Tail Bounds in Probabilistic Algorithms: From Markov to Chernoff

Agniva Chowdhury

Lecture 3: COMP 480/580
09/04/2025

Agenda

1. Intro and general research theme

2. Til bounds: Motivations

3. Expectation and variance of a random variable

4. Some concentration inequalities
   - Markov
   - Chebyshev
   - Chernoff

My background and research

2025-Present — Rice University, Houston, TX

Intel Corporation, Hillsboro, OR — 2024-2025

2022-2024 — Oak Ridge National Laboratory, Oak Ridge, TN

2015-2021 — Purdue University, Ph.D. in Statistics

Spent some time in the data science and analytics:
- HSBC Decision Sciences, Kolkata, WB, India
- EXL Service, Gurugram, Haryana, India

2011 — Indian Institute of Technology Kanpur, M.Sc. in Statistics

University of Calcutta, B.Sc. in Statistics — 2009

At Purdue:

- Randomized methods for regression, low-rank approximation, discriminant analysis etc.
- Extended to general linear optimization problems.

At ORNL:

- Further extended to non-linear optimizations e.g., logistic regression.
- Explored various SciML applications such as inverse problems that require solving massive-scale linear systems.
- Physics-based deep neural networks (PINNs) and impact of regularization on them.

At Intel and currently:

- Designing sketching-based, dimension reduction tools to enable privacy-preserving, efficient federated machine learning for large-scale deep learning models.
- Using sketching to improve the performances of modern deep learning architecture e.g., LLMs, ViTs, or VLMs.

Motivations

- Suppose we insert n keys into a hash table with m slots.
- Each key is hashed independently and uniformly at random.
- With separate chaining, collisions are stored in a linked list (chain) at each slot.

Question: How long can a chain get?

- Average chain length is small: $\frac{n}{m}$
- **But in worst-case, a chain could (in theory) hold all n keys!**

Why average alone isn't enough?

- **Expected chain length** is $\mu = \frac{n}{m}$
- But expectation doesn't tell us about concentration:
    1. Is it always close to μ?
    2. Or can it be much larger, with non-negligible probability?

Where do tail bounds enter?

- We model chain length at a slot as:

$$X = \sum_{i=1}^{n} X_i ,$$

where $X_i = 1$ if key i hashes to the slot (Bernoulli with $p = \frac{1}{m}$).

- **Markov:** Tells us the chance that chain length is much bigger than μ. But it's loose.
- **Chebyshev:** Uses variance ($\sigma^2 = np(1-p)$) to say chain length rarely strays far from μ.
- **Chernoff:** Gives exponentially small tail probabilities *e.g.,* the probability of a chain being 10x longer than average shrinks exponentially in n.

Expectation and variance of random variables

- The expected value (also called the expectation or mean) of a random variable X is given by

$$\mathbb{E}(X) = \sum_x x \cdot p(x) \text{ (discrete)}$$

or

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) dx \text{ (continuous)}$$

- $\mathbb{E}\left(\sum_i X_i\right) = \sum_i \mathbb{E}(X_i)$
- $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$
- $\text{Var}(X) = \mathbb{E}\left(X - \mathbb{E}(X)\right)^2 = \mathbb{E}(X^2) - \mathbb{E}^2(X)$
- $\text{Var}(aX + b) = a^2 \cdot \text{Var}(X)$
- If X and Y are independent, $\mathbb{E}(XY) = \mathbb{E}(X) \cdot \mathbb{E}(Y)$
- If $X_1, \dots, X_n$ are pairwise independent, $\text{Var}\left(\sum_i X_i\right) = \sum_i \text{Var}(X_i)$

Our first bound: Markov's Inequality

Let X be a non-negative random variable. For any $a > 0$, we have

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$$

To apply this bound you only need to know:

1. it's non-negative
2. Its expectation.

Our first bound: Markov's Inequality

Let X be a non-negative random variable. For any $a > 0$, we have

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$$

Proof 1: Define an indicator function $\mathrm{I}_{(X \geq a)} = 1$ if $X \geq a$ and $\mathrm{I}_{(X \geq a)} = 0$ if $X < a$.

