Deck 4: Scalar concentration inequalities

Math 7870: Topics in Randomized Numerical Linear Algebra

Spring 2026

Akil Narayan

Outline

- Simple moment-based probabilities
  (Markov and Chebyshev inequalities)
- Distribution function-based arguments
  (Glivenko-Cantelli, Barry-Esseen)
- Moment generating function-based bounds
  (Chernoff bound)
- Bounded and sub-Gaussian random variables
  (Hoeffding inequality)
- Martingale models
  (Azuma inequality)
- Functions of bounded differences
  (McDiarmid's inequality)
- High-probability distribution bounds
  (Dvoretsky-Kiefer-Wolfowitz)

Motivation for concentration

Let X be a scalar (real) random variable. In practice we want to know things like:

$$\Pr(X \geqslant t), \qquad \Pr(|X - \mathbb{E}X| \geqslant t).$$

E.g: $X = \| AB - \sum_{i \in [n]} X_i \|_F$

Motivation for concentration

Let X be a scalar (real) random variable. In practice we want to know things like:

$$\Pr(X \geqslant t), \qquad \Pr(|X - \mathbb{E}X| \geqslant t).$$

These probabilities are computable if we can analytically manipulate the distribution function:

$$\Pr(X \geqslant t) = 1 - F_X(t^-), \qquad \Pr(|X - \mathbb{E}X| \geqslant t) = F_X(\mathbb{E}X - t) + 1 - F_X(\mathbb{E}X + t^+).$$

The problem is that we often don't have access to the exact distribution.
(E.g., X is a finite iid sum.)

$F_X(t) = Pr(X \leq t)$

$1 - F_X(t) = Pr(X > t)$

Motivation for concentration

Let X be a scalar (real) random variable. In practice we want to know things like:

$$\Pr(X \geqslant t), \qquad \Pr(|X - \mathbb{E}X| \geqslant t).$$

These probabilities are computable if we can analytically manipulate the distribution function:

$$\Pr(X \geqslant t) = 1 - F_X(t^-), \qquad \Pr(|X - \mathbb{E}X| \geqslant t) = F_X(\mathbb{E}X - t) + 1 - F_X(\mathbb{E}X + t^+).$$

The problem is that we often don't have access to the exact distribution.
(E.g., X is a finite iid sum.)

However, it generally is feasible to compute the first few moments of X.

The task of estimating probabilities from moments is the study of *concentration* (of X).

Markov and Chebyshev inequalities

We've already seen one of the simplest examples of concentration inequalities, *Markov's inequality*:

$$X \geqslant 0 \text{ wp1}, t > 0 \quad \Longrightarrow \quad \Pr(X \geqslant t) \leqslant \frac{\mathbb{E}X}{t} \quad \text{or} \quad \Pr(X \geqslant t\,\mathbb{E}X) \leqslant \frac{1}{t}$$

This latter form is only useful if $t > 1$.

Markov's inequality is quite useful:

$$\Pr(|X - \mathbb{E}X| \geq t) \leq ?$$

$$\Pr(|X-\mathbb{E}X| \geq t) = \Pr(|X-\mathbb{E}X|^2 \geq t^2) \overset{\text{Markov}}{\leq} \frac{\mathbb{E}|X-\mathbb{E}X|^2}{t^2} = \frac{\mathrm{Var}(X)}{t^2}$$

Markov and Chebyshev inequalities

We've already seen one of the simplest examples of concentration inequalities, *Markov's inequality*:

$$X \geqslant 0 \text{ wp1}, t > 0 \quad \Longrightarrow \quad \Pr(X \geqslant t) \leqslant \frac{\mathbb{E}X}{t} \quad \text{or} \quad \Pr(X \geqslant t\,\mathbb{E}X) \leqslant \frac{1}{t}$$

This latter form is only useful if $t > 1$.

Applying Markov's inequality to $Y = (X - \mathbb{E}X)^2 \geqslant 0$ yields *Chebyshev's inequality*:

$$\Pr(Y \geqslant t^2) \leqslant \frac{\mathbb{E}Y}{t^2} = \frac{\text{Var}X}{t^2} \quad \Longrightarrow \quad \Pr(|X - \mathbb{E}X| \geqslant t) \leqslant \frac{\text{Var}X}{t^2} \quad \text{or} \quad \Pr(|X - \mathbb{E}X| \geqslant t\sigma) \leqslant \frac{1}{t^2}$$

where $\sigma = \text{StDev}(X)$, and again the latter form is useful only when $t > 1$.

These are the simplest bounds for estimating probabilities for moments.
(They're also the most general: almost nothing is assumed about X.)

Sharpness of Markov and Chebyshev inequalities

Without any further assumptions, the Markov and Chebyshev inequalities cannot be improved.

Example 1 (Markov inequality sharpness). *Let X_s be a random variable parameterized by any $s > 0$, with mass function $p_{X_s}(0) = 1 - 1/s$, and $p_{X_s}(s) = 1/s$.*

$$\mathbb{E}X_s = 0 \cdot \left(1 - \frac{1}{s}\right) + s\frac{1}{s} = 1$$

Pick $t > 0$: is there a RV s.t. Markov's inequality is sharp for that t?

$s = t$

$$P(X_t \geq t) = \frac{1}{t} = \frac{\mathbb{E}X_t}{t}$$

Sometimes we expect better

If X_i, $i \in \mathbb{N}$ are centered and iid, then $Z_n := \frac{1}{n} \sum_{i \in [n]} X_i$ should approach $\frac{1}{\sqrt{n}} \mathcal{N}(0, \mathrm{Var} X)$.

What we expect, is that with $\sigma^2 = \mathrm{Var} X$, then if $Z \sim \mathcal{N}(0, \sigma^2)$, we have,

$$
\begin{aligned}
\Pr(Z_n \geqslant t\sigma) \text{“}\approx\text{”} \Pr\left(\frac{1}{\sqrt{n}} Z \geqslant t\sigma\right) &= \Pr(Z \geqslant t\sigma\sqrt{n}) = 1 - F_Z(t\sigma\sqrt{n}) \\
&= \cancel{\approx} \frac{1}{2} - \frac{1}{2} \mathrm{erf}\left(\frac{t\sqrt{n}}{\sqrt{2}}\right) \\
&\overset{n \gg 1}{\approx} \frac{1}{4} e^{-nt^2/2}
\end{aligned}
$$

Sometimes we expect better

If X_i, $i \in \mathbb{N}$ are centered and iid, then $Z_n := \frac{1}{n} \sum_{i \in [n]} X_i$ should approach $\frac{1}{\sqrt{n}} \mathcal{N}(0, \mathrm{Var} X)$.

