CSCI699: Theory of Machine Learning Fall 2021

Lecture 4: Concentration Inequalities

Instructor: Vatsal Sharan *Scribe: Ta-Yang Wang*

Today

- Finish concentration inequalities

Last time, we saw how applying Markov's inequality to the exponential function gives Chernoff-style bounds.

$$\log(\Pr[X - \mathbb{E}[X] \geq t]) \leq \inf_{\lambda \geq 0} \left(\log \left(\mathbb{E} \left[e^{\lambda(X - \mathbb{E}[X])} \right] \right) - \lambda t \right) \tag{1}$$

Today, we use this to derive concentration bounds for sums of random variables. We start with just deriving a tail bound for a Gaussian using (1).

1 Gaussian tail bounds

- Let $X \sim \mathcal{N}(\mu, \sigma^2)$
- Goal: Bound $\Pr[X - \mu \geq t]$.

By using (1) and applying the exponential function on both sides,

$$\Pr[X - \mu \geq t] \leq \inf_{\lambda \geq 0} \frac{e^{(\lambda(X - \mathbb{E}[X]))}}{e^{\lambda t}}.$$

Let $X - \mathbb{E}[X] = y$. Note that $y \sim \mathcal{N}(0, \sigma^2)$. We can now derive the moment generating function of the Gaussian,

$$\begin{aligned} \mathbb{E}(e^{\lambda(X - \mathbb{E}[X])}) &= \mathbb{E}(e^{\lambda y}) \\ &= \int_{\mathbb{R}} e^{\lambda y} \frac{e^{-y^2/2\sigma^2}}{\sqrt{2\pi}\sigma} \, \mathrm{d}y \\ &= \int_{\mathbb{R}} \frac{e^{-(y - \lambda/\sigma^2)^2 \cdot \frac{1}{2\sigma^2}} e^{\lambda^2 \sigma^2 / 2}}{\sqrt{2\pi}\sigma} \, \mathrm{d}y \\ &= e^{\lambda^2 \sigma^2 / 2} \int_{\mathbb{R}} \frac{e^{-(y - \lambda/\sigma^2) \cdot \frac{1}{2\sigma^2}}}{\sqrt{2\pi}\sigma} \, \mathrm{d}y \\ &= e^{\frac{\lambda^2 \sigma^2}{2}}. \end{aligned}$$

Using this, we can write,

$$\log(\Pr[X - \mathbb{E}[X] \geq t]) \leq \inf_{\lambda \geq 0} \left(\frac{\lambda^2 \sigma^2}{2} - \lambda t \right) = \frac{-t^2}{2\sigma^2} \quad \begin{cases} \text{[Exercise: minimize the quadratic} \\ \text{to show the equality]} \end{cases}$$

$$\implies \Pr[X - \mathbb{E}[X] \geq t] \leq e^{-t^2/2\sigma^2}.$$

For the case of a Gaussian distribution, we could have derived a tail bound directly without using the moment generating function or the Chernoff-bound. But we didn't loose much, the bound above is in fact tight up to polynomial factors, i.e., the best bound one could hope to show is $\Pr[X - \mathbb{E}[X] \geq t] \leq e^{-t^2/2\sigma^2} \cdot (\text{poly}(t, \sigma))^{-1}$.

Let us see what this implies for the sum of Gaussians. Let

$$\overline{X}_n = \sum_{i=1}^{n} X_i/n, \quad \mathbb{E}[X_i] = \mu, \quad \text{Var}[X_i] = \sigma^2.$$

Since the Gaussian distribution is a stable distribution, the sum of Gaussians is also a Gaussian, therefore $\overline{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$. Therefore using the tail bound we derived for a single Gaussian,

$$\Pr[\overline{X}_n - \mu \geq t] \leq e^{-nt^2/2\sigma^2}.$$

If we now want to set the failure probability to be δ, then the deviation expected from the true mean given n samples is about $\mathcal{O}\left(\sigma\sqrt{\frac{\log(1/\delta)}{n}}\right)$:

$$\Pr\left[\overline{X}_n - \mu \geq \sigma\sqrt{\frac{2\log(1/\delta)}{n}}\right] \leq \delta.$$

2 Sub-Gaussian random variables

We didn't really use too many special properties of the Gaussian to get the previous tail bound. All we really need is a bound on the moment generating function, and the notion of *sub-Gaussian random variables* formalizes this.

Definition 1. *A random variable X with mean $\mu = \mathbb{E}(X)$ is **sub-Gaussian** if there exists a positive number σ such that*

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\sigma^2\lambda^2/2} \quad \forall\, \lambda \in \mathbb{R}.$$

σ is known as the sub-Gaussian parameter (think of this as the variance proxy for the distribution).

