**Machine Learning Theory** **Spring 2021**

# Lecture 2: Concentration Inequalities

*Lecturer: Liwei Wang* *Scribe: Baihe Huang, Bi'an Du, Zichen Wu, Yuanchen Qiu*



## 2.1 Recap

Recall Chernoff inequality and Chebyshev inequality from last lecture.

**Theorem 2.1.1 (Chernoff Inequality).** *Let $X$ be a random variable that is non-negative with moment generating function $\mathbb{E}e^{tX}$. Then $\forall k > 0$,*

$$\mathbb{P}(X \geq k) \leq \inf_{t>0} e^{-tk}\mathbb{E}[e^{tX}].$$

**Theorem 2.1.2 (Chebyshev Inequality).** *Let random variables $X_1, X_2, \ldots, X_n \sim$ iid Bernoulli$(1, p)$. We have*

$$\mathbb{P}(\frac{1}{n}\sum_{i=1}^{n} X_i - p \geq \epsilon) \leq \frac{\mathrm{Var}(\sum_{i=1}^{n} X_i/n)}{\epsilon^2} = \frac{p(1-p)}{n\epsilon^2}$$

Notice that Chebyshev inequality only uses second moment information of random variables, therefore its convergence rate is only inversely proportional.

From law of large number, we naturally expect

$$\mathbb{P}(\frac{1}{n}\sum_{i=1}^{n} X_i - p \geq \epsilon) \leq e^{-O(n)}.$$

## 2.2 Concentration Inequalities

### 2.2.1 Backgrounds of information theory

**Definition 2.2.1 (Entropy).** *Let $X$ be a random variable with probability mass function $p = (p_1, p_2, \ldots)$. The entropy of $X$ is defined by*

$$H(X) := \begin{cases} \sum_i p_i \log_2 \frac{1}{p_i} \text{ (bits)} \\ \sum_i p_i \ln \frac{1}{p_i} \text{ (nats)} \end{cases}$$

**Remark 2.2.2.** *The entropy of a random variable is the average level of "information", "surprise", or "uncertainty" inherent in the variable's possible outcomes.*

**Definition 2.2.3** **(Relative Entropy).** *For two probability mass functions $P = (p_1, p_2, \dots)$ and $Q = (q_1, q_2, \dots)$, the relative entropy from $Q$ to $P$ is defined to be*

$$D(P||Q) := \begin{cases} \sum_i p_i \log_2 \frac{p_i}{q_i} \text{ (bits)} \\ \sum_i p_i \ln \frac{p_i}{q_i} \text{ (nats)} \end{cases}$$

*In particular, for two Bernoulli random variables $P = (p, 1-p), Q = (q, 1-q)$*

$$D_B^{(e)}(p||q) := p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q}$$

**Remark 2.2.4.** *Relative entropy measures the difference of two distributions, but this relation is asymmetric. Note that $D(P||Q) \geq 0$ for any $P, Q$ and usually $D(P||Q) \neq D(Q||P)$.*

### 2.2.2 Chernoff Bound

**Theorem 2.2.5.** *Let $X_1, X_2, \dots, X_n$ be $n$ iid Bernoulli random variables satisfying $\mathbb{E}[X_i] = p, \forall i \in [n]$. Then for all $\epsilon > 0$ we have*

$$\mathbb{P}(\frac{1}{n} \sum_{i=1}^n X_i - p \geq \epsilon) \leq e^{-n D_B^{(e)}(p+\epsilon||p)}.$$

*Proof.* By Chernoff inequality,

$$\mathbb{P}(\frac{1}{n} \sum_{i=1}^n X_i - p \geq \epsilon) \leq \inf_{t>0} e^{-t(p+\epsilon)} \mathbb{E}[e^{t \sum_{i=1}^n X_i}].$$

Notice that

$$\mathbb{E}[e^{t \sum_{i=1}^n X_i}] = \prod_{i=1}^n \mathbb{E}[e^{t X_i}] = (pe^t + 1 - p)^n. \tag{2.1}$$

It thus follows that

$$\begin{aligned} \mathbb{P}(\frac{1}{n} \sum_{i=1}^n X_i - p \geq \epsilon) &\leq \inf_{t>0} e^{-nt(p+\epsilon)} \cdot (pe^t + 1 - p)^n \\ &\leq e^{-n D_B^{(e)}(p+\epsilon||p)}. \end{aligned}$$

