Chapter 3

Concentration Inequalities

집중도 부등식(?)
ㄴ concentration on 평균
$X_1 \cdots X_n$의 "합"이 ↳ 집중도 ↑: 민감도 ↓ (변수 변화해도 차이 less)

In this chapter, we take a little diversion and develop the notion of *concentration inequalities*. Assume that we have independent random variables $X_1, \ldots, X_n$. We will develop tools to show results that formalize the intuition for these statements:

indep RVs

1. $\underbrace{X_1 + \ldots + X_n}_{X_1 \cdots X_n\text{의 "합" 혹은}}$ concentrates around $\mathbb{E}[X_1 + \ldots + X_n]$.
   $X_1, \cdots X_n$에 "어떠한 처리 f"를 한것
2. More generally, $f(X_1, \ldots, X_n)$ concentrates around $\mathbb{E}[f(X_1, \ldots, X_n)]$.

These inequalities will be used in subsequent chapters to bound several key quantities of interest.

As it turns out, the material from this chapter constitutes arguably the important mathematical tools in the entire course. No matter what area of machine learning one wants to study, if it involves sample complexity, some kind of concentration result will typically be required. Hence, concentration inequalities are some of the most important tools in modern statistical learning theory. Sample Complexity:

More precisely, the sample complexity is the number of training-samples that we need to supply to the algorithm, so that the function returned by the algorithm is within an arbitrarily small error of the best possible function, with probability arbitrarily close to 1.

3.1 The big-O notation

Throughout the rest of this course, we will use "big-O" notation in the following sense: every occurrence of $O(x)$ is a placeholder for some function $f(x)$ such that for every x, $|f(x)| \leq Cx$ for some absolute/universal constant C. In other words, when $O(n_1), \ldots, O(n_k)$ occur in a statement, it means that **there exist** absolute constants $C_1, \ldots, C_k > 0$ and functions $f_1, \ldots, f_k$ satisfying $|f_i(x)| \leq C_i x$ for all x, such that after replacing each occurrence $O(n_i)$ by $f_i(n_i)$, the statement is true. The difference from traditional "big-O" notation is that we do not need to send $n \to \infty$ in order to define "big-O". In nearly all cases, big-O notation is used to define an upper bound; then, the bound is identical if we simply substitute Cx in place of $O(x)$.

Note that the x in our definition of big-O is a surrogate (대역) for an arbitrary variable. For instance, later on in this chapter, we will encounter the term $O(\sigma\sqrt{\log n})$. The definition above, applied with $x = \sigma\sqrt{\log n}$, yields the following conclusion: $O(\sigma\sqrt{\log n}) = f(\sigma\sqrt{\log n})$ for some function f and constant C such that $|f(\sigma\sqrt{\log n})| \leq C\sigma\sqrt{\log n}$ for all values that $\sigma\sqrt{\log n}$ can take.

Lastly, for any $a, b \geq 0$, we will let $a \lesssim b$ mean that there is some absolute constant $c > 0$ such that $a \leq cb$.

$a \lesssim b$ 의미: $\exists C>0$ s.t. $a \leq Cb$

3.2 Chebyshev's inequality

RV "Z" with finite 분산: tail behavior가 어케됨? →

We begin by considering an arbitrary random variable Z with finite variance. One of the most famous results characterizing its tail behavior is the following theorem:

(quadratic decay rates)

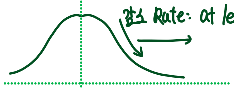
Theorem 3.1 (Chebyshev's inequality). *Let Z be a random variable with finite expectation and variance. Then*

Z: RV ① ②

$$\Pr[\underbrace{|Z - \mathbb{E}[Z]|}_{\text{평균과의 차}} \geq t] \leq \frac{\overbrace{\mathrm{Var}(Z)}^{\text{분산}}}{t^2}, \quad \forall t > 0. \tag{3.1}$$

$\Pr[|Z-\mathbb{E}[Z]| \le t] \ge 1 - \frac{\mathrm{Var}(Z)}{t^2}$ & $t = \text{standard dev}(Z)/\sqrt{\delta}$

Intuitively, this means that as we approach the tails of the distribution of Z, the density decreases at a rate of at least $1/t^2$. Moreover, for any $\delta \in (0, 1]$, by plugging in $t = \mathrm{sd}(Z)/\sqrt{\delta}$ to (3.1) we see that



감소 Rate: at least $1/t^2$.

$$\Pr\left[|Z - \mathbb{E}[Z]| \leq \frac{\mathrm{sd}(Z)}{\sqrt{\delta}}\right] \geq 1 - \delta. \tag{3.2}$$

quadratic rate(△) ($1/t^2$ rate)

Unfortunately, it turns out that Chebyshev's inequality is a rather weak concentration inequality. To illustrate this, assume $Z \sim \mathcal{N}(0, 1)$. We can show (using the Gaussian tail bound derived in Problem 3(c) in Homework 0) that mean: 0, SD: 1, Var: 1.

