CS 388R: Randomized Algorithms, Fall 2023 October 23rd, 2023

Lecture 17: Concentration Inequalities Revisited

Prof. Eric Price *Scribe: Ruichen Jiang*

NOTE: THESE NOTES HAVE NOT BEEN EDITED OR CHECKED FOR CORRECTNESS

1 Overview

In the last lecture, we discussed matrix concentration inequalities as a preliminary for the graph sparsification problem.

In this lecture, we continue our journey in concentration inequalities. Specifically, we show a useful technique that derives tail probability bounds from moment generating functions, introduce subgaussian and subgamma random variables, and finally discuss some applications.

2 Moment Generating Function

Recall **Markov's inequality**: if a random variable X is nonnegative, then we have

$$\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}[X]}{t}.$$

One important corollary is **Chebyshev's inequality**. Let X be a random variable with mean μ and variance σ^2. By applying Markov's inequality to $(X-\mu)^2 \geq 0$, we can obtain

$$\mathbb{P}[X-\mu \geq t] \leq \mathbb{P}[(X-\mu)^2 \geq t^2] \leq \frac{\mathbb{E}[(X-\mu)^2]}{t^2} = \frac{\sigma^2}{t^2}.$$

As a result, we can conclude that $X \leq \mu + \frac{\sigma}{\sqrt{\delta}}$ with probability at least $1-\delta$.

Is Chebyshev's inequaltiy tight enough? On the one hand, if δ is a constant, this is probably the best we can get: with a constant probability, we would expect X to deviate from the mean by σ. On the other hand, when δ tends to 0, the upper bound $\mu + \frac{\sigma}{\sqrt{\delta}}$ grows polynoimally, which can be undesirable.

To obtain tighter concentration bounds, we will rely on the **moment generating function (MGF)**. The MGF of a random variable X is defined as

$$\phi_X(\lambda) := \mathbb{E}[e^{\lambda(X-\mu)}].$$

To obtain a tail bound inequality, we can use a similar argument as in the derivation of Chebyshev's inequality. Specifically, for $\lambda \geq 0$, since $x \mapsto e^{\lambda x}$ is an increasing function, we have

$$\mathbb{P}[X-\mu \geq t] = \mathbb{P}[e^{\lambda(X-\mu)} \geq e^{\lambda t}] \leq \frac{\mathbb{E}[e^{\lambda(X-\mu)}]}{e^{\lambda t}} = \frac{\phi_X(\lambda)}{e^{\lambda t}}.$$

Note that this holds for any $\lambda \geq 0$. Hence, to get the best bound, we can try to minimize the right-hand side w.r.t. λ, leading to

$$\mathbb{P}[X - \mu \geq t] \leq \min_{\lambda \geq 0} \frac{\phi_X(\lambda)}{e^{\lambda t}}. \tag{1}$$

Example: Let's see what the above implies when X is a Gaussian random variable. Let $X \sim N(\mu, \sigma^2)$. Then we can compute its MGF explicitly as follows:

$$\begin{aligned}
\Phi_X(\lambda) = \mathbb{E}[e^{\lambda(X-\mu)}] &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{t^2}{2\sigma^2}} e^{\lambda t} \, dt \\
&= e^{\frac{\sigma^2\lambda^2}{2}} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\sigma^2\lambda)^2}{2\sigma^2}} \, dt \\
&= e^{\frac{\sigma^2\lambda^2}{2}},
\end{aligned}$$

where we used $-\frac{t^2}{2\sigma^2} + \lambda t = -\frac{(t-\sigma^2\lambda)^2}{2\sigma^2} + \frac{\sigma^2\lambda^2}{2}$ in the third equality. Hence, from (1) we further have

$$\mathbb{P}[X - \mu \geq t] \leq \min_{\lambda \geq 0} e^{\frac{\sigma^2\lambda^2}{2}} e^{-\lambda t} = \min_{\lambda \geq 0} e^{\frac{1}{2}(\sigma\lambda - \frac{t}{\sigma})^2} \cdot e^{-\frac{t^2}{2\sigma^2}} = e^{-\frac{t^2}{2\sigma^2}},$$

where the minimum is achieved by $\lambda = \frac{t}{\sigma^2}$. In fact, following similar arguments we can also prove that

$$\mathbb{P}[X - \mu \leq -t] \leq e^{-\frac{t^2}{2\sigma^2}}.$$

3 Subgaussian Random Variables

Notice that in the example above, Gaussianity is not essential: the same concentration inequalities still hold so long as $\phi_X(\lambda) \leq e^{-\frac{t^2}{2\sigma^2}}$. This motivates the definition of subgaussian random variables.