The last step is a simple calculation and left as homework. □

**Theorem 2.2.6.** *Let $X_1, \dots, X_n$ be $n$ random variables satisfying $X_i \in [0, 1]$ and $\mathbb{E}[X_i] = p, \forall i \in [n]$. Then for all $\epsilon > 0$, we have*

$$\mathbb{P}(\frac{1}{n} \sum_{i=1}^n X_i - p \geq \epsilon) \leq e^{-n D_B^{(e)}(p+\epsilon||p)}.$$

*Proof.* Notice that exponent function is convex. By Jensen's inequality, we have

$$\mathbb{E}[e^{tX}] \leq \mathbb{E}[X e^t] + \mathbb{E}[(1-X)e^0] = pe^t + 1 - p. \tag{2.2}$$

It thus follows that

$$\mathbb{E}[e^{t \sum_{i=1}^n X_i}] \leq (pe^t + 1 - p)^n.$$

Replacing Eq (2.1) by this inequality, the rest of the proof is the same as Theorem 2.2.5. □

**Theorem 2.2.7.** *Let $X_1, \dots, X_n$ be $n$ random variables satisfying $X_i \in [0,1]$ and $\mathbb{E}[X_i] = p_i, \forall i \in [n]$. Mark $p = \frac{1}{n}\sum_{i=1}^n p_i$, then for all $\epsilon > 0$ we have*

$$\mathbb{P}(\frac{1}{n}\sum_{i=1}^n X_i - p \geq \epsilon) \leq e^{-nD_B^{(e)}(p+\epsilon||p)}.$$

*Proof.* Notice that logarithmic function is concave. By Jensen's inequality, we have

$$\frac{\sum_{i=1}^n \ln(1-p_i+p_ie^t)}{n} \leq \ln(1-p+pe^t),$$

then combining this with Eq (2.2)

$$\begin{aligned}\mathbb{E}[e^{t\sum_{i=1}^n X_i}] &\leq \prod_{i=1}^n (1-p_i+p_ie^t) \\ &\leq (1-p+pe^t)^n.\end{aligned}$$

Replacing Eq (2.1) by this inequality, the rest of the proof is the same as Theorem 2.2.5. □

**Remark 2.2.8.** *The other side of tail bound can be proved similarly.*

**Lemma 2.2.9.** *(left as homework, find when the gap reaches infimum) $D_B^{(e)}(p+\epsilon||p) \geq 2\epsilon^2$.*

Plugging this lemma into Theorem 2.2.7, we have the following bound.

**Theorem 2.2.10 (Additive Chernoff Bound).** *Let $X_1, \dots, X_n$ be $n$ random variables satisfying $X_i \in [0,1]$ and $\mathbb{E}[X_i] = p_i, \forall i \in [n]$. Let $p = \frac{1}{n}\sum_{i=1}^n p_i$, then for all $\epsilon > 0$ we have*

$$\mathbb{P}(\frac{1}{n}\sum_{i=1}^n X_i - p \geq \epsilon) \leq e^{-2n\epsilon^2}.$$

**Remark 2.2.11.** *Note that Chernoff inequality requires $X_i$ are mutually independent, while pairwise independence suffices for Chebyshev inequality.*

### 2.2.3 Hoeffding's inequality

**Theorem 2.2.12 (Hoeffding's inequality).** *Let $X_1, X_2, ...X_n$ be $n$ independent random variables in $[a_i, b_i]$. Let $\mu = \frac{\sum_{i=1}^n E[X_i]}{n}$, then we have*

$$\mathbb{P}(\frac{1}{n}\sum_{i=1}^n X_i - \mu \geq \epsilon) \leq e^{\frac{-2n^2\epsilon^2}{\sum_{i=1}^n (b_i-a_i)^2}}.$$

### 2.2.4 Draw with/without replacement in a population

For $N$ numbers $a_1, a_2, ..., a_N \in \{0,1\}$,let $p = \frac{1}{N}\sum_{i=1}^N a_i$. We consider the following cases.

**Draw with replacement** $x_1, x_2, ..., x_n$ are randomly drawn with replacement from $\{a_1, a_2, ..., a_N\}$. Then $X_i$ are iid Bernoulli random variables with $\mathbb{E}[X_i] = p$. This case is essentially the same as Theorem 2.2.5.

