Some Basic Concentration Inequalities

Andrew Nobel

March, 2023

Concentration Inequalities

Recall: For a random variable X

- $\mathbb{E}X$ tells us about the center of its distribution
- $\mathrm{Var}(X)$ tells us about the spread of its distribution

Concentration Inequalities: Bounds on the probability that a random variable is far from its expectation

$$\mathbb{P}(X \geq \mathbb{E}X + t) \qquad \mathbb{P}(X \leq \mathbb{E}X - t) \qquad \mathbb{P}(|X - \mathbb{E}X| \geq t)$$

- Often $X = U_1 + \cdots + U_n$ sum of independent random variables
- More generally, $X =$ function of independent random variables
- Many applications in statistics, machine learning, probability

Markov and Chebyshev

Markov's and Chebyshev's Inequalities

Markov's inequality: If $X \geq 0$ and $t > 0$ then

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

Chebyshev's Inequality: If $\mathbb{E}X^2 < \infty$ then for each $t > 0$

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

- Upper bound may be larger than 1 (not useful)
- Upper bound is less than 1 if $t >$ SD(X)

Extending Chebyshev

Applying same proof idea we can show that for each $t > 0$,

$$\mathbb{P}(|X - \mathbb{E}X| \geq t) \;\leq\; \min_{s>0} \frac{\mathbb{E}|X - \mathbb{E}X|^s}{t^s}$$

Upshot: smaller central moments yield better upper bounds

Application: Weak Law of Large Numbers

WLLN: Let $U_1, U_2, \ldots, U$ be iid with $\mathrm{Var}(U)$ finite. Then for each $t > 0$,

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} U_i \, - \, \mathbb{E}(U) \right| \geq t \right) \to 0$$

Proof: Apply Chebyshev's inequality to $X = n^{-1} \sum_{i=1}^{n} U_i$

Order of Magnitude

Note: If $X_1, X_2, \ldots$ are iid with $\mathbb{E}X_i = \mu$ and $0 < \mathrm{Var}(X_i) = \sigma^2 < \infty$ then by CLT

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \approx \mathcal{N}(0,1)$$

Corollaries

1. The centered sum $\sum_{i=1}^n X_i - n\mu$ is of order $\sigma\sqrt{n}$
2. The centered average $n^{-1}\sum_{i=1}^n X_i - \mu$ is of order $\sigma/\sqrt{n}$

Upshot: Probability $\mathbb{P}(\sum_{i=1}^n X_i - n\mu \geq t)$ can be small only if $t \gtrsim \sigma\sqrt{n}$

MGFs and Chernoff Bound

Moment Generating Functions

Recall: The moment generating function (MGF) of a rv X is defined by

$$M_X(s) = \mathbb{E}\left[e^{sX}\right] \quad \text{for } s \in \mathbb{R}$$

Note that $M_X(s) \geq 0$, and that $M_X(s)$ may be $+\infty$.

Fact: if $X_1, \ldots, X_n$ are independent and MGFs $M_{X_i}(s)$ are finite in a neighborhood of 0 then $S_n = X_1 + \cdots + X_n$ has MGF

$$M_{S_n}(s) \;=\; \prod_{i=1}^{n} M_{X_i}(s)$$

MGFs are a useful tool in the study of sums of independent random variables

MGF Examples

1. Normal: If $X \sim \mathcal{N}(0, \sigma^2)$ then $M_X(s) = e^{s^2\sigma^2/2}$
2. Poisson: If $X \sim \text{Poiss}(\lambda)$ then $M_X(s) = e^{\lambda(e^s - 1)}$
3. Chi-squared: If $X \sim \chi^2_k$ then $M_X(s) = (1 - 2s)^{-k/2}$ for $s < 1/2$
4. Sign: If $X = 1, -1$ with probability $1/2$ then $M_X(s) = (e^s + e^{-s})/2$

Chernoff's Bound

Chernoff Bound: For any random variable X and $t \in \mathbb{R}$

$$\mathbb{P}(X \geq t) \;\leq\; \min_{s>0} e^{-st}\,\mathbb{E}e^{sX} \;=\; \min_{s>0} e^{-st}\,M_X(s)$$

Corollary: If MGF of $(X - \mathbb{E}X)$ is bounded by $M(s)$ for $s \geq 0$, then for $t > 0$

$$\mathbb{P}(X \geq \mathbb{E}X + t) \;\leq\; \inf_{s>0} e^{-st}\,M(s)$$

- Inequalities for left tail $\mathbb{P}(X \leq \mathbb{E}X - t)$ established in same way
- Bound on $\mathbb{P}(|X - \mathbb{E}X| \geq t)$ can be obtained by adding L/R tail bounds