(and also using Chebyshev's inequality though)

$$\Pr\left[|Z - \mathbb{E}[Z]| \leq \mathrm{sd}(Z)\sqrt{2\log(2/\delta)}\right] \geq 1 - \delta. \tag{3.3}$$

↳ Sub-Gaussian dist의 정의. $P(|X-E(X)| \ge t) \le 2e^{-t^2/2\sigma^2}$

$* \ t = \sigma\sqrt{2\log(2/\delta)}$

$P(|X-E(X)| \ge \sigma\sqrt{2\log(2/\delta)}) \le \delta$

$P(|X-E(X)| \le \sigma\sqrt{2\log(2/\delta)}) \ge 1-\delta$ ← exponential Rate (△)

for any $\delta \in (0, 1]$. In other words, the density at the tails of the normal distribution is decreasing at an exponential rate, while Chebyshev's inequality only gives a quadratic rate. The discrepancy between (3.2) and (3.3) is made more apparent when we consider inverse-polynomial $\delta = \frac{1}{n^c}$ for some parameter n and degree c (we will see concrete instances of this setup in future chapters). Then the tail bound for the normal distribution (3.3) implies that

$O\left(\sqrt{2\log(2n^c)}\right) \to O\left(\sqrt{\log n}\right)$ 꼴..

$$|Z - \mathbb{E}[Z]| \leq \mathrm{sd}(Z) \cdot \sqrt{\log O(n^c)} = \mathrm{sd}(Z) \cdot O\left(\sqrt{\log n}\right) \quad \underbrace{w.p.\ 1 - \delta}_{\text{with prob.}}, \tag{3.4}$$

while Chebyshev's inequality gives us the weaker result

$O\left(\frac{1}{\sqrt{1/n^c}}\right)$ 꼴 → $O(n^{c/2})$ 꼴..

$$|Z - \mathbb{E}[Z]| \leq \mathrm{sd}(Z) \cdot \sqrt{O(n^c)} = \mathrm{sd}(Z) \cdot O(n^{c/2}) \quad w.p.\ 1 - \delta. \tag{3.5}$$

Chebyshev's inequality is actually optimal without further assumptions, in the sense that there exist distributions with finite variance for which the bound is tight. However, in many cases, we will be able to improve the $1/t^2$ rate (quadratic) of tail decay in Chebyshev's inequality to an e^{-t} rate. In the next two sections, we will demonstrate how to construct tail bounds with exponential decay rates.

3.3 Hoeffding's inequality (exponential decay rates)

We next provide a brief overview of Hoeffding's inequality, a concentration inequality for bounded random variables with an exponential tail bound:

Theorem 3.2 (Hoeffding's inequality). *Let $X_1, X_2, \ldots, X_n$ be independent real-valued random variables drawn from some distribution, such that $a_i \leq X_i \leq b_i$ almost surely. Define $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, and let $\mu = \mathbb{E}[\bar{X}]$. Then for any $\varepsilon > 0$,*

"happens" with prob 1.

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ Sample mean

$\mu = E[\bar{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]$ Population mean

$$\Pr\left[|\bar{X} - \mu| \leq \varepsilon\right] \geq 1 - 2\exp\left(\frac{-2n^2\varepsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right). \tag{3.6}$$

Note that the demoninator within the exponential term, $\sum_{i=1}^{n}(b_i - a_i)^2$, can be thought of as an upper bound (or proxy) for the variance $\mathrm{Var}(X_i)$. In fact, under the independence assumption, we can show

including... $\mathrm{Var}(\bar{X})$

$$\mathrm{Var}\left(\bar{X}\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) \leq \frac{1}{n^2}\sum_{i=1}^{n}(b_i - a_i)^2. \tag{3.7}$$

def

$= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$

$\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$

$= \frac{1}{n^2}\left(\sum \mathrm{Var}\, X_i\right)$

$\leq \frac{1}{n^2}\sum(b_i - a_i)^2$

Let $\sigma^2 = \frac{1}{n^2}\sum_{i=1}^{n}(b_i - a_i)^2$. If we take $\varepsilon = O(\sigma\sqrt{\log n}) = \sigma\sqrt{c\log n}$ so that ε is bounded above by some large (i.e., $c \geq 10$) multiple of the standard deviation of the X_i's times $\sqrt{\log n}$, we can substitute this value of ε into (3.6) to reach the following conclusion:

(3.6): $\Pr\left[|\bar{X}-\mu| \leq \varepsilon\right] \geq 1 - 2\exp\left(\frac{-2n^2\varepsilon^2}{\sum(b_i-a_i)^2}\right)$

$$\Pr\left[|\bar{X} - \mu| \leq \varepsilon\right] \geq 1 - 2\exp\left(\frac{-2\varepsilon^2}{\sigma^2}\right) \tag{3.8}$$

$\sigma^2 = \frac{\sum(b_i-a_i)^2}{n^2}$ bnd above by S.D of X_i times of H.

$\varepsilon^2 = \sigma^2 c\log n$ ($\varepsilon = \sigma\sqrt{c\log n}$)

$$= 1 - 2\exp(-2c\log n) \tag{3.9}$$

$$= 1 - 2n^{-2c} \xrightarrow{\lambda\to\text{grows}} 1 \quad (c\text{: large}) \tag{3.10}$$

We can see that as n grows, the right-most term tends to zero such that $\Pr[|\bar{X} - \mu| \leq \varepsilon]$ very quickly approaches 1. Intuitively, this result tells us that, with high probability, the sample mean $\bar{X}$ will not be "much farther" from the population mean μ by more than some sublogarithmic ($\sqrt{c\log n}$) factor of the standard deviation.[1] Thus, we can restate the above claim we reached as follows:

Remark 3.3. For sufficiently large n, $|\bar{X} - \mu| \leq \overset{\varepsilon:}{O}(\sigma\sqrt{\log n})$ with high probability.