What we expect, is that with $\sigma^2 = \mathrm{Var} X$, then if $Z \sim \mathcal{N}(0, \sigma^2)$, we have,

$$
\begin{aligned}
\Pr(Z_n \geqslant t\sigma) \text{``} \approx \text{''} \Pr\left(\frac{1}{\sqrt{n}} Z \geqslant t\sigma \right) &= \Pr(Z \geqslant t\sigma\sqrt{n}) = 1 - F_Z(t\sigma\sqrt{n}) \\
&= \cancel{\approx} \frac{1}{2} - \frac{1}{2}\mathrm{erf}\left(\frac{t\sqrt{n}}{\sqrt{2}} \right) \\
&\overset{n \gg 1}{\approx} \frac{1}{4} e^{-nt^2/2}
\end{aligned}
$$

By comparison, Chebyshev's inequality yields,

$$
\Pr(Z_n \geqslant t\sigma) \leqslant \Pr(|Z_n| \geqslant t\sigma) \leqslant \frac{\mathrm{Var} Z_n}{\sigma^2 t^2} = \frac{1}{nt^2}.
$$

The point: $e^{-nt^2} \ll \frac{1}{nt^2}$ when n and/or t are large.

Can the CLT help?

The only sketchy "$\approx$" we employed is

$$\Pr(Z_n \geqslant t\sigma) \text{``} \approx \text{''} \Pr\left(\frac{1}{\sqrt{n}} Z \geqslant t\sigma\right),$$

which appeals to the CLT argument that $\sqrt{n}Z_n$ is an n-asymptotic normal random variable Z.

We hope that a "quantitative" CLT can make this precise.

Can the CLT help?

The only sketchy "$\approx$" we employed is

$$\Pr(Z_n \geqslant t\sigma) \text{``}\approx\text{''} \Pr\left(\frac{1}{\sqrt{n}} Z \geqslant t\sigma\right),$$

which appeals to the CLT argument that $\sqrt{n} Z_n$ is an n-asymptotic normal random variable Z.

We hope that a "quantitative" CLT can make this precise.

The question about the value of $\Pr(Z_n \geqslant t)$ is equivalent to understanding how well the distribution functions converge:

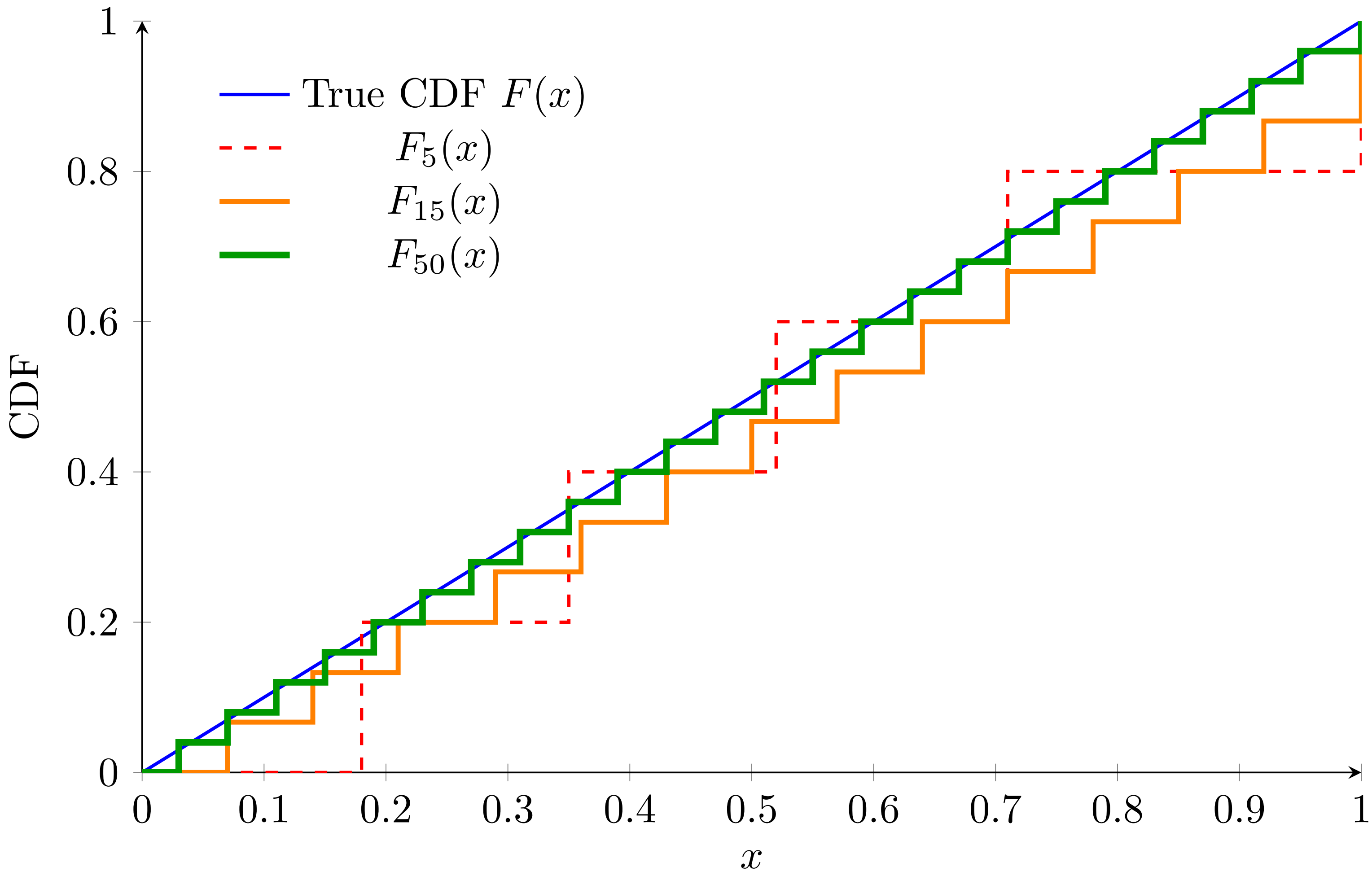
$$F_n(t) := \Pr(Z_n \leqslant t) \qquad\qquad F(t) := \Pr\left(\frac{Z}{\sqrt{n}} \leqslant t\right)$$

Theorem (Glivenko-Cantelli)**.** *With the above setup, then with probability 1:*

$$\lim_{n\to\infty} \|F - F_n\|_{L^\infty(\mathbb{R})} = \lim_{n\to\infty} \sup_{t\in\mathbb{R}} |F(t) - F_n(t)| = 0.$$

We therefore expect our idea about working through distribution functions is possible.

Glivenko-Cantelli, visualized



More precise convergence

Something stronger than Glivenko-Cantelli is true.