Definition 1. *A random variable X is* ***subgaussian*** *with variance proxy σ^2 if*

$$\forall \lambda: \quad \phi_X(\lambda) \leq e^{\frac{\sigma^2\lambda^2}{2}}. \tag{2}$$

By following the exact same argument as in the Gaussian case, we obtain the following tail probability bounds.

Proposition 2. *If X is subgaussian with variance proxy σ^2, then we have $\mathbb{P}[X \geq \mu + t] \leq e^{-\frac{t^2}{2\sigma^2}}$ and $\mathbb{P}[X \leq \mu - t] \leq e^{-\frac{t^2}{2\sigma^2}}$.*

As a corollary of Proposition 2, we have $|X - \mu| \leq \sigma\sqrt{2\log\frac{2}{\delta}}$ with probability at least $1 - \delta$.

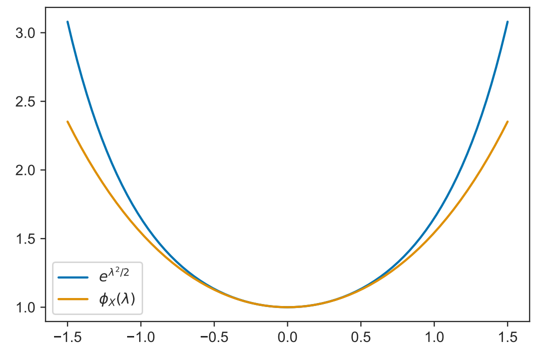

Figure 1: MGF of the Bernoulli random variable.

Example: Let X be a Bernoulli random variable with $\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = \frac{1}{2}$. We can compute that $\phi_X(\lambda) = \frac{1}{2}(e^{\lambda} + e^{-\lambda})$ As we observe in Fig. 1, the MGF of X is upper bounded by $e^{\frac{\sigma^2\lambda^2}{2}}$. Hence, by definition, X is subgaussian with $\sigma^2 = 1$. More generally, one can show that:

Lemma 3. *If $X \in [-1, 1]$ almost surely, then X is subgaussian with $\sigma^2 = 1$. Moreover, if $X \in [-a, b]$ almost surely, then X is subgaussian with $\sigma^2 = (\frac{b-a}{2})^2$.*

A particular convenient property of subgaussian random variables is the following composition rule.

Proposition 4. *Suppose X_1 and X_2 are subgaussian with variance proxy σ_1^2 and σ_2^2, respectively.*

- *If X_1 and X_2 are independent, then $X_1 + X_2$ is subgaussian with variance proxy $\sigma_1^2 + \sigma_2^2$.*
- *If X_1 and X_2 are not independent, then X_1+X_2 is subgaussian with variance proxy $(\sigma_1+\sigma_2)^2$.*

Proof. We only prove the first result as the second one is not very useful in practice. Using independence, we can compute the MGF of $X_1 + X_2$ by

$$\mathbb{E}[e^{\lambda(X_1+X_2)}] = \mathbb{E}[e^{\lambda X_1}]\,\mathbb{E}[e^{\lambda X_2}] \leq e^{\frac{\lambda^2\sigma_1^2}{2}} e^{\frac{\lambda^2\sigma_2^2}{2}} = e^{\lambda^2\frac{\sigma_1^2+\sigma_2^2}{2}}.$$

Thus, by definition, $X_1 + X_2$ is subgaussian with variance proxy $\sigma_1^2 + \sigma_2^2$. □

With the results above, we can derive the additive Chernoff bound covered in Lecture 2.

