Chapter 3

Concentration of Measure Inequalities

Concentration of measure inequalities are one of the main tools for analyzing learning algorithms. This chapter is devoted to a number of concentration of measure inequalities that form the basis for the results discussed in later chapters.

3.1 Markov's Inequality

Markov's Inequality is the simplest and relatively weak concentration inequality. Nevertheless, it forms the basis for many much stronger inequalities that we will see in the sequel, and for some distributions it is actually tight (see Exercise 3.1).

Theorem 3.1 (Markov's Inequality)**.** *For any non-negative random variable X and $\varepsilon > 0$:*

$$\mathbb{P}(X \geq \varepsilon) \leq \frac{\mathbb{E}[X]}{\varepsilon}.$$

Proof. Define a random variable $Y = \mathbb{1}(X \geq \varepsilon)$ to be the indicator function of whether X exceeds ε. Then $Y \leq \frac{X}{\varepsilon}$ (see Figure 3.1). Since Y is a Bernoulli random variable, $\mathbb{E}[Y] = \mathbb{P}(Y = 1)$ (see Appendix B). We have:

$$\mathbb{P}(X \geq \varepsilon) = \mathbb{P}(Y = 1) = \mathbb{E}[Y] \leq \mathbb{E}\left[\frac{X}{\varepsilon}\right] = \frac{\mathbb{E}[X]}{\varepsilon}.$$

Check yourself: where in the proof do we use non-negativity of X and strict positiveness of ε? □

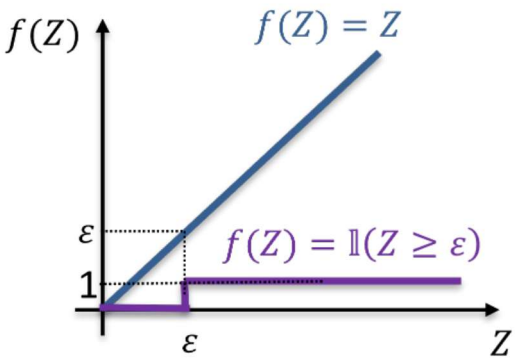

Figure 3.1: Relation between the identity function and the indicator function.

By denoting the right hand side of Markov's inequality by δ we obtain the following equivalent statement. For any non-negative random variable X:

$$\mathbb{P}\left(X \geq \frac{1}{\delta}\mathbb{E}\left[X\right]\right) \leq \delta.$$

Example. We would like to bound the probability that we flip a fair coin 10 times and obtain 8 or more heads. Let $X_1, \ldots, X_{10}$ be i.i.d. Bernoulli random variables with bias $\frac{1}{2}$. The question is equivalent to asking what is the probability that $\sum_{i=1}^{10} X_i \geq 8$. We have $\mathbb{E}\left[\sum_{i=1}^{10} X_i\right] = 5$ (the reader is invited to prove this statement formally) and by Markov's inequality

$$\mathbb{P}\left(\sum_{i=1}^{10} X_i \geq 8\right) \leq \frac{\mathbb{E}\left[\sum_{i=1}^{10} X_i\right]}{8} = \frac{5}{8}.$$

We note that even though Markov's inequality is weak, there are situations in which it is tight. We invite the reader to construct an example of a random variable for which Markov's inequality is tight.

3.2 Chebyshev's Inequality

Our next stop is Chebyshev's inequality, which exploits variance to obtain tighter concentration.

Theorem 3.2 (Chebyshev's inequality)**.** *For any $\varepsilon > 0$*

$$\mathbb{P}(|X - \mathbb{E}\left[X\right]| \geq \varepsilon) \leq \frac{\mathbb{V}\left[X\right]}{\varepsilon^2}.$$

Proof. The proof uses a transformation of a random variable. We have that $\mathbb{P}(|X - \mathbb{E}\left[X\right]| \geq \varepsilon) = \mathbb{P}\Big((X - \mathbb{E}\left[X\right])^2 \geq \varepsilon^2\Big)$, because the first statement holds if and only if the second holds. In addition, using Markov's inequality and the fact that $(X - \mathbb{E}\left[X\right])^2$ is a non-negative random variable we have

$$\mathbb{P}(|X - \mathbb{E}\left[X\right]| \geq \varepsilon) = \mathbb{P}\Big((X - \mathbb{E}\left[X\right])^2 \geq \varepsilon^2\Big) \leq \frac{\mathbb{E}\left[(X - \mathbb{E}\left[X\right])^2\right]}{\varepsilon^2} = \frac{\mathbb{V}\left[X\right]}{\varepsilon^2}.$$

Check yourself: where in the proof did we use the positiveness of ε? □

In order to illustrate the relative advantage of Chebyshev's inequality compared to Markov's consider the following example. Let $X_1, \ldots, X_n$ be n independent identically distributed Bernoulli random variables and let $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be their average. We would like to bound the probability that $\hat{\mu}_n$ deviates from $\mathbb{E}\left[\hat{\mu}_n\right]$ by more than ε (this is the central question in machine learning). We have $\mathbb{E}\left[\hat{\mu}_n\right] = \mathbb{E}\left[X_1\right] = \mu$ and by independence of X_i-s and Theorem B.26 we have $\mathbb{V}\left[\hat{\mu}_n\right] = \frac{1}{n^2}\mathbb{V}\left[n\hat{\mu}_n\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\left[X_i\right] = \frac{1}{n}\mathbb{V}\left[X_1\right]$. By Markov's inequality

$$\mathbb{P}(\hat{\mu}_n - \mathbb{E}\left[\hat{\mu}_n\right] \geq \varepsilon) = \mathbb{P}(\hat{\mu}_n \geq \mathbb{E}\left[\hat{\mu}_n\right] + \varepsilon) \leq \frac{\mathbb{E}\left[\hat{\mu}_n\right]}{\mathbb{E}\left[\hat{\mu}_n\right] + \varepsilon} = \frac{\mathbb{E}\left[X_1\right]}{\mathbb{E}\left[X_1\right] + \varepsilon}.$$

Note that as n grows the inequality stays the same. By Chebyshev's inequality we have

$$\mathbb{P}(\hat{\mu}_n - \mathbb{E}\left[\hat{\mu}_n\right] \geq \varepsilon) \leq \mathbb{P}(|\hat{\mu}_n - \mathbb{E}\left[\hat{\mu}_n\right]| \geq \varepsilon) \leq \frac{\mathbb{V}\left[\hat{\mu}_n\right]}{\varepsilon^2} = \frac{\mathbb{V}\left[X_1\right]}{n\varepsilon^2}.$$

Note that as n grows the right hand side of the inequality decreases at the rate of $\frac{1}{n}$. Thus, in this case Chebyshev's inequality is much tighter than Markov's and it illustrates that as the number of random variables grows the probability that their average significantly deviates from the expectation decreases. In the next section we show that this probability actually decreases at an exponential rate.

