Machine Learning Notes

This note is based on Machine Learning at DIKU

December 5, 2021

Chapter 2. Concentration of Measure Inequalities

2.3 Hoeffding's Inequality

Hoeffding's inequality is a much more powerful concentration result.

Theorem 2.3 (Hoeffding's Inequality). *Let $X_1, ..., X_n$ be independent real-valued random variables, such that for each $i \in \{1, ..., n\}$ there exist $a_i \leqslant b_i$, such that $X_i \in [a_i, b_i]$. Then for every $\varepsilon > 0$,*

$$\mathbf{Pr}\left(\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] \geqslant \varepsilon\right) \leqslant \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right), \tag{2.11}$$

and

$$\mathbf{Pr}\left(\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] \leqslant -\varepsilon\right) \leqslant \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right). \tag{2.12}$$

Corollary 2.4. *Under the assumptions of Theorem 2.3,*

$$\mathbf{Pr}\left(\left|\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right]\right| \geqslant \varepsilon\right) \leqslant 2\exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right). \tag{2.13}$$

By taking a union bound of the events in Eqs. (2.11) and (2.12) we obtain Corollary 2.4. Note that Eqs. (2.11) and (2.12) are known as "one-sided Hoeffding's inequalities" and (2.13) is known as "two-sided Hoeffding's inequality". If we assume that X_i-s are identically distributed and belong to the $[0, 1]$ interval we obtain Corollary 2.5.

NB. Let the event $\mathcal{A}$ represent $\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right]$, then we have $\mathbf{Pr}(\mathcal{A}_+) \stackrel{\text{def}}{=} \mathbf{Pr}(\mathcal{A} \geqslant \varepsilon)$ and $\mathbf{Pr}(\mathcal{A}_-) \stackrel{\text{def}}{=} \mathbf{Pr}(\mathcal{A} \leqslant -\varepsilon)$. Therefore,

$$\mathbf{Pr}(|\mathcal{A}| \geqslant \varepsilon) = \mathbf{Pr}\big((\mathcal{A} \geqslant \varepsilon) \vee (\mathcal{A} \leqslant -\varepsilon)\big) = \mathbf{Pr}(\mathcal{A} \geqslant \varepsilon) + \mathbf{Pr}(\mathcal{A} \leqslant -\varepsilon) \leqslant 2\exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right). \tag{2.14}$$

Corollary 2.5. *Let $X_1, ..., X_n$ be independent random variables, such that $X_i \in [0, 1]$ and $\mathbf{E}[X_i] = \mu$ for all i, then for every $\varepsilon > 0$,*

$$\mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geqslant \varepsilon\right) \leqslant e^{-2n\varepsilon^2}, \tag{2.15}$$

and

$$\mathbf{Pr}\left(\mu - \frac{1}{n}\sum_{i=1}^{n} X_i \geqslant \varepsilon\right) \leqslant e^{-2n\varepsilon^2}. \tag{2.16}$$

Recall that by Chebyshev's inequality $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$ converges to μ at the rate of n^{-1}. Hoeffding's inequality demonstrates that the convergence is actually much faster, at least at the rate of e^{-n}. The proof of Hoeffding's inequality is based on Hoeffding's lemma.

NB. To be specific, if $X_i \in [0,1]$ and $\mathbf{E}[X_i] = \mu$ for all i. Let $Y_i = \frac{1}{n}X_i$ for all i, then the event

$$\mathcal{A} \stackrel{\text{def}}{=} \sum_{i=1}^n Y_i - \mathbf{E}\left[\sum_{i=1}^n Y_i\right] = \frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{n}\mathbf{E}\left[\sum_{i=1}^n X_i\right] = \frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{n}\sum_{i=1}^n \mathbf{E}[X_i] = \frac{1}{n}\sum_{i=1}^n X_i - \mu\,, \tag{2.17}$$

where $\mathbf{E}\left[\sum_{i=1}^n \left(\frac{1}{n}X_i\right)\right] = \frac{1}{n}\sum_{i=1}^n \mathbf{E}[X_i] = \mu$. Thus $Y_i \in [0, \frac{1}{n}]$ for all i, and we could obtain that

$$\mathbf{Pr}(\mathcal{A} \geqslant \varepsilon) = \mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mu \geqslant \varepsilon\right) \leqslant \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^n (\frac{1}{n}-0)^2}\right) = \exp\left(\frac{-2\varepsilon^2}{\frac{1}{n^2}\sum_{i=1}^n 1}\right) = e^{-2n\varepsilon^2}\,, \tag{2.18a}$$

$$\mathbf{Pr}(\mathcal{A} \leqslant -\varepsilon) = \mathbf{Pr}\left(\mu - \frac{1}{n}\sum_{i=1}^n X_i \geqslant \varepsilon\right) \leqslant \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^n (\frac{1}{n}-0)^2}\right) = \exp\left(\frac{-2\varepsilon^2}{\frac{1}{n}}\right) = \exp(-2n\varepsilon^2)\,. \tag{2.18b}$$