Remark 3.4. If, in addition, we have $a_i = -O(1)$ and $b_i = O(1)$, then $\sigma^2 = O\left(\frac{1}{n}\right)$, and $|\bar{X} - \mu| \leq O\left(\sqrt{\frac{\log n}{n}}\right) = \widetilde{O}\left(\frac{1}{\sqrt{n}}\right)$.[2] #

tilde: "ignores" logarithmic factors

Remark 3.4 provides a compact form of the Hoeffding bound that we can use when the X_i are bounded almost surely.

So far, we have only shown how to construct exponential tail bounds for bounded random variables. Since requiring boundedness in $[0, 1]$ (or $[a, b]$ more generally) is limiting, it is worth asking what types of distributions permit such an exponential tail bound. The following section will explore such a class of random variables: *sub-Gaussian* random variables.

3.4 Sub-Gaussian random variables

What Dist.s permit such (3.3) types of Expo. tail bounds? Boundedness!

We begin by defining the class of sub-Gaussian random variables by way of a bound on their moment generating functions. After establishing this definition, we will see how this bound guarantees the exponential tail decay we desire.

Definition 3.5 (Sub-Gaussian Random Variables)**.** A random variable X with finite mean μ is *sub-Gaussian* with parameter σ if

정의:

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\sigma^2\lambda^2/2}, \quad \forall\lambda \in \mathbb{R}. \tag{3.11}$$

We say that X is σ-sub-Gaussian and say it has *variance proxy* σ^2. variance upper-bnd

Remark 3.6. As it turns out, (3.11) is quite a strong condition, requiring that infinitely many moments of X exist and do not grow too quickly. To see why, assume without loss of generality that $\mu = 0$ and take a power series expansion of the moment generating function:

power series expansion

$$\mathbb{E}[\exp(\lambda X)] = \mathbb{E}\left[\sum_{k=0}^{\infty}\frac{(\lambda X)^k}{k!}\right] = \sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\mathbb{E}[X^k]. \tag{3.12}$$

→ converge.. 하려면. ($\lesssim$ ~)

A bound on the moment generating function then is a bound on infinitely many moments of X, i.e. a requirement that the moments of X are all finite and grow slowly enough to allow the power series to converge. Though a proof of this result is beyond the scope of this monograph, Proposition 2.5.2 in [Vershynin, 2018] shows that (3.11) is equivalent to $\mathbb{E}\left[|X|^p\right]^{1/p} \lesssim \sqrt{p}$ for all $p \geq 1$.

[1] This is with the caveat, of course, that σ is not exactly the standard deviation but a loose upper bound on standard deviation.

[2] $\widetilde{O}$ is analogous to Big-O notation, but $\widetilde{O}$ hides logarithmic factors. That is; if $f(n) = O(\log n)$, then $f(n) \underset{\text{eq.}}{=} \widetilde{O}(1)$.

Although (3.11) is not a particularly intuitive definition, it turns out to imply exactly the type of exponential tail bound we want:

Sub-Gaussian RV의 tail bnd.

Theorem 3.7 (Tail bound for sub-Gaussian random variables). *If a random variable X with finite mean μ is σ-sub-Gaussian, then*

Necessary condition for Sub-Gaussianity
& Also Suff. cond (up to constant factor)

$\Pr[|X-\mu| \geq t] \leq 2e^{-t^2/2\sigma^2}$

$$\Pr[|X-\mu| \geq t] \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right), \quad \forall t \in \mathbb{R}. \tag{3.13}$$

Proof. Fix $t > 0$. For any $\lambda > 0$,

*Def. Sub-Gaussian RV "X".
(By its MGF)
$\mathbb{E}[e^{\lambda(x-\mu)}] \leq e^{\sigma^2\lambda^2/2}$ $\forall \lambda \in \mathbb{R}$

$$\begin{aligned}
\Pr[X-\mu \geq t] &= \Pr[\exp(\lambda(X-\mu)) \geq \exp(\lambda t)] && (3.14)\\
&\leq \exp(-\lambda t)\,\mathbb{E}[\exp(\lambda(X-\mu))] && \text{(by Markov's inequality)} \quad (3.15)\\
&\leq \exp(-\lambda t)\exp(\sigma^2\lambda^2/2) && \text{(by (3.11))} \quad (3.16)\\
&= \exp(-\lambda t + \sigma^2\lambda^2/2). && (3.17)
\end{aligned}$$

Because the bound (3.17) holds for any choice of $\lambda > 0$ and $\exp(\cdot)$ is monotonically increasing, we can optimize the bound (3.17) by finding λ which minimizes the exponent $-\lambda t + \sigma^2\lambda^2/2$. Differentiating and setting the derivative equal to zero, we find that the optimal choice is $\lambda = t/\sigma^2$, yielding the one-sided tail bound

diff. w.r.t. λ & set to 0.