3.3 Hoeffding's Inequality

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

Theorem 3.3 (Hoeffding's Inequality)**.** *Let $X_1, \ldots, X_n$ be independent real-valued random variables, such that for each $i \in \{1, \ldots, n\}$ there exist $a_i \leq b_i$, such that $X_i \in [a_i, b_i]$. Then for every $\varepsilon > 0$:*

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \geq \varepsilon\right) \leq e^{-2\varepsilon^2 / \sum_{i=1}^{n}(b_i - a_i)^2} \tag{3.1}$$

and

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \leq -\varepsilon\right) \leq e^{-2\varepsilon^2 / \sum_{i=1}^{n}(b_i - a_i)^2}. \tag{3.2}$$

By taking a union bound of the events in (3.1) and (3.2) we obtain the following corollary.

Corollary 3.4. *Under the assumptions of Theorem 3.3:*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right]\right| \geq \varepsilon\right) \leq 2e^{-2\varepsilon^2 / \sum_{i=1}^{n}(b_i - a_i)^2}. \tag{3.3}$$

Equations (3.1) and (3.2) are known as "one-sided Hoeffding's inequalities" and (3.3) 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 the following corollary (see Exercise 3.2).

Corollary 3.5. *Let $X_1, \ldots, X_n$ be independent random variables, such that $X_i \in [0, 1]$ and $\mathbb{E}[X_i] = \mu$ for all i, then for every $\varepsilon > 0$:*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq \varepsilon\right) \leq e^{-2n\varepsilon^2} \tag{3.4}$$

and

$$\mathbb{P}\left(\mu - \frac{1}{n}\sum_{i=1}^{n} X_i \geq \varepsilon\right) \leq e^{-2n\varepsilon^2}. \tag{3.5}$$

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.

Lemma 3.6 (Hoeffding's Lemma)**.** *Let X be a random variable, such that $X \in [a, b]$. Then for any $\lambda \in \mathbb{R}$:*

$$\mathbb{E}\left[e^{\lambda X}\right] \leq e^{\lambda \mathbb{E}[X] + \frac{\lambda^2 (b-a)^2}{8}}.$$

The function $f(\lambda) = \mathbb{E}\left[e^{\lambda X}\right]$ is known as the *moment generating function* of X, since $f'(0) = \mathbb{E}[X]$, $f''(0) = \mathbb{E}\left[X^2\right]$, and, more generally, $f^{(k)}(0) = \mathbb{E}\left[X^k\right]$. We provide a proof of the lemma immediately after the proof of Theorem 3.3.

Proof of Theorem 3.3. We prove the first inequality in Theorem 3.3. The second inequality follows by applying the first inequality to $-X_1, \ldots, -X_n$. The proof is based on Chernoff's bounding technique. For any $\lambda > 0$ the following holds:

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \geq \varepsilon\right) = \mathbb{P}\left(e^{\lambda\left(\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right]\right)} \geq e^{\lambda\varepsilon}\right) \leq \frac{\mathbb{E}\left[e^{\lambda\left(\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right]\right)}\right]}{e^{\lambda\varepsilon}},$$

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

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda\left(\sum_{i=1}^n X_i - \mathbb{E}\left[\sum_{i=1}^n X_i\right]\right)}\right] &= \mathbb{E}\left[e^{\left(\sum_{i=1}^n \lambda(X_i - \mathbb{E}[X_i])\right)}\right] \\
&= \mathbb{E}\left[\prod_{i=1}^n e^{\lambda(X_i - \mathbb{E}[X_i])}\right] \\
&= \prod_{i=1}^n \mathbb{E}\left[e^{\lambda(X_i - \mathbb{E}[X_i])}\right] \qquad (3.6) \\
&\leq \prod_{i=1}^n e^{\lambda^2 (b_i - a_i)^2/8} \qquad (3.7) \\
&= e^{(\lambda^2/8)\sum_{i=1}^n (b_i - a_i)^2},
\end{aligned}$$

where (3.6) holds since $X_1, \ldots, X_n$ are independent and (3.7) holds by Hoeffding's lemma applied to a random variable $Z_i = X_i - \mathbb{E}[X_i]$ (note that $\mathbb{E}[Z_i] = 0$ and that $Z_i \in [a_i - \mu_i, b_i - \mu_i]$ for $\mu_i = \mathbb{E}[X_i]$). *Pay attention to the crucial role that independence of $X_1, \ldots, X_n$ 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! And it is not just that the proof would break down, but it is actually possible to construct examples of dependent random variables for which the empirical mean does not converge to its expectation, see Exercise 3.4.* To complete the proof we substitute the bound on the expectation into the previous calculation and obtain:

$$\mathbb{P}\left(\sum_{i=1}^n X_i - \mathbb{E}\left[\sum_{i=1}^n X_i\right] \geq \varepsilon\right) \leq e^{(\lambda^2/8)\left(\sum_{i=1}^n (b_i - a_i)^2\right) - \lambda\varepsilon}.$$

This expression is minimized by

$$\lambda^* = \arg\min_\lambda e^{(\lambda^2/8)\left(\sum_{i=1}^n (b_i - a_i)^2\right) - \lambda\varepsilon} = \arg\min_\lambda \left((\lambda^2/8)\left(\sum_{i=1}^n (b_i - a_i)^2\right) - \lambda\varepsilon\right) = \frac{4\varepsilon}{\sum_{i=1}^n (b_i - a_i)^2}.$$

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

Proof of Lemma 3.6. Note that

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

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

$$\mathbb{E}\left[e^{\lambda Z}\right] \leq e^{\lambda^2 (b-a)^2/8}.$$

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

$$e^{\lambda z} \leq \frac{z - a}{b - a} e^{\lambda b} + \frac{b - z}{b - a} e^{\lambda a}.$$

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

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

where $u = \lambda(b-a)$ and $\phi(u) = -pu + \ln\left(1 - p + pe^u\right)$ and we used the fact that $\mathbb{E}[Z] = 0$. It is easy to verify that the derivative of ϕ is

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

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

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

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:

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

□

3.3.1 Understanding Hoeffding's Inequality

Hoeffding's inequality involves three interconnected terms: n, ε, and $\delta = 2e^{-2n\varepsilon^2}$, which is the bound on the probability that the event under $\mathbb{P}()$ holds (for the purpose of the discussion we consider two-sided Hoeffding's inequality for random variables bounded in $[0, 1]$). We can fix any two of the three terms n, ε, and δ and then the relation $\delta = e^{-2n\varepsilon^2}$ provides the value of the third. Thus, we have

$$\begin{aligned}
\delta &= 2e^{-2n\varepsilon^2}, \\
\varepsilon &= \sqrt{\frac{\ln\frac{2}{\delta}}{2n}}, \\
n &= \frac{\ln\frac{2}{\delta}}{2\varepsilon^2}.
\end{aligned}$$

Overall, Hoeffding's inequality tells by how much the empirical average $\frac{1}{n}\sum_{i=1}^n X_i$ can deviate from its expectation μ, but the interplay between the three parameters provides several ways of seeing and using Hoeffding's inequality. For example, if the number of samples n is fixed (we have made a fixed number of experiments and now analyze what we can get from them), there is an interplay between the precision ε and confidence δ. We can request higher precision ε, but then we have to compromise on the confidence δ that the desired bound $\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right| \leq \varepsilon$ holds. And the other way around: we can request higher confidence δ, but then we have to compromise on precision ε, i.e., we have to increase the allowed range $\pm\varepsilon$ around μ, where we expect to find the empirical average $\frac{1}{n}\sum_{i=1}^n X_i$.