Theorem (Barry-Esseen inequality)**.** *Let X have finite second and third moments, and let $\{X_i\}_{i\in\mathbb{N}}$ be centered and iid with $X_1 \sim X$, and $\mathrm{Var}(X_1) = \sigma^2 > 0$. Let*

$$Z_n = \frac{1}{n}\sum_{i\in[n]} X_i,$$

and let $Z \sim \mathcal{N}(0,1)$. Then:

$$\left|F_{\sqrt{n}\,Z_n/\sigma}(x) - F_Z(x)\right| \leqslant \frac{C}{\sqrt{n}}\frac{\mathbb{E}|X_1|^3}{\sigma^3}$$

The constant C is absolute.

More precise convergence

Something stronger than Glivenko-Cantelli is true.

Theorem (Barry-Esseen inequality)**.** *Let X have finite second and third moments, and let $\{X_i\}_{i\in\mathbb{N}}$ be centered and iid with $X_1 \sim X$, and $\mathrm{Var}(X_1) = \sigma^2 > 0$. Let*

$$Z_n = \frac{1}{n} \sum_{i \in [n]} X_i,$$

and let $Z \sim \mathcal{N}(0,1)$. Then:

$$\left| F_{\sqrt{n}\, Z_n/\sigma}(x) - F_Z(x) \right| \leqslant \frac{C}{\sqrt{n}} \frac{\mathbb{E}|X_1|^3}{\sigma^3}$$

The constant C is absolute.

It turns out this is *bad* for us: This rate of convergence is in general sharp.
If we can only replace $F_{\sqrt{n}\, Z_n/\sigma}$ with F_Z with a $1/\sqrt{n}$ mistake, then this will reflect in probability estimates. $(1/\sqrt{n} \gg e^{-n})$

Higher order moments

We seem to have concluded that for general X, a deviation-from-mean probability for $\frac{1}{n}\sum_{i\in[n]} X_i$ can't be derived, at least not using the arguments we've explored.

Goal: If $X_i \sim X$ for $i \in \mathbb{N}$ are iid, we seek to bound $\Pr(Z_n \geqslant t)$.
(As before: $Z_n = \frac{1}{n}\sum_{i\in[n]} X_i$.)

(Probably X is centered)

Higher order moments

We seem to have concluded that for general X, a deviation-from-mean probability for $\frac{1}{n}\sum_{i\in[n]} X_i$ can't be derived, at least not using the arguments we've explored.

Goal: If $X_i \sim X$ for $i \in \mathbb{N}$ are iid, we seek to bound $\Pr(Z_n \geqslant t)$.
(As before: $Z_n = \frac{1}{n}\sum_{i\in[n]} X_i$.)

We can try to see if this is possible in a simpler case going back to Markov's inequality:
Let's assume X is centered: $\mathbb{E}X = 0$. Using the same idea as for Chebyshev's inequality, for any integer $k \in \mathbb{N}$:

$$\Pr(|Z_n| \geqslant t) = \Pr(|Z_n|^{2k} \geqslant t^{2k}) \leqslant \frac{\mathbb{E}|Z_n|^{2k}}{t^{2k}}.$$

$$\mathbb{E}|Z_n|^{2k} = \mathbb{E}\, Z_n^{2k} = \mathbb{E}\left(\frac{1}{n}\sum_{i\in[n]} X_i\right)^{2k}$$

Higher order moments

We seem to have concluded that for general X, a deviation-from-mean probability for $\frac{1}{n}\sum_{i\in[n]} X_i$ can't be derived, at least not using the arguments we've explored.

Goal: If $X_i \sim X$ for $i \in \mathbb{N}$ are iid, we seek to bound $\Pr\left(Z_n \geqslant t\right)$.
(As before: $Z_n = \frac{1}{n}\sum_{i\in[n]} X_i$.)

We can try to see if this is possible in a simpler case going back to Markov's inequality:
Let's assume X is centered: $\mathbb{E}X = 0$. Using the same idea as for Chebyshev's inequality, for any integer $k \in \mathbb{N}$:

$$\Pr(|Z_n| \geqslant t) = \Pr(|Z_n|^{2k} \geqslant t^{2k}) \leqslant \frac{\mathbb{E}|Z_n|^{2k}}{t^{2k}}.$$

Using the multinomial theorem:

$$\Pr(|Z_n| \geqslant t) \leqslant \frac{1}{t^{2k}}\left[\frac{1}{n^{2k}}\sum_{\boldsymbol{j}\in\mathbb{N}_0^{2n}, |\boldsymbol{j}|=2k} \binom{2k}{\boldsymbol{j}} \prod_{\ell=1}^{n} \mu_{j_\ell}\right], \qquad \mu_j = \mathbb{E}X^j.$$

The point: for all k, $\Pr(|Z| \geqslant t) \lesssim t^{-2k} g(\mu_2, \ldots, \mu_{2k})$ for some function g.
This achieves $t^{-2k} \ll t^{-1}$ for large t.

Using moment generating functions

The problem: Not only is the previous expression unwieldly, it would essentially require estimation/computation of high-order moments.

However, the general idea here is valuable: higher-order moments can give us better estimation. We'd like to more elegantly build them into an estimate.

For simplicity, we'll also assume X is centered: $\mathbb{E}X = 0$, ~~so that $a < 0 < b$.~~

Using moment generating functions

The problem: Not only is the previous expression unwieldly, it would essentially require estimation/computation of high-order moments.

However, the general idea here is valuable: higher-order moments can give us better estimation. We'd like to more elegantly build them into an estimate.

For simplicity, we'll also assume X is centered: $\mathbb{E}X = 0$, so that $a < 0 < b$.

Our Markov inequality strategies have revolved around using the monotone functions $x \mapsto x^p$ (e.g., $p = 2, 2k$).

Through Markov's inequality, we compute $\mathbb{E}X^p$, which results in pth-order moments. Ideally, we'd use information from *all* moments.

Given a random variable X, its *moment generating function* $M_X(s) := \mathbb{E}e^{sX}$ encodes all moments of X. (E.g., $M_X^{(n)}(0) = \mathbb{E}X^n$.)

Using MGF's

With all of the above, the following strategy is fairly generic:

Let's use Markov's inequality in the same way as before, but instead of the function $x \mapsto x^2$ with image on $[0, \infty)$, we'll use $x \mapsto e^x$.

$$\Pr(X \geq t) = \Pr(e^X \geq e^t)$$

Using MGF's

With all of the above, the following strategy is fairly generic:

Let's use Markov's inequality in the same way as before, but instead of the function $x \mapsto x^2$ with image on $[0, \infty)$, we'll use $x \mapsto e^x$.

$$
\begin{aligned}
x \mapsto e^x \text{ monotone increasing, and } e^x \geqslant 0 \;\; \forall \;\; x \in \mathbb{R} \implies \Pr(Z_n \geqslant t) &\overset{s \geq 0}{=} \Pr(sZ_n \geqslant st) \\
&= \Pr(e^{sZ_n} \geqslant e^{st}) \\
&\leqslant e^{-st}\mathbb{E}e^{sZ_n}.
\end{aligned}
$$

The last inequality is Markov's inequality.