Lemma 2.6 (Hoeffding's Lemma)**.** *Let X be a random variable, such that $X \in [a,b]$. Then for any $\lambda \in \mathbb{R}$,*

$$\mathbf{E}\left[e^{\lambda X}\right] \leqslant \exp\left(\lambda \mathbf{E}[X] + \frac{\lambda^2(b-a)^2}{8}\right)\,. \tag{2.19}$$

The function $f(\lambda) = \mathbf{E}[e^{\lambda X}]$ is known as the *moment generating function* of X, since $f'(0) = \mathbf{E}[X]$, $f''(0) = \mathbf{E}[X^2]$, and, more generally, $f^{(k)}(0) = \mathbf{E}[X^k]$. We provide the proof of the lemma immediately after the proof of Theorem 2.3.

Proof of Theorem 2.3. We prove the first inequality in Theorem 2.3. The second inequality follows by applying the first inequality to $-X_1, ..., -X_n$. The proof is based on Chernoff's bounding technique. For any $\lambda > 0$, the following holds:

$$\begin{aligned}\mathbf{Pr}\left(\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] \geqslant \varepsilon\right) &= \mathbf{Pr}\left(\exp\left(\lambda\left(\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right]\right)\right) \geqslant e^{\lambda\varepsilon}\right) \\ &\leqslant \frac{\mathbf{E}\left[\exp\left(\lambda\left(\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right]\right)\right)\right]}{e^{\lambda\varepsilon}}, \end{aligned} \tag{2.20}$$

where the first step holds since $e^{\lambda x}$ is a monotonously increasing function for $\lambda > 0$ and the second step holds by Markov's inequality. We now take a closer look at the nominator:

$$\begin{aligned}\mathbf{E}\left[\exp\left(\lambda\left(\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right]\right)\right)\right] &= \mathbf{E}\left[\exp\left(\sum_{i=1}^{n} \lambda\big(X_i - \mathbf{E}[X_i]\big)\right)\right] = \mathbf{E}\left[\prod_{i=1}^{n} e^{\lambda\left(X_i - \mathbf{E}[X_i]\right)}\right] \\ &= \prod_{i=1}^{n} \mathbf{E}\left[e^{\lambda\left(X_i - \mathbf{E}[X_i]\right)}\right] \qquad (2.21) \\ &\leqslant \prod_{i=1}^{n} e^{\lambda^2 (b_i - a_i)^2/8} \qquad (2.22) \\ &= \exp\left(\frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2\right), \end{aligned}$$

where (2.21) holds since $X_1, ..., X_n$ are independent and (2.22) holds by Hoeffding's lemma applied to a random variable $Z_i = X_i - \mathbf{E}[X_i]$ (note that $\mathbf{E}[Z_i] = 0$ and that $Z_i \in [a_i - \mu_i, b_i - \mu_i]$ for $\mu_i = \mathbf{E}[X_i]$). *Put attention to the crucial role that independence of $X_1, ..., X_n$ plays in the proof! Without independence we would not have been able to exchange the expectation with the product and the proof would break down!*

NB. Note that $e^x > 0$ for any $x \in \mathbb{R}$ and $e^0 = 1$. For a random variable X, let $Y = X - \mathbf{E}[X]$ such that $Y \in [a, b]$. Thus we get $\mathbf{E}[Y] = \mathbf{E}[X - \mathbf{E}[X]] = \mathbf{E}[X] - \mathbf{E}[X] = 0$. Therefore, for any $\lambda \in \mathbb{R}$,

$$\mathbf{E}\left[e^{\lambda\left(X - \mathbf{E}[X]\right)}\right] \leqslant \exp\left(\lambda\mathbf{E}[Y] + \tfrac{\lambda^2(b-a)^2}{8}\right) = e^0 \exp\left(\tfrac{\lambda^2(b-a)^2}{8}\right) = \exp\left(\tfrac{\lambda^2(b-a)^2}{8}\right). \tag{2.23}$$