As another example, we may have target precision ε and confidence δ and then the inequality provides us the number of experiments n that we have to perform in order to achieve the target.

It is often convenient to write the inequalities (3.4) and (3.5) with a fixed confidence in mind, thus we have

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mu \geq \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right) &\leq \delta, \\
\mathbb{P}\left(\mu - \frac{1}{n}\sum_{i=1}^n X_i \geq \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right) &\leq \delta, \\
\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right| \geq \sqrt{\frac{\ln\frac{2}{\delta}}{2n}}\right) &\leq \delta.
\end{aligned}$$

(Pay attention that the $\ln 2$ factor in the last inequality comes from the union bound over the first two inequalities: if we want to keep the same confidence we have to compromise on precision.)

In many situations we are interested in the complimentary events. Thus, for example, we have

$$\mathbb{P}\left(\mu - \frac{1}{n}\sum_{i=1}^{n} X_i \leq \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right) \geq 1 - \delta.$$

Careful reader may point out that the inequalities above should be strict ("$<$" and "$>$"). This is true, but if it holds for strict inequalities it also holds for non-strict inequalities ("$\leq$" and "$\geq$"). Since strict inequalities provide no practical advantage we will use the non-strict inequalities to avoid the headache of remembering which inequalities should be strict and which should not.

The last inequality essentially says that with probability at least $1 - \delta$ we have

$$\mu \leq \frac{1}{n}\sum_{i=1}^{n} X_i + \sqrt{\frac{\ln\frac{1}{\delta}}{2n}} \tag{3.8}$$

and this is how we will occasionally use it. Note that the random variable is $\frac{1}{n}\sum_{i=1}^{n} X_i$ and the right way of interpreting the above inequality is actually that with probability at least $1 - \delta$

$$\frac{1}{n}\sum_{i=1}^{n} X_i \geq \mu - \sqrt{\frac{\ln\frac{1}{\delta}}{2n}},$$

i.e., the probability is over $\frac{1}{n}\sum_{i=1}^{n} X_i$ and not over μ. However, many generalization bounds that we study in Chapter 4 are written in the first form in the literature and we follow the tradition.

3.4 Sampling Without Replacement

Let $X_1, \ldots, X_n$ be a sequence of random variables *sampled without replacement* from a finite set of values $\mathcal{X} = \{x_1, \ldots, x_N\}$ of size N. The random variables $X_1, \ldots, X_n$ are *dependent*. For example, if $\mathcal{X} = \{-1, +1\}$ and we sample two values then $X_1 = -X_2$. Since $X_1, \ldots, X_n$ are dependent, the concentration results from previous sections do not apply directly. However, the following result by Hoeffding (1963, Theorem 4), which we cite without a proof, allows to extend results for sampling with replacement to sampling without replacement.

Lemma 3.7. *Let $X_1, \ldots, X_n$ denote a random sample without replacement from a finite set $\mathcal{X} = \{x_1, \ldots, x_N\}$ of N real values. Let $Y_1, \ldots, Y_n$ denote a random sample with replacement from $\mathcal{X}$. Then for any continuous and convex function $f : \mathbb{R} \to \mathbb{R}$*

$$\mathbb{E}\left[f\left(\sum_{i=1}^{n} X_i\right)\right] \leq \mathbb{E}\left[f\left(\sum_{i=1}^{n} Y_i\right)\right].$$

In particular, the lemma can be used to prove Hoeffding's inequality for sampling without replacement.

Theorem 3.8 (Hoeffding's inequality for sampling without replacement)**.** *Let $X_1, \ldots, X_n$ denote a random sample without replacement from a finite set $\mathcal{X} = \{x_1, \ldots, x_N\}$ of N values, where each element x_i is in the $[0, 1]$ interval. Let $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ be the average of the values in $\mathcal{X}$. Then for all $\varepsilon > 0$*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq \varepsilon\right) \leq e^{-2n\varepsilon^2},$$
$$\mathbb{P}\left(\mu - \frac{1}{n}\sum_{i=1}^{n} X_i \geq \varepsilon\right) \leq e^{-2n\varepsilon^2}.$$

The proof is a minor adaptation of the proof of Hoeffding's inequality for sampling with replacement using Lemma 3.7 and is left as an exercise. (Note that it requires a small modification inside the proof, because Lemma 3.7 cannot be applied directly to the statement of Hoeffding's inequality.)

While formal proof requires a bit of work, intuitively the result is quite expected. Imagine the process of sampling without replacement. If the average of points sampled so far starts deviating from the mean of the values in $\mathcal{X}$, the average of points that are left in $\mathcal{X}$ deviates in the opposite direction and "applies extra force" to new samples to bring the average back to μ. In the limit when $n = N$ we are guaranteed to have the average of X_i-s being equal to μ.

3.5 Basics of Information Theory: Entropy, Relative Entropy, and the Method of Types

In this section we briefly introduce a number of basic concepts from information theory that are very useful for deriving concentration inequalities. Specifically, we introduce the notions of entropy and relative entropy (Cover and Thomas, 2006, Chapter 2) and some basic tools from the method of types (Cover and Thomas, 2006, Chapter 11).

3.5.1 Entropy

We start with the definition of entropy.

Definition 3.9 (Entropy)**.** *Let $p(x)$ be a distribution of a discrete random variable X taking values in a finite set $\mathcal{X}$. We define the* entropy *of p as:*

$$\mathrm{H}(p) = -\sum_{x \in \mathcal{X}} p(x) \ln p(x).$$

We use the convention that $0 \ln 0 = 0$ (which is justified by continuity of $z \ln z$, since $z \ln z \to 0$ as $z \to 0$).

We have special interest in Bernoulli random variables.

Definition 3.10 (Bernoulli random variable)**.** *X is a Bernoulli random variable with bias p if X accepts values in $\{0, 1\}$ with $\mathbb{P}(X = 0) = 1 - p$ and $\mathbb{P}(X = 1) = p$.*

Note that expectation of a Bernoulli random variable is equal to its bias:

$$\mathbb{E}[X] = 0 \times \mathbb{P}(X = 0) + 1 \times \mathbb{P}(X = 1) = \mathbb{P}(X = 1) = p.$$

With a slight abuse of notation we specialize the definition of entropy to Bernoulli random variables.

Definition 3.11 (Binary entropy)**.** *Let p be a bias of Bernoulli random variable X. We define the entropy of p as*

$$\mathrm{H}(p) = -p \ln p - (1 - p) \ln(1 - p).$$

Note that when we talk about Bernoulli random variables p denotes the bias of the random variable and when we talk about more general random variables p denotes the complete distribution.

