CSC 696H: Topics in Bandits and Reinforcement Learning Theory **Fall 2024**

Lecture 2 - Concentration of Measure; Generalization in ML

Lecturer: Chicheng Zhang *Scribe: Junfeng Xu*

1 Concentration of Measure

- Concentration of Measure: basically provides a way to quantify how close is the sample mean to the population
  - Factors: the distribution of the original random variable, sample size, unlucky sample draw
  - Example of the concentration quality reduces all three factors above:

    Theorem 1. *Suppose $X_1 \ldots X_n$ are iid, $\mathbb{E}[X_i] = \mu$, if X_i's are $b^2 - SG$ (all random variables are sub-Gaussian), then*

    $$P(|\overline{X_n} - \mu| \geq \epsilon) \leq 2exp(-\frac{n\epsilon^2}{2b^2}) \tag{1}$$

  - Let ϵ to be such that $2exp(-\frac{2n\epsilon^2}{b^2}) = \delta \Rightarrow \epsilon = b\sqrt{\frac{\ln \frac{2}{\delta}}{2n}}$
  - To prove equation 1: $LHS = P(\overline{X_n} - \mu \geq \epsilon) + P(\overline{X_n} - \mu \leq -\epsilon)$
    $(\square) = P(\overline{X_n} - \mu \geq \epsilon), (\triangle) = P(\overline{X_n} - \mu \leq -\epsilon)$

    $$\begin{aligned}(\square) &= P(\overline{X_n} - \mu \geq \epsilon) \\ &= P(\overline{X_n} - \mu \geq \epsilon) \\ &= P(\sum_{i=1}^{n} X_i - n\mu \geq n\epsilon)\end{aligned}$$

    First we choose a free parameter λ greater than zero and scale both sides by the factor of λ

    $$= P(\lambda(\sum_{i=1}^{n} X_i - n\mu) \geq \lambda n\epsilon), \lambda > 0$$

    We need to use sub-Gaussian distribution property here. The sub-gaussianness is about the deviation of the random variable to its mean, but exponentiated. It is nature that we can exponentiate both sides of the equation.

    $$= P(e^{\lambda(\sum_{i=1}^{n} X_i - n\mu)} \geq e^{\lambda n\epsilon})$$

    Denote $e^{\lambda(\sum_{i=1}^{n} X_i - n\mu)}$ as Z, this random variable Z has non-negativity property, based on markov inequality, the probability that it deviates is greater than the threshold w cannot be too large if w is already very large. The tail probability will be smaller if my original random variable has a smaller expectation or my threshold is chosen to be large. i.e. $P(Z \geq w) \leq \frac{\mathbb{Z}}{w}$.

$$
\begin{aligned}
&\leq e^{-\lambda n\epsilon}\mathbb{E}[e^{\lambda(\sum\limits_{i=1}^{n} X_i - n\mu)}] \\
&= e^{-\lambda n\epsilon}\phi_{\sum\limits_i X_i - n\mu}(\lambda) \\
&= e^{-\lambda n\epsilon}\phi_{\sum\limits_i X_i - \mathbb{E}[\sum\limits_i X_i]}(\lambda)
\end{aligned}
$$

Is $\sum\limits_i X_i$ a SG random variable? Yes. if so, what is various proxy? nb^2. $n = 2, x_1 + x_2$ is $2b^2 - SG$. Given a sub-gaussian random variable, it is scaled by a constant factor. The result is still a sub-Gaussian random variable. These various proxy will be scaled by a factor of the scaling square. If we have two independent random variables, both of which are sub-gaussian, their summation must be sub-gaussian. The variance proxy of the new random variable is the summation over each individual random variable's variance proxy. Then applying the definition of the sub-gaussian:

$$
\leq e^{-\lambda n\epsilon + \frac{\lambda^2 nb^2}{2}}
$$

for any $\lambda > 0$, choosing λ that minimizes that bound $\Rightarrow \lambda = \frac{\epsilon^2}{b^2}$

$$
\Rightarrow (\square) \leq e^{-\frac{n\epsilon^2}{2b^2}}
$$

Similarly

$$
(\triangle) \leq e^{-\frac{n\epsilon^2}{2b^2}}
$$

$\Rightarrow$ Theorem 1

Theorem 2: (Bernstein Inequality) Let X_1 ... X_n be iid random variables, $|X_i - \mathbb{E}X_i| \leq R$, $\mu = \mathbb{E}X_i$, $\sigma^2 = var(X_i)$, then for any $\epsilon > 0$:

$$
P(|\overline{X_n} - \mu| \geq \epsilon) \leq 2exp(-\frac{n\epsilon^2}{2\sigma^2 + \frac{2}{3}R\epsilon}) \tag{2}
$$