Theorem 5. *Suppose $X_1, \ldots, X_n \in [0, 1]$ are independent and let $\mu = \sum_{i=1}^n \mathbb{E}[X_i]$. Then*

$$\mathbb{P}\left[\sum_{i=1}^n X_i \geq \mu + t\right] \leq e^{-\frac{2t^2}{n}}.$$

Proof. Note that X_i is subgaussian with $\sigma_i^2 = \frac{1}{4}$ by Lemma 3. Thus, by Proposition 4, $\sum_{i=1}^n X_i$ is subgaussian with $\sigma^2 = \frac{1}{4}n$. The theorem now directly follows from Proposition 2. □

Finally, we mention some other characterizations of subgaussian random variables. Up to constant factors, the following statements are equivalent:

- **(MGF bound)** X is subgaussian with variance proxy σ^2, i.e., $\phi_X(\lambda) \le e^{\frac{\sigma^2\lambda^2}{2}}$;
- **(Tail probability bound)** $\mathbb{P}[|X-\mu| \ge t] \le e^{-\frac{t^2}{2\sigma^2}}$;
- **(Moment bound)** $\mathbb{E}[|X-\mu|^k] \le \sigma^k k^{k/2}$ for any positive integer k.

4 Subgamma Random Variables

Not all random variables are subgaussian. As a motivating example, let $X \sim N(0,1)$ and consider the random variable X^2. Note that $\mathbb{E}[X^2] = 1$, and we can also compute its MGF explicitly by

$$\begin{aligned}
\phi_{X^2}(\lambda) = \mathbb{E}[e^{\lambda(X^2-1)}] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{\lambda(x^2-1)} e^{-\frac{x^2}{2}} \, dx \\
&= \frac{1}{\sqrt{2\pi}} e^{-\lambda} \int_{-\infty}^{+\infty} e^{(\lambda-1/2)x^2} \, dx \\
&= \frac{e^{-\lambda}}{\sqrt{1-2\lambda}}.
\end{aligned}$$

Notice that $\phi_{X^2}(\lambda) \to \infty$ when $\lambda \to 1/2$, and hence it cannot satisfy the condition in (2). On the other hand, around the origin 0, the MGF does not grow too fast. In fact, we can numerically observe that

$$\phi_{X^2}(\lambda) = \frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \le e^{4\cdot\frac{\lambda^2}{2}} \quad \forall|\lambda| < \frac{1}{3}.$$

To generalize this observation, we introduce the definition of subgamma random variables.

Definition 6. *A random variable X is **subgamma** with parameters (σ^2, c) if*

$$\forall|\lambda| \le \frac{1}{c}: \quad \phi_X(\lambda) \le e^{\frac{\sigma^2\lambda^2}{2}}.$$

Some examples:

- X^2 where $X \sim N(0,1)$ is $(4,3)$-subgamma;
- σ^2-subgaussian is $(\sigma^2, 0)$-subgamma.

Similar to Proposition 2, we can derive the following concentration result for subgamma random variables.

Proposition 7. *If the random variable X is (σ^2, c)-subgamma, then we have*

$$\mathbb{P}[X-\mu \ge t] \le \max\left\{e^{-\frac{t^2}{2\sigma^2}}, e^{-\frac{t}{2c}}\right\} \quad \textit{and} \quad \mathbb{P}[X-\mu \le -t] \le \max\left\{e^{-\frac{t^2}{2\sigma^2}}, e^{-\frac{t}{2c}}\right\}.$$

Proof. We use the similar MGF trick. For any $\lambda \in (0, 1/c)$, we can bound

$$\mathbb{P}[X - \mu \geq t] \leq \frac{\mathbb{E}[e^{\lambda(t-\mu)}]}{e^{\lambda t}} \leq e^{\frac{\lambda^2\sigma^2}{2} - \lambda t} = e^{\frac{1}{2}(\lambda\sigma - \frac{t}{\sigma})^2} e^{-\frac{t^2}{2\sigma^2}}.$$

If there were no constraints on λ, then the bound above would be minimized by $\lambda = t/\sigma^2$. However, we need to ensure that $0 \leq \lambda \leq 1/c$. To this end, we consider two cases:

1. If $t/\sigma^2 \leq 1/c$, then we can set $\lambda = t/\sigma^2$ and obtain $\mathbb{P}[X - \mu \geq t] \leq e^{-\frac{t^2}{2\sigma^2}}$.
2. Otherwise, if $t/\sigma^2 > 1/c$, then we set $\lambda = 1/c$. By using $\sigma^2/c^2 < t/c$, we get

$$\mathbb{P}[X - \mu \geq t] \leq e^{\frac{\sigma^2}{2c^2} - \frac{t}{c}} \leq e^{-\frac{t}{2c}}.$$

