Chapter 6. Concentration Inequalities

6.3: Even More Inequalities

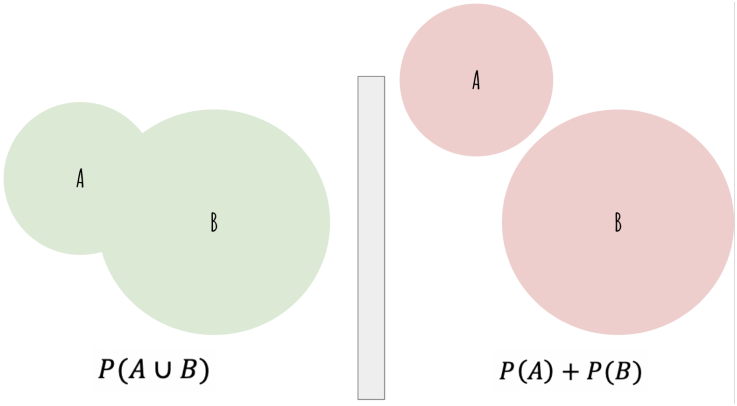
Slides (Google Drive) Alex Tsun Video (YouTube)

In this section, we will talk about a potpourri of remaining concentration bounds. More specifically, the union bound, Jensen's inequality for convex functions, and Hoeffding's inequality.

6.3.1 The Union Bound

Suppose there are many bad events $B_1, \ldots, B_n$, and we don't want any of them to happen. They may or may not be independent. Can we bound the probability that any (at least one) bad event occurs?

The intuition for the union bound is fairly simple. Suppose we have two events A and B. Then $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$ since the event space of A and B may overlap:



We will now define the Union Bound more formally.

Theorem 6.3.1: The Union Bound

Let $E_1, E_2, ..., E_n$ be a collection of events. Then:

$$\mathbb{P}\left(\bigcup_{i=1}^{n} E_i\right) \leq \sum_{i=1}^{n} \mathbb{P}(E_i)$$

Additionally, if $E_1, E_2, ...$ is a (countably) infinite collection of events, then:

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i\right) \leq \sum_{i=1}^{\infty} \mathbb{P}(E_i)$$

We can prove the union bound using induction.

Proof of Union Bound by Induction.
Base Case: For $n = 2$ events, by inclusion-exclusion, we know

$$\begin{aligned}\mathbb{P}(A \cup B) &= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \\ &\leq \mathbb{P}(A) + \mathbb{P}(B) \qquad [\text{since } \mathbb{P}(A \cap B) \geq 0]\end{aligned}$$

Inductive Hypothesis: Suppose it's true for n events, $\mathbb{P}(E_1 \cup ... \cup E_n) \leq \sum_{i=1}^{n} \mathbb{P}(E_i)$.
Inductive Step: We will show it for $n + 1$.

$$\begin{aligned}\mathbb{P}(E_1 \cup ... \cup E_n \cup E_{n+1}) &= \mathbb{P}((E_1 \cup ... \cup E_n) \cup E_{n+1}) && [\text{associativity of } \cup] \\ &= \mathbb{P}(E_1 \cup ... \cup E_n) + \mathbb{P}(E_{n+1}) && [\text{base case}] \\ &\leq \sum_{i=1}^{n} \mathbb{P}(E_i) + \mathbb{P}(E_{n+1}) && [\text{inductive hypothesis}] \\ &= \sum_{i=1}^{n+1} \mathbb{P}(E_i)\end{aligned}$$

□

The union bound, though seemingly trivial, can actually be quite useful.

Example(s)

This will relate to the earlier question of bounding the probability of at least one bad event happening.

Suppose the probability Alex is late to teaching class on a given day is at most 0.01. Bound the probability that Alex is late at least once over a 30-class quarter. Do **not** make any independence assumptions.