(no worse, can be much better than hoeffding equality. Because the term $\frac{2}{3}R\epsilon$ can be ignored and the denominator has the actual variance of the random variable rather than the variance proxy of the random variable. Generally for a random variable, the variance of it can be much smaller than the variance proxy which makes this bound significantly better.)

Choosing ϵ so that RHS $= \delta$, choosing a tight enough ϵ, such that RHS $\leq \delta$

$$
2exp(-\frac{n\epsilon^2}{2\sigma^2 + \frac{2}{3}R\epsilon}) \leq \delta
$$

Dividing by two on both sides of the equation and taking a logarithmic on both sides, then moving the denominator.

$$
\Leftrightarrow \; n\epsilon^2 \geq (2\sigma^2 + \frac{2}{3}R\epsilon)\ln\frac{2}{\delta}
$$

$X > A + B \Leftarrow X \geq 2A$ and $x \geq 2B$. X must be greater than the average of the two, which is A+B. Then applying this:

$$\Leftarrow \ n\epsilon^2 \geq 2 * 2\sigma^2 \ln\frac{2}{\delta} \text{ and } n\epsilon^2 \geq 2 * \frac{2}{3}R\epsilon\ln\frac{2}{\delta}$$
$$\Leftrightarrow \ \epsilon \geq \sqrt{\frac{4\sigma^2\ln\frac{2}{\delta}}{n}} \text{ and } \epsilon \geq \frac{4R\ln\frac{2}{\delta}}{3n}$$
$$\Leftarrow \ \epsilon \geq \sqrt{\frac{4\sigma^2\ln\frac{2}{\delta}}{n}} + \frac{4R\ln\frac{2}{\delta}}{3n}$$

In summary:

$$P(|\overline{X_n} - \mu| \geq \sqrt{\frac{4\sigma^2\ln\frac{2}{\delta}}{n}} + \frac{4R\ln\frac{2}{\delta}}{3n}) \leq \delta \tag{3}$$

2 Generalization in supervised machine learning

- Instance space X (e.g. $[0,1]^{W\times H}$ camera images in pixel representation)
- Label space Y (e.g. Y $= \{L, R\}$)
- Loss function, $\ell(\hat{y}, g) \in [0, B]$, $\hat{y}$: prediction, g: ground truth label (e.g. $\ell(\hat{y}, g) = I(\hat{y} \neq g)$)
- distribution D cover X, Y $(X, Y) \sim D$ (e.g. Camera capture demonstrated by expert steering)
- prediction rule $f : X \to Y$, quality measure: generalization loss:

$$L_D(f) = \mathbb{E}_{(X,Y)\sim D}[\ell(f(X), Y)] \quad \text{(smaller the better)} \tag{4}$$

- $\mathcal{F}$: a predictor class to learn $\hat{f}$ from. (e.g. neural network with a fixed architecture)
- can we design a general approach to find a good $\hat{f}$ for any F? $\hat{f}$ approximately minimizes $L_D(f)$

Idea:

$$\forall f : L_n(f) = \frac{1}{n}\sum_{i=1}^{n}\ell(f(x_i), y_i) \to L_D(f) \tag{5}$$

(Concentration of measure)

Algorithm: Empirical risk minimization (ERM)

$$return \ \hat{f} = \arg\min_{f\in\mathcal{F}} L_n(f) \tag{6}$$

Theorem: (ERM) Suppose $|F| \leq \infty$, then $\forall\delta > 0$, with probability $1 - \delta$:

$$\forall f \in F \text{ simultaneously, } |L_n(f) - L_D(f)| \leq B\sqrt{\frac{\ln\frac{|F|}{\delta}}{2n}} =: \epsilon_n \tag{7}$$

and therefore

$$L_D(\hat{f}) \leq L_D(f^*) + 2\epsilon_n \tag{8}$$

(notation: $f^* = \text{argmin}_{f\in F} L_D(f)$)