Let $Y = \sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right]$ and $Y \in [a, b]$, then

$$\mathbf{E}[Y] = \mathbf{E}\left[\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right]\right] = \mathbf{E}\left[\sum_{i=1}^{n} X_i - \sum_{i=1}^{n}\mathbf{E}[X_i]\right] = \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] - \sum_{i=1}^{n}\mathbf{E}[X_i] = 0, \tag{2.24}$$

due to the independence of $X_1, ..., X_n$. Then for any $\lambda \in \mathbb{R}$,

$$\begin{aligned}\mathbf{E}[e^{\lambda Y}] = \mathbf{E}\left[e^{\lambda\left(\sum_{i=1}^{n} X_i - \mathbf{E}[\sum_{i=1}^{n} X_i]\right)}\right] &= \mathbf{E}\left[\exp\left(\lambda\left(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n}\mathbf{E}[X_i]\right)\right)\right] \\ &= \mathbf{E}\left[\exp\left(\sum_{i=1}^{n}\lambda\big(X_i - \mathbf{E}[X_i]\big)\right)\right] = \mathbf{E}\left[\prod_{i=1}^{n} e^{\lambda\left(X_i - \mathbf{E}[X_i]\right)}\right] \\ &= \prod_{i=1}^{n}\mathbf{E}\left[e^{\lambda\left(X_i - \mathbf{E}[X_i]\right)}\right] \leqslant \prod_{i=1}^{n}\exp\left(\lambda\mathbf{E}\left[X_i - \mathbf{E}[X_i]\right] + \tfrac{\lambda^2(b_i - a_i)^2}{8}\right) \qquad (2.25a) \\ &= \prod_{i=1}^{n} e^{\lambda\cdot 0}\cdot\exp\left(\frac{\lambda^2\left((b_i - \mu_i) - (a_i - \mu_i)\right)^2}{8}\right) = \prod_{i=1}^{n} e^{\lambda^2(b_i - a_i)^2/8} \qquad (2.25b) \\ &= \exp\left(\sum_{i=1}^{n}\tfrac{\lambda^2(b_i - a_i)^2}{8}\right) = \exp\left(\frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2\right), \end{aligned}$$

where $\mu_i \overset{\text{def}}{=} \mathbf{E}[X_i]$ and $X_i - \mathbf{E}[X_i] \in [a_i - \mu_i, b_i - \mu_i]$ for all i.

To complete the proof we substitute the bound on the expectation into the previous calculation and obtain:

$$\mathbf{Pr}\left(\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] \geqslant \varepsilon\right) \leqslant \exp\left(\frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2 - \lambda\varepsilon\right). \tag{2.26}$$

This expression is minimised by

$$\lambda^* = \operatorname*{argmin}_{\lambda} \exp\left(\frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2 - \lambda\varepsilon\right) = \operatorname*{argmin}_{\lambda}\left(\frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2 - \lambda\varepsilon\right) = \frac{4\varepsilon}{\sum_{i=1}^{n}(b_i - a_i)^2}. \tag{2.27}$$

It is important to note that the best choice of λ does not depend on the sample. In particular, it allows to fix λ before observing the sample.

NB. According to Theorem **??**,

$$\mathbf{Pr}\left(\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] \geqslant \varepsilon\right) = \mathbf{Pr}\left(e^{\lambda\left(\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right]\right)} \geqslant e^{\lambda\varepsilon}\right) \leqslant \frac{\exp\left(\lambda\left(\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right]\right)\right)}{e^{\lambda\varepsilon}} \tag{2.28a}$$

$$\leqslant \frac{1}{e^{\lambda\varepsilon}}\exp\left(\frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2\right) = \exp\left(\frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2 - \lambda\varepsilon\right). \tag{2.28b}$$

Let $f(\lambda) \overset{\text{def}}{=} \frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2 - \lambda\varepsilon$ and $AB \overset{\text{def}}{=} \sum_{i=1}^{n}(b_i - a_i)^2$ for brevity, then

$$8f(\lambda) = \sum_{i=1}^{n}(b_i - a_i)^2\lambda^2 - 8\varepsilon\lambda = \sum_{i=1}^{n}(b_i - a_i)^2\left(\lambda^2 - \frac{8\varepsilon}{AB}\lambda\right) = AB\left(\lambda^2 - \frac{8\varepsilon}{AB}\lambda + \frac{16\varepsilon^2}{AB^2} - \frac{16\varepsilon^2}{AB^2}\right)$$

$$= \sum_{i=1}^{n}(b_i - a_i)^2\left(\lambda - \frac{4\varepsilon}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)^2 - \frac{16\varepsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}, \tag{2.29a}$$

$$\lambda^* = \operatorname*{argmin}_{\lambda} f(\lambda) = \frac{4\varepsilon}{\sum_{i=1}^{n}(b_i - a_i)^2} \quad \text{where} \quad \min_{\lambda} f(\lambda) = \frac{-2\varepsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}. \tag{2.29b}$$

Finally,

$$\mathbf{Pr}\left(\sum_{i=1}^{n} X_i - \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] \geqslant \varepsilon\right) \leqslant \exp\left(\frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2 - \lambda\varepsilon\right) = \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right). \tag{2.30}$$