Solution
Let A_i be the event Alex is late to class on day i for $i = 1, \ldots, 30$. Then, by the union bound,

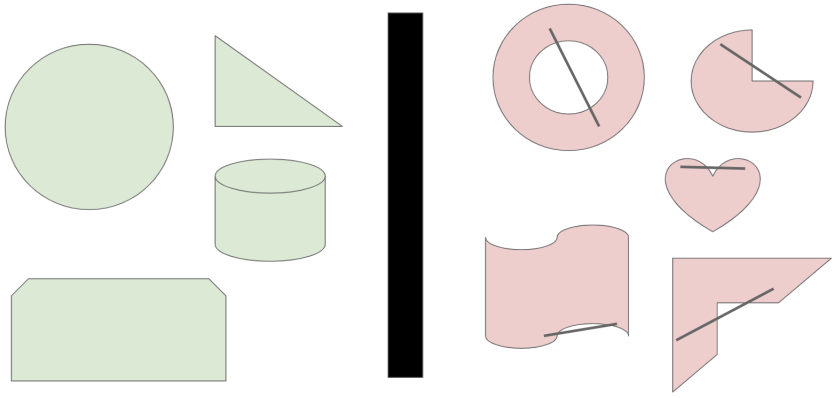
$$\begin{aligned}\mathbb{P}(\text{late at least once}) &= \mathbb{P}\left(\bigcup_{i=1}^{30} A_i\right) \\ &\leq \sum_{i=1}^{30} \mathbb{P}(A_i) && [\text{union bound}] \\ &\leq \sum_{i=1}^{30} 0.01 && [\mathbb{P}(A_i) \leq 0.01] \\ &= 0.30\end{aligned}$$

Sometimes it may be useless though; imagine I asked instead about over a 200-day period. Then the union bound would've given me a bound of 2.0 which is not helpful since probabilities have to be at most 1 already...
□

6.3.2 Convex Sets and Functions

Our next inequality is called Jensen's inequality, and deals with convex functions. So first, we need to define what that means. Before convex functions though, we need to discuss convex sets.

Let's look at some examples of convex (left) and non-convex (right) sets:



The sets on the left hand side are said to be **convex** because if you take any two points in the set and draw the line segment between them, it is always contained in the set. The sets on the right hand side are non-convex because I found two endpoints in the set, but the line segment connecting them is not completely contained in the set.
How can we describe this mathematically? Well for *any* two points $x, y \in S$, the set of points between them must be entirely contained in S. The set of points making up the line segment between two points x, y can be described as a weighted average $(1-p)x + py$ for $p \in [0,1]$. If $p = 0$, we just get x; if $p = 1$, we just get y, and if $p = 1/2$, we get the midpoint $(x+y)/2$. So p controls the fraction of the way we are from x to y.

Definition 6.3.1: Convex Sets

A set $S \subseteq \mathbb{R}^n$ is a **convex set** if for any $x, y \in S$, the entire line segment between them is contained in S. That is, for any two points $x, y \in S$,