The free parameter $s > 0$ (the location where we evaluate the MGF), can be tuned to achieve optimal results.

Chernoff bounds

$$Z_n = \frac{1}{n} \sum_{i \in [n]} X_i \qquad \mathbb{E} e^{(X_1 + X_2)} = (\mathbb{E} e^{X_1})(\mathbb{E} e^{X_2})$$

Since X_i are iid, then $\mathbb{E}e^{sZ_n} = \left(\mathbb{E}e^{sX/n}\right)^n$. So we have concluded:

$$M_X(s) = \mathbb{E} e^{sX}$$

$$\Pr(Z_n \geqslant t) \;\cancel{=}\; \leq e^{-st}\left(\mathbb{E}e^{sX/n}\right)^n = e^{-st}\left(M_X\left(\frac{s}{n}\right)\right)^n.$$

The remaining question is how to estimate the MGF $M_X(s/n)$.

This generic strategy of bounding this tail probability through an MGF is called a <u>Chernoff bound</u>.

There are a few ways to estimate the MGF.

Chernoff for Rademacher

Suppose X has a Rademacher distribution:

$$p_X(+1) = p_X(-1) = \frac{1}{2}.$$

$$\mathbb{E} e^{sX} = \frac{1}{2} e^{+s} + \frac{1}{2} e^{-s}$$

Estimating the MGF is straightforward:

$$\mathbb{E}e^{sX} = \cosh s$$

Recall that the MGF behavior around 0 is important, so we want a tight MGF bound there.

Chernoff for Rademacher

Suppose X has a Rademacher distribution:

$$p_X(+1) = p_X(-1) = \frac{1}{2}.$$

Estimating the MGF is straightforward:

$$\mathbb{E}e^{sX} = \cosh s$$

Recall that the MGF behavior around 0 is important, so we want a tight MGF bound there.

$$\begin{aligned}
\cosh s &= 1 + \frac{s^2}{2} + \frac{s^4}{24} + \ldots + \frac{s^{2n}}{(2n)!} + \ldots \\
&\leqslant 1 + \frac{s^2}{2} + \frac{1}{2!}\left(\frac{s^2}{2}\right)^2 + \ldots + \frac{1}{n!}\left(\frac{s^2}{2}\right)^n + \ldots \\
&= e^{s^2/2}
\end{aligned}$$

$$\frac{1}{(2n)!} = \frac{1}{n! \cdot (n+1) \cdots (2n)} \leq \frac{1}{n! \cdot 2 \cdot 2 \cdots 2} = \frac{1}{n!(2^n)}$$

Chernoff for Rademacher

Suppose X has a Rademacher distribution:

$$p_X(+1) = p_X(-1) = \frac{1}{2}.$$

Estimating the MGF is straightforward:

$$\mathbb{E}e^{sX} = \cosh s$$

Recall that the MGF behavior around 0 is important, so we want a tight MGF bound there.

$$\begin{aligned}
\cosh s &= 1 + \frac{s^2}{2} + \frac{s^4}{24} + \ldots + \frac{s^{2n}}{(2n)!} + \ldots \\
&\leqslant 1 + \frac{s^2}{2} + \frac{1}{2!}\left(\frac{s^2}{2}\right)^2 + \ldots + \frac{1}{n!}\left(\frac{s^2}{2}\right)^n + \ldots \\
&= e^{s^2/2}
\end{aligned}$$

Therefore: $M_X(s) \leqslant e^{s^2/2}$, so that resuming our iid sum Chernoff bound:

$$\Pr(Z_n \geqslant t) = e^{-st}\left(M_X\left(\frac{s}{n}\right)\right)^n = \exp\left(-st + \frac{s^2}{2\,\cancel{4}n}\right).$$

The next step could be to optimize s. Before doing that, let's generalize beyond Rademacher.

Hoeffding's Lemma

Here's another, more general way to estimate an MGF:

Suppose now that X is a bounded (nontrivial) random variable: $X \in [a, b]$ wp1.

Lemma (Hoeffding's Lemma). *Suppose $Y \in [a, b]$ is a centered random variable. Then $\mathbb{E}e^{sY} \leqslant e^{\frac{1}{8}s^2(b-a)^2}$ for any $s \in \mathbb{R}$.*

$$\mathbb{E}Y^p = \int_a^b y^p f(y)\,dy$$

NB: Rademacher $\Rightarrow$ $a=-1$, $b=+1$

$$\Rightarrow M_Y(s) \leq e^{\frac{1}{8}s^2 4} = e^{s^2/2}$$

Hoeffding's Lemma

Here's another, more general way to estimate an MGF:

Suppose now that X is a bounded (nontrivial) random variable: $X \in [a, b]$ wp1.

Lemma (Hoeffding's Lemma). *Suppose* $Y \in [a, b]$ *is a centered random variable. Then* $\mathbb{E}e^{sY} \leqslant e^{\frac{1}{8}s^2(b-a)^2}$ *for any* $s \in \mathbb{R}$.

Proof sketch:

- $x \mapsto e^x$ is convex. Therefore,

$$e^{sx} \leqslant \frac{b-x}{b-a}e^{sa} + \frac{x-a}{b-a}e^{sb}$$

Hoeffding's Lemma

Here's another, more general way to estimate an MGF:

Suppose now that X is a bounded (nontrivial) random variable: $X \in [a, b]$ wp1.

Lemma (Hoeffding's Lemma)**.** *Suppose* $Y \in [a, b]$ *is a centered random variable. Then* $\mathbb{E}e^{sY} \leqslant e^{\frac{1}{8}s^2(b-a)^2}$ *for any* $s \in \mathbb{R}$.

Proof sketch:

- $x \mapsto e^x$ is convex. Therefore,

$$e^{sx} \leqslant \frac{b-x}{b-a}e^{sa} + \frac{x-a}{b-a}e^{sb}$$

- This implies

$$\mathbb{E}e^{sY} \leqslant \frac{1}{b-a}\left(be^{sa} - ae^{sb}\right).$$

Hoeffding's Lemma

Here's another, more general way to estimate an MGF:

Suppose now that X is a bounded (nontrivial) random variable: $X \in [a, b]$ wp1.