Key observation: $a\, \mathrm{I}_{(X \geq a)} \leq X$

Take expectation on the both side.

Proof 2: $\mathbb{E}(X) = \int_0^\infty x\, p(x) dx = \int_0^a x\, p(x) dx + \int_a^\infty x\, p(x) dx \geq \int_a^\infty x\, p(x) dx$

$\geq \int_a^\infty a\, p(x) dx = a\, \mathbb{P}(X \geq a).$

Example 1: X ~ Geometric distribution

Suppose you roll a fair (6-sided) die until you see a 6. Let X be the number of rolls. Bound the probability that $X \geq 12$

Here X = # of rolls until the first success (*i.e.,* rolling a 6). Clearly, X follows a geometric distribution with success probability $p = \frac{1}{6}$. Therefore,

$$\mathbb{P}(X = k) = \left(\frac{5}{6}\right)^{k-1} \cdot \frac{1}{6} \quad \text{, for } k = 1,2, \ldots$$

We can prove that $\mathbb{E}(X) = \frac{1}{p} = 6$. Using Markov's, $\mathbb{P}(X \geq 12) \leq \frac{6}{12} = 0.5$

- Exact probability $\mathbb{P}(X \geq 12) = 1 - \mathbb{P}(X \leq 11) = 0.1346$

So, Markov's inequality gives a very loose (but valid) upper bound.

Example 2

Suppose the average number of ads you see on a website is 25. Give an upper bound on the probability of seeing a website with 75 or more ads.

$$\mathbb{P}(X \geq 75) \leq \frac{\mathbb{E}(X)}{75} = \frac{25}{75} = \frac{1}{3}$$

Example 3: Useless!

Suppose the average number of ads you see on a website is 25. Give an upper bound on the probability of seeing a website with **20** or more ads.

$$\mathbb{P}(X \geq 20) \leq \frac{\mathbb{E}(X)}{20} = \frac{25}{20} = 1.25$$

Well, that's…true. Technically.
But without more information we couldn't hope to do much better. What if every page gives exactly 25 ads? Then the probability really is 1.

So…what do we do?

A better inequality!

We’re trying to bound the tails of the distribution.
What parameter of a random variable describes the tails?
The variance!

Our second bound: Chebyshev's Inequality

Let X be a random variable. For any $a > 0$, we have

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2}$$

To apply this bound you need to know:

1. ~~it's non-negative.~~
2. Its expectation.
3. Its variance.

Our second bound: Chebyshev's Inequality

Let X be a random variable. For any $a > 0$, we have

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2}$$

Proof:

Let $Z = X - \mathbb{E}(X)$. We know $|Z| \geq a \Leftrightarrow Z^2 \geq a^2$

So, $\mathbb{P}(|X - \mathbb{E}(X)| \geq a) = \mathbb{P}(|Z| \geq a) = \mathbb{P}(Z^2 \geq a^2) \leq \frac{\mathbb{E}(Z^2)}{a^2} = \frac{\mathrm{Var}(X)}{a^2}$

Example 1 again!

Suppose you roll a fair (6-sided) die until you see a 6. Let X be the number of rolls. Bound the probability that $X \geq 12$.

$$\mathbb{P}(X \geq 12) \leq \mathbb{P}(|X - 6| \geq 6) \leq \frac{(5/6)/(1/36)}{6^2} = \frac{5}{6}$$

- Not any better than Markov's!

Example 1 again!

Let X be a geometric r.v. with parameter p. Bound the probability that $X \geq \frac{2}{p}$.

- Chebyshev: $\mathbb{P}\left(X \geq \frac{2}{p}\right) \leq \mathbb{P}\left(\left|X - \frac{1}{p}\right| \geq \frac{1}{p}\right) \leq \frac{\frac{1-p}{p^2}}{\frac{1}{p}^2} = 1 - p$

- While Markov gives
$$\mathbb{P}\left(X \geq \frac{2}{p}\right) = \frac{\mathbb{E}(X)}{2/p} = \frac{1/p}{2/p} = \frac{1}{2}$$

- For large p, Chebyshev is better!