Bound for Chi-squared Distribution

Fact: Let $X \sim \chi^2_k$. Then

1. $X \stackrel{d}{=} \sum_{i=1}^k Z_i^2$ where Z_i are iid $\sim \mathcal{N}(0,1)$
2. $\mathbb{E}X = k$ and $\mathrm{Var}(X) = 2k$
3. $M_X(s) = (1-2s)^{-k/2}$ for $s < 1/2$

Fact: For $x \geq 0$, $1 + x \leq \exp\{x - (x^2 - x^3)/2\}$

Proposition: If $X \sim \chi^2_k$ then for $0 \leq \epsilon \leq 1$

1. $\mathbb{P}(X \geq (1+\epsilon)k) \ \leq \ \exp\{-k(\epsilon^2 - \epsilon^3)/4\}$
2. $\mathbb{P}(X \leq (1-\epsilon)k) \ \leq \ \exp\{-k(\epsilon^2 - \epsilon^3)/4\}$

Application: Low Dimensional Euclidean Embeddings

Basic Embedding Problem

Question: Can we embed given vectors $x_1, \ldots, x_n \in \mathbb{R}^d$ in a lower dimensional space while preserving their pairwise distances?

Definition: Let $\epsilon \in (0, 1)$. A function $f : \mathbb{R}^d \to \mathbb{R}^k$ is an ϵ-embedding of $x_1, \ldots, x_n$ if for all $1 \le i, j \le n$

$$(1-\epsilon)\,||x_i - x_j||^2 \;\le\; ||f(x_i) - f(x_j)||^2 \;\le\; (1+\epsilon)\,||x_i - x_j||^2$$

Upshot

- Establish *existence* of linear embeddings using probabilistic arguments
- Existence requires $k \gtrsim \log n/\epsilon^2$, independent of dimension d

Random Projections via Gaussian Random Matrices

GRM: Let W be a $k \times d$ matrix with iid $\mathcal{N}(0,1)$ entries

Fact: Fix $u \in \mathbb{R}^d$ and define the random vector $V = (V_1, \ldots, V_k)^t = k^{-1/2}\, Wu$

1. $V_1, \ldots, V_k$ are iid $\mathcal{N}(0, ||u||^2/k)$
2. If $k \geq 8(\epsilon^2 - \epsilon^3)^{-1} \log n$ then

$$\mathbb{P}(||V||^2 \leq (1-\epsilon)||u||^2) \;\leq\; \frac{1}{n^2} \quad \text{and} \quad \mathbb{P}(||V||^2 \geq (1+\epsilon)||u||^2) \;\leq\; \frac{1}{n^2}$$

Johnson-Lindenstrauss Lemma

Recall: Function $f : \mathbb{R}^d \to \mathbb{R}^k$ is an ϵ-embedding of $x_1, \ldots, x_n \in \mathbb{R}^d$ if for $1 \leq i, j \leq n$

$$(1 - \epsilon)\,||x_i - x_j||^2 \;\leq\; ||f(x_i) - f(x_j)||^2 \;\leq\; (1 + \epsilon)\,||x_i - x_j||^2$$

Theorem: Let W be a $k \times d$ matrix with iid $\mathcal{N}(0,1)$ entries. Define $f_W : \mathbb{R}^d \to \mathbb{R}^k$ by

$$f_W(x) = k^{-1/2}\, W x$$

If $k \geq 8(\epsilon^2 - \epsilon^3)^{-1} \log n$ then for each fixed sequence $x_1, \ldots, x_n \in \mathbb{R}^d$

$$\mathbb{P}\big(f_W \text{ is an } \epsilon\text{-embedding of } x_1, \ldots, x_n\big) \;\geq\; 1/n$$

Upshot: An ϵ-embedding of $x_1, \ldots, x_n$ exists. In practice, we can generate GRMs W until we find one that works

Hoeffding’s Inequality

Hoeffding's MGF Bound and Hoeffding's Inequality

MGF bound: If $X \in [a, b]$ then for every $s \geq 0$

$$\mathbb{E}e^{s(X-\mathbb{E}X)} \leq e^{s^2(b-a)^2/8}$$

Hoeffding's Inequality: Let $X_1, \ldots, X_n$ be independent with $a_i \leq X_i \leq b_i$ and let $S_n = X_1 + \cdots + X_n$. For every $t \geq 0$,

$$\mathbb{P}(S_n - \mathbb{E}S_n \geq t) \ \leq \ \exp\left\{\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right\}$$