$-t + \sigma^2\lambda = 0 \to \lambda = t/\sigma^2$

$$\Pr[X-\mu \geq t] \leq \exp\left(-\frac{t^2}{2\sigma^2}\right). \tag{3.18}$$

Going through the same line of reasoning but for $-X$ and $-t$, we can also show that for any $t > 0$,

$$\Pr[X-\mu \leq -t] \leq \exp\left(-\frac{t^2}{2\sigma^2}\right). \tag{3.19}$$

We can then obtain (3.13) by applying the union bound:

union bnd

$$\Pr[|X-\mu| \geq t] = \underbrace{\Pr[X-\mu \geq t]}_{(3.18)} + \underbrace{\Pr[X-\mu \leq -t]}_{(3.19)} \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right). \tag{3.20}$$

□

Remark 3.8 (Tail bound implies sub-Gaussianity). In addition to being a necessary condition for sub-Gaussianity (Theorem 3.7), the tail bound (3.13) for sub-Gaussian random variables is also a sufficient condition up to a constant factor. In particular, if a random variable X with finite mean μ satisfies (3.13) for some $\sigma > 0$, then X is $O(\sigma)$-sub-Gaussian. Unfortunately, the proof of this reverse direction is somewhat more involved, so we refer the interested reader to Theorem 2.6 and its proof in Section 2.4 of [Wainwright, 2019] and Proposition 2.5.2 in [Vershynin, 2018] for details. While the tail bound is the property we ultimately care about most when studying sub-Gaussian random variables, the definition in (3.11) is a more technically convenient characterization, as we will see in the proof of Theorem 3.10.

Gaussian tail Bnd

Remark 3.9. Note that in light of Remark 3.6, the tail bound (3.3) requires all central moments of X to exist and not grow too quickly. In contrast, Chebyshev's inequality (and more generally any polynomial variant of Markov's inequality $\Pr[|X-\mu| \geq t] = \Pr[|X-\mu|^k \geq t^k] \leq t^{-k}\,\mathbb{E}[|X-\mu|^k]$) only requires that the second central moment $\mathbb{E}[(X-\mu)^2]$ (more generally, the kth central moment $\mathbb{E}[|X-\mu|^k]$) is finite to yield a tail bound. If infinite moments exist, however, it turns out that $\inf_{k\in\mathbb{N}} t^{-k}\,\mathbb{E}[|X-\mu|^k] \leq \inf_{\lambda>0} \exp(-\lambda t)\,\mathbb{E}[\exp(\lambda(X-\lambda))]$, i.e. the optimal polynomial tail bound is tighter than the optimal exponential tail bound (see Exercise 2.3 in [Wainwright, 2019]). As we will see shortly though, using exponential functions of random variables allows us to prove results about sums of random variables more conveniently. This "tensorization" property is why most researchers use exponential tail bounds in practice.

tail bnd yield by Chebyshev.

Gaussian tail bnd

using expo.func. (expo.tail bnds)가 이거랑 뭔상관?

exp를 들여와서 Π랑 Σ의 성질을 사용하는거.

Having defined and derived exponential tail bounds for sub-Gaussian random variables, we can now accomplish the first of the goals we set out at the beginning of the chapter: show that under certain conditions, namely independence and sub-Gaussianity of $X_1, \ldots, X_n$, the sum $Z = \sum_{i=1}^n X_i$ concentrates around $\mathbb{E}[Z] = \mathbb{E}[\sum_{i=1}^n X_i]$.

Theorem 3.10 (Sum of sub-Gaussian random variables is sub-Gaussian)**.** *If $X_1, \ldots, X_n$ are independent sub-Gaussian random variables with variance proxies $\sigma_1^2, \ldots, \sigma_n^2$, then $Z = \underbrace{\sum_{i=1}^n X_i}_{\text{Sum of Sub Gaussian RVs}}$ is sub-Gaussian with variance proxy $\sum_{i=1}^n \sigma_i^2$. As a consequence, we have the tail bound*

Variance upper bnd

Variance upper bnd of $Z = \sum_{i=1}^n X_i$

$$\Pr[|Z - \mathbb{E}[Z]| \geq t] \leq 2 \exp\left(-\frac{t^2}{2 \underbrace{\sum_{i=1}^n \sigma_i^2}_{\text{Var. Proxy of } Z}}\right), \tag{3.21}$$

for all $t \in \mathbb{R}$.

Proof. Using the independence of $X_1, \ldots, X_n$, we have that for any $\lambda \in \mathbb{R}$:

$$\mathbb{E}\left[\exp\{\lambda(\underset{\uparrow \atop \Sigma X_i}{Z} - \mathbb{E}[\underset{\uparrow \atop \Sigma X_i}{Z}])\}\right] = \mathbb{E}\left[\prod_{i=1}^n \exp\{\lambda(X_i - \mathbb{E}[X_i])\}\right] \tag{3.22}$$

X_i's indep

$$= \prod_{i=1}^n \mathbb{E}\left[\exp\{\lambda(X_i - \mathbb{E}[X_i])\}\right] \tag{3.23}$$

정의 of Sub-Gaussian RVs. (3.11)

$$\leq \prod_{i=1}^n \exp\left(\frac{\lambda^2 \sigma_i^2}{2}\right) \tag{3.24}$$

$$= \exp\left(\frac{\lambda^2 \sum_{i=1}^n \sigma_i^2}{2}\right), \tag{3.25}$$

다시 정의 of Sub-Gaussian RVs. (3.11)

so Z is sub-Gaussian with variance proxy $\sum_{i=1}^n \sigma_i^2$. The tail bound then follows immediately from (3.13). $\square$ ness. condi of Sub Gaussian RVs.