$$\overline{xy} = \{(1-p)x + py : p \in [0,1]\} \subseteq S$$

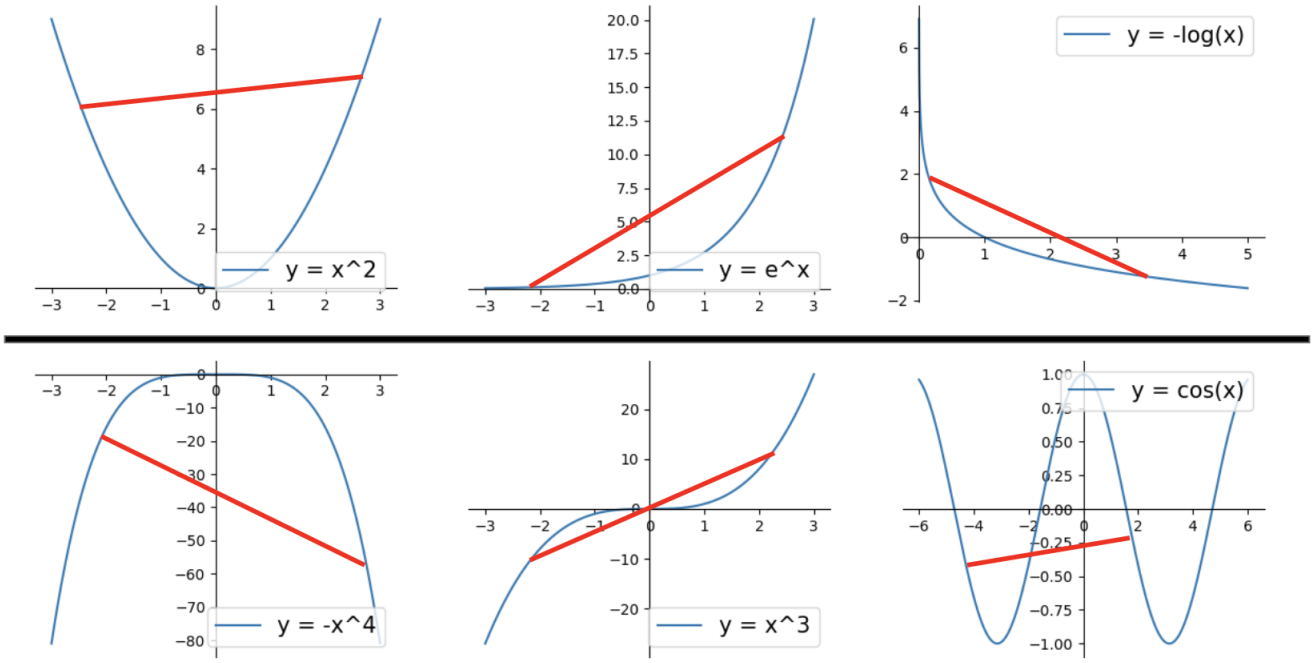
Equivalently, for any points $x_1, ..., x_m \in S$, convex polyhedron formed by the "corners" is contained in S. (This sounds complicated, but if $m = 3$, it just says the triangle formed by the 3 corners completely lies in the set S. If $m = 4$, the quadrilateral formed by the 4 corners completely lies in the set S.) The points in the convex polyhedron are described by taking weighted average of the points, where the weights are non-negative and sum to 1. (This should remind you of a probability distribution!)

$$\left\{ \sum_{i=1}^{m} p_i x_i : p_1, ..., p_m \geq 0 \text{ and } \sum_{i=1}^{m} p_i = 1 \right\} \subseteq S$$

Here are some examples of convex sets:

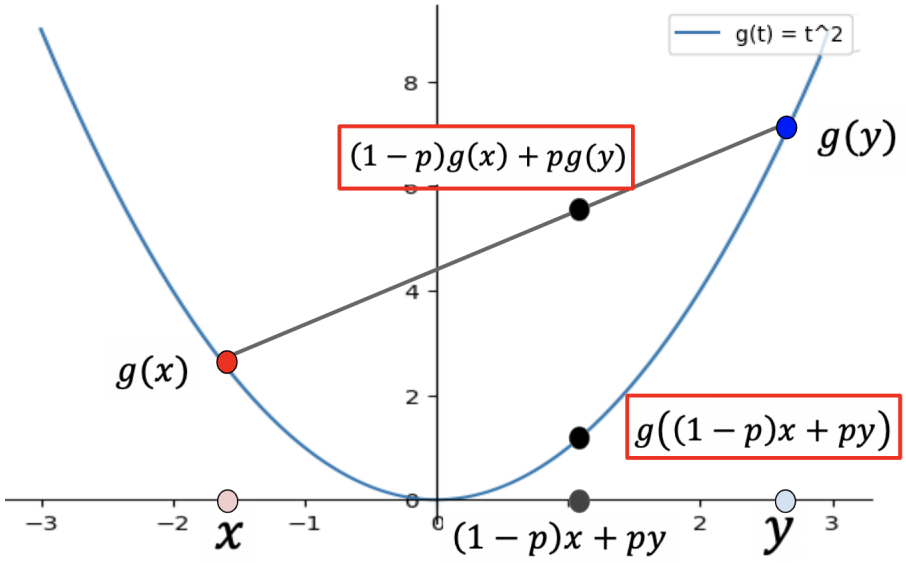
1. Any interval ($[a,b], (a,b)$, etc.) in $\mathbb{R}$ is a convex set (and the only convex sets in $\mathbb{R}$ are intervals).
2. The circle $C = \{(x,y) : x^2 + y^2 \leq 1\}$ in $\mathbb{R}^2$ is a convex set.
3. Any n-dimensional box $B = [a_1, b_1] \times [a_2, b_2] \times ... \times [a_n, b_n]$ is a convex set.

