \leq \frac{D^2}{2\alpha_K} + \frac{M^2}{2}\alpha_k$. And notice $R(\bar{\theta}_K) \leq \frac{1}{K}\sum_{k=1}^K R(\theta_k)$, we would get the result. □

Corollary 2. *If we choose the step size $\alpha_k = \frac{D}{M\sqrt{k}}$, then $\mathbb{E}R(\bar{\theta}_K) - R(\theta^*) \leq \frac{3DM}{2\sqrt{K}}$.*

Proof. Noticing $\sum_{k=1}^{K} \frac{1}{\sqrt{k}} \leq \int_0^K \frac{1}{\sqrt{t}} dt = 2\sqrt{K}$, therefore applying the result in Theorem 2, we have:

$$\mathbb{E}R(\bar{\theta}_K) - R(\theta^*) \leq \frac{3DM}{2\sqrt{K}} \leq \frac{DM}{2\sqrt{K}} + \frac{DM}{\sqrt{K}}.$$

□

Remark 2. *We can think of K as the number of access to the gradient oracle. If $G(\theta) = \nabla_\theta \ell(\theta; \ell(\theta; Z_i))$, then K is the number of samples.*

Remark 3. *Often, we iterate through data C times. This gives gains on the empirical loss. But the population loss-wise, theory doesn't give gains as C grows. In fact, we cannot do better and we will show this through information theoretical minimax bound next class.*

We refer the detailed notes of minimax analysis of stochastic optimization to Chapter 5 in Duchi (2018).

References

John C Duchi. Introductory lectures on stochastic optimization. ***In The Mathematics of Data, IAS/Park City Mathematics Series. American Mathematical Society***, 25:99–186, 2018.

Martin J Wainwright. ***High-dimensional statistics: A non-asymptotic viewpoint***, volume 48. Cambridge university press, 2019.