By substituting λ^* into the calculation we obtain the result of the theorem. ■

Proof of Lemma 2.6. Note that

$$\mathbf{E}\left[e^{\lambda X}\right] = \mathbf{E}\left[e^{\lambda\left(X-\mathbf{E}[X]\right)+\lambda\mathbf{E}[X]}\right] = e^{\lambda\mathbf{E}[X]} \times \mathbf{E}\left[e^{\lambda\left(X-\mathbf{E}[X]\right)}\right]. \tag{2.31}$$

Hence, it is sufficient to show that for any random variable Z with $\mathbf{E}[Z] = 0$ and $Z \in [a,b]$ we have:

$$\mathbf{E}\left[e^{\lambda Z}\right] \leqslant e^{\lambda^2(b-a)^2/8}. \tag{2.32}$$

NB. Because X is a random variable such that $X \in [a,b]$, let $\mu \stackrel{\text{def}}{=} \mathbf{E}[X]$ for brevity . Then $\mu \in [a,b]$, and for any $\lambda \in \mathbb{R}$,

$$\mathbf{E}\left[e^{\lambda X}\right] = \mathbf{E}\left[e^{\lambda\left(X-\mathbf{E}[X]+\mathbf{E}[X]\right)}\right] = \mathbf{E}\left[e^{\lambda\left(X-\mathbf{E}[X]\right)} \cdot e^{\lambda\mathbf{E}[X]}\right] = e^{\lambda\mathbf{E}[X]}\mathbf{E}\left[e^{\lambda\left(X-\mathbf{E}[X]\right)}\right]. \tag{2.33}$$

Let $Z \stackrel{\text{def}}{=} X - \mathbf{E}[X]$, then $\mathbf{E}[Z] = \mathbf{E}\left[X - \mathbf{E}[X]\right] = \mathbf{E}[X] - \mathbf{E}[X] = 0$. As $X \in [a,b]$, we have $Z \in [a-\mu, b-\mu]$, and $\lambda Z \in [\lambda(a-\mu), \lambda(b-\mu)]$. Therefore, let $\lambda z \stackrel{\text{def}}{=} t\lambda(a-\mu) + (1-t)\lambda(b-\mu)$, that is, $z = t(a-\mu) + (1-t)(b-\mu) = t(a-b) + (b-\mu)$ we could obtain that $t = \frac{b-\mu-z}{b-a} = \frac{b-\mu-z}{b-a}$ and that $1-t = \frac{b-\mu-(a-\mu)}{b-a} - \frac{b-\mu-z}{b-a} = \frac{z-(a-\mu)}{b-a}$. Therefore, let $\hat{a} = a-\mu$ and $\hat{b} = b-\mu$, then

$$e^{\lambda z} = e^{\lambda\left(t(a-\mu)+(1-t)(b-\mu)\right)} \leqslant te^{\lambda(a-\mu)} + (1-t)e^{\lambda(b-\mu)} = \tfrac{b-\mu-z}{b-a}e^{\lambda(a-\mu)} + \tfrac{z-(a-\mu)}{b-a}e^{\lambda(b-\mu)}, \tag{2.34a}$$

$$\mathbf{E}\left[e^{\lambda z}\right] = \tfrac{\hat{b}-z}{\hat{b}-\hat{a}}e^{\lambda\hat{a}} + \tfrac{z-\hat{a}}{\hat{b}-\hat{a}}e^{\lambda\hat{b}} = \tfrac{\hat{b}-\mathbf{E}[z]}{b-a}e^{\lambda\hat{a}} + \tfrac{\mathbf{E}[z]-\hat{a}}{b-a}e^{\lambda\hat{b}} = \tfrac{b-\mu}{b-a}e^{\lambda(a-\mu)} + \tfrac{-(a-\mu)}{b-a}e^{\lambda(b-\mu)}. \tag{2.34b}$$

Let $p = \frac{-\hat{a}}{b-a}$ and $u = \lambda(b-a)$, then $1-p = \frac{\hat{b}}{b-a}$, and

$$\begin{aligned}\mathbf{E}\left[e^{\lambda z}\right] &= (1-p)e^{\lambda\hat{a}} + pe^{\lambda\hat{b}} = e^{\lambda\hat{a}}\left(1-p+pe^{\lambda\hat{b}-\lambda\hat{a}}\right) = e^{\lambda\frac{-\hat{a}}{b-a}(-1)(b-a)}\left(1-p+pe^{\lambda(\hat{b}-\hat{a})}\right)\\ &= e^{-p\lambda(b-a)}\left(1-p+pe^{\lambda(b-a)}\right) = e^{-pu}\left(1-p+pe^{u}\right) = e^{-pu+\ln(1-p+pe^u)} = e^{\phi(u)} \leqslant e^{\frac{\lambda^2(b-a)^2}{8}}.\end{aligned} \tag{2.35}$$

To summarise, $\mathbf{E}\left[e^{\lambda X}\right] = e^{\lambda\mathbf{E}[X]}\mathbf{E}\left[e^{\lambda Z}\right] \leqslant e^{\lambda\mathbf{E}[X]+\lambda^2(b-a)^2/8}$. Q.E.D.

