CMPUT 654 Fa 23: Theoretical Foundations of Machine Learning **Fall 2023**

Lecture 3: Measure concentration: MGFs, SG, Hoeffding (Sept. 12)

Lecturer: Csaba Szepesvári *Scribes: Shivam Garg*

Note: *LaTeX template courtesy of UC Berkeley EECS dept. (link to directory)*
Disclaimer: *These notes have **not** been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

As before, our goal is to answer the question of how to evaluate a given algorithm. Recall that $\mathcal{X}$ and $\mathcal{Y}$ represent the space of the inputs and outputs from which we sample n data points $(X_1, Y_1), \dots, (X_n, Y_n) \sim P$ iid. Given a function $f : \mathcal{X} \to \mathcal{Y}$ and a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$, we define the empirical loss for f to be

$$L_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i)$$

and its expected loss (the quantity we finally care about) to be

$$L(f) = \int \ell(f(x), y) P(dx, dy).$$

Then our goal reduces to answering whether (and when) $L_n(f)$ is a good estimate of $L(f)$? One desirable property of L_n is that it is unbiased: $\mathbb{E}[L_n(f)] = L(f)$. Another desirable property to have could be

$$\mathbb{P}(|L_n(f) - L(f)| \geq \varepsilon) \leq \delta(n, \varepsilon), \qquad \text{for all } \varepsilon > 0,$$

where $\delta(n, \varepsilon)$ is a small quantity, say for instance, $\exp(-n\varepsilon^2/(2\sigma^2))$. To obtain such guarantees, we will use the concentration of measure phenomenon.

3.1 Concentration of Measure

In this section, we discuss the concentration of measure phenomenon for subgaussian random variables. Before doing that, let us recall the definition of the moment generating function (MGF) and list some of its properties.
Moment generating function: For a random variable X, its moment generating function is defined as

$$M_X(\lambda) := \mathbb{E}[\exp(\lambda X)],$$

for all $\lambda \in D$, where $D := \{\lambda \in \mathbb{R} : \text{ the expectation } \mathbb{E}[\exp(\lambda X)] \text{ exists}\}$.[1] The following properties hold true for a random variable drawn from any distribution:

(a) D is a convex subset of $\mathbb{R}$,

(b) $M_X(0) = 1$, which also implies that zero belongs to the set D,[2]

(c) (**should we mention something like this theorem is valid:** differentiating under the integral sign) $M'_X(\lambda) = \frac{d}{d\lambda}\mathbb{E}[\exp(\lambda X)] = \mathbb{E}[\frac{d}{d\lambda}\exp(\lambda X)] = \mathbb{E}[X \exp(\lambda X)]$, and

[1] Note that the expectation $\mathbb{E}[X]$ does not hold for all random variables X; for instance, the mean of heavy tailed distributions, such as the Cauchy distribution, does not exist.

[2] ??? **Were there some additional conditions here?**

(d) $M_X^{(k)}(\lambda) = \mathbb{E}[X^k \exp(\lambda X)]$, which directly implies that $\mathbb{E}[M_X^{(k)}(\lambda)] = \mathbb{E}[X^k]$ (hence the name "moment generating" function).

The logarithm of MGF is known as the cumulant generating function defined as

$$\psi_X := \log M_X(\lambda),$$

and is convex. Let us come back to discussing subgaussian random variables now. First, recall the definition:

Definition 3.1. The random variable X is said to be σ-subgaussian, if

$$M_X(\lambda) := \mathbb{E}[\exp(\lambda X)] \le \exp(\lambda^2\sigma^2/2), \qquad \text{for all } \lambda \in \mathbb{R}$$

or equivalently

$$\psi_X(\lambda) := \log M_x(\lambda) \le \lambda^2\sigma^2/2, \qquad \text{for all } \lambda \in \mathbb{R}.$$

If X is a σ-subgaussian distribution, it can be shown that $\mathbb{E}[X] = 0$ and $\mathbb{V}[X] \le \sigma^2$. Further, for all $\varepsilon > 0$,

$$\mathbb{P}(X \ge \varepsilon) \le \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right) \qquad \text{and} \qquad \mathbb{P}(X \le -\varepsilon) \le \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right).$$