**Draw without replacement** $y_1, y_2, ..., y_n$ are randomly drawn without replacement from $\{a_1, a_2, ..., a_N\}$. Now $y_1, \ldots, y_n$ are dependent. However, we can also show that:

$$\mathbb{P}(\frac{1}{n}\sum_{i=1}^{n} y_i - p \geq \epsilon) \leq e^{-2n\epsilon^2}.$$

*Proof.* It suffices to proof

$$\mathbb{E}[e^{t\sum_{i=1}^{n} y_i}] \leq \mathbb{E}[e^{t\sum_{i=1}^{n} x_i}], \tag{2.3}$$

namely the moment generation function is consistently less than the case where we draw with replacement. To prove this we expand moment generation functions into polynomials

$$\mathbb{E}[e^{t\sum_{i=1}^{n} x_i}] = 1 + t\mathbb{E}[\sum_{i=1}^{n} x_i] + \frac{t^2}{2}\mathbb{E}[(\sum_{i=1}^{n} x_i)^2] + ...$$

Notice that every polynomial terms look like $f(t)\mathbb{E}[\prod_{i\in I} x_i] = f(t)\mathbb{P}(\prod_{i\in I} x_i = 1)$ where $f(t)$ is a polynomial function of $t$. For the case where numbers are drawn without replacement, we have $f(t)\mathbb{E}[\prod_{i\in I} y_i] = f(t)\mathbb{P}(\prod_{i\in I} y_i = 1)$. Now $\prod_{i\in I} y_i = 1$ holds only when $y_i = 1, \forall i \in T$, which happens with less probability when drawn without replacement. Then we have

$$\mathbb{E}(\prod_{i\in I} y_i) \leq \mathbb{E}(\prod_{i\in I} x_i)$$

and thus Eq (2.3) holds. □

### 2.2.5 McDiarmid Inequality

Chernoff bound is a special case of McDiarmid inequality.

**Theorem 2.2.13 (McDiarmid's inequality).** *Let $X_1, \ldots, X_n \in \mathcal{X}$ be $n$ independent random variables and there exists constant $c_1, \ldots, c_n$ such that $f : \mathcal{X} \mapsto \mathrm{R}$ satisfies*

$$|f(x_1, ..., x_i, ..., x_n) - f(x_1, ..., x_i^{'}, ..., x_n)\ | \leq c_i$$

*for all $i \in [n]$ and $\forall x_1, x_2, ..., x_n, x_i{}' \in \mathcal{X}$. Then for all $\epsilon > 0$ we have*

$$\mathbb{P}(|f(x_1, ..., x_n) - \mathbb{E}[f(x_1, ..., x_n)]\ | \geq \epsilon) \leq \exp(\frac{-2\epsilon^2}{\sum_{i=1}^{n} c_i^2}).$$

## 2.3 VC Theory (Uniform Convergence Theory for ERM)

**Binary classification** We consider learning a hypothesis $f$ from $n$ data points $(x_1, y_1), \ldots, (x_n, y_n)$ sampled from $\mathcal{D}$, where $x_i \in \mathbb{R}^d, y_i \in \{\pm 1\}$. Training error can thus be written as $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[y_i \neq f(x_i)]$. We can also represent test error as $\mathbf{P}_{(x,y)\sim\mathcal{D}}(y \neq f(x))$.

Notice that $\mathbb{E}[\mathbb{1}(y_i \neq f(x_i))] = \mathbf{P}_{(x,y)\sim\mathcal{D}}(y \neq f(x))$. Generalization gap measures the gap between training loss and population loss. Fix $f$, $\mathbb{1}[y_i \neq f(x_i)]$ are iid Bernoulli variables. Thus we can show by Theorem 2.2.5 that for any $\epsilon > 0$

$$\mathbb{P}\left(\mathbf{P}_{(x,y)\sim\mathcal{D}}(y \neq f(x)) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[y_i \neq f(x_i)] \geq \epsilon\right) \leq e^{-2n\epsilon^2}.$$

This inequality seems to be conflicted with overfitting phenomenon. Actually, the function $\widehat{f}$ is learned depending on the training data so that $\mathbb{1}[y_i = \widehat{f}(x_i)]$ are not independent. We therefore cannot bound the error and may suffer from overfitting.

## References

[1] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[2] Hang Li. Statistical learning method. *Tsinghua university press*, pages 95–115, 2012.

[3] Rostamizadeh.A Mohri.M and TALWALKAR.A. *Foundations of Machine Learning*. MIT Press, 2012.

[4] Wikipedia. Concentration inequality. `https://en.wikipedia.org/wiki/Concentration_inequality`, 2021. [Online].

[5] Wikipedia. Inequalities in information theory. `https://en.wikipedia.org/wiki/Inequalities_in_information_theory`, 2021. [Online].

[6] Wikipedia. Vapnik–Chervonenkis theory. `https://en.wikipedia.org/wiki/Vapnik%E2%80%93Chervonenkis_theory`, 2021. [Online].