- Any Gaussian random variable with σ^2 is sub-Gaussian with parameter σ.
- Many non-Gaussian random variables also have this property! Let's see one example.

Rademacher random variable

Let X be the random variable which is $\{-1, +1\}$ with equal probability. This random variable is known as a *Rademacher random variable.*

Claim 2. *The Rademacher random variable is sub-Gaussian with parameter $\sigma = 1$.*

Proof. Note that $\mathbb{E}(X) = 0$. We can write,

$$\begin{aligned}
\mathbb{E}[e^{\lambda X}] &= \frac{1}{2}\left(e^{-\lambda} + e^{\lambda}\right) \\
&= \frac{1}{2}\left(\sum_{k=0}^{\infty}\frac{(-\lambda)^k}{k!} + \sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\right) \\
&= \sum_{k=0}^{\infty}\frac{\lambda^{2k}}{(2k)!} \\
&\leq 1 + \sum_{k=1}^{\infty}\frac{\lambda^{2k}}{2^k k!} \quad ((2k)! \geq 2^k k!) \\
&= e^{\lambda^2/2}.
\end{aligned}$$

□

2.1 Concentration bound for sub-Gaussian random variables

We now show that a sub-Gaussian random variable with sub-Gaussian parameter σ has similar tail upper bounds as a Gaussian random variable with standard deviation σ,

Lemma 3. *If X is a sub-Gaussian with parameter σ, then for any $t > 0$,*

(1) $\Pr[X > \mathbb{E}[X] + t] \leq e^{-t^2/2\sigma^2}$.

(2) $\Pr[X < \mathbb{E}[X] - t] \leq e^{-t^2/2\sigma^2}$.

(3) $\Pr[|X - \mathbb{E}[X]| \geq t] \leq 2e^{-t^2/2\sigma^2}$.

Proof. (1) Using (1),

$$\log(\Pr[X - \mathbb{E}[X] \geq t]) \leq \inf_{\lambda \geq 0}(\log \mathbb{E}(e^{\lambda(X - \mathbb{E}(X))}) - \lambda t).$$

By sub-Gaussianity,

$$\mathbb{E}[e^{\lambda(X - \mathbb{E}[X])}] \leq e^{\sigma^2\lambda^2/2}.$$

Therefore,

$$\log(\Pr[X - \mathbb{E}[X] \geq t]) \leq \inf_{\lambda \geq 0}\left(\frac{\sigma^2\lambda^2}{2} - \lambda t\right) = \frac{-t^2}{2\sigma^2}$$

and part (1) follows.

(2) Take $X' = -X$. Note that X' is also sub-Gaussian with parameter σ (by symmetry of the definition with respect to λ, i.e. the definition requires the inequality to be satisfied both for $\lambda \geq 0$ and $\lambda \leq 0$). We can now write,

$$\Pr[X < \mathbb{E}[X] - t] = \Pr[X' > \mathbb{E}[X'] + t].$$

The result now follows by part (1).

(3) Follows from a union bound.

□

2.2 Concentration bound for sum of sub-Gaussian random variable

As we did for Gaussians, we can now get a tail-bound for sums of independent sub-Gaussian random variables. The crucial property we will use here is that *sums of independent sub-Gaussian random variables are sub-Gaussian.*

Theorem 4. *Suppose that the random variable $\{X_i\}_{i=1}^n$ are independent and $\mathbb{E}[X_i] = \mu_i$ and X_i is sub-Gaussian with parameter σ_i. Then for all $t \geq 0$, we have*

$$\Pr\left[\left|\sum_{i=1}^{n}(X_i - \mu_i)\right| \geq t\right] \leq 2\exp\left(\frac{-t^2}{2\sum_{i=1}^{n}\sigma_i^2}\right)$$

Proof. We show that $Z = \sum_{i=1}^{n} X_i$ is sub-Gaussian with parameter $\sigma_Z^2 = \sum_{i=1}^{n} \sigma_i^2$.

$$\begin{aligned}
\mathbb{E}\left(e^{\lambda(Z-\mathbb{E}[Z])}\right) &= \mathbb{E}\left(e^{\lambda(\sum X_i - \mathbb{E}[\sum X_i])}\right) \\
&= \prod_{i=1}^{n} \mathbb{E}\left(e^{\lambda(X_i - \mathbb{E}[X_i])}\right) \quad \text{(independence)} \\
&\leq \prod_{i=1}^{n} e^{\sigma_i^2\lambda^2/2} = e^{(\sum \sigma_i^2)\lambda^2/2}.
\end{aligned}$$

The Theorem now follows from Lemma 3. □

We state the following direct Corollary of the above theorem for the case where the random variables X_i have the same mean and variance.