Also $\mathbb{P}(S_n - \mathbb{E}S_n \leq -t) \leq$ RHS and $\mathbb{P}(|S_n - \mathbb{E}S_n| \geq t) \leq$ 2 RHS

Note: Hoeffding bound does *not* use information about the variance of the X_is

Example: Bernoulli Random Variables

Let $X_1, \ldots, X_n$ be iid Bern(p). Note that $\mathbb{E}(\sum_{i=1}^n X_i) = np$

Chebyshev: Uses $\mathrm{Var}(X_i) = p(1-p)$. For each $t \geq 0$

$$\mathbb{P}\left(\sum_{i=1}^n X_i - np \geq t\right) \leq \frac{n\,p(1-p)}{t^2} \leq \frac{n}{4t^2}$$

Hoeffding: Uses $0 \leq X_i \leq 1$. For each $t \geq 0$

$$\mathbb{P}\left(\sum_{i=1}^n X_i - np \geq t\right) \leq \exp\left\{\frac{-2t^2}{n}\right\}$$

Note: Bounds meaningful only when $t \gtrsim \sqrt{n}$. Hoeffding bound independent of p!

Bernoulli Example, cont.

Compare bounds of Chebyshev and Hoeffding when $n = 100$ and $p = 1/2$

t	Chebyshev	Hoeffding
5	1	.607
10	.250	.135
12	.173	.0561
14	.128	.0198
16	.0977	.0060
20	.0625	.000335

Upshot: Once bounds kick in, Hoeffding is better

Bernoulli Example, cont.

Bounds for sums can be converted into bounds for averages, and vice versa

Chebyshev: For each $t \geq 0$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - p \geq t\right) \leq \frac{p(1-p)}{n\,t^2} \leq \frac{1}{4\,n\,t^2}$$

Hoeffding: For each $t \geq 0$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - p \geq t\right) \leq \exp\left\{-2\,n\,t^2\right\}$$

Note: Upper bounds useful only when $t \gtrsim 1/\sqrt{n}$

Other Examples of Hoeffding's Inequality

Ex: Let $X_1, \ldots, X_n \in \mathcal{X}$ be iid with distribution P and let $A \subseteq \mathcal{X}$. For $t \geq 0$,

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i \in A) - P(A) \right| \geq t \right) \leq 2 \exp\left\{ -2nt^2 \right\}$$

Ex: Let $X_1, \ldots, X_n$ iid $\sim$ U$(-\theta, \theta)$. Note that $\mathbb{E}X = 0$. For $t \geq 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq t \right) \leq \exp\left\{ \frac{-t^2}{2n\theta^2} \right\}$$

Bennett and Bernstein Inequalities

Bennett and Bernstein Inequalities

MGF bound: If $\mathbb{E}X = 0$, $\mathrm{Var}(X) = \sigma^2$, and $|X| \leq c$ then

$$M_X(s) \leq \exp\{c^{-2}\sigma^2(e^{sc} - 1 - sc)\}$$

Bennett's Inequality: If $X_1, \ldots, X_n$ are independent with $\mathbb{E}X = 0$, $\mathrm{Var}(X_i) = \sigma_i^2$, and $|X_i| \leq c$, then for every $t \geq 0$,

$$\mathbb{P}(S_n \geq t) \;\leq\; \exp\left\{\frac{-n\sigma^2}{c^2} \cdot h\left(\frac{ct}{n\sigma^2}\right)\right\}$$

where $\sigma^2 = n^{-1}\sum_{i=1}^n \sigma_i^2$ and $h(u) = (1+u)\log(1+u) - u$.

