Intro to generalization; Chernoff method

Generalization?

Let's summarize where we are in the course. To put everything together, let's define a few quantities.

$$\begin{aligned}
\mathcal{R}(f) &:= \mathbb{E}(\ell(f(X), Y)), \\
\widehat{\mathcal{R}}(f) &:= \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i)), \\
\bar{f} &:= \text{"argmin}_{f\in\mathcal{F}}\text{"}\mathcal{R}(f), \\
\bar{f}_n &:= \text{"argmin}_{f\in\mathcal{F}}\text{"}\widehat{\mathcal{R}}(f), \\
\hat{f} &\quad \text{output of optimization algorithm}, \\
\bar{g} &:= \text{"argmin}_{g \text{ measurable}}\text{"}\mathcal{R}(g).
\end{aligned}$$

(As usual, argmin has technical issues we are avoiding, hence the quotes.)

The goal in a machine learning problem is to make an algorithm that outputs $\hat{f}$ so that $\mathcal{R}(\hat{f}) - \mathcal{R}(\bar{g})$ is small. We can decompose this error into the following pieces:

$$\begin{aligned}
\mathcal{R}(\hat{f}) - \mathcal{R}(\bar{g}) = \mathcal{R}(\hat{f}) - \widehat{\mathcal{R}}(\hat{f}) \qquad & (\triangle) \\
+ \widehat{\mathcal{R}}(\hat{f}) - \widehat{\mathcal{R}}(\bar{f}_n) \quad & (\square) \\
+ \widehat{\mathcal{R}}(\bar{f}_n) - \widehat{\mathcal{R}}(\bar{f}) \quad & (\spadesuit) \\
+ \widehat{\mathcal{R}}(\bar{f}) - \mathcal{R}(\bar{f}) \quad & (\triangle) \\
+ \widehat{\mathcal{R}}(\bar{f}) - \mathcal{R}(\bar{f}) \quad & (\diamond).
\end{aligned}$$

These terms can be controlled as follows. (This course was designed around this decomposition!)

- $(\triangle)$. Third part of this course: generalization/statistics.
- $(\square)$. Second part of this course: optimization.
- $(\spadesuit)$. This one is ≤ 0 by choice of $\bar{f}_n$.
- $(\diamond)$. First part of this course: representation/approximation.

So the rest of this course is part 3, the statistical problem above (comparing $\widehat{\mathcal{R}}$ and $\mathcal{R}$), and then we'll leave a few lectures for some advanced/miscellaneous topics.

Chernoff's bounding method

We will work towards bounds of roughly the following form: with probability at least $1-\delta$ over the draw of n i.i.d. samples $(X_1, \ldots, X_n)$ with $\mathbb{E}(X_1) = 0$,

$$\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}(X_1) \leq \mathcal{O}\left(\text{scaling} \cdot \sqrt{\frac{\ln(1/\delta)}{n}}\right).$$

The key thing to note is that the confidence parameter δ appears as $\ln(1/n)$; said another way, the number of samples scales *linearly* in the number of bits of desired precision. For some of our uses, this type of scaling will be essential.

Remark. All of our bounds will be of the preceding form: an upper bound on $n^{-1}X_i$, where $\mathbb{E}(X_1) = 0$. To convert this into a lower bound, we can swap X_i with $-X_i$, and to work with random variables which are not zero mean, we can use $X_i - \mathbb{E}(X_i)$. There are, however, relevant bounds (which we will *not* cover) where the right hand side depends on whether it is a lower or upper bound; see for instance "multiplicative Chernoff bounds". *[future matus: cite please.]* (Another choice, followed by some authors (for instance Maxim Raginsky's book), is to work with $|X_i - \mathbb{E}(X_i)|$ and union bound over both directions of the inequality, gettin $\ln(2/\delta)$ rather than $\ln(1/\delta)$.)

The first step along this path is the Markov inequality.

