16. Concentration of Measures

Generalization

- Starting today, we discuss the topic of generalization
  - Most intriguing, yet imperfect

- **Target of analysis.** The gap between the training risk and the test risk
  - Approximation. How small can our test risk (potentially) be?
  - Optimization. How small can our training risk be?

- In a sense, generalization links what we have learned so far:
  - How can we find the solution for approximation, by solving the optimization?

Why do we expect generalization?

- The training risk can be re-written as:

$$\hat{R}(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, g(x_i; w)) = \mathbb{E}_{P_n}[\ell(Y, g(X; w)]$$

- On the other hand, the test risk is:

$$R(w) = \mathbb{E}_{P}[\ell(Y, g(X; w))]$$

- By "the law of large numbers" type of results, we have:

$$P_n \to P$$

  - Very rough; various senses of convergence

- Thus, we expect that $\hat{R}(w) \to R(w)$, for any FIXED w
  - **Question.** But how fast is this convergence?

Concentration of measure

- "Concentration of measure" bounds provide quantitative versions of LLN

- **Goal.** Nonasymptotic version of the following claim:
  - Suppose that we have many independent-ish random variables

$$X_{1:n} = (X_1, X_2, \ldots, X_n)$$

  - If our function $f(x_1, x_2, \ldots, x_n)$ is not overly sensitive to one coordinate, then the random variable

$$f(X_1, X_2, \ldots, X_n)$$

    behaves sufficiently deterministically
    - i.e., close to $\mathbb{E}[f]$.

- Does this sound too vague?

First tool: Markov's inequality

- Let us focus on the probability of excess deviations

$$\mathbf{Pr}\Big[\,|f(X_1, X_2, \cdots, X_N) - \mathbb{E}[f]\,| > \epsilon\Big] \leq \delta$$

Theorem 12.1 (Markov's inequality).

For any nonnegative random variable X, we have

$$\mathbf{Pr}[X > \epsilon] \leq \frac{\mathbb{E}[X]}{\epsilon} \qquad \forall \epsilon > 0$$

- **Proof idea.** Think about the random variable $\epsilon \cdot \mathbf{1}\{\mathbf{X} > \epsilon\}$

Applying the Markov's inequality

$$\mathbf{Pr}[X > \epsilon] \leq \frac{\mathbb{E}[X]}{\epsilon} \qquad \forall \epsilon > 0$$

- Markov's inequality works for any nonnegative random variable X
  - How can we use this to analyze the following?

$$\mathbf{Pr}\Big[\,|f(X_1, X_2, \cdots, X_N) - \mathbb{E}[f]\,| > \epsilon\Big] \leq \delta$$

- **Naïve choice.** Plug in $X = |f(X_{1:n}) - \mathbb{E}f|$
  - Requires a bounded absolute mean
  - If f is an i.i.d. sample mean, difficult to get a good n-dependency

Applying the Markov's inequality

$$\mathbf{Pr}[X > \epsilon] \leq \frac{\mathbb{E}[X]}{\epsilon} \qquad \forall \epsilon > 0$$

- Target:

$$\mathbf{Pr}\Big[\,|f(X_1, X_2, \cdots, X_N) - \mathbb{E}[f]\,| > \epsilon\Big] \leq \delta$$

- **Chebyshev's choice.** Plug in $X = (f - \mathbb{E}f)^2$
  - Requires a bounded variance
  - If f is an i.i.d. sample mean, RHS decays with $1/n$ dependency

- **Question.** How far can we push this requirement-dependency tradeoff?

Applying the Markov's inequality

$$\mathbf{Pr}[X > \epsilon] \leq \frac{\mathbb{E}[X]}{\epsilon} \qquad \forall \epsilon > 0$$

$$\mathbf{Pr}\Big[\,|f(X_1, X_2, \cdots, X_N) - \mathbb{E}[f]\,| > \epsilon\Big] \leq \delta$$

- **Chernoff's choice.** Plug in $X = \exp\Big(t \cdot (f - \mathbb{E}f)\Big)$, for some $t > 0$ (tunable)
  - Then, we get something like

$$\mathbf{Pr}[\exp(t \cdot (f - \mathbb{E}f)) \geq \exp(t\epsilon)] \leq \frac{\mathbb{E}[\exp(t \cdot (f - \mathbb{E}f))]}{\exp(t\epsilon)}$$

  - Tidying up a bit, we get:

$$\mathbf{Pr}[f - \mathbb{E}f \geq \epsilon] \leq \frac{\mathbb{E}[\exp(t \cdot (f - \mathbb{E}f))]}{\exp(t\epsilon)}$$

  - Note. To get an absolute value, we can use the union bound

Subgaussian random variable

$$\mathbf{Pr}[f - \mathbb{E}f \geq \epsilon] \leq \frac{\mathbb{E}[\exp(t \cdot (f - \mathbb{E}f))]}{\exp(t\epsilon)}$$

- For RHS to be finite, the following quantity need to be bounded:

$$\mathbb{E}[\exp(t \cdot (f - \mathbb{E}f))]$$

Definition (subgaussian).