Lemma (Hoeffding's Lemma)**.** *Suppose $Y \in [a, b]$ is a centered random variable. Then $\mathbb{E} e^{sY} \leqslant e^{\frac{1}{8}s^2(b-a)^2}$ for any $s \in \mathbb{R}$.*

Proof sketch:

- $x \mapsto e^x$ is convex. Therefore,

$$e^{sx} \leqslant \frac{b-x}{b-a} e^{sa} + \frac{x-a}{b-a} e^{sb}$$

- This implies

$$\mathbb{E} e^{sY} \leqslant \frac{1}{b-a} \left(b e^{sa} - a e^{sb} \right).$$

- Define $\phi(z)$ as,

$$\frac{1}{b-a} \left(b e^{sa} - a e^{sb} \right) = e^{\phi(s(b-a))}$$

Hoeffding's Lemma

Here's another, more general way to estimate an MGF:
Suppose now that X is a bounded (nontrivial) random variable: $X \in [a, b]$ wp1.

Lemma (Hoeffding's Lemma)**.** *Suppose $Y \in [a, b]$ is a centered random variable. Then $\mathbb{E}e^{sY} \leqslant e^{\frac{1}{8}s^2(b-a)^2}$ for any $s \in \mathbb{R}$.*

Proof sketch:

- $x \mapsto e^x$ is convex. Therefore,
$$e^{sx} \leqslant \frac{b-x}{b-a}e^{sa} + \frac{x-a}{b-a}e^{sb}$$
- This implies
$$\mathbb{E}e^{sY} \leqslant \frac{1}{b-a}\left(be^{sa} - ae^{sb}\right).$$
- Define $\cancel{\phi(z)}$ $\phi(\cdot)$ as,
$$\frac{1}{b-a}\left(be^{sa} - ae^{sb}\right) = e^{\phi(s(b-a))}$$
- Taylor series estimation: $\phi(z) \leqslant Cz^2$, with $C = 1/8$.

Sewing it together

By positivity and monotonicity of $x \mapsto e^x$, Markov's inequality yields:

$$\Pr(Z_n \geqslant t) = e^{-st}\left(\mathbb{E}e^{sX/n}\right)^n$$

By Hoeffding's Lemma:

$$\left(\mathbb{E}e^{sX/n}\right)^n \leqslant \left(\exp\left(\frac{1}{8}\frac{s^2}{n^2}(b-a)^2\right)\right)^n = \exp\left(\frac{1}{8}\frac{s^2}{n}(b-a)^2\right)$$

Therefore:

$$\Pr(Z_n \geqslant t) \leqslant \exp\left(-st + \frac{s^2(b-a)^2}{8n}\right)$$

minimize $\underbrace{-st + \frac{s^2(b-a)^2}{8n}}_{f(s)}$ by choosing s:

$f'(s) = -t + \frac{s(b-a)^2}{4n} \Rightarrow s = \frac{4nt}{(b-a)^2}$

Sewing it together

By positivity and monotonicity of $x \mapsto e^x$, Markov's inequality yields:

$$\Pr(Z_n \geqslant t) = e^{-st} \left(\mathbb{E} e^{sX/n}\right)^n$$

By Hoeffding's Lemma:

$$\left(\mathbb{E} e^{sX/n}\right)^n \leqslant \left(\exp\left(\frac{1}{8}\frac{s^2}{n^2}(b-a)^2\right)\right)^n = \exp\left(\frac{1}{8}\frac{s^2}{n}(b-a)^2\right)$$

Therefore:

$$\Pr(Z_n \geqslant t) \leqslant \exp\left(-st + \frac{s^2(b-a)^2}{8n}\right)$$

Now we can choose s to minimize this probability:

$$s_* = \arg\min_{s>0} \exp\left(-st + \frac{s^2(b-a)^2}{8n}\right) = \frac{4nt}{(b-a)^2} \implies \Pr(Z_n \geqslant t) \leqslant \exp\left(-2nt^2/(b-a)^2\right).$$

NB: this behaves *exactly* like e^{-nt^2} that we "expect" from the CLT!
The $(b-a)^2$ factor is "essentially" $\mathrm{Var} X$.

Hoeffding's inequality

A particular Chernoff bound for concentration is the Hoeffding inequality, which bounds the MGF of X using Hoeffding's Lemma for bounded random variables.

The result has slightly more generality than we've presented:

- The X_i need not be centered: $\mathbb{E}X_i \neq 0$ is ok.
- The X_i must be independent, but *not* identically distributed. We do require boundedness: $X_i \in [a_i, b_i]$ wp1 for all i.

$$Z_n = \frac{1}{n}\sum_i X_i$$

$$M_{Z_n}(s) = \mathbb{E}\exp\left(s\frac{1}{n}\sum_i X_i\right)$$

$$= \prod_i \mathbb{E}\exp\left(s\frac{1}{n}X_i\right)$$

$$= \prod_i M_{X_i}(s/n) \leq \exp\left(\frac{1}{8}\frac{s^2}{n^2}\sum_i (b_i - a_i)^2\right)$$

Hoeffding's inequality

A particular Chernoff bound for concentration is the Hoeffding inequality, which bounds the MGF of X using Hoeffding's Lemma for bounded random variables.

The result has slightly more generality than we've presented:

- The X_i need not be centered: $\mathbb{E}X_i \neq 0$ is ok.
- The X_i must be independent, but *not* identically distributed. We do require boundedness: $X_i \in [a_i, b_i]$ wp1 for all i.

Theorem (Hoeffding's inequality)**.** *Suppose* $\{X_i\}_{i \in \overset{[n]}{\cancel{\mathbb{N}}}}$ *is a sequence of independent random variables, with* $X_i \in [a_i, b_i]$ *wp1 for all* $i \in \underset{[n]}{\cancel{\mathbb{N}}}$. *Then:*

$$\Pr\left(|S_n - \mathbb{E}S_n| \geqslant t\right) \leqslant 2\exp\left(-\frac{2t^2}{\underbrace{\sum_{i\in[n]} (b_i - a_i)^2}_{\approx \sum_i \mathrm{Var}(X_i)}}\right), \qquad S_n := \sum_{i\in[n]} X_i.$$

Observations about Hoeffding's inequality

$$\Pr\left(|S_n - \mathbb{E}S_n| \geqslant t\right) \leqslant 2\exp\left(-\frac{2t^2}{\sum_{i\in[n]}(b_i - a_i)^2}\right), \qquad S_n := \sum_{i\in[n]} X_i.$$

- $\mathbb{E}S_n$ can depend on n.
- We've shown the $S_n - \mathbb{E}S_n \geqslant t$ bound proof. The other direction, $S_n - \mathbb{E}S_n \leqslant t$ is a minor variant. (Work with $\Pr(-S_n \geq t)$)
- This above is a two-sided bound: $|S_n - \mathbb{E}S_n| \geqslant t$. The price paid is a multiplicative 2, from a union bound.
- When $X_i \sim X$ are iid, with $X \in [a, b]$, and $b_i - a_i = \frac{1}{n}(b - a)$, this reduces to

$$\Pr\left(\left|\frac{1}{n}S_n - \mathbb{E}X\right| \geqslant t\right) \leqslant 2\exp\left(-\frac{2nt^2}{(b-a)^2}\right).$$

Another Chenoff bound

One can derive various Chernoff-type bounds from the basic idea of a Chernoff bound.