Entropy is one of the central quantities in information theory and it has numerous applications. We start by using binary entropy to bound binomial coefficients.

Lemma 3.12.

$$\frac{1}{n+1} e^{n \,\mathrm{H}\left(\frac{k}{n}\right)} \leq \binom{n}{k} \leq e^{n \,\mathrm{H}\left(\frac{k}{n}\right)}.$$

(Note that $\frac{k}{n} \in [0, 1]$ and $\mathrm{H}\left(\frac{k}{n}\right)$ in the lemma is the binary entropy.)

Proof. By the binomial formula we know that for any $p \in [0, 1]$:

$$\sum_{i=0}^{n} \binom{n}{i} p^i (1 - p)^{n-i} = 1. \tag{3.9}$$

We start with the upper bound. Take $p = \frac{k}{n}$. Since the sum is larger than any individual term, for the k-th term of the sum we get:

$$\begin{aligned}
1 &\geq \binom{n}{k} p^k (1-p)^{n-k} \\
&= \binom{n}{k} \left(\frac{k}{n}\right)^k \left(1 - \frac{k}{n}\right)^{n-k} \\
&= \binom{n}{k} \left(\frac{k}{n}\right)^k \left(\frac{n-k}{n}\right)^{n-k} \\
&= \binom{n}{k} e^{k \ln \frac{k}{n} + (n-k) \ln \frac{n-k}{n}} \\
&= \binom{n}{k} e^{n\left(\frac{k}{n} \ln \frac{k}{n} + \frac{n-k}{n} \ln \frac{n-k}{n}\right)} \\
&= \binom{n}{k} e^{-n\, \mathrm{H}\left(\frac{k}{n}\right)}.
\end{aligned}$$

By changing sides of the inequality we obtain the upper bound.

For the lower bound it is possible to show that if we fix $p = \frac{k}{n}$ then $\binom{n}{k} p^k (1-p)^{n-k} \geq \binom{n}{i} p^i (1-p)^{n-i}$ for any $i \in \{0, \ldots, n\}$, see Cover and Thomas (2006, Example 11.1.3) for details. We also note that there are $n+1$ elements in the sum in equation (3.9). Again, take $p = \frac{k}{n}$, then

$$1 \leq (n+1) \max_i \binom{n}{i} \left(\frac{k}{n}\right)^i \left(\frac{n-k}{n}\right)^{n-i} = (n+1) \binom{n}{k} \left(\frac{k}{n}\right)^k \left(\frac{n-k}{n}\right)^{n-k} = (n+1) \binom{n}{k} e^{-n\, \mathrm{H}\left(\frac{k}{n}\right)},$$

where the last step follows the same steps as in the derivation of the upper bound. □

Lemma 3.12 shows that the number of configurations of chosing k out of n objects is directly related to the entropy of the imbalance $\frac{k}{n}$ between the number of objects that are selected (k) and the number of objects that are left out ($n-k$).

3.5.2 The Kullback-Leibler (KL) Divergence (Relative Entropy)

We now introduce an additional quantity, the *Kullback-Leibler (KL) divergence*, also known as *Kullback-Leibler distance* and as *relative entropy.*

Definition 3.13 (Relative entropy or Kullback-Leibler divergence)**.** *Let $p(x)$ and $q(x)$ be two probability distributions of a random variable X (or two probability density functions, if X is a continuous random variable), the* Kullback-Leibler divergence *or* relative entropy *is defined as:*

$$\mathrm{KL}(p\|q) = \mathbb{E}_p\left[\ln \frac{p(X)}{q(X)}\right] = \begin{cases} \sum_{x \in \mathcal{X}} p(x) \ln \frac{p(x)}{q(x)}, & \text{if } \mathcal{X} \text{ is discrete} \\ \int_{x \in \mathcal{X}} p(x) \ln \frac{p(x)}{q(x)} dx, & \text{if } \mathcal{X} \text{ is continuous} \end{cases}.$$

We use the convention that $0 \ln \frac{0}{0} = 0$ and $0 \ln \frac{0}{q} = 0$ and $p \ln \frac{p}{0} = \infty$.

We specialize the definition to Bernoulli distributions.

Definition 3.14 (Binary kl-divergence)**.** *Let p and q be biases of two Bernoulli random variables. The* binary kl divergence *is defined as:*

$$\mathrm{kl}(p\|q) = \mathrm{KL}([1-p, p]\|[1-q, q]) = p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q}.$$

KL divergence is the central quantity in information theory. Although it is not a distance measure, because it does not satisfy the triangle inequality, it is the right way of measuring distances between probability distributions. This is illustrated by the following example.

Example 3.15. Let $X_1, \ldots, X_n$ be an i.i.d. sample of n Bernoulli random variables with bias p and let $\frac{1}{n}\sum_{i=1}^{n} X_i$ be the empirical bias of the sample. (Note that $\frac{1}{n}\sum_{i=1}^{n} X_i \in \left\{0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}\right\}$.) Then for $p \in (0,1)$

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{k}{n}\right) &= \binom{n}{k} p^k (1-p)^{n-k} \\
&= \binom{n}{k} e^{-n\,\mathrm{H}\left(\frac{k}{n}\right)} e^{n\,\mathrm{H}\left(\frac{k}{n}\right)} e^{n\left(\frac{k}{n}\ln p + \frac{n-k}{n}\ln(1-p)\right)} \\
&= \binom{n}{k} e^{-n\,\mathrm{H}\left(\frac{k}{n}\right)} e^{-n\,\mathrm{kl}\left(\frac{k}{n}\|p\right)} \qquad (3.10)
\end{aligned}$$

By Lemma 3.12 we have $\frac{1}{n+1} \leq \binom{n}{k} e^{-n\,\mathrm{H}\left(\frac{k}{n}\right)} \leq 1$, which gives

$$\frac{1}{n+1} e^{-n\,\mathrm{kl}\left(\frac{k}{n}\|p\right)} \leq \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{k}{n}\right) \leq e^{-n\,\mathrm{kl}\left(\frac{k}{n}\|p\right)}. \qquad (3.11)$$

Thus, $\mathrm{kl}\left(\frac{k}{n}\|p\right)$ governs the probability of observing empirical bias $\frac{k}{n}$ when the true bias is p. It is easy to verify that $\mathrm{kl}(p\|p) = 0$, and it is also possible to show that $\mathrm{kl}(\hat{p}\|p)$ is convex in $\hat{p}$, and that $\mathrm{kl}(\hat{p}\|p) \geq 0$ (Cover and Thomas, 2006). And so, the probability of empirical bias is maximized when it coincides with the true bias.

Properties of the KL and kl Divergences

The KL divergence between two probability distributions is always non-negative.

Theorem 3.16 (Nonnegativity of KL (Cover and Thomas, 2006, Theorem 2.3.6))**.** *Let $p(x)$ and $q(x)$ be two probability distributions. Then*

$$\mathrm{KL}(p\|q) \geq 0$$

with equality if and only if $p(x) = q(x)$ for all x.