By convexity of the exponential function, for $z \in [a,b]$ we have:

$$e^{\lambda z} \leqslant \frac{z-a}{b-a}e^{\lambda b} + \frac{b-z}{b-a}e^{\lambda a}. \tag{2.36}$$

Let $p = -a/(b-a)$. Then:

$$\begin{aligned}\mathbf{E}\left[e^{\lambda Z}\right] &\leqslant \mathbf{E}\left[\frac{Z-a}{b-a}e^{\lambda b} + \frac{b-Z}{b-a}e^{\lambda a}\right] = \frac{\mathbf{E}[Z]-a}{b-a}e^{\lambda b} + \frac{b-\mathbf{E}[Z]}{b-a}e^{\lambda a}\\ &= \frac{-a}{b-a}e^{\lambda b} + \frac{b}{b-a}e^{\lambda a} = \left(1-p+pe^{\lambda(b-a)}\right)e^{-p\lambda(b-a)} = e^{\phi(u)},\end{aligned} \tag{2.37a}$$

where $u = \lambda(b-a)$ and $\phi(u) = -pu + \ln(1-p+pe^u)$ and we used the fact that $\mathbf{E}[Z] = 0$.

NB. By convexity[1] of the exponential function e^z, for $z \in [a,b]$, we have: for $t \in (0,1)$, let $z \stackrel{\text{def}}{=} ta + (1-t)b = t(a-b) + b$, then $t = \frac{b-z}{b-a}$ and $1-t = \frac{b-a}{b-a} - \frac{b-z}{b-a} = \frac{z-a}{b-a}$, thus,

$$e^z = e^{ta+(1-t)b} \leqslant te^a + (1-t)e^b = \tfrac{b-z}{b-a}e^a + \tfrac{z-a}{b-a}e^b. \tag{2.38}$$

[1] Wikipedia (Convex function). Let X be a convex subset of a real vector space and let $f : X \to \mathbb{R}$ be a function. Then f is called *convex* if and only if any of the following equivalent conditions hold:

1) For all $0 \leqslant t \leqslant 1$ and all $x_1, x_2 \in \mathcal{X}$: $f\left(tx_1 + (1-t)x_2\right) \leqslant tf(x_1) + (1-t)f(x_2)$. The right hand side represents

Let $z' = \lambda z \in [\lambda a, \lambda b]$, then $z' \overset{\text{def}}{=} t(\lambda a) + (1-t)(\lambda b) = \lambda\big(t(a-b)+b\big) = \lambda z$. Thus, $t = \frac{b-\frac{1}{\lambda}z'}{b-a} = \frac{b-z}{b-a}$, and $1-t = \frac{\frac{1}{\lambda}z'-a}{b-a} = \frac{z-a}{b-a}$. Then we obtain that

$$e^{z'} = e^{\lambda z} = e^{t\lambda a+(1-t)\lambda b} \leqslant te^{\lambda a} + (1-t)e^{\lambda b} = \tfrac{b-z}{b-a}e^{\lambda a} + \tfrac{z-a}{b-a}e^{\lambda b}\,, \tag{2.39a}$$

$$\begin{aligned}\mathbf{E}\left[e^{\lambda z}\right] &\leqslant \mathbf{E}\left[\tfrac{b-z}{b-a}e^{\lambda a}\right] + \mathbf{E}\left[\tfrac{z-a}{b-a}e^{\lambda b}\right] = \mathbf{E}\left[\tfrac{b-z}{b-a}\right]e^{\lambda a} + \mathbf{E}\left[\tfrac{z-a}{b-a}\right]e^{\lambda b} \\ &= \tfrac{\mathbf{E}[b-z]}{b-a}e^{\lambda a} + \tfrac{\mathbf{E}[z-a]}{b-a}e^{\lambda b} = \tfrac{b-\mathbf{E}[z]}{b-a}e^{\lambda a} + \tfrac{\mathbf{E}[z]-a}{b-a}e^{\lambda b} = \tfrac{b}{b-a}e^{\lambda a} + \tfrac{-a}{b-a}e^{\lambda b}\,,\end{aligned} \tag{2.39b}$$

because of $\mathbf{E}[Z] = 0$. Let $p = -a/(b-a)$, then $1-p = 1-\frac{-a}{b-a} = b/(b-a)$, and

$$\mathbf{E}\left[e^{\lambda z}\right] \leqslant \frac{b}{b-a}e^{\lambda a} + \frac{-a}{b-a}e^{\lambda b} = (1-p)e^{\lambda a} + pe^{\lambda b} = e^{\lambda a}\left((1-p) + pe^{\lambda(b-a)}\right) \tag{2.40a}$$

$$= e^{-\frac{-a}{b-a}\lambda(b-a)}\left(1-p+pe^{\lambda(b-a)}\right) = e^{-p\lambda(b-a)}\left(1-p+pe^{\lambda(b-a)}\right)\,. \tag{2.40b}$$