E.g., if $X_i \sim \text{Bernoulli}(p_i)$ are independent, then the following is a popular "multiplicative" form of a Chernoff bound:

$$\Pr\left(S_n \geqslant (1+\delta)\mu_n\right) \leqslant \left[\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right]^{\mu_n}, \qquad \delta > 0, \qquad \mu_n := \mathbb{E} S_n$$

This is derived:

- Using a generic Chernoff bound strategy: $\Pr(S_n \geqslant t) \leqslant e^{-st} \prod_{i \in [n]} \mathbb{E} e^{sX_i}$.
- Use $t = (1+\delta)\mathbb{E} S_n$, compute $\mathbb{E} e^{sX_i}$ explicitly, and bound the result.

$$\Pr(S_n \geq t) \leq \inf_{s>0} e^{-st} \prod_{i \in [n]} M_{X_i}(s)$$

How general is Hoeffding's inequality?

The iid sum version of Hoeffding's inequality required $X_i \in [a, b]$, are bounded with probability 1. Recall:

$$\Pr\left(\frac{1}{n}S_n - \mathbb{E}X_1 \geqslant t\right) \leqslant \exp\left(-\frac{2nt^2}{(b-a)^2}\right),$$

i.e., $|a|, |b| < \infty$. (E.g., without this we can't use Hoeffding's lemma.)

How general is Hoeffding's inequality?

The iid sum version of Hoeffding's inequality required $X_i \in [a, b]$, are bounded with probability 1. Recall:

$$\Pr\left(\frac{1}{n}S_n - \mathbb{E}X_1 \geqslant t\right) \leqslant \exp\left(-\frac{2nt^2}{(b-a)^2}\right),$$

i.e., $|a|, |b| < \infty$. (E.g., without this we can't use Hoeffding's lemma.)

However, we expect that this result should hold for at least some unbounded random variables as well. E.g., if $X_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, then:

$$\frac{1}{n}S_n \sim \mathcal{N}\left(0, \frac{1}{n}\right) \quad \Longrightarrow \quad \Pr\left(\frac{1}{n}S_n \geqslant t\right) \leqslant \exp\left(-2nt^2\right).$$

How general is Hoeffding's inequality?

The iid sum version of Hoeffding's inequality required $X_i \in [a, b]$, are bounded with probability 1. Recall:

$$\Pr\left(\frac{1}{n}S_n - \mathbb{E}X_1 \geqslant t\right) \leqslant \exp\left(-\frac{2nt^2}{(b-a)^2}\right),$$

i.e., $|a|, |b| < \infty$. (E.g., without this we can't use Hoeffding's lemma.)

However, we expect that this result should hold for at least some unbounded random variables as well. E.g., if $X_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, then:

$$\frac{1}{n}S_n \sim \mathcal{N}\left(0, \frac{1}{n}\right) \quad \Longrightarrow \quad \Pr\left(\frac{1}{n}S_n \geqslant t\right) \leqslant \exp\left(-2nt^2\right).$$

This estimate for unbounded random variables behaves in the Hoeffding-type way, but we don't have a way to analyze the corresponding MGF's.

$\exp\left(\frac{1}{2}s^2\sigma^2\right)$

In particular, if $Y \sim \mathcal{N}(0, \sigma^2)$, then $M_Y(s) = \exp(\cancel{s^2/(2\sigma^2)})$, and this $\sim \exp(s^2)$ MGF behavior is exactly what we needed for the Hoeffding-type Chernoff bound.

What kinds of random variables have MGF's behaving like $\exp(s^2)$?

Random variables dominated by Gaussians

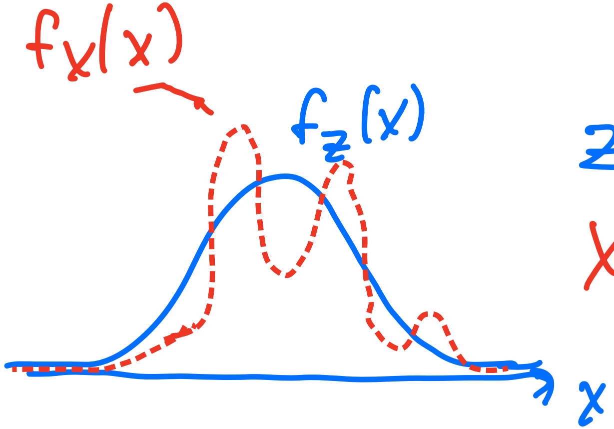
The following is a common class of random variables in probability theory.

Definition 1 (Sub-Gaussian random variables)**.** *A random variable X is called* <u>*sub-Gaussian*</u> *if there is a $c \geqslant 0$ and a centered normal random variable Y such that for all $t > 0$:*

$$\Pr(|X| \geqslant t) \leqslant c \Pr(|Y| \geqslant t).$$

This is equivalent to requiring that there exists a $C > 0$ such that,

$$\Pr\left(|X| \geqslant t\right) \leqslant 2\exp(-t^2/C^2)$$



$Z \sim$ Normal

$X \sim$ subGaussian

Random variables dominated by Gaussians

The following is a common class of random variables in probability theory.

Definition 1 (Sub-Gaussian random variables)**.** *A random variable X is called sub-Gaussian if there is a $c \geqslant 0$ and a centered normal random variable Y such that for all $t > 0$:*

$$\Pr(|X| \geqslant t) \leqslant c \Pr(|Y| \geqslant t).$$

This is equivalent to requiring that there exists a $C > 0$ such that,

$$\Pr\left(|X| \geqslant t\right) \leqslant 2 \exp(-t^2/C^2)$$

There are lots of non-sub-Gaussian random variables. E.g., Cauchy distributions, Poisson distributions, exponential distributions, ...

Equivalent definitions of sub-Gaussians

There are several well-known equivalent definitions of a sub-Gaussian random variable. Here are a few of relevance.

Theorem (Equivalent sub-Gaussian distribution definitions)**.** *Let X be a centered random variable. The following statements are equivalent:*

- *There is a positive C_1 such that,* $\Pr(|X| \geqslant t) \leqslant 2\exp(-t^2/C_1^2)$.
- *There is a positive C_2 such that* $\mathbb{E}|X|^p \lesssim C_2^p p^{p/2}$.
- *There is a positive C_3 such that* $M_X(s) \leqslant \exp\left(\frac{C_3^2 s^2}{2}\right)$.