Corollary 3.17 (Nonnegativity of kl)**.** *For $p, q \in [0,1]$*

$$\mathrm{kl}(p\|q) \geq 0$$

with equality if and only if $p = q$.

The KL divergence is also convex.

Theorem 3.18 (Convexity of KL (Cover and Thomas, 2006, Theorem 2.7.2))**.** $\mathrm{KL}(p\|q)$ *is convex in the pair (p, q); that is, if (p_1, q_1) and (p_2, q_2) are two pairs of probability mass functions, then*

$$\mathrm{KL}(\lambda p_1 + (1-\lambda)p_2 \| \lambda q_1 + (1-\lambda)q_2) \leq \lambda\,\mathrm{KL}(p_1\|q_1) + (1-\lambda)\,\mathrm{KL}(p_2\|q_2)$$

for all $0 \leq \lambda \leq 1$.

Corollary 3.19 (Convexity of kl)**.** *For $p_1, q_1, p_2, q_2 \in [0,1]$*

$$\mathrm{kl}(\lambda p_1 + (1-\lambda)p_2 \| \lambda q_1 + (1-\lambda)q_2) \leq \lambda\,\mathrm{kl}(p_1\|q_1) + (1-\lambda)\,\mathrm{kl}(p_2\|q_2)$$

for all $0 \leq \lambda \leq 1$.

3.6 The kl Inequality

Example 3.15 shows that kl can be used to bound the empirical bias when the true bias is known. But in machine learning we are usually interested in the inverse problem - how to infer the true bias p when the empirical bias $\hat{p}$ is known. Next we demonstrate that this is also possible and that it leads to an inequality, which is tighter than Hoeffding's inequality. We start with a simple version of kl lemma based on one-line derivation. Then we provide a tight version of the kl lemma and a lower bound showing that it cannot be improved any further. We then use the kl lemma to derive a kl inequality, and also provide a tighter version of the kl inequality, which is not based on the kl lemma. And we finish with relaxations of the kl inequality, which provide an intuitive interpretation of its implication.

3.6.1 A Simple Version of the kl Lemma

Lemma 3.20 (Simple kl Lemma)**.** *Let $X_1, \dots, X_n$ be i.i.d. Bernoulli with bias p and let $\hat{p} = \frac{1}{n}\sum_{i=1}^n X_i$ be the empirical bias. Then*

$$\mathbb{E}\left[e^{n\,\mathrm{kl}(\hat{p}\|p)}\right] \le n+1.$$

Proof. For $p \in (0,1)$

$$\mathbb{E}\left[e^{n\,\mathrm{kl}(\hat{p}\|p)}\right] = \sum_{k=0}^{n} \mathbb{P}\left(\hat{p} = \frac{k}{n}\right) e^{n\,\mathrm{kl}\left(\frac{k}{n}\|p\right)} \le \sum_{k=0}^{n} e^{-n\,\mathrm{kl}\left(\frac{k}{n}\|p\right)} e^{n\,\mathrm{kl}\left(\frac{k}{n}\|p\right)} = n+1,$$

where the inequality was derived in equation 3.11. For $p \in \{0,1\}$ we have $\mathbb{E}\left[e^{n\,\mathrm{kl}(\hat{p}\|o)}\right] = 1$, so the inequality is satisfied trivially. □

3.6.2 A Tight Version of the kl Lemma

In this section we provide a tight versions of the kl lemma. The improvement is based on a tighter control of the binomial coefficients, which is achieved by using Stirling's approximation of the factorial, $\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \le n! \le \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}$. The result is due to Wozengraft and Reiffen (1961).

Lemma 3.21. *For $1 \le k \le n-1$*

$$\frac{1}{2}\sqrt{\frac{n}{2k(n-k)}}e^{n\,\mathrm{H}\left(\frac{k}{n}\right)} \le \binom{n}{k} \le \frac{e^{\frac{1}{12n}}}{\sqrt{2\pi}}\sqrt{\frac{n}{k(n-k)}}e^{n\,\mathrm{H}\left(\frac{k}{n}\right)}.$$

The upper bound can be simplified using $\frac{e^{\frac{1}{12n}}}{\sqrt{2\pi}} < \frac{1}{2}$ for $n \ge 1$.

Proof. By Stirling's approximation

$$\begin{aligned}\binom{n}{k} &\le \frac{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}}{\sqrt{2\pi k}\left(\frac{k}{e}\right)^k \sqrt{2\pi(n-k)}\left(\frac{n-k}{e}\right)^{n-k}} \\ &= \frac{e^{\frac{1}{12n}}}{\sqrt{2\pi}}\sqrt{\frac{n}{k(n-k)}}\frac{1}{\left(\frac{k}{n}\right)^k\left(\frac{n-k}{n}\right)^{n-k}} \\ &= \frac{e^{\frac{1}{12n}}}{\sqrt{2\pi}}\sqrt{\frac{n}{k(n-k)}}e^{n\,\mathrm{H}\left(\frac{k}{n}\right)}.\end{aligned}$$

The lower bound is derived in a similar way, see Cover and Thomas (2006, Lemma 17.5.1). □

By combining Theorem 3.21 with Equation (3.10), for $1 \le k \le n-1$

$$\frac{1}{2}\sqrt{\frac{n}{2k(n-k)}}e^{-n\,\mathrm{kl}\left(\frac{k}{n}\|p\right)} \le \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i = \frac{k}{n}\right) \le \frac{e^{\frac{1}{12n}}}{\sqrt{2\pi}}\sqrt{\frac{n}{k(n-k)}}e^{-n\,\mathrm{kl}\left(\frac{k}{n}\|p\right)}.$$

The refinement can be used to tighten Theorem 3.20.

Lemma 3.22 (kl Lemma (Maurer, 2004))**.** *Let $X_1, \dots, X_n$ be i.i.d. with $X_1 \in [0,1]$, $\mathbb{E}[X_1] = p$, and $\hat{p} = \frac{1}{n}\sum_{i=1}^n X_i$. Then*

$$\mathbb{E}\left[e^{n\,\mathrm{kl}(\hat{p}\|p)}\right] \le 2\sqrt{n}.$$

The proof of Theorem 3.22 is based on two auxiliary results. The first is a technical bound on a summation.

Lemma 3.23 ((Maurer, 2004, Lemma 4))**.** *For $n \ge 2$*

$$1 \le \sum_{k=1}^{n-1} \frac{1}{\sqrt{k(n-k)}} \le \pi.$$

The second result allows to extend results for Bernoulli random variables to random variables bounded in the $[0,1]$ interval.

Lemma 3.24 ((Maurer, 2004, Lemma 3))**.** *Let $X_1, \dots, X_n$ be i.i.d. with $X_1 \in [0,1]$, and let $Y_1, \dots, Y_n$ be i.i.d. Bernoulli, such that $\mathbb{E}[Y_1] = \mathbb{E}[X_1]$. Then for any convex function $f : [0,1]^n \to \mathbb{R}$*

$$\mathbb{E}[f(X_1, \dots, X_n)] \le \mathbb{E}[f(Y_1, \dots, Y_n)].$$

The proof of Theorem 3.22 acts the same way as the proof of Theorem 3.20, just using the tighter bound on the binomial coefficients.