(Sub-Gaussian RV의 특징.)

The proof above demonstrates the value of the moment generating functions of sub-Gaussian random variables: they factorize conveniently when dealing with sums of independent random variables.

3.4.1 Examples of sub-Gaussian random variables

We now provide several examples of classes of random variables that are sub-Gaussian, some of which will appear repeatedly throughout the remainder of the course.

Example 3.11 (Rademacher random variables)**.** A *Rademacher random variable* ϵ takes a value of 1 with probability 1/2 and a value of -1 with probability 1/2. To see that ϵ is 1-sub-Gaussian, we follow Example 2.3 in [Wainwright, 2019] and upper bound the moment generating function of ϵ by way of a power series expansion of $\exp(\cdot)$:

MGF of ϵ

$$\mathbb{E}[\exp(\lambda\epsilon)] = \frac{1}{2}\{\exp(-\lambda) + \exp(\lambda)\} \tag{3.26}$$

expansion

$$= \frac{1}{2}\left\{\sum_{k=0}^{\infty} \frac{(-\lambda)^k}{k!} + \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}\right\} \tag{3.27}$$

only even k's survive

$$= \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{(2k)!} \qquad \text{(for odd } k, (-\lambda)^k + \lambda^k = 0) \tag{3.28}$$

$$\leq \underset{k=0}{1} + \sum_{k=1}^{\infty} \frac{(\lambda^2)^k}{2^k k!} \qquad (2^k k! \text{ is every other term of } (2k)!) \tag{3.29}$$

except for $k=0$.

$$= \exp(\lambda^2/2), \tag{3.30}$$

정의 of Sub-Gaussian RVs. (3.11)

which is exactly the moment generating function bound (3.11) required for 1-sub-Gaussianity.

Variance proxy: $\sigma^2 = 1$.

Example 3.12 (Random variables with bounded distance to mean). Suppose a random variable X satisfies $|X - \mathbb{E}[X]| \le M$ almost surely for some constant M. Then X is $O(M)$-sub-Gaussian.

평균으로부터 bounded 거리.

↳ variance proxy $= O(M)^2$

We now provide an even more general class of sub-Gaussian random variables that subsume the random variables in Example 3.12:

포함하다

Example 3.13 (Bounded random variables). If X is a random variable such that $a \le X \le b$ almost surely for some constants $a, b \in \mathbb{R}$, then

$\frac{(b-a)}{2}$-sub-Gaussianity: $E[e^{\lambda(X-E[X])}] \le e^{\lambda^2(b-a)^2/8}$

Thm 3.10의 반대 버전 : $Pr[|Z-E[Z]| \le t] \ge 1-2e^{-t^2/2\sum\sigma_i^2}$

흠... 왜지?

$$\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \le \exp\left[\frac{\lambda^2(b-a)^2}{8}\right],$$

$\sigma^2 = (b-a)^2/4$ 즉 $\frac{|b-a|}{2}$-sub-Gaussian

i.e., X is sub-Gaussian with variance proxy $(b-a)^2/4$. (We will prove this in Question 2(a) of Homework 1.) Note that combining the $(b-a)/2$-sub-Gaussianity of i.i.d. bounded random variables $X_1, \ldots, X_n$ and Theorem 3.10 yields a proof of Hoeffding's inequality. $\left(\Pr\left[|\bar{X} - \mu| \le \varepsilon\right] \ge 1 - 2\exp\left(\frac{-2n^2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)\right)$

$\Pr[|Z - \mathbb{E}[Z]| \ge t] \le 2\exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right)$

Example 3.14 (Gaussian random variables). If X is Gaussian with variance σ^2, then X satisfies (3.11) with equality. In this special case, the variance and the variance proxy are the same.

upper bnd.

3.5 Concentrations of functions of random variables

일반화!

$X_1 + \cdots + X_n$ concentrates around $\mathbb{E}[X_1 + \cdots + X_n]$ 뿐 아니라, $f(X_1, \ldots, X_n)$도 〃 〃 $\mathbb{E}[f(X_1, \ldots, X_n)]$.

독립적인 RV들 X_1, ..., X_n과 특정 함수 f에 대해, f(X_1, ..., X_n)은 걔의 기댓값인 E[f(X_1, ..., X_n)]에 '집중'한다.

We now introduce some important inequalities related to the second of our two goals, namely, showing that for independent $X_1, \ldots, X_n$ and certain functions f, $f(X_1, \ldots, X_n)$ concentrates around $\mathbb{E}[f(X_1, \ldots, X_n)]$.

f의 차(차이)분 제한 조건: 변수 하나가 바뀌어도 함수 값은 많이 움직이지 X

Theorem 3.15 (McDiarmid's inequality). *Suppose $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the bounded difference condition: there exist constants $c_1, \ldots, c_n \in \mathbb{R}$ such that for $\boxed{\text{all}}$ real numbers $x_1, \ldots, x_n$ and x'_i,*

Universally..