Bernstein's Inequality: Under the same conditions, for every $t \geq 0$,

$$\mathbb{P}(S_n \geq t) \;\leq\; \exp\left\{\frac{-t^2}{2n\sigma^2 + 2ct/3}\right\}$$

Bernstein vs. Hoeffding

Let $X_1, \ldots, X_n$ be independent with $\mathbb{E}X = 0$ and $|X_i| \leq c$. If $t \geq n^{-1} \sum_{i=1}^n \mathrm{Var}(X_i)$ then Bernstein's inequality yields

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n X_i \geq t \right) \leq \exp\left\{ \frac{-nt}{2 + 2c/3} \right\}$$

while Hoeffding's inequality yields

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n X_i \geq t \right) \leq \exp\left\{ \frac{-nt^2}{2c^2} \right\}$$

Note: If $X_1, \ldots, X_n$ are Bern(p) with $p \geq 1/2$, Bernstein's inequality shows for $t \geq 0$

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n X_i - p \geq t \right) \leq \exp\left\{ \frac{-3nt^2}{8p(1-p)} \right\}$$

which is (up to constants) what one expects from the CLT

Bounds on Expectations

Idea: Bounds on tail probabilities yield bounds on expectations

Fact: Let X be a random variable, $a \geq 1$, and $b > 0$

1. If $\mathbb{P}(|X| \geq t) \leq a\, e^{-bt}$ for $t \geq 0$ then $\mathbb{E}|X| \;\leq\; (1 + \log a)/b$

2. If $\mathbb{P}(|X| \geq t) \leq a\, e^{-bt^2}$ for $t \geq 0$ then $\mathbb{E}|X| \;\leq\; \sqrt{(1 + \log a)/b}$

General Concentration Inequalities

Hoeffding, Bennett, and Bernstein inequalities show that a sum $\sum_{i=1}^{n} X_i$ of bounded, independent random variables is close to its mean

Goal: Inequalities for functions $f(X_1, \ldots, X_n)$ of independent random variables

- Chernoff inequality and upper bounds on the MGF of $f(X_1, \ldots, X_n)$
- Martingale differences and Gaussian smart-path argument
- Key assumption: the value of $f(x_1, \ldots, x_n)$ does not change too much if we make a small changes to any single argument x_i

Azuma-Hoeffding Inequality

Martingale Differences

Setting: Random variables $X_1, \ldots, X_n$ defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and nested sigma fields $\{\emptyset, \Omega\} = \mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_n \subseteq \mathcal{F}$

Definition: $X_1, \ldots, X_n$ is a *martingale difference* with respect to $\mathcal{F}_0, \ldots, \mathcal{F}_n$ if

1. X_i is measurable $\mathcal{F}_i$
2. $\mathbb{E}|X_i| < \infty$
3. $\mathbb{E}(X_i \,|\, \mathcal{F}_{i-1}) = 0$

In many cases $\mathcal{F}_i$ is the sigma field $\sigma(X_1^i)$ generated by $X_1, \ldots, X_i$

Martingale Differences

Fact: If $X_1, \ldots, X_n$ is a martingale difference with respect to $\mathcal{F}_0, \ldots, \mathcal{F}_n$ then

1. $\mathbb{E}X_i = 0$ for $i = 1, \ldots, n$
2. $\mathbb{E}(X_i X_j) = 0$ if $i \neq j$
3. $\mathrm{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathrm{Var}(X_i)$

Fact: Let X be a random variable and $\mathcal{G} \subseteq \mathcal{F}$ a sigma field such that

1. $\mathbb{E}(X \,|\, \mathcal{G}) = 0$
2. There exists $\mathcal{G}$-measurable U and $c \geq 0$ such that $U \leq X \leq U + c$ wp1

Then $\mathbb{E}(\exp(sX) \,|\, \mathcal{G}) \leq \exp(s^2 c^2 / 8)$

Azuma-Hoeffding Inequality

Fact: Let $X_1, \ldots, X_n$ be a martingale difference with respect to $\mathcal{F}_0, \ldots, \mathcal{F}_n$. Suppose that for each $1 \leq i \leq n$ there is a rv U_{i-1} measurable $\mathcal{F}_{i-1}$ and $c_{i-1} \geq 0$ such that

$$U_{i-1} \ \leq \ X_i \ \leq \ U_{i-1} + c_{i-1}$$

with probability one. Then for each $t > 0$

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq t\right) \ \leq \ \exp\left\{\frac{-2t^2}{\sum_{i=1}^{n} c_i^2}\right\}$$

Note: The same upper bound holds for $\mathbb{P}\left(\sum_{i=1}^{n} X_i \leq -t\right)$

Bounded Difference Inequality

Bounded Difference Inequality

Setting: Let $\mathcal{X}$ be a set, possibly finite

- Function $f : \mathcal{X}^n \to \mathbb{R}$
- $X_1, \ldots, X_n \in \mathcal{X}$ independent, not necessarily identically distributed