Proof of Theorem 3.22. We prove the result for Bernoulli random variables. The extension to general random variables bounded in $[0,1]$ follows by Theorem 3.24. For Bernoulli random variables and $p \in (0,1)$ we have

$$\mathbb{E}\left[e^{n\,\mathrm{kl}(\hat{p}\|p)}\right] = \sum_{k=0}^{n} \mathbb{P}\left(\hat{p} = \frac{k}{n}\right) e^{n\,\mathrm{kl}\left(\frac{k}{n}\|p\right)} = \sum_{k=0}^{n} \binom{n}{k} e^{-n\,\mathrm{H}\left(\frac{k}{n}\right)} e^{-n\,\mathrm{kl}\left(\frac{k}{n}\|p\right)} e^{n\,\mathrm{kl}\left(\frac{k}{n}\|p\right)} = \sum_{k=0}^{n} \binom{n}{k} e^{-n\,\mathrm{H}\left(\frac{k}{n}\right)}, \tag{3.12}$$

where the middle equality is by (3.10). By Theorem 3.21, for $n \ge 3$ we have

$$\sum_{k=0}^{n} \binom{n}{k} e^{-n\,\mathrm{H}\left(\frac{k}{n}\right)} = 2 + \sum_{k=1}^{n-1} \binom{n}{k} e^{-n\,\mathrm{H}\left(\frac{k}{n}\right)} \le 2 + e^{\frac{1}{12n}} \sqrt{\frac{n}{2\pi}} \sum_{k=1}^{n-1} \frac{1}{\sqrt{k(n-k)}} \le 2 + e^{\frac{1}{12n}} \sqrt{\frac{\pi n}{2}},$$

where the last inequality is by Theorem 3.23. For $n \ge 8$ we have $2 + e^{\frac{1}{12n}} \sqrt{\frac{\pi n}{2}} \le 2\sqrt{n}$, whereas for $n \in \{1, \dots, 7\}$ a direct calculation confirms that we also have $\sum_{k=0}^{n} \binom{n}{k} e^{-n\,\mathrm{H}\left(\frac{k}{n}\right)} \le 2\sqrt{n}$. For $p \in \{0,1\}$ we have $\mathbb{E}\left[e^{n\,\mathrm{kl}(\hat{p}\|p)}\right] = 1$, so the lemma holds trivially. □

The next result shows that the kl Lemma (Theorem 3.22) cannot be improved much further.

Lemma 3.25 (kl Lemma - Lower Bound (Maurer, 2004))**.** *Let $X_1, \dots, X_n$ be i.i.d. Bernoulli with $\mathbb{E}[X_1] = p$ and $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then for $p \in (0,1)$*

$$\mathbb{E}\left[e^{n\,\mathrm{kl}(\hat{p}\|p)}\right] \ge \sqrt{n}.$$

Proof. Starting from (3.12) and applying Theorem 3.21, for $p \in (0,1)$ and $n \ge 3$ we have

$$\mathbb{E}\left[e^{n\,\mathrm{kl}(\hat{p}\|p)}\right] = \sum_{k=0}^{n} \binom{n}{k} e^{-n\,\mathrm{H}\left(\frac{k}{n}\right)} = 2 + \sum_{k=1}^{n-1} \binom{n}{k} e^{-n\,\mathrm{H}\left(\frac{k}{n}\right)} \ge \frac{1}{2}\sqrt{\frac{n}{2}} \sum_{k=1}^{n-1} \frac{1}{\sqrt{k(n-k)}}.$$

The function $f(n) = \sum_{k=1}^{n-1} \frac{1}{\sqrt{k(n-k)}}$ is monotonically increasing with n. For $n \ge 88$ we have $f(n) > \sqrt{8}$, thus $f(n)\sqrt{\frac{n}{8}} \ge \sqrt{n}$. For $n \in \{1, \dots, 87\}$ a direct calculation confirms that $\sum_{k=0}^{n} \binom{n}{k} e^{-n\,\mathrm{H}\left(\frac{k}{n}\right)} \ge \sqrt{n}$. □

3.6.3 kl Inequality

By combining the kl lemma with Markov's inequality, we obtain the kl inequality.

Theorem 3.26 (kl Inequality via kl Lemma)**.** *Let $X_1, \dots, X_n$ be i.i.d. with $X_1 \in [0,1]$, $\mathbb{E}[X_1] = p$, and $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then*

$$\mathbb{P}\left(\mathrm{kl}(\hat{p}\|p) \ge \frac{\ln \frac{2\sqrt{n}}{\delta}}{n}\right) \le \delta.$$

Proof.

$$\mathbb{P}\left(\mathrm{kl}(\hat{p}\|p) \ge \frac{\ln \frac{2\sqrt{n}}{\delta}}{n}\right) = \mathbb{P}\left(e^{n\,\mathrm{kl}(\hat{p}\|p)} \ge \frac{2\sqrt{n}}{\delta}\right) \le \frac{\delta}{2\sqrt{n}} \mathbb{E}\left[e^{n\,\mathrm{kl}(\hat{p}\|p)}\right] \le \delta,$$

where the first inequality is by Markov's inequality, and the second inequality is by the kl lemma (Theorem 3.22). □

Even though Theorem 3.22 cannot be improved much further, it is possible to improve the kl inequality through a direct derivation that does not go through $\mathbb{E}\left[e^{n\,\mathrm{kl}(\hat{p}\|p)}\right]$.

Theorem 3.27 (kl Inequality (Langford, 2005, Foong et al., 2021, 2022)). *Let $X_1, \dots, X_n$ be i.i.d. with $X_1 \in [0,1]$, $\mathbb{E}[X_1] = p$, and $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then, for any $\delta \in (0,1)$:*

$$\mathbb{P}\left(\mathrm{kl}(\hat{p}\|p) \geq \frac{\ln\frac{1}{\delta}}{n}\right) \leq \delta.$$

We note that the direct derivation of the kl inequality that is behind Theorem 3.27 cannot be combined with PAC-Bayesian analysis that we study in Section 4.8. There we need to use Theorem 3.22 and pay the cost of $\ln 2\sqrt{n}$, as in Theorem 3.26. But in direct applications of the kl inequality, for example, in combination of the kl inequality with the Occam's razor, this cost can be avoided (see Exercise 4.7).

3.6.4 Relaxations of the kl-inequality: Pinsker's and refined Pinsker's inequalities

Theorem 3.27 implies that with probability at least $1-\delta$

$$\mathrm{kl}(\hat{p}\|p) \leq \frac{\ln\frac{1}{\delta}}{n}. \tag{3.13}$$

This leads to an implicit bound on p, which is not very intuitive and not always convenient to work with. In order to understand the behavior of the kl inequality better we use a couple of its relaxations. The first relaxation is known as Pinsker's inequality, see Cover and Thomas (2006, Lemma 11.6.1).