Now, onto convex *functions*. Let's take a look at some convex (top) and non-convex (bottom) functions:



The functions on the top (convex) have the property that, for **any** two points on the function curve, the line segment connecting them lies **above** the function always. The functions on the bottom don't have this property: you can see that some or all of the line segment is below the function.

Let's try to formalize what this means. For the convex function $g(t) = t^2$ below, we can see that any line drawn connecting 2 points of the function clearly lies above the function itself and so it is convex. Look at any two points on the curve $g(x)$ and $g(y)$. Pick a point on the x-axis between x and y, call it $(1-p)x+py$ where $p \in [0,1]$. The function value at this point is $g((1-p)x+py)$. The corresponding point above it on the line segment connecting $g(x)$ and $g(y)$ is actually the weighted average $(1-p)g(x)+pg(y)$. Hence, a function g is convex if it satisfies the following for any x, y and $p \in [0,1]$: $g((1-p)x+py) \leq (1-p)g(x)+pg(y)$



Definition 6.3.2: Convex Functions

Let $S \subseteq \mathbb{R}^n$ be a *convex set* (a convex function must have the domain being a convex set). A function $g : S \to \mathbb{R}$ is a **convex function** if for any line segment connecting $g(x)$ and $g(y)$, the function g lies entirely below the line. Mathematically, for any $p \in [0, 1]$ and $x, y \in \mathbb{R}$,

$$g((1-p)x + py) \leq (1-p)g(x) + pg(y)$$

Equivalently, for any m points $x_1, ..., x_m \in S$, and $p_1, ..., p_m \geq 0$ such that $\sum_{i=1}^{m} p_i = 1$,

$$g\left(\sum_{i=1}^{m} p_i x_i\right) \leq \sum_{i=1}^{m} p_i g(x_i)$$

Here are some examples of convex functions:

1. $g(x) = x^2$
2. $g(x) = x$
3. $g(x) = -\log(x)$
4. $g(x) = e^x$

6.3.3 Jensen's Inequality

Now after learning about convex sets and functions, we can learn Jensen's inequality, which relates $\mathbb{E}[g(X)]$ and $g(\mathbb{E}[X])$ for convex functions. Remember we said many times that these two quantities were never equal (use LOTUS to compute $\mathbb{E}[g(X)]$)!

Theorem 6.3.2: Jensen's Inequality

Let X be any random variable, and $g : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Then,

$$g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)]$$

Proof of Jensen's Inequality. We will only prove it in the case X is a discrete random variable (not a random vector), and with finite range (not countably infinite). However, this inequality does hold for any random variable.

The proof follows immediately from the definition of a convex function. Since X has finite range, let $\Omega_X = \{x_1, ..., x_n\}$ and $p_X(x_i) = p_i$. By definition of a convex function (see above),

$$\begin{aligned} g(\mathbb{E}[X]) &= g\left(\sum_{i=1}^{n} p_i x_i\right) && \text{[def of expectation]} \\ &\leq \sum_{i=1}^{n} p_i g(x_i) && \text{[def of convex function]} \\ &= \mathbb{E}[g(X)] && \text{[LOTUS]} \end{aligned}$$

□

Example(s)

Show that variance of any random variable X is always non-negative using Jensen's inequality.

Solution We already know that $\mathsf{Var}(X) = \mathbb{E}\left[(X-\mu)^2\right] \geq 0$ since $(X-\mu)^2$ is a non-negative RV, but let's prove it a different way.