Let $u = \lambda(b-a)$ and $\phi(u) = -pu + \ln(1-p+pe^u)$, then

$$\phi'(u) = -p + \tfrac{1}{1-p+pe^u}(pe^u) = -p + \tfrac{p}{p+(1-p)e^{-u}}\,, \tag{2.41a}$$

$$\phi''(u) = p(-1)\big(p+(1-p)e^{-u}\big)^{-2}(1-p)e^{-u}(-1) = \tfrac{p(1-p)e^{-u}}{\left(p+(1-p)e^{-u}\right)^2}\,, \tag{2.41b}$$

$$\phi(u) = -p\lambda(b-a) + \ln\left(1-p+pe^{\lambda(b-a)}\right)\,, \tag{2.41c}$$

$$\begin{aligned}e^{\phi(u)} &= e^{-p\lambda(b-a)}\cdot\left(1-p+pe^{\lambda(b-a)}\right) = e^{-\frac{-a}{b-a}\lambda(b-a)}\cdot\left(1-\tfrac{-a}{b-a}+\tfrac{-a}{b-a}e^{\lambda(b-a)}\right) \\ &= e^{\lambda a}\left(\tfrac{b}{b-a}+\tfrac{-a}{b-a}e^{\lambda(b-a)}\right) = \tfrac{b}{b-a}e^{\lambda a} + \tfrac{-a}{b-a}e^{\lambda b} = e^{\lambda a}\left(1-p+pe^{\lambda b-\lambda a}\right) = (1-p)e^{\lambda a} + pe^{\lambda b}\,.\end{aligned} \tag{2.41d}$$

---

the straight line between $(x_1, f(x_1))$ and $(x_2, f(x_2))$ in the graph of f as a function of t; increasing t from 0 to 1 or decreasing t from 1 to 0 sweeps this line. Similarly, the argument of the function f in the left hand side represents the straight line between x_1 and x_2 in X or the x-axis of the graph of f. So, this condition requires that the straight line between any pair of points on the curve of f to be above or just meets the graph.

2) For all $0 < t < 1$ and all $x_1, x_2 \in \mathcal{X}$ such that $x_1 \neq x_2$:

$$f\big(tx_1 + (1-t)x_2\big) \leqslant tf(x_1) + (1-t)f(x_2)\,.$$

The difference of this second condition with respect to the first condition above is that this condition does not include the intersection points (e.g., $(x_1, f(x_1))$ and $(x_2, f(x_2))$) between the straight line passing through a pair of points on the curve of f (the straight line is represented by the right hand side of this condition) and the curve of f; the first condition includes the intersection points as it becomes $f(x_1) \leqslant f(x_1)$ or $f(x_2) \leqslant f(x_2)$ at $t = 0$ or 1, or $x_1 = x_2$. In fact, the intersection points do not need to be considered in a condition of convex using $f(tx_1 + (1-t)x_2) \leqslant tf(x_1) + (1-t)f(x_2)$ because $f(x_1) \leqslant f(x_1)$ and $f(x_2) \leqslant f(x_2)$ are always true (so not useful to be a part of a condition).

The second statement characterising convex functions that are valued in the real line $\mathbb{R}$ is also the statement used to define *convex functions* that are valued in the extended real number line $[-\infty, \infty] = \mathbb{R} \cup \{\pm\infty\}$, where such a function f is allowed to (but is not required to) take $\pm\infty$ as a value. The first statement is not used because it permits t to take 0 or 1 as a value, in which case, if $f(x_1) = \pm\infty$ or $f(x_2) = \pm\infty$, respectively, then $tf(x_1) + (1-t)f(x_2)$ would be undefined (because the multiplications $0 \cdot \infty$ and $0 \cdot (-\infty)$ are undefined). Then sum $-\infty + \infty$ is also undefined so a convex extended real-valued function is typically only allowed to take exactly one of $-\infty$ and $+\infty$ as a value.

The second statement can also be modified to get the definition of *strict convexity*, where the latter is obtained by replacing $\leqslant$ with the strict inequality $<$. Explicitly, the map f is called *strictly convex* if and only if for all real $0 < t < 1$ and all $x_1, x_2 \in \mathcal{X}$ such that $x_1 \neq x_2$: $f\big(tx_1 + (1-t)x_2\big) < tf(x_1) + (1-t)f(x_2)$. A strictly convex function f is a function that the straight line between any pair of points on the curve f is above the curve f except for the intersection points between the straight line and the curve. The function f is said to be *concave* (resp. *strictly concave*) if $-f$ (f multiplied by -1) is convex (resp. strictly convex).