A random variable Z is subgaussian with mean μ and proxy σ^2, whenever

$$\mathbb{E}[\exp(\lambda(Z - \mu))] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right)$$

holds for any $\lambda > 0$.

Warm-up

$$\mathbb{E}[\exp(\lambda(Z-\mu))] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right)$$

- Suppose that we have a Gaussian random variable $Z \sim \mathcal{N}(0,\rho^2)$.
- **Question.** Is this subgaussian?

Toward the Chernoff's bound

- Now, consider the case

$$f(X_{1:n}) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

  - where all X_i are mutually independent and are $(0,\sigma^2)$-subgaussian.

- Then, the previous bound becomes:

$$\mathbf{Pr}[f \geq \epsilon] \leq \frac{\mathbb{E}[\exp(t \cdot f)]}{\exp(t\epsilon)} = \frac{\mathbb{E}[\prod_{i=1}^{n} \exp(tX_i/n)]}{\exp(t\epsilon)} = \frac{\prod_{i=1}^{n} \mathbb{E}[\exp(tX_i/n)]}{\exp(t\epsilon)}$$

$$\leq \frac{\prod_{i=1}^{n} \exp\left(\frac{t^2\sigma^2}{2n^2}\right)}{\exp(t\epsilon)} = \exp\left(\frac{t^2\sigma^2}{2n} - t\epsilon\right)$$

Toward the Chernoff's bound

$$\mathbf{Pr}[f \geq \epsilon] \leq \exp\left(\frac{t^2\sigma^2}{2n} - t\epsilon\right)$$

- To minimize the RHS, choose $t = n\epsilon/\sigma^2$.
  - Then, we get:

$$\mathbf{Pr}[f \geq \epsilon] \leq \exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right)$$

- More generally, if all variables have different subgaussianity σ_i^2, we have:

$$\mathbf{Pr}[f > \epsilon] \leq \exp\left(-\frac{n^2\epsilon^2}{\sum_{i=1}^{n}\sigma_i^2}\right)$$

Summary so far

- We were interested in controlling the excess deviation probability:

$$\mathbf{Pr}[\,|f(X_{1:n}) - \mathbb{E}[f]\,| \geq \epsilon] \leq \delta$$

  - With particular interests in the case

$$f(X_{1:n}) = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad X_i \sim P \quad \text{i.i.d.}$$

- We started from the Markov's inequality:

$$\mathbf{Pr}[Z \geq \epsilon] \leq \frac{\mathbb{E}Z}{\epsilon}$$

- We saw its various applications:
  - Under bounded mean: $\delta = \mathcal{O}(1)$
  - Under bounded variance: $\delta = \mathcal{O}(1/n)$
  - Under subgaussian tail: $\delta = \mathcal{O}(\exp(-n \cdot C))$

Bounded case

- Now, consider a bounded case — i.e., no tail

- Think about a bounded random variable, i.e.,

$$\mathbf{Pr}\left[X \in [a,b]\right] = 1$$

Property.

$$\mathrm{Var}(X) \leq \frac{1}{4}(b-a)^2, \qquad \mathrm{Var}(X) \leq \mathbb{E}[(X-a)^2]$$

- **Proof idea.** Note that

$$\mathrm{Var}(X) = \mathrm{Var}(X-a) \leq \mathbb{E}[(X-a)^2]$$

  - Plug in the right quantity

Hoeffding's lemma

- With this property, we are now ready to prove the Hoeffding's lemma:

Lemma (Hoeffding).

Let $X \in [a, b]$ almost surely. Then, we know that

$$\mathbb{E}\left[\exp(\lambda(X - \mathbb{E}X))\right] \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right), \qquad \forall \lambda > 0$$

- In other words, X is $(\mathbb{E}X, (b-a)^2/4)$-subgaussian

Proof sketch

- Assume WLOG that $\mathbb{E}X = 0$.
  - Then, we want to show that:

$$\mathbb{E}\left[\exp(\lambda X)\right] \leq \exp\left(\frac{\lambda^2 (b-a)^2}{8}\right), \qquad \forall \lambda > 0$$

- Think about the log moment-generating function

$$\psi(\lambda) = \log \mathbb{E}[\exp(\lambda X)]$$

  - If suffices to show that

$$\psi(\lambda) \leq \lambda^2 (b-a)^2 / 8$$

Proof sketch

show $\psi(\lambda) \leq \lambda^2(b-a)^2/8$, for $\psi(\lambda) = \log \mathbb{E}[\exp(\lambda X)]$,