Equivalent definitions of sub-Gaussians

There are several well-known equivalent definitions of a sub-Gaussian random variable. Here are a few of relevance.

Theorem (Equivalent sub-Gaussian distribution definitions)**.** *Let X be a centered random variable. The following statements are equivalent:*

- *There is a positive C_1 such that,* $\Pr\left(|X| \geqslant t\right) \leqslant 2\exp(-t^2/C_1^2)$.
- *There is a positive C_2 such that* $\mathbb{E}|X|^p \lesssim C_2^p p^{p/2}$.
- *There is a positive C_3 such that* $M_X(s) \leqslant \exp\left(\frac{C_3^2 s^2}{2}\right)$.

Therefore, sub-Gaussian random variables are *precisely* those random variables whose MGF's behave in a Hoeffding-lemma-type way.

The connection is actually stronger than suggested above: The (smallest) constant C_3^2 above is called the *variance proxy* of X, and often properties involving the variance σ^2 of a Gaussian random variable hold for sub-Gaussian ones by replacing σ^2 with the variance proxy.

General Hoeffding inequality

With a fairly good understanding of MGF's that behave like e^{s^2}, we can state a quite general form of Hoeffding's inequality.

Theorem 1 (Sub-Gaussian Hoeffding inequality)**.** *Let* $\{X_i\}_{i\in\cancel{\mathbb{N}}\,[n]}$ *be independent sub-Gaussian random variables, and let* σ_i^2 *be the variance proxy of* X_i. *Then:*

$$\Pr\left(|S_n - \mathbb{E}S_n| \geqslant t\right) \leqslant 2\exp\left(\frac{-t^2}{2\sum_{i\in[n]}\sigma_i^2}\right).$$

Beyond Chernoff bounds

Several concentration results use Chernoff-like ideas to construct bounds. The payoff is that one can move beyond strictly independent sums. For example: if X_i is a sequence of centered independent random variables, we have,

$$\mathbb{E}\left[S_{n+1} \mid S_0, S_1, \ldots, S_n\right] = \mathbb{E}\left[X_{n+1} + S_n \mid S_0, S_1, \ldots, S_n\right] = S_n, \qquad S_n = \sum_{i \in [n]} X_i$$

This is perhaps the simplest example of a *martingale*: a sequence whose expectation conditioned on some history equals the most recent value in that history.

The Chernoff-like bounds we've derived can be generalized to general martingales beyond the simple example above.

The Azuma-Hoeffding inequality

For completeness, we'll state a more general version of the inequality. The general version operates on a *supermartingale*, which is a sequence $\{X_i\}_i$ satisfying,

$$\mathbb{E}\left[X_{n+1} \,\middle|\, X_0, \ldots, X_n\right] \leqslant X_n.$$

For supermartingales, the conditional expectation is *non-increasing* relative to the provided history.

The Azuma-Hoeffding inequality

For completeness, we'll state a more general version of the inequality. The general version operates on a *supermartingale*, which is a sequence $\{X_i\}_i$ satisfying,

$$\mathbb{E}\left[X_{n+1} \,\middle|\, X_0, \ldots, X_n\right] \leqslant X_n.$$

For supermartingales, the conditional expectation is *non-increasing* relative to the provided history.

Theorem 2 (Azuma-Hoeffding inequality). *Let $\{X_i\}_{i \in \mathbb{N}_0}$ be a supermartingale, and assume that the increments are bounded:*

$$A_{i+1} \leqslant X_{i+1} - X_i \leqslant B_{i+1}, \qquad\qquad B_{i+1} - A_{i+1} \leqslant c_{i+1} \in (0, \infty),$$

for a deterministic sequence c_i. Then for every $t > 0$:

$$\Pr\left(X_n - X_0 \geqslant t\right) \leqslant \exp\left(-\frac{2t^2}{\sum_{i \in [n]} c_i^2}\right).$$

where

NB: For submartingales (non-decreasing conditional expectation), a bound on the deviation below X_0 can be derived.
For martingales, a two-sided bound can be derived.

Azuma-Hoeffding proof idea

The sketch of the proof is as follows:

- (Doob decomposition) Decompose the supermartingale X_i into $Y_i + Z_i$, where Y_i is a martingale, and Z_i is decreasing wp1.
- Since Z_i is decreasing, then the event $X_n - X_0 \geqslant t$ implies the event $Y_{n+1} - Y_0 \geqslant t$.

Azuma-Hoeffding proof idea

The sketch of the proof is as follows:

- (Doob decomposition) Decompose the supermartingale X_i into $Y_i + Z_i$, where Y_i is a martingale, and Z_i is decreasing wp1.
- Since Z_i is decreasing, then the event $X_n - X_0 \geqslant t$ implies the event $Y_{n+1} - Y_0 \geqslant t$.
- Write $Y_n - Y_0 = \sum_{i \in [n]} (Y_i - Y_{i-1})$.
- Write a Chernoff bound for $Y_n - Y_0$, and compute the MGF of Y_n.

Azuma-Hoeffding proof idea

The sketch of the proof is as follows:

- (Doob decomposition) Decompose the supermartingale X_i into $Y_i + Z_i$, where Y_i is a martingale, and Z_i is decreasing wp1.
- Since Z_i is decreasing, then the event $X_n - X_0 \geqslant t$ implies the event $Y_{n+1} - Y_0 \geqslant t$.
- Write $Y_n - Y_0 = \sum_{i \in [n]} (Y_i - Y_{i-1})$.
- Write a Chernoff bound for $Y_n - Y_0$, and compute the MGF of Y_n.
- Bound each term in the telescoping sum of the MGF via the tower property:

$$\mathbb{E} \exp\left(s \sum_{i \in [n]} (Y_i - Y_{i-1}) \right) = \mathbb{E}\left[\exp\left(s \sum_{i \in [n-1]} (Y_i - Y_{i-1}) \right) \mathbb{E}\left(s(Y_n - Y_{n-1}) \,\middle|\, Y_0, \ldots, Y_{n-1} \right) \right]$$

- Use $X_{n+1} - X_n \leqslant B_{n+1} - A_{n+1} \leqslant c_{n+1}$ to bound the conditional difference of $Y_{n+1} - Y_n$.

