Machine Learning Theory (CSC 482A/581A) - Lecture 35

Nishant Mehta

1 Stochastic convex optimization

A stochastic convex optimization problem is specified by a probability distribution P over a set $\mathcal{Z}$, a convex set $\mathcal{W}$, and a function $f\colon \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$ that is convex in its first argument. The goal is to find some $w \in \mathcal{W}$ which minimizes the objective

$$F(w) = \mathsf{E}_{Z \sim P}\left[f(w, Z)\right].$$

We will use $w^* \in \mathcal{W}$ to denote an arbitrary minimizer of F, so that $F(w^*) = \min_{w \in \mathcal{W}} F(w)$.

Supervised learning with linear predictors can be recovered as follows:

- take $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, so that $Z = (X, Y)$;
- define $f(w, z) = f(w, (x, y)) = \ell(\langle w, x\rangle, y)$ for some loss function $\ell\colon \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ that is convex in its first argument.

In order to approximately minimize the objective $F(w)$, a learning algorithm will be presented with i.i.d. samples $Z_1, \ldots, Z_T$ distributed according to P, similar to the statistical learning setting.

We will study algorithms for solving the stochastic optimization problem based on online convex optimization followed by a technique known as an *online-to-batch conversion*. The idea will be to first frame an online version of the above problem as an online convex optimization problem, to use online gradient descent to obtain low regret for this problem, and finally to obtain a single recommended prediction $\hat{w}$ whose excess risk $F(w^*) - F(w)$ is bounded by the regret (averaged over rounds) of online gradient descent.

First, observe that for each $t \in [T]$, we may define the cost function $c_t(w) = f(w, Z_t)$. We may thus use the online gradient descent algorithm to obtain low regret, i.e., to ensure that

$$\sum_{t=1}^{T} f(w_t, Z_t) - \inf_{w \in \mathcal{W}} \sum_{t=1}^{T} f(w, Z_t)$$

is not too large.

Stochastic gradients and connection to gradient descent. Note that the cost functions themselves are thus i.i.d., and this implies that for any fixed $w \in \mathcal{W}$, the gradient $\nabla c_t(w) = \nabla f(w, Z_t)$ is stochastic and satisfies

$$\mathsf{E}[\nabla c_t(w)] = \mathsf{E}[\nabla f(w, Z_t)] = \nabla\, \mathsf{E}[f(w, Z_t)] = \nabla F(w).$$

Thus, for fixed a fixed action w, the stochastic gradient $\nabla c_t(w)$ is an unbiased estimator of the gradient $\nabla F(w)$, and taking a step of the right size in the direction of the negative gradient should in expectation move us towards the optimum w^*.

However, online gradient descent uses gradients evaluated at the played action w_t, which can depend on $Z_1^{t-1} = (Z_1, \ldots, Z_{t-1})$ and hence which itself is stochastic. Conditioning on Z_1^{t-1}, w_t becomes fixed and hence

$$\begin{aligned}
\mathsf{E}[\nabla c_t(w_t) \mid Z_1^{t-1}] &= \mathsf{E}[\nabla f(w_t, Z_t) \mid Z_1^{t-1}] \\
&= \nabla \mathsf{E}[f(w_t, Z_t) \mid Z_1^{t-1}] \\
&= \nabla \mathsf{E}[F(w_t) \mid Z_1^{t-1}] \\
&= \nabla F(w_t).
\end{aligned}$$

Thus, conditional on the past, the stochastic gradient $\nabla c_t(w_t)$ is an unbiased estimator of the gradient $\nabla F(w_t)$. Intuitively, stochastic gradient descent (online gradient descent with stochastic gradients as above) should make progress towards minimizing F.

2 Online-to-batch conversion

Suppose that an online learning algorithm that plays $w_1, \ldots, w_T$ against the sequence $Z_1, \ldots, Z_T$ obtains regret R_T.[1] We will prove that the simple average $\bar{w}_T := \frac{1}{T}\sum_{t=1}^T$ obtains low excess risk whenever R_T is small.

We will derive an in-expectation bound using elementary arguments and then a high probability bound using a more sophisticated martingale-based argument. For both parts, our starting point will be the simple inequality

$$F(\bar{w}_T) \le \frac{1}{T}\sum_{t=1}^T F(w_t),$$

which holds from the convexity of f in its first argument, since

$$F(\bar{w}_T) = \mathsf{E}_{Z\sim P}\left[f\left(\frac{1}{T}\sum_{t=1}^T w_t, Z\right)\right] \le \mathsf{E}_{Z\sim P}\left[\frac{1}{T}\sum_{t=1}^T f\left(w_t, Z\right)\right] = \frac{1}{T}\sum_{t=1}^T F(w_t).$$