Hence, we obtain the desired result by combining both cases. The other tail probability bound follows similarly. □

As a corollary of Proposition 7, we have $X \leq \mu + \sigma\sqrt{2\log\frac{1}{\delta}} + c\log\frac{1}{\delta}$ with probability at least $1-\delta$. The term $\sigma\sqrt{2\log\frac{1}{\delta}}$ corresponds to the Gaussian tail, while the term $c\log\frac{1}{\delta}$ corresponds to the exponential tail. When δ is sufficiently small, the second term is the dominant term.

Next, we turn to the composition rule for subgamma random variables.

Proposition 8. *Suppose X_1 and X_2 are independent subgamma random variables with parameter (σ_1^2, c_1) and (σ_2^2, c_2), respectively. Then $X_1 + X_2$ is $(\sigma_1^2 + \sigma_2^2, \max(c_1, c_2))$.*

Using this result, we can derive one of the multiplicative Chernoff bounds covered in Lecture 2. But before that, we first need to introduce the following lemma.

Lemma 9 (Bernstein)**.** *If $|X - \mu| \leq M$ almost surely, then X is $(2\mathrm{Var}(X), 2M)$-subgamma.*

It is interesting to contrast Lemma 9 with Lemma 3. At first glance, it might appear that Lemma 9 is strictly weaker: the tail probability of a subgamma random variable decays at a rate of e^{-t}, whereas the tail probability of a subgaussian random variable decays at a faster rate of e^{-t^2}. The catch is that the parameter σ^2 in Lemma 9 depends on the *actual variance* of the random variable X, while the parameter σ^2 in Lemma 3 is given by the range of X, regardless of its distribution. In particular, if the distribution is skewed (i.e., has a low variance), then probability bounds derived from Lemma 9 could potentially lead to a tighter result.

Now we prove a version of the multiplicative Chernoff bound using Proposition 8 and Lemma 9.

Theorem 10. *Suppose $X_1, \ldots, X_n \in [0, 1]$ are independent and let $\mu = \sum_{i=1}^n \mathbb{E}[X_i]$. Then*

$$\mathbb{P}\left[\sum_{i=1}^n X_i \geq (1+\epsilon)\mu\right] \leq e^{-\frac{\mu}{4}\min\{\epsilon, \epsilon^2\}}.$$

Proof. Let $p_i = \mathbb{E}[X_i]$. Since $X_i \in [0, 1]$, we also have $\mathrm{Var}[X_i] \leq \mathbb{E}[X_i^2] \leq \mathbb{E}[X_i] = p_i$. Thus, by Lemma 9, the random variable X_i is $(2p_i, 2)$-subgamma. Since $X_1, \ldots, X_n$ are independent, we obtain from Proposition 8 that $\sum_{i=1}^n X_i$ is $(2\sum_{i=1}^n p_i = 2\mu, 2)$-subgamma. Using Proposition 7, we conclude that

$$\mathbb{P}[X \geq \mu + t] \leq \max\left\{e^{-\frac{t^2}{4\mu}}, e^{-\frac{t}{4}}\right\}.$$

By taking $t = \epsilon\mu$, we obtain

$$\mathbb{P}[X \geq (1+\epsilon)\mu] \leq e^{-\frac{\mu}{4}\min\{\epsilon,\epsilon^2\}}.$$

□

5 Application: Johnson-Lindenstrauss Transform

Suppose that we are given n points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ in a space of large dimension d. Sometimes, we would like to to reduce the dimension by projecting these points to a smaller subspace, while preserving the relative positions between any two points. The celebrated JL lemma shows that this can be achieved by projecting the points to a random subspace of dimension $m = \mathcal{O}(\log n)$.