Azuma-Hoeffding proof idea

The sketch of the proof is as follows:

- (Doob decomposition) Decompose the supermartingale X_i into $Y_i + Z_i$, where Y_i is a martingale, and Z_i is decreasing wp1.
- Since Z_i is decreasing, then the event $X_n - X_0 \geqslant t$ implies the event $Y_{n+1} - Y_0 \geqslant t$.
- Write $Y_n - Y_0 = \sum_{i \in [n]} (Y_i - Y_{i-1})$.
- Write a Chernoff bound for $Y_n - Y_0$, and compute the MGF of Y_n.
- Bound each term in the telescoping sum of the MGF via the tower property:

$$\mathbb{E} \exp \left(s \sum_{i \in [n]} (Y_i - Y_{i-1}) \right) = \mathbb{E} \left[\exp \left(s \sum_{i \in [n-1]} (Y_i - Y_{i-1}) \right) \mathbb{E} \left(s(Y_n - Y_{n-1}) \,\middle|\, Y_0, \ldots, Y_{n-1} \right) \right]$$

- Use $X_{n+1} - X_n \leqslant B_{n+1} - A_{n+1} \leqslant c_{n+1}$ to bound the conditional difference of $Y_{n+1} - Y_n$.
- Use Hoeffding's lemma on each telescoping term.

I.e.: Azuma's inequality rests heavily on Chernoff/Hoeffding arguments.

Functions with bounded differences

Here's a well-known application of Azuma's inequality:

A function $f : D_1 \times D_2 \times \cdots \times D_n \to \mathbb{R}$ satisfies the *bounded difference* property if

$$\sup_{y_i \in D_i} |f(\boldsymbol{x}) - f(\boldsymbol{x}_{i,y_i})| \leqslant c_i \in (0, \infty), \qquad i \in [n], \qquad \boldsymbol{x}_{i,y_i} = (x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_n).$$

I.e., replacing the value in one coordinate has bounded impact on the function.

Functions with bounded differences

Here's a well-known application of Azuma's inequality:

A function $f : D_1 \times D_2 \times \cdots \times D_n \to \mathbb{R}$ satisfies the *bounded difference* property if

$$\sup_{y_i \in D_i} |f(\boldsymbol{x}) - f(\boldsymbol{x}_{i,y_i})| \leqslant c_i \in (0, \infty), \qquad i \in [n], \qquad \boldsymbol{x}_{i,y_i} = (x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_n).$$

I.e., replacing the value in one coordinate has bounded impact on the function.

Now suppose that $\{X_i\}_{i \in [n]}$ are independent random variables, $X_i \in D_i$ wp1. Define:

$$Y_i := \mathbb{E}\left[f(\boldsymbol{X}) \,\middle|\, X_1, \ldots, X_i\right].$$

One can show that Y_i is a martingale, and in particular that

$$|Y_i - Y_{i-1}| \leqslant c_i.$$

McDiarmid's inequality

Hence, Azuma's inequality yields the following result.

Theorem (McDiarmid's Inequality). *Suppose* $f : \times_{i \in [n]} D_i \to \mathbb{R}$ *is a function satisfying the bounded differences property with constants* $(c_i)_{i \in [n]}$*. Let* $\{X_i\}_{i \in [n]}$ *be a sequence of independent random variables, with* $X_i \in D_i$*. Then for all* $t > 0$*:*

$$\Pr\left(|f(\boldsymbol{X}) - \mathbb{E}f(\boldsymbol{X})| \geqslant t\right) \leqslant \exp\left(-\frac{2t^2}{\sum_{i \in [n]} c_i^2}\right).$$

Note that f can be a fairly general function.

There are generalizations to "sub-Gaussian differences", or to differences that are bounded with reasonably high probability.

Distribution functions

Here's a nice application of McDiarmid's inequality:

Let X be a random variable, and let $\{X_i\}_{i \in \mathbb{N}} \stackrel{\text{iid}}{\sim} X$.
Let X have distribution function F, and let F_n be the n-sample empirical distribution function:

$$F_n(z) := \frac{1}{n} \sum_{i \in [n]} \mathbb{1}(X_i \leqslant z)$$

Distribution functions

Here's a nice application of McDiarmid's inequality:

Let X be a random variable, and let $\{X_i\}_{i \in \mathbb{N}} \stackrel{\text{iid}}{\sim} X$.
Let X have distribution function F, and let F_n be the n-sample empirical distribution function:

$$F_n(z) := \frac{1}{n} \sum_{i \in [n]} \mathbb{1}(X_i \leqslant z)$$

Now fix n and z, and define $f : \mathbb{R}^n \to \mathbb{R}$ as:

$$f_z(\boldsymbol{x}) := \frac{1}{n} \sum_{i \in [n]} \mathbb{1}(x_i \leqslant z)$$

Note that:

$$\mathbb{E} f_z(\boldsymbol{X}) = \frac{1}{n} \sum_{i \in [n]} \mathbb{E}\mathbb{1}(X_i \leqslant z) = \frac{1}{n} \sum_{i \in [n]} F_X(z) = F_X(z).$$

Distribution functions

Here's a nice application of McDiarmid's inequality:

Let X be a random variable, and let $\{X_i\}_{i \in \mathbb{N}} \overset{\text{iid}}{\sim} X$.
Let X have distribution function F, and let F_n be the n-sample empirical distribution function:

$$F_n(z) := \frac{1}{n} \sum_{i \in [n]} \mathbb{1}(X_i \leqslant z)$$

Now fix n and z, and define $f : \mathbb{R}^n \to \mathbb{R}$ as:

$$f_z(\boldsymbol{x}) := \frac{1}{n} \sum_{i \in [n]} \mathbb{1}(x_i \leqslant z)$$

Note that:

$$\mathbb{E} f_z(\boldsymbol{X}) = \frac{1}{n} \sum_{i \in [n]} \mathbb{E}\mathbb{1}(X_i \leqslant z) = \frac{1}{n} \sum_{i \in [n]} F_X(z) = F_X(z).$$

Also note that f_z is a function of bounded differences:

$$\left| f_z(\boldsymbol{x}) - f_z(\boldsymbol{x}_{i,y_i}) \right| \leqslant \frac{1}{n} = c_i$$

The Dvoretsky-Kiefer-Wolfowitz-Massart theorem

The previous analysis suggests that there is a concentration inequality for empirical distribution functions.

Theorem (Dvoretsky-Kiefer-Wolfowitz-Massart). *Let* $\{X_i\}_{i \in \mathbb{N}} \overset{\text{iid}}{\sim} X$. *Let* X *have distribution function* F_X, *and let* F_n *be the empirical distribution function of* $\{X_i\}_{i \in [n]}$. *Then for every* $t > 0$*:*

$$\Pr\left(\|F_X(\cdot) - F_n(\cdot)\|_{L^\infty(\mathbb{R})} \leqslant t\right) \leqslant 2\exp(-2nt^2).$$

This ensures convergence of distribution functions in high probability.