Of interest: bounds on the probability that the random variable

$$Z = f(X_1, \ldots, X_n)$$

is far from its mean $\mathbb{E}Z$

Bounded Difference Inequality

Definition: The ith *difference coefficient* c_i of f is the maximum possible change in the value of f if we change the value of the ith coordinate,

$$c_i \; = \; \sup |f(x_1^n) - f(x_1^{i-1}, x_i', x_{i+1}^n)|$$

where the supremum is over all sequences $x_1, \ldots, x_i, x_i', x_{i+1}, \ldots, x_n \in \mathcal{X}$

Theorem (McDiarmid): If $X_1, \ldots, X_n \in \mathcal{X}$ are independent, then for every $t \geq 0$

$$\mathbb{P}\left(|f(X_1^n) - \mathbb{E}f(X_1^n)| \geq t\right) \; \leq \; 2 \, \exp\left\{ \frac{-2t^2}{\sum_{i=1}^n c_i^2} \right\}$$

Moreover $\mathrm{Var}(f(X_1^n)) \leq \sum_{i=1}^n c_i^2/4$

Examples

Bin Packing: Fix $n \geq 1$. The bin-packing function $f : [0,1]^n \to \mathbb{N}$ is defined by

$$f_n(x_1^n) \;=\; \text{min \# size 1 bins needed to hold objects of size } x_1^n$$

Uniform LLN: Let $X_1, \ldots, X_n \in \mathcal{X}$ be iid and let $\mathcal{G}$ be a family of functions $g : \mathcal{X} \to [-c, c]$. Define $f : \mathcal{X}^n \to \mathbb{R}$ by

$$f_n(x_1^n) \;=\; \sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^{n} g(x_i) - \mathbb{E} g(X) \right|$$

Gaussian Concentration Inequality

Gaussian Concentration Inequality

Definition: A function $F : \mathbb{R}^n \to \mathbb{R}$ is *Lipschitz continuous* with Lipschitz constant L if for every $x, y \in \mathbb{R}^n$

$$|F(x) - F(y)| \ \leq \ L \, ||x - y||$$

Theorem: Let $X_1, \ldots, X_n$ be iid $\sim \mathcal{N}(0,1)$. If $F : \mathbb{R}^n \to \mathbb{R}$ is Lipschitz continuous with constant L then for every $t > 0$

$$\mathbb{P}\left(F(X_1^n) - \mathbb{E}F(X_1^n) \geq t\right) \ \leq \ \exp\left\{\frac{-t^2}{2L^2}\right\}$$

The same bound holds for $\mathbb{P}\left(F(X_1^n) - \mathbb{E}F(X_1^n) \leq -t\right)$

Examples

Ex: maximum of multinormal: Let $Y \sim \mathcal{N}_d(0, \Sigma)$. Find concentration inequality for

$$U = \max(Y_1, \ldots, Y_d)$$

Ex: ℓ_p-norm of multinormal: Let $Y \sim \mathcal{N}_d(0, \Sigma)$. Find concentration inequality for

$$U = ||Y||_{\ell_p} = \left(\sum_{i=1}^{d} |Y_i|^p \right)^{1/p}$$

Association Inequalities for Expectations

Definition: A function $f : \mathbb{R} \to \mathbb{R}$ is

- *non-decreasing* if $x \leq y$ implies $f(x) \leq f(y)$
- *non-increasing* if $x \leq y$ implies $f(x) \geq f(y)$

Theorem: Let X be a random variable and let $f, g : \mathbb{R} \to \mathbb{R}$. Assuming all expectations are well-defined,

(a) f, g non-decreasing implies $\mathbb{E}(f(X)\, g(X)) \geq \mathbb{E}f(X)\, \mathbb{E}g(X)$

(b) f, g non-increasing implies $\mathbb{E}(f(X)\, g(X)) \geq \mathbb{E}f(X)\, \mathbb{E}g(X)$

(c) f non-decreasing and g non-increasing implies $\mathbb{E}(f(X)\, g(X)) \leq \mathbb{E}f(X)\, \mathbb{E}g(X)$

Association Inequality Examples

1. $\mathbb{E}(X^4) \geq \mathbb{E}(X)\,\mathbb{E}(X^3)$

2. $\mathbb{E}(Xe^{-X}) \leq \mathbb{E}(X)\,\mathbb{E}(e^{-X})$.

3. $\mathbb{E}[X\,\mathbb{I}(X \geq a)] \geq \mathbb{E}(X)\,\mathbb{P}(X \geq a)$