Better Example!

Suppose the average number of ads you see on a website is 25. And the variance of the number of ads is 16. Give an upper bound on the probability of seeing a website with 30 or more ads.

Chebyshev: $\mathbb{P}(X \geq 30) = \mathbb{P}(X - 25 \geq 5) \leq \mathbb{P}(|X - 25| \geq 5) \leq \frac{16}{25} \approx 0.64$

- While Markov gives

$$\mathbb{P}(X \geq 30) = \frac{\mathbb{E}(X)}{30} = \frac{25}{30} \approx 0.83$$

- Chebyshev gives a tighter bound here because it uses variance information.

A tighter bound: Chernoff's

Let $X_1, X_2, \ldots, X_n$ be independent Bernoulli random variables, and define $X = \sum_{i=1}^{n} X_i$ and $\mu = \mathbb{E}(X)$. Then, we have

$$\mathbb{P}(X \geq (1+\delta)\mu) \leq \exp\left(-\frac{\delta^2 \mu}{2+\delta}\right), \text{ for any } \delta \geq 0$$

and

$$\mathbb{P}(X \leq (1-\delta)\mu) \leq \exp\left(-\frac{\delta^2 \mu}{2}\right), \text{ for any } 0 \leq \delta \leq 1$$

For Chernoff, we need:

- expectation + sum of independent rvs (gives exponentially small tail bounds).

Chernoff bound: proof??

Let $X_1, X_2, \dots, X_n$ be independent Bernoulli random variables, and define $X = \sum_{i=1}^{n} X_i$ and $\mu = \mathbb{E}(X)$. Then, we have

$$\mathbb{P}(X \geq (1+\delta)\mu) \leq \exp\left(-\frac{\delta^2 \mu}{2+\delta}\right), \text{ for any } \delta \geq 0$$

and

$$\mathbb{P}(X \leq (1-\delta)\mu) \leq \exp\left(-\frac{\delta^2 \mu}{2}\right), \text{ for any } 0 \leq \delta \leq 1$$

- **Exponential Markov trick**: $\mathbb{P}(X \geq a) = \mathbb{P}(e^{tX} \geq e^{ta}) \leq \frac{\mathbb{E}(e^{tX})}{e^{ta}}$
- Numerator of the RHS is the definition of a moment generation function (MGF) of X.
- Decompose + optimize over t

Going back to the first example

- Suppose $n = m = 1000$.
- Expected chain length: $\mu = 1$.
- What is $\mathbb{P}(X \geq 10)$?
- **Markov:**$\mathbb{P}(X \geq 10) \leq \frac{\mu}{10} = 0.1$
- **Chebyshev:** $\mathbb{P}(X \geq 10) = \mathbb{P}(X - 1 \geq 9) \leq \mathbb{P}(|X - 1| \geq 9) \leq \frac{1}{9^2} \approx 0.0123$
- **Chernoff**: Take $\delta = 9$. $\mathbb{P}(X \geq 10) \leq \exp\left(-\frac{9^2.1}{2+9}\right) \approx 0.00063$

Conclusions

1. **Why tail bounds matter:**
   - Randomized algorithms and data structures rely on probabilistic guarantees.
   - Expectation alone is not enough: we need to know how far outcomes deviate.
   - Tail bounds quantify how unlikely large deviations are.
2. **The big three:**
   - **Markov's Inequality:** Requires only expectation. Very general but often loose.
   - **Chebyshev's Inequality:** Uses variance for sharper bounds and useful when variance is small. Still only gives polynomial decay.
   - **Chernoff Bounds:** Strongest in practice for sums of independent variables. Provide exponentially small tail probabilities.

Tail bounds let us move from average-case analysis to reliable guarantees with high probability.

Thank you

Questions?