Theorem (Markov's inequality). For any random variable X and any $a > 0$,

$$\Pr[|X| \geq a] \leq \frac{\mathbb{E}|X|}{a}.$$

Proof. It suffices to note $\mathbf{1}[|X| \geq a] \leq |X|/a$ and apply $\mathbb{E}(\cdot)$ to both sides. □

Corollary. For any nondecreasing $f : \mathbb{R} \to \mathbb{R}_+$ and any $a \in \mathbb{R}$ with $f(a) > 0$,

$$\Pr[X \geq a] \leq \frac{\mathbb{E}(f(X))}{f(a)}.$$

Proof. Since f is nondecreasing, $\Pr[X \geq a] \leq \Pr[f(X) \geq f(a)]$, and the bound follows by Markov's inequality. (If f were strictly increasing then the first inequality would be an equality.) □

Example.

- Let's first try out $f(x) := |x|^p$. For any $\epsilon > 0$, this gives

$$\Pr[X \geq \epsilon] \leq \frac{\mathbb{E}|X|^p}{\epsilon^p}.$$

- Now let's apply that to the Goal we motivated this section with. Namely, let's consider an i.i.d. draw of $(X_1, \ldots, X_n)$ with $\mathbb{E}(X_1) = 0$, and consider the random variable $n^{-1}\sum_i X_i$; by the preceding bound, we get

$$\Pr[n^{-1}\sum_i X_i \geq \epsilon] \leq \frac{\mathbb{E}(|n^{-1}\sum_i X_i|^p}{\epsilon^p} = \frac{\mathbb{E}(|\sum_i X_i|^p}{n^p\epsilon^p}.$$

  For instance, setting $M := \max_{2 \leq j \leq p} \mathbb{E}(|X_1|^j)$, then $\mathbb{E}(|\sum_i X_i|^p) = \mathcal{O}((Mpn)^{p/2})$, which gives

$$\Pr[n^{-1}\sum_i X_i \geq \epsilon] \leq \frac{\mathcal{O}(1)}{\epsilon^p n^{p/2}}$$

  (where this last $\mathcal{O}$ has treated M, p as constants). To convertthis into the type of bound we motivated this section with, we set the right hand side to δ and solve for ϵ, which gives that with probability at least $1 - \delta$,

$$\frac{1}{n}\sum_i X_i \leq \frac{1}{\sqrt{n}}\left(\frac{1}{\delta}\right)^{2/p}.$$

  If $p \geq \ln(1/\delta)/(2\ln(\ln(1/\delta)))$, then this last term is smaller than $\ln(1/\delta)$, which means we get what we wanted. So in general, it seems that higher moments help. *[future matus: give a ref for the above.]*

- As a more concrete example of the preceding, suppose X_i is drawn uniformly from $\{-1, +1\}$, thus $\sum_i X_i$ is the standard random walk on the integers. Applying the preceding,

$$\Pr[n^{-1}\sum_i X_i \geq \epsilon] \leq \frac{\mathrm{Var}(\sum_i X_i)}{n^2\epsilon^2}.$$

  To analyze the variance, let Y_i be uniform on $\{0, 1\}$, and note $\sum_i Y_i \sim \mathrm{Bin}(n, 1/2)$, and thus $\mathrm{Var}(\sum_i Y_i) \leq n/4$. Since $X_i = 2Y_i - 1$, then $\mathrm{Var}(\sum_i X_i) = 4\mathrm{Var}(\sum_i Y_i) \leq n$. Consequently,

$$\Pr[n^{-1}\sum_i X_i \geq \epsilon] \leq \frac{1}{n\epsilon^2};$$

  Setting the right hand side to δ and solving for *epsilon* gives $\epsilon = 1/\sqrt{n\delta}$: with probability at least $1 - \delta$ over the draw of $(X_1, \ldots, X_n)$,

$$\frac{1}{n}\sum_i X_i \leq \frac{1}{\sqrt{n\delta}}.$$

  In other words, the standard random walk seems to be at distance $\mathcal{O}(\sqrt{n})$ from the origin after n steps. This estimate is roughly correct; for instance we can use the law of the iterated logarithm to really pin this down. *[future matus: maybe that's a weird comment, we haven't discussed asymptotics..]*

So far, it has seemed that higher moments help; at least, with controls on enough moments, we get the $\ln(1/\delta)$ that we desired. Unfortunately, the number of moments we need depends on the desired confidence δ, which is awkward.