We know $g(t) = t^2$ is a convex function, so by Jensen's inequality,

$$\mathbb{E}[X]^2 = g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)] = \mathbb{E}\left[X^2\right]$$

Hence $\mathsf{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \geq 0$. □

6.3.4 Hoeffding's Inequality

One final inequality that is commonly used is called Hoeffding's inequality. We'll state it without proof since it is quite complicated. The proof uses Jensen's inequality and ideas from the proof of the Chernoff bound (MGFs)!

Definition 6.3.3: Hoeffding's Inequality

Let $X_1, ...X_n$ be independent random variables, where each X_i is bounded: $a_i \leq X_i \leq b_i$ and let $\overline{X}_n$ be their sample mean. Then,

$$\mathbb{P}\left(|\overline{X}_n - \mathbb{E}\left[\overline{X}_n\right]| \geq t\right) \leq 2\ \exp\left(\frac{-2n^2t^2}{\sum_{i-1}^{n}(b_i - a_i)^2}\right)$$

where $\exp(x) = e^x$.

In the case $X_1, ..., X_n$ are iid (so $a \leq X_i \leq b$ for all i) with mean μ, then

$$\mathbb{P}\left(|\overline{X}_n - \mu| \geq t\right) \leq 2\ \exp\left(\frac{-2n^2t^2}{n(b-a)^2}\right) = 2\ \exp\left(\frac{-2nt^2}{(b-a)^2}\right)$$

Example(s)

Suppose an email company ColdMail is responsible for delivering 100 emails per day. ColdMail has a bad day if it takes longer than 190 seconds to deliver all 100 emails, and a bad week if there is even one bad day in the week.

The time it takes to send an email on average is 1 second, with a worst-case time of 5 seconds; independently of other emails. (Note we don't know anything else like its PDF).

1. Give an upper bound for the probability that ColdMail has a bad day.
2. Give an upper bound for the probability that ColdMail has a bad week.

Solution

1. In this scenario, we may use Hoeffding's inequality since we have $X_1, \ldots, X_{100}$ the (independent) times to send each email bounded in the interval $[0, 5]$ seconds, with $\mathbb{E}\left[\overline{X}_{100}\right] = 1$. Asking that the total time to be at least 190 seconds is the same as asking the mean time to be at least 1.9 seconds.

   Like we did for Chebyshev, we have to massage (and weaken) a little bit to get in the same form as required for Hoeffding's:

   $$\mathbb{P}\left(\overline{X}_{100} \geq 1.9\right) \leq \mathbb{P}\left(\overline{X}_{100} \geq 1.9 \cup \overline{X}_{100} \leq 0.1\right) = \mathbb{P}\left(|\overline{X}_{100} - 1| \geq 0.9\right)$$

   Applying Hoeffding's (since $\mathbb{E}\left[\overline{X}_n = 1\right]$):

   $$\mathbb{P}\left(\overline{X}_{100} \geq 1.9\right) \leq \mathbb{P}\left(|\overline{X}_{100} - 1| \geq 0.9\right) \leq 2\exp\left(\frac{-2 \cdot 100 \cdot 0.9^2}{(5-0)^2}\right) \approx 0.0031$$

2. For $i = 1, \ldots, 7$, let B_i be the event we had a bad day on day i. Then,

$$\begin{aligned}
\mathbb{P}\left(\text{bad week}\right) &= \mathbb{P}\left(\bigcup_{i=1}^{7} B_i\right) \\
&\leq \sum_{i=1}^{7} \mathbb{P}\left(B_i\right) && \text{[union bound]} \\
&\leq \sum_{i=1}^{7} 0.0031 && \text{[Hoeffding in previous part]} \\
&\approx 0.0215
\end{aligned}$$

You might be tempted to use the CLT (and you should when you can), as it would probably give a better bound than Hoeffding's. But we didn't know the variances, so we wouldn't know which Normal to use. Hoeffding's gives us a way! □