여기서 함수 f는 Lipschitz 연속:: 하나의 변수가 바뀔 때, 함수 값의 변화 크기가 c_i로 제한됨

$$|f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x'_i, x_{i+1}, \ldots, x_n)| \le c_i. \tag{3.31}$$

고작 하나의 좌표가 바뀐다고 해서, f가 민감하게 바뀌지 않는다는 의미 (3.31)

c_i는 각 변수에 대한 변화폭(최대 영향)

(Intuitively, (3.31) states that f is not overly sensitive to arbitrary changes in a single coordinate.) Then, for any independent random variables $X_1, \ldots, X_n$,

$Pr[Z_n - Z_0 \ge t] \le \exp\left(-\frac{2t^2}{\sum c_i^2}\right)$

$$\Pr\left[f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)] \ge t\right] \le \exp\left(-\frac{2t^2}{\underbrace{\sum_{i=1}^n c_i^2}_{\sigma}}\right). \tag{3.32}$$

Moreover, $f(X_1, \ldots, X_n)$ is $O\left(\overset{\sigma}{\sqrt{\sum_{i=1}^n c_i^2}}\right)$-sub-Gaussian.

Remark 3.16. Note that McDiarmid's inequality is a generalization of Hoeffding's inequality with $a_i \le x_i \le b_i$ and

Mcdiarmid의 부등식은 호프딩 부등식의 일반화식임 with a_i <= x_i <= b_i

$$f(x_1, \ldots, x_n) = \sum_{i=1}^n x_i. \tag{3.33}$$

Proof. The idea of this proof is to take the quantity $f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)]$ and break it into manageable components by conditioning on portions of the sample. To this end, we begin by defining:

$Z_0 = \mathbb{E}[f(X_1, \ldots, X_n)]$	(Given)	constant
$Z_1 = \mathbb{E}[f(X_1, \ldots, X_n) \mid X_1]$	(Depends on X_1)	a function of X_1
$\ldots$		
$Z_i = \mathbb{E}[f(X_1, \ldots, X_n) \mid X_1, \ldots, X_i]$	(Depends on $X_1, \cdots, X_i$)	a function of $X_1, \ldots, X_i$
$\ldots$		
$Z_n = f(X_1, \ldots, X_n)$	← (Clearly a function of $X_1, \cdots, X_n$)	

LTE: $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$ X, Y in a same prob. space

$\mathbb{E}[X] = \int_x x f_X(x)\,dx$ (f: pdf of X)

$= \int_x x \int_y f_Y(x,y)\,dy\,dx$

$= \int_x x \int_y f_{X|Y}(x|y) f_Y(y)\,dy\,dx$

$= \int_y \left(\int_x x f_{X|Y}(x,y)\,dx\right) f_Y(y)\,dy$

$= \int_y \mathbb{E}[X|Y] f_Y(y)\,dy$

$= \mathbb{E}[\mathbb{E}[X|Y]]$.

$\mathbb{E}\left[f(X_1,\cdots,X_n)\,|\,X_1,\cdots,X_i\right]$

Using the law of total expectation, we show also that the expectation of Z_i equals Z_0 for all i.

$$\begin{aligned}
\mathbb{E}[Z_i] &= \mathbb{E}\left[\mathbb{E}\left[f(X_1,\ldots,X_n)|X_1,\ldots,X_i\right]\right] \quad \text{LTE.} \\
&= \mathbb{E}[f(X_1,\ldots,X_n)] \quad \text{by def.} \\
&= Z_0
\end{aligned}$$

The fact that $\mathbb{E}[D_i] = 0$, where $D_i = Z_i - Z_{i-1}$, is an immediate corollary of this result. Next, we observe that we can rewrite the quantity of interest, $Z_n - Z_0$, as a telescoping sum in the increments $Z_i - Z_{i-1}$:

$$\begin{aligned}
Z_n - Z_0 &= (Z_n - Z_{n-1}) + (Z_{n-1} - Z_{n-2}) + \cdots + (Z_1 - Z_0) \\
&= \sum_{i=1}^{n} D_i
\end{aligned}$$

Next, we show that conditional on $X_1,\ldots,X_{i-1}$, D_i is a bounded random variable. First, observe that:

$B_i - A_i \le c_i$

$$\begin{aligned}
A_i &= \inf_x \mathbb{E}\left[f(X_1,\ldots,X_n)|X_1,\ldots,X_{i-1},X_i = x\right] - \mathbb{E}\left[f(X_1,\ldots,X_n)|X_1,\ldots,X_{i-1}\right] \\
B_i &= \sup_x \mathbb{E}\left[f(X_1,\ldots,X_n)|X_1,\ldots,X_{i=1},X_i = x\right] - \mathbb{E}\left[f(X_1,\ldots,X_n)|X_1,\ldots,X_{i-1}\right]
\end{aligned}$$

$A_i = \inf_x Z_i - Z_{i-1}$

$B_i = \sup_x Z_i - Z_{i-1}$

It is clear from their definition that $A_i \le D_i \le B_i$. Furthermore, by independence of the X_i's, we have that:

$$\begin{aligned}
B_i - A_i &\le \sup_{x_{1:i-1}} \sup_{x,x'} \int \left(f(x_1,\ldots,x_{i-1},x,x_{i+1},\ldots,x_n) - f(x_1,\ldots,x_{i-1},x',x_{i+1},\ldots,x_n)\right) dP(x_{i+1},\ldots,x_n) \\
&\le c_i
\end{aligned}$$

Using this bound, the properties of conditional expectation, and Example 3.13, we can now prove that that $Z_n - Z_0$ is $O\left(\sqrt{\sum_{i=1}^n c_i^2}\right)$-sub-Gaussian.