Corollary 5 (Corollary of Theorem 4 for identical random variables)**.** *Consider n i.i.d. random variable $X_1, \ldots, X_n$ each with $\mathbb{E}[X_i] = \mu$ and sub-Gaussian parameter σ. Then for all $t \geq 0$,*

$$\Pr\left[\left|\frac{\sum_{i=1}^{n} X_i}{n} - \mu\right| \geq \epsilon\right] \leq \exp\left(\frac{-n\epsilon^2}{2\sigma^2}\right).$$

2.3 Hoeffding's inequality

We now use the tools we have developed to show Hoeffding's inequality, which is a concentration bound for bounded random variables. We first show that bounded random variables are sub-Gaussian.

Claim 6 (Bounded random variable is sub-Gaussian)**.** *Let X be a random variable with mean zero and supported on the interval $[a, b]$. Then X is sub-Gaussian with sub-Gaussian parameter $(b - a)$ (in fact, it is possible to show sub-Gaussianity with parameter $(b - a)/2$, but we won't do that here).*

Proof. The proof uses the idea of *symmetrization*, which we'll see again in future lectures. Let X' be an independent copy of X. Note that since $\mathbb{E}[X'] = 0$, we can write,

$$\mathbb{E}_X[e^{\lambda X}] = \mathbb{E}_X[\exp(\lambda(X - \mathbb{E}[X']))] \leq \mathbb{E}_{X,X'}[\exp(\lambda(X - X'))]$$

where the last step uses Jensen's inequality, $f(\mathbb{E}[X]) \leq \mathbb{E}(f(X))$ for any convex f. Note that $X - X'$ and $X' - X$ have the same distribution. This implies

$$\mathbb{E}_{X,X'}[\exp(\lambda(X - X'))] = \mathbb{E}_{X,X'}[\exp(\lambda(X' - X))].$$

Therefore, we can introduce a Rademacher random variable ϵ, which $\{\pm 1\}$ with probability 1/2 each, without changing the expectation,

$$\mathbb{E}_{X,X'}[\exp(\lambda(X - X'))] = \mathbb{E}_{X,X',\epsilon}[\exp(\lambda\epsilon(X - X'))].$$

Now fixing X, X' and just taking the expectation over ϵ, we can use the moment generating function bound for Rademacher random variables in Claim 2 to write,

$$\mathbb{E}_\epsilon[\exp(\lambda\epsilon(X - X'))] \leq \exp(\lambda^2(X - X')^2/2).$$

Plugging this back,

$$\mathbb{E}[e^{\lambda X}] \leq \mathbb{E}_{X,X'}[\exp(\lambda^2(X - X')^2/2)] \leq \exp(\lambda^2(b - a)^2/2)$$

where last inequality follows since X, X' lie in the interval $[a, b]$. □

Using Theorem 4 we now get a slightly weaker version of Hoeffding's inequality.

Lemma 7 (weaker version of Hoeffding's)**.** *Let $X_1, X_2, \ldots, X_n$ be independent random variables such that $a_i \leq x_i \leq b_i$ for each $i \in [n]$. Then for any $\epsilon > 0$,*

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

Remark: This bound is only loose compared to Hoeffding's inequality stated earlier in terms of a factor of 2 in the denominator of the exponent, which should instead be in the numerator. If we use the right sub-Gaussian parameter $(b_i - a_i)/2$ for the random variables, we'll recover Hoeffding's inequality as stated earlier.

So far, we've only worked with sums of random variables. There is a very rich and vast literature on showing that various other functions of random variables are also close to their expectation with high probability. A statement of this form is McDiarmid's inequality, which we will use in a few lectures.

Theorem 8 (McDiarmid's inequality)**.** *Let $X_1, X_2, \ldots, X_n$ be independent random variables taking values over some domain $\mathcal{X}$. A function $f : \mathcal{X}^n \to \mathbb{R}$ satisfies the* ***bounded differences property*** *if:*

$$\forall\, i \in [n], \quad \forall\, x_1, \ldots, x_n, x_i' \in \mathcal{X},$$

$$|f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \leq c_i.$$

If f satisfies this property, then

$$\Pr(f(x_1, \ldots, x_n) - \mathbb{E}[f(x_1, \ldots, x_n)] > \epsilon) \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right).$$

Remark: Note that this gives us back Hoeffding's inequality by taking $f = \dfrac{\sum X_i}{n}$, and realizing that f satisfies the bounded differences property with $c_i = \dfrac{b_i - a_i}{n} \quad \forall\, i$.