It is easy to verify that the derivative of ϕ is

$$\phi'(u) = -p + \frac{p}{p + (1-p)e^{-u}}, \tag{2.42}$$

and, therefore, $\phi(0) = \phi'(0) = 0$. Furthermore,

$$\phi''(u) = \frac{p(1-p)e^{-u}}{\left(p + (1-p)e^{-u}\right)^2} \leqslant \frac{1}{4}. \tag{2.43}$$

By Taylor's theorem, $\phi(u) = \phi(0) + u\phi'(0) + \frac{u^2}{2}\phi''(\theta)$ for some $\theta \in [0, u]$. Thus, we have:

$$\phi(u) = \phi(0) + u\phi'(0) + \frac{u^2}{2}\phi''(\theta) = \frac{u^2}{2}\phi''(\theta) \leqslant \frac{u^2}{8} = \frac{\lambda^2(b-a)^2}{8}. \tag{2.44}$$

NB. As $u = \lambda(b-a)$, we have known that if let $u = 0$, then

$$\phi(0) = -pu + \ln(1 - p + pe^u)\big|_{u=0} = 0 + \ln(1 - p + p) = 0, \tag{2.45a}$$

$$\phi'(0) = -p + \frac{p}{p+(1-p)e^{-u}}\bigg|_{u=1} = -p + \frac{p}{p+(1-p)} = 0, \tag{2.45b}$$

$$\phi''(0) = \frac{p(1-p)e^{-u}}{(p+(1-p)e^{-u})^2}\bigg|_{u=0} = \frac{p(1-p)}{(p+(1-p))^2} = p(1-p), \tag{2.45c}$$

$$\begin{aligned}\phi''(u) = \frac{p(1-p)e^{-u}}{(p+(1-p)e^{-u})^2} &= \frac{p(1-p)}{\left((1-p)\sqrt{e^{-u}} + \frac{p}{\sqrt{e^{-u}}}\right)^2} \leqslant \frac{p(1-p)}{\left(2\sqrt{(1-p)\sqrt{e^{-u}}\frac{p}{\sqrt{e^{-u}}}}\right)^2} \\ &= \frac{p(1-p)}{\left(2\sqrt{(1-p)p}\right)^2} = \frac{p(1-p)}{4(1-p)p} = \frac{1}{4},\end{aligned} \tag{2.45d}$$

due to $a + b \geqslant 2\sqrt{ab}$ for any $a, b \geqslant 0$.[2] At last, by Taylor's theorem[3,4] we have

$$\phi(u) = \phi(0) + u\phi'(0) + \tfrac{u^2}{2}\phi''(0) + \tfrac{u^3}{3!}\phi'''(\theta'), \quad \text{for some } \theta' \in [0, u] \tag{2.51a}$$

$$= \phi(0) + u\phi'(0) + \tfrac{u^2}{2}\phi''(\theta), \quad \text{for some } \theta \in [0, u] \tag{2.51b}$$

$$= 0 + u \cdot 0 + \tfrac{u^2}{2}\phi''(\theta) \leqslant 0 + \tfrac{u^2}{2} \cdot \tfrac{1}{4} = \tfrac{u^2}{8} = \tfrac{(\lambda(b-a))^2}{8} = \tfrac{\lambda^2(b-a)^2}{8}. \tag{2.51c}$$

[2] Note that $a + b = (\sqrt{a})^2 + (\sqrt{b})^2 = (\sqrt{a} + \sqrt{b})^2 - 2\sqrt{ab} = (\sqrt{a} - \sqrt{b})^2 + 2\sqrt{ab}$, therefore, $a + b \geqslant 2\sqrt{ab}$ for any $a, b \geqslant 0$.

[3] Wikipedia (Taylor's theorem). Let $k \geqslant 1$ be an integer and let the function $f : \mathbb{R} \to \mathbb{R}$ be k-times differentiable at the point $a \in \mathbb{R}$. Then there exists a function $h_k : \mathbb{R} \to \mathbb{R}$ such that

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2 + \cdots + \frac{f^{(k)}(a)}{k!}(x-a)^k + h_k(x)(x-a)^k, \tag{2.46}$$

and $\lim_{x \to a} h_k(x) = 0$. This is called the Peano form of the remainder.