(example 3.13)

$\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \le e^{\lambda^2(b-a)^2/8}$

with $a \le X \le b$

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda(Z_n - Z_0)}\right] &= \mathbb{E}\left[e^{\lambda\sum_{i=1}^n (Z_i - Z_{i-1})}\right] \\
&\overset{\text{LTE}}{=} \mathbb{E}\left[\mathbb{E}\left[e^{\lambda(\overset{\le c_n}{Z_n - Z_{n-1}})}\middle|X_1,\ldots,X_{n-1}\right]e^{\lambda\sum_{i=1}^{n-1}(Z_i - Z_{i-1})}\right] \\
&\le e^{\lambda^2 c_n^2/8}\,\mathbb{E}\left[e^{\lambda\sum_{i=1}^{n-1}(Z_i - Z_{i-1})}\right] \\
&\le e^{\lambda^2(\sum_{i=1}^n c_i^2)/8}
\end{aligned}$$

이걸 반복.

사실 $\frac{\sqrt{\Sigma c_i^2}}{2}$-sub-Gaussian

By Def, $Z_n - Z_0$ is $\sqrt{\Sigma c_i^2}$-Sub-Gaussian.

Thm 3.7: $\Pr[|X-\mu| \ge t] \le 2e^{-t^2/2\sigma^2}$ for σ-Sub-Gaussian RV X.

yes 절댓값.

(3.32): $\Pr[Z_n - Z_0 \ge t] \le \exp\left(-\frac{2t^2}{\Sigma c_i^2}\right)$

No 절댓값.

$* \; 2e^{\left(-t^2/2\cdot\frac{\sqrt{\Sigma c_i^2}}{2}\cdot\frac{\sqrt{\Sigma c_i^2}}{2}\right)} = 2e^{-2t^2/\Sigma c_i^2}$

$\Pr[Z_n - Z_0 \ge t] \le e^{-2t^2/\Sigma c_i^2}$

The final inequality given in (3.32) follows by Theorem 3.7.

A more general version of McDiarmid's inequality comes from Theorem 3.18 in [van Handel, 2016]. The setup for this theorem requires defining the *one-sided differences* of a function $f : \mathbb{R}^n \to \mathbb{R}$:

현재 coordinate에서 i번째 변수만 임의로 다른 값인 z로 바꾸었을 때, 함수 값이 가장 많이 작아지는 정도 // 가장 크게 감소할 수 있는 함수 값의 폭

$$D_i^- f(x) = f(x_1,\ldots,x_n) - \inf_z f(x_1,\ldots,x_{i-1},z,x_{i+1},\ldots,x_n) \tag{3.34}$$

$$D_i^+ f(x) = \sup_z f(x_1,\ldots,x_{i-1},z,x_{i+1},\ldots,x_n) - f(x_1,\ldots,x_n). \tag{3.35}$$

현재 coordinate에서 i번째 변수만 임의로 다른 값인 z로 바꾸었을 때, 함수 값이 가장 많이 커지는 정도 // 가장 크게 증가할 수 있는 함수 값의 폭

These two quantities are functions of $x \in \mathbb{R}^n$, and hence can be interpreted as describing the sensitivity of f *at a particular point.* (Contrast this with the bounded difference condition (3.31), which bounds the sensitivity of f universally over all points.) For convenience, define

"특정 지점"에서의 f의 민감도를 나타낼 수 있음

(3.31)은 universally하게 f의 민감도를 측정

$$d^+ = \left\|\sum_{i=1}^n |D_i^+ f|^2\right\|_\infty = \sup_{x_1,\ldots,x_n} \sum_{i=1}^n [D_i^+ f(x_1,\ldots,x_n)]^2 \tag{3.36}$$

$$d^- = \left\|\sum_{i=1}^n |D_i^- f|^2\right\|_\infty = \sup_{x_1,\ldots,x_n} \sum_{i=1}^n [D_i^- f(x_1,\ldots,x_n)]^2. \tag{3.37}$$

d+ :: 함수의 모-든 coordinate의 값들 (i=1부터 ...n까지) 에 대해,

각 변수 (i=1~n)별 영향량을 제곱해서 모두 더한 뒤!! 그 값의 최댓값sup을 적는 것

즉, 어떤 입력에서든 [""모든 변수의 최대 증가 영향(D_i f)"" 제곱합]이 가장 크게 되는 케이스

이게 작을수록, 함수 f가 입력의 변화(i번째 변수가 바뀌는 등)가 있더라도, 평균 근처에 더 집중!! 하게 됨..

d+, d- 둘 다 작을수록! 함수가 변수 변화에 덜 민감하다는 소리

Theorem 3.17 (Bounded difference inequality, Theorem 3.18 in [van Handel, 2016]). *Let $f : \mathbb{R}^n \to \mathbb{R}$, and let $X_1, \ldots, X_n$ be independent random variables. Then, for all $t \geq 0$,* 여기서 (1) t가 클수록, (2) d-가 작을수록(함수가 변수 변화에 덜 민감할수록) 식의 확률값이 빠르게 작아짐.