Lemma 3.28 (Pinsker's inequality).

$$\mathrm{KL}(p\|q) \geq \frac{1}{2}\|p-q\|_1^2,$$

where $\|p-q\|_1 = \sum_{x\in\mathcal{X}} |p(x)-q(x)|$ is the L_1-norm.

Corollary 3.29 (Pinsker's inequality for the binary kl divergence).

$$\mathrm{kl}(p\|q) \geq \frac{1}{2}\left(|p-q| + |(1-p)-(1-q)|\right)^2 = 2(p-q)^2. \tag{3.14}$$

By applying Corollary 3.29 to inequality (3.13), we obtain that with probability at least $1-\delta$

$$p \leq \hat{p} + \sqrt{\frac{\mathrm{kl}(\hat{p}\|p)}{2}} \leq \hat{p} + \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}, \tag{3.15}$$

where the first inequality is a deterministic inequality following by (3.14), and the second inequality holds with probability at least $1-\delta$ by (3.13). Note that inequality (3.15) is exactly the same as Hoeffding's inequality in Equation (3.8) (in fact, one way of proving Hoeffding's inequality is by deriving it via the kl divergence). Therefore, the kl inequality is always at least as tight as Hoeffding's inequality. But since (3.15) was achieved by Pinsker's relaxation of the kl inequality, the unrelaxed kl inequality can be tighter than Hoeffding's inequality.

Next we show that for small values of $\hat{p}$ the kl inequality is significantly tighter than Hoeffding's inequality. For this we use refined Pinsker's inequality (Marton, 1996, 1997, Samson, 2000, Boucheron et al., 2013, Lemma 8.4).

Lemma 3.30 (Refined Pinsker's inequality).

$$\mathrm{kl}(p\|q) \geq \frac{(p-q)^2}{2\max\{p,q\}} + \frac{(p-q)^2}{2\max\{(1-p),(1-q)\}}.$$

Corollary 3.31 (Refined Pinsker's inequality). *If $q > p$ then*

$$\mathrm{kl}(p\|q) \geq \frac{(p-q)^2}{2q}.$$

Corollary 3.32 (Refined Pinsker's inequality - upper bound)**.** *If* $\mathrm{kl}(p\|q) \leq \varepsilon$ *then*

$$q \leq p + \sqrt{2p\varepsilon} + 2\varepsilon.$$

Corollary 3.33 (Refined Pinsker's inequality - lower bound)**.** *If* $\mathrm{kl}(p\|q) \leq \varepsilon$ *then*

$$q \geq p - \sqrt{2p\varepsilon}.$$

By applying Corollary 3.32 to inequality (3.13), we obtain that with probability at least $1-\delta$

$$p \leq \hat{p} + \sqrt{\frac{2\hat{p}\ln\frac{1}{\delta}}{n}} + \frac{2\ln\frac{1}{\delta}}{n}. \tag{3.16}$$

When $\hat{p}$ is close to zero, the latter inequality is significantly tighter than Hoeffding's inequality. It exhibits what is know as "fast convergence rate", where for small values of $\hat{p}$ it approaches p at the rate of $\frac{1}{n}$ rather than $\frac{1}{\sqrt{n}}$, as in Hoeffding's inequality. Similarly, Theorem 3.33 clearly illustrates that when $\hat{p}$ is close to zero, the convergence of $\hat{p}$ to p from below also has "fast convergence rate".

We note that the kl inequality is always at least as tight as any of its relaxations, and that although there is no analytic inversion of $\mathrm{kl}(\hat{p}\|p)$, it is possible to invert it numerically to obtain even tighter bounds than the relaxations above. We use $\mathrm{kl}^{-1^+}(\hat{p},\varepsilon) := \max\{p : p \in [0,1] \text{ and } \mathrm{kl}(\hat{p}\|p) \leq \varepsilon\}$ to denote the upper inverse of kl and $\mathrm{kl}^{-1^-}(\hat{p},\varepsilon) := \min\{p : p \in [0,1] \text{ and } \mathrm{kl}(\hat{p}\|p) \leq \varepsilon\}$ to denote the lower inverse of kl. Then under the conditions of Theorem 3.27

$$\mathbb{P}\left(p \geq \mathrm{kl}^{-1^+}\left(\hat{p}, \frac{1}{n}\ln\frac{1}{\delta}\right)\right) \leq \delta, \tag{3.17}$$

$$\mathbb{P}\left(p \leq \mathrm{kl}^{-1^-}\left(\hat{p}, \frac{1}{n}\ln\frac{1}{\delta}\right)\right) \leq \delta. \tag{3.18}$$

Since $\mathrm{kl}(\hat{p}\|p)$ is convex in p, the inverses can be found using binary search.

Finally, we remind the reader that the random variable in the kl inequality is $\hat{p}$. Therefore, the correct way to see all the inequalities above is as inequalities on $\hat{p}$ rather than p. I.e., it is $\hat{p}$ that does not deviate a lot from p with high probability, rather than p staying close to $\hat{p}$ with high probability.

3.7 Split-kl Inequality

The kl inequality in Theorem 3.27 is almost the tightest that can be achieved for sums of Bernoulli random variables. (It is possible to obtain a bit tighter bounds by analyzing the binomial distribution directly, but the extra gains are minor (Langford, 2005).). However, it can potentially be very lose for sums of random variables taking values within the $[0,1]$ interval. The reason is that the kl inequality first maps any random variable taking values in the $[0,1]$ interval to a Bernoulli random variable (a random variable taking values $\{0,1\}$) with identical expectation, and then bounds the concentration of the original random variables by concentration of the Bernoulli random variables, as in Theorem 3.24. Whenever the original random variables have small variance, and the corresponding Bernoulli random variables have large variance, this approach is unable to exploit the small variance. (See Exercise 3.6, where you are asked to prove that if we fix expectation of a random variable, then Bernoulli random variable has the highest variance out of all random variables taking values in the $[0,1]$ interval and having a target expectation.) As an extreme example, imagine random variables $X_1,\ldots,X_n$, which all take value $p=\frac{1}{2}$ with probability 1. Then for any $\varepsilon > 0$ we have $\mathbb{P}\left(\left|p - \frac{1}{n}\sum_{i=1}^n X_i\right| > \varepsilon\right) = 0$, whereas the kl inequality only guarantees convergence of $\hat{p} = \frac{1}{n}\sum_{i=1}^n X_i$ to p at the rate of $\sqrt{\frac{\ln\frac{1}{\delta}}{n}}$.