The above display is equivalent to the following: for all $\delta > 0$,

$$\mathbb{P}\Big(X \le \sigma\sqrt{2\log(1/\delta)}\Big) \ge 1-\delta \qquad \text{and} \qquad \mathbb{P}\Big(X \ge -\sigma\sqrt{2\log(1/\delta)}\Big) \ge 1-\delta.$$

The above equations are called "one-tailed" bounds, and can be combined (using union bound) to obtain the following "two-tailed" bound:

$$\text{w.p. } 1-\delta: \quad X \in \Big[-\sigma\sqrt{2\log(2/\delta)}, +\sigma\sqrt{2\log(2/\delta)}\Big]. \tag{3.1}$$

Proposition 3.2. *Let X be σ-subgaussian, X_1 be σ_1-subgaussian, and X_2 be σ_2-subgaussian random variables. Also assume $X_1 \perp X_2$, i.e. they are independent of each other. Then*

(a) for all $c \in \mathbb{R}$, cX is $(|c|\sigma)$-subgaussian, and

(b) the random variable $X_1 + X_2$ is $\big(\sqrt{\sigma_1^2+\sigma_2^2}\big)$-subgaussian.[3]

Proof. We will bound the MGFs of the random variables cX and $X_1 + X_2$ to obtain the desired results. For part (a), note that

$$M_{cX}(\lambda) = \mathbb{E}[\exp(\lambda(cX))] = \mathbb{E}[\exp((\lambda c)X)] \le \exp((\lambda c)^2\sigma^2/2) = \exp(\lambda^2(|c|\sigma)^2/2),$$

where the inequality follows from the σ-subgaussianity of X. For part (b), note that

$$\begin{aligned} M_{X_1+X_2}(\lambda) &= \mathbb{E}[\exp(\lambda(X_1+X_2))] = \mathbb{E}[\exp(\lambda X_1)\cdot\exp(\lambda X_2)] \\ &= \mathbb{E}[\exp(\lambda X_1)]\cdot\mathbb{E}[\exp(\lambda X_2)] && \text{(since } X_1 \perp X_2\text{)} \\ &\le \exp(\lambda^2\sigma_1^2/2)\cdot\exp(\lambda\sigma_2^2/2) = \exp(\lambda^2(\sigma_1^2+\sigma_2^2)/2), \end{aligned}$$

and the result follows. □

Using the above result n times, we obtain the following corollary:

[3] How would this result change if $X_1 \not\perp X_2$ (for instance, say, $X_1 = X_2$)?

Corollary 3.3. *For n iid σ-subgaussian random variables $X_1, \dots, X_n$, their mean $\bar{X}_n := (\sum_{i=1}^n X_i)/n$ is $(\sigma/\sqrt{n})$-subgaussian.*

Remark 3.4. We can use the above results to answer our question of evaluating an algorithm by assuming the loss of a function to be subgaussian. For instance, consider the problem of comparing K different functions $f_1, \dots, f_n$ using n different data points $\{(X_i, Y_i)\}_{i\in[n]}$. Define $X_i^{(j)} := \ell(f_j(X_i), Y_i) - L(f_j)$, for $j \in [K]$. Assume $X_i^{(j)}$s to be σ-subgaussian random variables. Also, let $\bar{X}_n^{(j)} := (\sum_{i=1}^n X_i^{(j)})/n$. Then, using the previous corollary and Eq. 3.1, we get $\mathbb{P}\big(|\bar{X}_n^{(j)}| \geq \sigma\sqrt{2\log(2K/\delta)/n}\big) \leq \delta/K$. Combining this inequality for all the K functions (by using union bound) then gives us:

$$\mathbb{P}\left(\max_{j\in[K]} |L_n(f_j) - L(f_j)| \geq \sigma\sqrt{\frac{2\log(2/\delta) + \log K}{n}}\right) \leq \delta.$$

This inequality says that the empirical losses of all these functions are close to their true means. Note that the factor K comes inside a logarithm, whereas n comes outside of it. From this result, we observe that the sample size n doesn't need to grow too fast as the number of functions being compared K grows.

Note that the subgaussianity assumption is not necessarily too restrictive in practice. For instance, a bounded zero-mean random variable is subgaussian, and thus all the above results apply to bounded random variables as well.