함수 값이 평균보다 t만큼 더 클 확률이!! <=

$$\Pr[f(X_1, \ldots, X_n) \geq \mathbb{E}[f(X_1, \ldots, X_n)] + t] \leq \exp\left(-\frac{t^2}{4d^-}\right) \tag{3.38}$$

즉 여기서의 확률은 "평균을 벗어날" 확률

$$\Pr[f(X_1, \ldots, X_n) \leq \mathbb{E}[f(X_1, \ldots, X_n)] - t] \leq \exp\left(-\frac{t^2}{4d^+}\right). \tag{3.39}$$

3.5.1 Bounds for Gaussian random variables

Unfortunately, the bounded difference condition (3.31) is often only satisfied by bounded random variables or a bounded function. To get similar concentration inequalities for unbounded random variables, we need some other special conditions. The following inequalities assume that the random variables have the standard normal distribution.

\# of continuous derivatives..

예) C^k : function of smoothness at least k 의 class.

C^∞ : infinitively differentiable functions.

Theorem 3.18 (Gaussian Poincaré inequality, Corollary 2.27 in [van Handel, 2016]). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be smooth. If $X_1, \ldots, X_n$ are independently sampled from $\mathcal{N}(0, 1)$, then*

* 기울기가 크다는 건 함수가 특정 입력값에 민감하다는 것
** 함수가 입력 값에 아주 민감하지 않으면 (기울기가 크지 않으면), f(X)도 평균 근처에 매우 집중된다 (즉, 분산의 upper bound가 작다.)

RV들 X_1, ..., X_n이 독립적으로 표준정규분포에서 뽑혔다고 가정할 때,

$$\mathrm{Var}(f(X_1, \ldots, X_n)) \leq \mathbb{E}\left[\|\nabla f(X_1, \ldots, X_n)\|_2^2\right]. \tag{3.40}$$

다변수 함수의 분산은 <= (상계) 그 함수의 기울기의 제곱의 평균

Smooth..의 빽중하나?

Before introducing the next theorem, we recall that a function $f : \mathbb{R}^n \to \mathbb{R}$ is *L-Lipschitz* with respect to the ℓ_2-norm if there exists a non-negative constant $L \in \mathbb{R}$ such that for all $x, y \in \mathbb{R}^n$,

L is universal.

만약 (3.41)처럼 두 함수 값의 차이를 upper bound하는 상수 L (non-neg 상수)가 존재한다면,
그 함수는 L-Lipschitz하다 - 고 한다.
즉, 입력이 조금만 변하더라도 함수 값이 크게 튀지 않는!

$$|f(x) - f(y)| \leq L\|x - y\|_2. \tag{3.41}$$

We emphasize that L is universal for all points in $\mathbb{R}^n$.

아 여기서 L은 모든 점에서 universal (같아야!)하므로, f의 전체 민감도의 "최댓값"이다.

L이 작다 = 함수가 입력에 둔감하다 (민감하지 않다!)

Theorem 3.19 (Theorem 2.26 in [Wainwright, 2019]). *Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz with respect to Euclidean distance, and let $X = (X_1, \ldots, X_n)$, where $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$. Then for all $t \in \mathbb{R}$,*

RV X = (X_1, ..., X_n)이, 각 성분이 표준정규분포에서 독립적으로 뽑힌 벡터라면!
그리고 f가 L-리프시츠 함수라면,
함수값이 평균에서 t이상으로 벗어날 확률은
RHS 이하로 제한된다.

$$\Pr[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2\exp\left(-\frac{t^2}{2L^2}\right). \tag{3.42}$$

Thm 3.7처럼.

L이 작을수록, 즉 f의 전체 민감도의 "최댓값"이 작을수록,
평균 근처에 더욱 강하게 (평균을 벗어나지 않을 확률이 크게) 집중된다.

In particular, $f(X)$ is sub-Gaussian.

어떤 RV가 sub-Gaussian이라는 건,
모든 실수 ㅅ에 대해
E[e^{ㅅ(X-mu)}] <= e^{sigma^2 ㅅ^2 / 2}를 만족하는 sigma(>0)가 존재한다는 것.
그니까 분산이 sigma^2인 정규분포와 비교했을 때, 꼬리가 "두껍지 않다(오히려 같거나 더 얇다)"는 것
--> 즉, 데이터의 극단값이 나올 확률이 매우 작아서 평균 근처의 집중도가 크다는 것!

즉 "가우시안 벡터 + Lipschitz 함수" 조합의 출력은
평균 주변에 매우 집중(small tail risk)되어 있고,
확률적으로 큰 변동이 발생할 가능성이 매우 낮다는 것을 보장