Lemma 11 (JL Lemma)**.** *Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be arbitrary n points in $\mathbb{R}^d$. For any $\epsilon \in (0,1)$, there exists $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \in \mathbb{R}^m$ with $m = \mathcal{O}(\frac{1}{\epsilon^2}\log n)$ such that*

$$\|\boldsymbol{y}_i - \boldsymbol{y}_j\|_2 = (1 \pm \epsilon)\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2, \quad \forall i, j. \tag{3}$$

Proof. Let $\mathbf{A} \in \mathbb{R}^{m \times d}$ be a matrix with entries drawn i.i.d. from $N(0, \frac{1}{m})$, and we will show that choosing $\boldsymbol{y}_i = \mathbf{A}\boldsymbol{x}_i$ for $i = 1, \ldots, n$ satisfies the condition in (3). To begin with, we will show that, for any $\boldsymbol{z} \in \mathbb{R}^d$,

$$\mathbb{P}(\|\mathbf{A}\boldsymbol{z}\|^2 \geq (1+\epsilon)\|\boldsymbol{z}\|^2) \leq \exp\left(-\frac{\epsilon^2 m}{8}\right). \tag{4}$$

Note that when $\boldsymbol{z}$ is fixed, we have $\mathbf{A}\boldsymbol{z} \sim N(0, \frac{\|\boldsymbol{z}\|_2}{\sqrt{m}}\mathbf{I}_m)$ and $\mathbb{E}[\|\mathbf{A}\boldsymbol{z}\|^2] = \|\boldsymbol{z}\|_2^2$. Thus, by rescaling, it is sufficient to consider a Gaussian random variable $X \sim N(0, \mathbf{I}_m)$ and prove that

$$\mathbb{P}\left[\|\boldsymbol{X}\|_2^2 \geq (1+\epsilon)\,\mathbb{E}[\|\boldsymbol{X}\|_2^2]\right] \leq \exp\left(-\frac{\epsilon^2 m}{8}\right).$$

Note that $\|\boldsymbol{X}\|_2^2 = \sum_{i=1}^m X_i^2$ and $\mathbb{E}[\|\boldsymbol{X}\|^2] = m$, where $X_i \sim N(0,1)$. Since X_i^2 is $(4,3)$-subgamma, by Proposition 8 we can obtain that $\|\boldsymbol{X}\|_2^2$ is $(4m, 3)$-subgamma. Hence, it follows from Proposition 7 that

$$\begin{aligned}
&\mathbb{P}[\|\boldsymbol{X}\|_2^2 \geq m + t],\ \mathbb{P}[\|\boldsymbol{X}\|_2^2 \leq m - t] \leq \exp\left\{-\min\left(\frac{t^2}{8m}, \frac{t}{6}\right)\right\} \\
\Rightarrow\quad &\mathbb{P}[\|\boldsymbol{X}\|_2^2 \geq (1+\epsilon)m],\ \mathbb{P}[\|\boldsymbol{X}\|_2^2 \leq (1-\epsilon)m] \leq \exp\left\{-\min\left(\frac{\epsilon^2 m}{8}, \frac{\epsilon m}{6}\right)\right\} = \exp\left(-\frac{\epsilon^2 m}{8}\right).
\end{aligned}$$

Now note that (3) is equivalent to $\|\mathbf{A}(\boldsymbol{x}_i - \boldsymbol{x}_j)\| = (1 \pm \epsilon)\|\boldsymbol{x}_i - \boldsymbol{x}_j\|$ for all $1 \leq i, j \leq n$. Since the total number of (i,j)-pairs is n^2, we can use the union bound to get

$$\mathbb{P}(\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2 = (1 \pm \epsilon)\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2) \geq 1 - 2n^2\exp\left(-\frac{\epsilon^2 m}{8}\right).$$

By choosing $m = \frac{8}{\epsilon^2} \log \frac{2n^2}{\delta}$, we obtain that $\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2 = (1 \pm \epsilon)\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2$ holds with probability at least $1 - \delta$. There is a minor detail: in (3) we have the unsquared Euclidean norm, but in the above inequality we have the squared Euclidean norm. But they are equivalent up to a constant, since $\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2 \le (1+\epsilon)\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 \Rightarrow \|\boldsymbol{y}_i - \boldsymbol{y}_j\| \le \sqrt{1+\epsilon}\|\boldsymbol{x}_i - \boldsymbol{x}_j\| \le (1+\frac{1}{2}\epsilon)\|\boldsymbol{x}_i - \boldsymbol{x}_j\|$. $\square$