Lemma 3.5 (Hoeffding's)**.** *Let $a, b \in \mathbb{R}$ and $b \geq a$. If $X \in [a, b]$ and $\mathbb{E}[X] = 0$, then X is $\big(\frac{b-a}{2}\big)$-subgaussian.*

Proof. We will show this by bounding the cumulant generating function ψ_X of X. Fix $\lambda \in \mathbb{R}$. Then by Taylor's theorem with remainder, there exists $\tilde{\lambda} \in [0, \lambda]$, such that

$$\psi_X(\lambda) = \psi_X(0) + \psi_X'(0)\cdot\lambda + \psi_X''(\tilde{\lambda})\cdot\frac{\lambda^2}{2}.$$

Note that $\psi_X(0) = \log M_X(0) = 0$. Also note that $\psi_X'(\lambda) = \frac{d}{d\lambda}M_X(\lambda) = \frac{M_X'(\lambda)}{M_X(\lambda)}$, which along with the zero mean assumption on X implies that $\psi_X'(0) = M_X'(0) = \mathbb{E}[X] = 0$. Therefore, the previous display reduces to

$$\psi_X(\psi) = \psi_X''(\tilde{\lambda})\cdot\frac{\lambda^2}{2}, \qquad \text{for some } \tilde{\lambda} \in [0, \lambda]. \tag{3.2}$$

All we need to do now is to bound $\psi_X''(\tilde{\lambda})$. To do this, let $X \sim P$, and define a distribution Q as follows:

$$Q(dx) := \frac{\exp(\tilde{\lambda}x)\cdot P(dx)}{\int \exp(\tilde{\lambda}x)P(dx)} = \frac{\exp(\tilde{\lambda}x)}{\exp(\psi_X(\tilde{\lambda}))}P(dx) = \exp\big(\tilde{\lambda}x - \psi_X(\tilde{\lambda})\big)\cdot P(dx).$$

Let the random variable $Z \sim Q$. Note that this means that Z would be bounded between $[a, b]$ a.s. (since $\int_a^b Q(dx) = 1$). Since, Z is bounded, its variance bounded by: $\mathbb{V}_Q[Z] \leq \big(\frac{b-a}{2}\big)^2$; indeed,

$$\mathbb{V}_Q[Z] = \mathbb{E}_Q[(Z - \mathbb{E}_Q[Z])^2] \overset{\text{(why?)}}{=} \arg\min_c \mathbb{E}_Q[(Z-c)^2] \leq \mathbb{E}_Q[(Z - (a+b)/2)^2] \leq \left(\frac{b-a}{2}\right)^2$$

(also see this). Finally, note that

$$\begin{aligned}\psi_X''(\tilde{\lambda}) &= \frac{d}{d\tilde{\lambda}}\psi_X'(\tilde{\lambda}) = \frac{d}{d\tilde{\lambda}}\frac{M_X'(\tilde{\lambda})}{M_X(\tilde{\lambda})} = \frac{M_X''(\tilde{\lambda})M_X(\tilde{\lambda}) - M_X'(\tilde{\lambda})^2}{M_X(\tilde{\lambda})^2} = \frac{M_X''(\tilde{\lambda})}{M_X(\tilde{\lambda})} - \left(\frac{M_X'(\tilde{\lambda})}{M_X(\tilde{\lambda})}\right)^2 \\ &= \frac{\mathbb{E}_P[X^2\exp(\tilde{\lambda}X)]}{M_X(\tilde{\lambda})} - \left(\frac{\mathbb{E}_P[X\exp(\tilde{\lambda}X)]}{M_X(\tilde{\lambda})}\right)^2 = \mathbb{E}_Q[Z^2] - \mathbb{E}_Q[Z]^2 = \mathbb{V}_Q[Z].\end{aligned}$$

(Why is $M_X(\tilde{\lambda}) \neq 0$? And we needed to differentiate under the integral sign twice, so need to verify some properties.) The above equation along with Eq. 3.2 implies that

$$\psi_X(\psi) = \psi_X''(\tilde{\lambda})\frac{\lambda^2}{2} = \mathbb{V}_Q[Z]\frac{\lambda^2}{2} \leq \left(\frac{b-a}{2}\right)^2\frac{\lambda^2}{2},$$

which means that X is $\big(\frac{b-a}{2}\big)$-subgaussian. □

3.2 Bibliography