Interpretation:

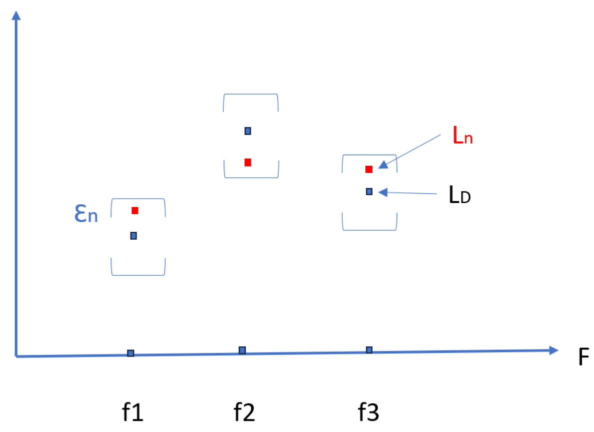

Figure 1: Thm 3 figure

- as long as $n >> \ln|F|$, $\hat{f}$'s performance is competitive with the best model in F. To get a sense of how large is $ln|F|$, suppose F has m parameters, each taking V values, $|F| = V^m \Rightarrow \ln|F| = m \ln V$
- instance space X can be enormous, the training set only has a tiny fraction of all possible instances, yet ERM has strong guarantee $\Rightarrow$ achieves generalization.
- $L_D(f^*)$: approximation error
- $2\epsilon_n$: estimation error
- Larger F $\Rightarrow$ estimation error increases, approximation error decreases.
- F can encode learner's inductive bias, which can help the learning in application-specific ways. For example, for image classification, $X = images$, $F = convolutional\ neural\ network$, then classifiers in F satisfies translational invariance, that is, $\forall f \in F$, $f(X) = f(X')$, if X is a translation of X'.
- In modern deep learning regime, $\ln|F| >> n$ is more common. The guarantees provided by this theorem is vacuous. Nevertheless, researchers found that popular learners (e.g. stochastic gradient-based learners) manage to converge to "simple" predictors, which implicitly uses a much smaller $\mathcal{F}$. Note: "Implicit Regularization" by Nati Srebro is a great video to watch.

Proof: (7) $\Rightarrow$ (8) refer to figure 1.

$$
\begin{aligned}
L_D(\hat{f}) &\leq L_n(\hat{f}) + \epsilon_n && \hat{f}\text{'s training and test loss are within } \epsilon_n \\
&\leq L_n(f^*) + \epsilon_n && \hat{f} \text{ is ERM} \\
&\leq (L_D(f^*) + \epsilon_n) + \epsilon_n && f^* \text{ is training in the test loss are within } \epsilon_n
\end{aligned}
$$

Proving (7): want to show: $P(E) \geq 1 - \delta$, equivalently, we want to show a statement like:

$$
\begin{aligned}
& P(\forall f \in F : |L_n(f) - L_D(f)| \leq \epsilon) \geq 1 - \delta \\
\Leftrightarrow \quad & P(\exists f \in F : |L_n(f) - L_D(f)| > \epsilon) \leq \delta
\end{aligned}
$$

$$\begin{aligned}
LHS &= P(\bigcup_{f\in F} |L_n(f) - L_D(f)| > \epsilon) \\
&\leq \sum_{f\in F} P(|L_n(f) - L_D(f)| > \epsilon) \\
&\leq exp(-\frac{2n\epsilon^2}{B^2}) \\
&= |F| 2exp(-\frac{2n\epsilon^2}{B^2})
\end{aligned}$$

Setting ϵ such that this bounds $\delta \Rightarrow 2|F|exp(-\frac{2n\epsilon^2}{B^2}) = \delta \Rightarrow \epsilon = B\sqrt{\frac{\ln\frac{2F}{B}}{2n}} = \epsilon_n$

$$\Rightarrow P(\exists f : |L_n(f) - L_D(f)| > \epsilon_n) \leq \delta \tag{9}$$

Follow up question: what would we get for ϵ_n if we instead using Chebyshev's inequality for bounding the deviation probability? (Hint: $\ln \frac{|F|}{\delta}$ will become without $\frac{|F|}{\delta}$, can you see why?)