2.1 Warm-up: In-expectation bound

Observe that

$$\begin{aligned}
\mathsf{E}\left[\frac{1}{T}\sum_{t=1}^T F(w_t)\right] &= \frac{1}{T}\sum_{t=1}^T \mathsf{E}\left[f(w_t, Z)\right] \\
&= \frac{1}{T}\sum_{t=1}^T \mathsf{E}\left[f(w_t, Z_t)\right] && (Z \text{ and } Z_t \text{ are i.i.d. and } w_t \text{ depends only on } Z_1^{t-1}) \\
&= \mathsf{E}\left[\frac{1}{T}\sum_{t=1}^T f(w_t, Z_t)\right] \\
&= \mathsf{E}\left[\inf_{w\in\mathcal{W}} \frac{1}{T}\sum_{t=1}^T f(w, Z_t)\right] + \frac{1}{T}\mathsf{E}\left[R_T\right] \\
&\le \inf_{w\in\mathcal{W}} \mathsf{E}\left[\frac{1}{T}\sum_{t=1}^T f(w, Z_t)\right] + \frac{1}{T}\mathsf{E}\left[R_T\right] \\
&= F(w^*) + \frac{1}{T}\mathsf{E}\left[R_T\right].
\end{aligned}$$

[1] Note that R_T is a random variable by way of its dependence on $Z_1, \ldots, Z_T$.

Thus, an upper bound on the expected regret implies an upper bound on the expected excess risk of $\bar{w}_T$.

With a little more work, we can establish a similar guarantee that holds with high probability with respect to $Z_1, \ldots, Z_T$.

2.2 High probability bound

In order to obtain a high probability bound, we will develop some machinery to analyze stochastic processes. The development here will be somewhat informal to keep things accessible. Let $X_1, X_2, \ldots, X_T$ be a stochastic process, and let each X_t be measurable with respect to a history H_t. Informally, this means that X_t is deterministic given H_t. We say that the process $X_1, \ldots, X_T$ is a *martingale difference sequence* if, for all $t \in [T]$, both of the following hold:

- $\mathsf{E}[|X_t|] < \infty$;
- $\mathsf{E}\left[X_t \mid H_{t-1}\right] = 0$.

The next concentration inequality is known as Hoeffding-Azuma's inequality, also commonly referred to as Azuma's inequality.

Theorem 1. *Let $X_1, X_2, \ldots, X_T$ be a martingale difference sequence such that, for all $t \in [T]$, $|X_t| \leq B$ with probability one. Then for all $\varepsilon > 0$,*

$$\Pr\left(\sum_{t=1}^{T} X_t \geq \varepsilon\right) \leq \exp\left(-\frac{\varepsilon^2}{2B^2T}\right).$$

We now begin proving a high probability excess risk bound for $\bar{w}_T$. Define for each $t \in [T]$,

$$\begin{aligned} X_t &= f(w^*, Z_t) - f(w_t, Z_t) - \mathsf{E}\left[f(w^*, Z_t) - f(w_t, Z_t) \mid Z_1^{t-1}\right] \\ &= f(w^*, Z_t) - f(w_t, Z_t) - (F(w^*) - F(w_t)) \end{aligned}$$

As we will see, $X_1, \ldots, X_T$ is a martingale difference sequence and will play a key role in our analysis.

Observe that

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^{T} F(w_t) &= F(w^*) + \frac{1}{T}\sum_{t=1}^{T}\left(f(w_t, Z_t) - f(w^*, Z_t)\right) + \frac{1}{T}\sum_{t=1}^{T} X_t \\ &\leq F(w^*) + \frac{1}{T}\sum_{t=1}^{T} f(w_t, Z_t) - \inf_{w \in \mathcal{W}} \frac{1}{T}\sum_{t=1}^{T} f(w, Z_t) + \frac{1}{T}\sum_{t=1}^{T} X_t \\ &= F(w^*) + \frac{R_T}{T} + \frac{1}{T}\sum_{t=1}^{T} X_t \end{aligned}$$

Now, for each $t \in [T]$ it holds that $\mathsf{E}\left[X_t \mid Z_1^{t-1}\right] = 0$. Moreover, if we assume that $|f(w, Z)| \leq B$ for all $w \in \mathcal{W}$ and $Z \in \mathcal{Z}$, then it holds that $|X_t| \leq 2B$. Therefore, under this boundedness assumption, $X_1, \ldots, X_T$ is a martingale difference sequence.

Applying Theorem 1, we see that with probability at least $1 - \delta$,

$$\frac{1}{T}\sum_{t=1}^{T} F(w_t) \leq F(w^*) + \frac{R_T}{T} + 2B\sqrt{\frac{2\log\frac{1}{\delta}}{T}}.$$