Along these lines, recall the Taylor expansion

$$e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!}.$$

In the first homework, we discussed the **moment generating function** $\mathbb{E}(\exp(tX))$. Based on the above Taylor expansion, this function should be sensitive to all the moments, though in a strange way: moment p is rescaled by $p!$. Let's see what happens if we plug this into the earlier corollary which combined Markov with a nondecreasing function: for any $\epsilon > 0$,

$$\Pr[X \geq \epsilon] = \inf_{t \geq 0} \Pr[X \geq \epsilon] \leq \inf_{t \geq 0} \frac{\mathbb{E}(\exp(tX))}{\exp(t\epsilon)}.$$

If instead we have the random variable $n^{-1}\sum_i X_i$, then we get

$$\Pr[n^{-1}\sum_i X_i \geq \epsilon] \leq \inf_{t \geq 0} \frac{\mathbb{E}(\prod_i \exp(tX_i/n))}{\exp(t\epsilon)},$$

where so far we have neither assumed i.i.d., nor have we assumed X_i has mean zero!

Recall (from the homework) that a random variable Z is **c-subgaussian** if $\mathbb{E}(\exp(tZ)) \leq \exp(t^2c/2)$ for every $t \in \mathbb{R}$. We can use this to complete the preceding proof.

Theorem. Suppose $(X_1, \ldots, X_n)$ are independent, and X_i is c-subgaussian. Then

$$\Pr[n^{-1}\sum_i X_i \geq \epsilon] \leq \exp\left(-\frac{n\epsilon^2}{2c}\right).$$

In particular, with probability at least $\mathbf{1} - \delta$,

$$\frac{1}{n}\sum_i X_i < \sqrt{\frac{c}{2n}\ln(1/\delta)}.$$

Proof. Continuing from the earlier derivation, using independence and the definition of c-subgaussian

$$\Pr[n^{-1}\sum_i X_i \geq \epsilon] \leq \inf_{t\geq 0} \frac{\mathbb{E}(\prod_i \exp(tX_i/n))}{\exp(t\epsilon)} = \inf_{t\geq 0} \frac{\prod_i \mathbb{E}(\exp(tX_i/n))}{\exp(t\epsilon)} \leq \inf_{t\geq 0} \exp\left(\frac{ct^2}{2n^2} - t\epsilon\right).$$

The quantity with the $\exp(\cdot)$ is a convex quadratic minimized at $t := \epsilon n^2/c$. Plugging this in gives the first bound, and the second bound follows from setting the right hand side to δ and solving for ϵ. □

Example.

- **Gaussians.** If $X_i \sim \mathcal{N}(0, \sigma^2)$, then X_i is σ^2-subgaussian (as computed in the homework), thus with probability at least $1-\delta$,
$$\frac{1}{n}\sum_i X_i \leq \sigma\sqrt{12n\ln(1/\delta)}.$$

- **Bounded random variables.** Suppose X_i has mean zero and $X_i \in [a,b]$ with probability 1. Then X_i is $(b-a)^2/4$-subgaussian; this bound is usually called **Hoeffding's Lemma**, and together it gives **Hoeffding's inequality**: if $(X_1, \ldots, X_n)$ are i.i.d., then
$$\Pr[n^{-1}\sum_i X_i \leq \mathbb{E}(X_1) + \epsilon] \leq \exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right).$$