The polynomial appearing in Taylor's theorem is the *k-th order Taylor polynomial*

$$P_k(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(k)}(a)}{k!}(x-a)^k, \tag{2.47}$$

of the function f at the point a. The Taylor polynomial is the unique "asymptotic best fit" polynomial in the sense that if there exists a function $h_k : \mathbb{R} \to \mathbb{R}$ and a k-th order polynomial p such that $f(x) = p(x) + h_k(x)(x-a)^k$, $\lim_{x \to a} h_k(x) = 0$,then $p = P_k$. Taylor's theorem describes the asymptotic behavior of the *remainder term* $R_k(x) = f(x) - P_k(x)$,which is the approximation error when approximating f with its Taylor polynomial. Using the little-o notation, the statement in Taylor's theorem reads as $R_k(x) = \mathbf{o}(|x-a|^k)$, $x \to a$.

[4] **Explicit formulas for the remainder** Under stronger regularity assumptions on f there are several precise formulas for the remainder term R_k of the Taylor's polynomial, the most common ones being the following.

Remark (Mean-value forms of the remainder). *Let $f : \mathbb{R} \to \mathbb{R}$ be $(k+1)$ times differentiable on the open interval with*

To sum up, we have

$$\mathbf{E}\left[e^{\lambda\left(X-\mathbf{E}[X]\right)}\right] = \mathbf{E}[e^{\lambda z}] \leqslant e^{\phi(u)} \leqslant e^{\frac{\lambda^2(b-a)^2}{8}}, \tag{2.52a}$$

$$\mathbf{E}\left[e^{\lambda X}\right] = e^{\lambda \mathbf{E}[X]} \cdot \mathbf{E}\left[e^{\lambda\left(X-\mathbf{E}[X]\right)}\right] \leqslant e^{\lambda \mathbf{E}[X]} e^{\frac{\lambda^2(b-a)^2}{8}} = e^{\lambda \mathbf{E}[X] + \frac{\lambda^2(b-a)^2}{8}}, \tag{2.52b}$$

that is, the inequality in Lemma 2.6 is demonstrated. ■

$f^{(k)}$ continuous on the closed interval between a and x. Then

$$R_k(x) = \frac{f^{(k+1)}(\xi_L)}{(k+1)!}(x-a)^{k+1}, \tag{2.48}$$

for some real number ξ_L between a and x. This is the Lagrange form *of the remainder. Similarly,*

$$R_k(x) = \frac{f^{(k+1)}(\xi_C)}{k!}(x-\xi_C)^k(x-a), \tag{2.49}$$

for some real number ξ_C between a and x. This is the Cauchy form *of the remainder.*

These refinements of Taylor's theorem are usually proved using the mean value theorem, whence the name. Also other similar expressions can be found. For example, if $G(t)$ is continuous on the closed interval and differentiable with a non-vanishing derivative on the open interval between a and x, then

$$R_k(x) = \frac{f^{(k+1)}(\xi)}{k!}(x-\xi)^k \frac{G(x)-G(a)}{G'(\xi)},$$

for some number ξ between a and x. This version covers the Lagrange and Cauchy forms of the remainder as special cases, and is proved below using Cauchy's mean value theorem. The statement for the integral form of the remainder is more advanced than the previous ones, and requires understanding of Lebesgue integration theory for the full generality. However, it holds also in the sense of Riemann integral provided the $(k+1)$-th derivative of f is continuous on the closed interval $[a, x]$.

Remark (Integral form of the remainder). *Let $f^{(k)}$ be absolutely continuous on the closed interval between a and x. Then*

$$R_k(x) = \int_a^x \frac{f^{(k+1)}(t)}{k!}(x-t)^k dt. \tag{2.50}$$

Due to absolute continuity of $f^{(k)}$ on the closed interval between a and x, its derivative $f^{(k+1)}$ exists as an L^1-function, and the result can be proven by a formal calculation using fundamental theorem of calculus and integration by parts.

Estimmates for the remainder It is often useful in practice to be able to estimate the remainder term appearing in the Taylor approximation, rather than having an exact formula for it. Suppose that f is $(k+1)$-times continuously differentiable in an interval ℓ containing a. Suppose that there are real constants q and Q such that $q \leqslant f^{(k+1)}(x) \leqslant Q$ throughout ℓ. Then the remainder term satisfies the inequality

$$q\frac{(x-a)^{k+1}}{(k+1)!} \leqslant R_k(x) \leqslant Q\frac{(x-a)^{k+1}}{(k+1)!},$$

if $x > a$, and a similar estimate if $x < a$. This is a simple consequence of the Lagrange form of the remainder. In particular, if $|f^{(k+1)}(x)| \leqslant M$ on an interval $\ell = (a-r, a+r)$ with some $r > 0$, then

$$|R_k(x)| \leqslant M\frac{|x-a|^{k+1}}{(k+1)!} \leqslant M\frac{r^{k+1}}{(k+1)!},$$

for all $x \in (a-r, a+r)$. The second inequality is called a uniform estimate, because it holds uniformly for all x on the interval $(a-r, a+r)$.