- Derivatives can be written as:

$$\psi'(\lambda) = \frac{\mathbb{E}[X \exp(\lambda X)]}{\mathbb{E}[\exp(\lambda X)]}$$

$$\psi''(\lambda) = \frac{\mathbb{E}[X^2 \exp(\lambda X)]}{\mathbb{E}[\exp(\lambda X)]} - \left(\frac{\mathbb{E}[X \exp(\lambda X)]}{\mathbb{E}[\exp(\lambda X)]} \right)^2$$

- Consider a random variable Z, distributed as

$$p_Z(z) = \frac{\exp(\lambda z)}{\mathbb{E}[\exp(\lambda X)]} p_X(z)$$

  - Check by yourself whether this is a valid probability measure
  - This trick is called “exponential tilting”
- Then, we know that $\mathbf{Pr}\left[Z \in [a,b]\right] = 1$

Proof sketch

show $\psi(\lambda) \le \lambda^2(b-a)^2/8$, for $\psi(\lambda) = \log \mathbb{E}[\exp(\lambda X)]$,

$$p_Z(z) = \frac{\exp(\lambda z)}{\mathbb{E}[\exp(\lambda X)]} p_X(z)$$

- Think about the mean and variance of Z

$$\mathbb{E}Z = \frac{\mathbb{E}[X \cdot \exp(\lambda X)]}{\mathbb{E}[\exp(\lambda X)]}$$

$$\begin{aligned}
\mathrm{Var}(Z) &= \mathbb{E}[Z^2] - (\mathbb{E}Z)^2 \\
&= \frac{\mathbb{E}[X^2 \cdot \exp(\lambda X)]}{\mathbb{E}[\exp(\lambda X)]} - \left(\frac{\mathbb{E}[X \cdot \exp(\lambda X)]}{\mathbb{E}[\exp(\lambda X)]} \right)^2 \\
&= \psi''(\lambda) \\
&\le \frac{1}{4}(b-a)^2
\end{aligned}$$

Proof sketch

show $\psi(\lambda) \leq \lambda^2(b-a)^2/8$, for $\psi(\lambda) = \log \mathbb{E}[\exp(\lambda X)]$,

$$\psi''(\lambda) \leq (b-a)^2/4$$

- We have:

$$\psi(0) = \log \mathbb{E}[\exp(0 \cdot X)] = \log \mathbb{E}[1] = 0$$

$$\psi'(0) = \psi'(0) = \frac{\mathbb{E}[X \exp(0 \cdot X)]}{\mathbb{E}[\exp(0 \cdot X)]} = \frac{\mathbb{E}[X]}{1} = 0$$

- Thus,

$$\psi(\lambda) = \int_0^\lambda \left(\int_0^\tau \psi''(t)\, \mathrm{d}t \right) \mathrm{d}\tau \leq \int_0^\lambda \left(\int_0^\tau \frac{(b-a)^2}{4}\, \mathrm{d}t \right) \mathrm{d}\tau$$

$$\leq \frac{(b-a)^2}{4} \cdot \left(\int_0^\lambda \tau\, \mathrm{d}\tau \right) = \frac{\lambda^2(b-a)^2}{8}$$

Hoeffding's theorem

Theorem 12.3 (Hoeffding).

Given independent $X_1, \ldots, X_n$ with $X_i \in [a_i, b_i]$, we have

$$\mathbf{Pr}\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X_i \geq \epsilon\right] \leq \exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

- **Application.** Let f be some fixed predictor.
  - Let $Z = \mathbf{1}\left[f(X) \neq Y\right]$
    - i.e., a random variable indicating the "error" of the predictor on a random sample
  - Then, with probability $1 - \delta$, we have

$$\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}Z\right| \leq \sqrt{\frac{\log(2/\delta)}{2n}}$$

Other sophisticated results

Theorem (McDiarmid).

Let $f(\cdot)$ have the bounded difference property, i.e.,

$$|f(x_1, \cdots, x_{i-1}, x_i, x_{i+1}, \cdots, x_n) - f(x_1, \cdots, x_{i-1}, \tilde{x}_i, x_{i+1}, \cdots, x_n)| \leq c_i, \qquad \forall \ldots$$

Then, for independent $X_1, \ldots, X_n$, we have

$$\mathbf{Pr}[f - \mathbb{E}f \geq \epsilon] \leq \exp\left(\frac{2\epsilon^2}{\sum c_i^2}\right)$$

- Also note the Bernstein's inequality
  - Boundedness + Variance bound for a tighter bound

Next up

- Uniform convergence