- **Classification.** Just from Hoeffding's inequality, we can already control the error of a single classifier. Let $f : \mathbb{R}^d \to \mathbb{R}$ be any predictor, and let $((X_i, Y_i))_{i=1}^n$ be i.i.d. copies of some random variable pair (X, Y). For every i, the random variables $Z_i := \mathbf{1}[\mathrm{sgn}(f(X_i)) \neq Y_i]$ and $W_i := \mathbb{E}(Z_i) - Z_i$; by these choices, each W_i is i.i.d., has zero mean, and lies in an interval of length 1; thus we can apply Hoeffding's inequality to $\sum_i W_i$, and provide that with probability at least $1-\delta$ over the draw of $((X_i, Y_i))_{i=1}^n$,
$$\Pr[\mathrm{sgn}(f(X)) \neq Y] \leq \frac{1}{n}\sum_{i=1}^n \mathbf{1}[\mathrm{sgn}(f(X_i)) \neq Y_i] + \sqrt{\frac{2}{n}\ln\left(\frac{1}{\delta}\right)}.$$

Before we close this topic, there is a final useful inequality for us, which goes beyond the pure i.i.d. setting.

Theorem (Azuma's inequality). Let martingale sequence $(X_0, \ldots, X_n)$ be given, meaning $\mathbb{E}(X_i - X_{i-1}|X_{i-1}, \ldots, X_1) = 0$. Suppose further that $|X_i - X_{i-1}| \leq c_i$ with probability 1 (for some c_i). Then
$$\Pr\left[X_n - X_0 \geq \epsilon\right] \leq \exp\left(-\frac{\epsilon^2}{\sum_i c_i^2}\right).$$

Remark.

- Some people call this Azuma-Hoeffding; we'll use "Hoeffding" for the i.i.d. case and "Azuma" for the Martingale case to keep things simple.
- Let i.i.d., zero-mean random variabls $(Y_i)_{i=1}^n$ be given with $Y_i \in [a,b]$ with probability 1. We can recover Hoeffding (applied to $n^{-1}\sum_i Y_i$) from Azuma by the choice $X_i = n^{-1}\sum_{j\leq i} X_j$ (with $X_0 = 0$), whereby $\mathbb{E}(X_i - X_{i-1}|X_{i-1}, \ldots, X_0) = \mathbb{E}(n^{-1}Y_i) = 0$, and $|X_i - X_{i-1}| = |Y_i|/n \leq (b-a)/n$.

[future matus: maybe only use martingale difference sequences?]

Proof (of Azuma)**.** For convenience, define $Z_i := X_i - X_{i-1}$, whereby $X_n - X_0 = \sum_i Z_i$. Stepping through the proof of Hoeffding's inequality, independence was not used to obtain the inequality
$$\Pr[X_n - X_0 \geq \epsilon] = \Pr[\sum_i Z_i \geq \epsilon] \leq \inf_{t\geq 0} \frac{\mathbb{E}(\prod_{i=1}^n \exp(t(Z_i)))}{\exp(t\epsilon)}.$$

By properties of conditional expectation and martingales, we can manipulate this expression to conditionally apply the Hoeffding lemma:

$$\begin{aligned}
\mathbb{E}(\prod_i \exp(t(Z_i))) &= \mathbb{E}\left(\mathbb{E}\left(\prod_{i=1}^{n} \exp(t(Z_i))|X_{n-1},\ldots,X_0\right)\right) \\
&= \mathbb{E}\left(\mathbb{E}\left(\exp(t(Z_n))|X_{n-1},\ldots,X_0\right)\right)\mathbb{E}\left(\mathbb{E}\left(\prod_{i=1}^{n-1} \exp(t(Z_i))|X_{n-1},\ldots,X_0\right)\right) \\
&= \exp(t^2/(8c_i^2))\mathbb{E}\left(\mathbb{E}\left(\prod_{i=1}^{n-1} \exp(t(Z_i))|X_{n-1},\ldots,X_0\right)\right).
\end{aligned}$$

Proceeding with this derivation recursively, we obtain

$$\Pr[X_n - X_0 \geq \epsilon] \leq \inf_{t\geq 0} \frac{\exp(-\sum_i \epsilon^2/c_i^2)}{\exp(t\epsilon)}.$$

and finish as in the subgaussian case. □

References