3.7.1 Split-kl Inequality for Discrete Random Variables

In order to address the issue above, Wu et al. (2024) have proposed a way to represent discrete random variables as a superposition of Bernoulli random variables, and then apply the kl inequality to the Bernoulli elements in the decomposition. This approach preserves the kl tightness for the decomposition

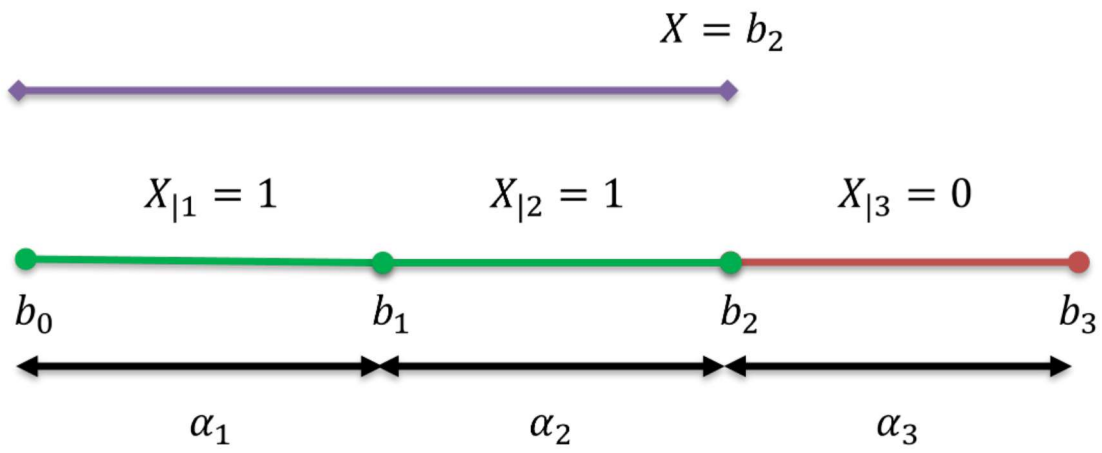

Figure 3.2: **Decomposition of a discrete random variable into a superposition of binary random variables.** The figure illustrates a decomposition of a discrete random variable X with domain of four values $b_0 < b_1 < b_2 < b_3$ into a superposition of three binary random variables, $X = b_0 + \sum_{j=1}^{3} \alpha_j X_{|j}$. A way to think about the decomposition is to compare it to a progress bar. In the illustration X takes value b_2, and so the random variables $X_{|1}$ and $X_{|2}$ corresponding to the first two segments "light up" (take value 1), whereas the random variable $X_{|3}$ corresponding to the last segment remains "turned off" (takes value 0). The value of X equals the sum of the lengths α_j of the "lighted up" segments. (The figure is borrowed from Wu et al. (2024).)

elements, and through it provides the combinatorial tightness of kl for general discrete random variables. The approach builds on an earlier work by Wu and Seldin (2022) for ternary random variables.

We now describe the decomposition. Let $X \in \{b_0, \dots, b_K\}$ be a $(K+1)$-valued random variable, where $b_0 < b_1 < \cdots < b_K$. For $j \in \{1, \dots, K\}$ define $X_{|j} = \mathbb{1}(X \geq b_j)$ and $\alpha_j = b_j - b_{j-1}$. Then $X = b_0 + \sum_{j=1}^{K} \alpha_j X_{|j}$, see Figure 3.2 for an illustration.

For a sequence $X_1, \dots, X_n$ of $(K+1)$-valued random variables with the same support, we let $X_{i|j} = \mathbb{1}(X_i \geq b_j)$ denote the elements of binary decomposition of X_i.

Theorem 3.34 (Split-kl inequality for discrete random variables (Wu et al., 2024)). *Let $X_1, \dots, X_n$ be i.i.d. random variables taking values in $\{b_0, \dots, b_K\}$ with $\mathbb{E}[X_i] = p$ for all i. Let $\hat{p}_{|j} = \frac{1}{n}\sum_{i=1}^{n} X_{i|j}$. Then for any $\delta \in (0,1)$:*

$$\mathbb{P}\left(p \geq b_0 + \sum_{j=1}^{K} \alpha_j \operatorname{kl}^{-1,+}\left(\hat{p}_{|j}, \frac{1}{n}\ln\frac{K}{\delta}\right)\right) \leq \delta.$$

Proof. Let $p_{|j} = \mathbb{E}[\hat{p}_{|j}]$, then $p = b_0 + \sum_{j=1}^{K} \alpha_j p_{|j}$ and

$$\mathbb{P}\left(p \geq b_0 + \sum_{j=1}^{K} \alpha_j \operatorname{kl}^{-1,+}\left(\hat{p}_{|j}, \frac{1}{n}\ln\frac{K}{\delta}\right)\right) \leq \mathbb{P}\left(\exists j : p_{|j} \geq \operatorname{kl}^{-1,+}\left(\hat{p}_{|j}, \frac{1}{n}\ln\frac{K}{\delta}\right)\right) \leq \delta,$$

where the first inequality is by the decomposition of p and the second inequality is by the union bound and (3.17). □

Since the kl inequalities provide almost the tightest bounds on the deviations of $\hat{p}_{|j}$ from $p_{|j}$ for each j individually, the split-kl inequality is almost the tightest that can be achieved for discrete random variables overall, as long as K and the corresponding $\ln K$ cost in the bound is not too large.

3.7.2 Split-kl Inequality for Bounded Continuous Random Variables

The split-kl inequality can also be applied to continuous random variables. Let $b_0 < b_1 < \cdots < b_K$ be an arbitrary split of an interval $[b_0, b_K]$ into K segments with $\alpha_j = b_j - b_{j-1}$ being the length of segment

j, and let $X \in [b_0, b_K]$ be a continuous random variable. Let

$$X_{|j} = \begin{cases} 0, & \text{if } X < b_{j-1}, \\ \frac{X - b_{j-1}}{\alpha_j}, & \text{if } b_{j-1} \le X \le b_j, \\ 1, & \text{if } X > b_j. \end{cases}$$

Then $X = b_0 + \sum_{j=1}^{K} \alpha_j X_{|j}$, and the split-kl inequality can be applied in exactly the same way as in the discrete case. The tightness of split-kl for continuous random variables depends on whether the probability mass is concentrated on or between the segment boundaries $b_0, \dots, b_K$ and on the magnitude of the $\ln K$ cost of the union bound.

3.8 Bernstein's Inequality

Bernstein's inequality is one of the most broadly known tools that exploit small variance to obtain tighter concentration. As most concentration of measure inequalities we have seen so far, it is based on a bound on a moment generating function.

Lemma 3.35 (Bernstein's Lemma)**.** *Let Z be a random variable, such that $\mathbb{E}[Z] = 0$, $\mathbb{E}\left[Z^2\right] \le \nu$, and $Z \le b$. Then for any $\lambda \in \left(0, \frac{3}{b}\right)$*

$$\mathbb{E}\left[e^{\lambda Z}\right] \le \exp\left(\frac{\lambda^2 \nu}{2\left(1 - \frac{b\lambda}{3}\right)}\right).$$

Proof. For $x \le 0$ we have $e^x \le 1 + x + \frac{1}{2}x^2$ and, therefore, for $Z \le 0$ we have

$$e^{\lambda Z} \le 1 + \lambda Z + \frac{\lambda^2 Z^2}{2} \le 1 + \lambda Z + \frac{\lambda^2 Z^2}{2\left(1 - \frac{b\lambda}{3}\right)},$$

where the last inequality holds because $\lambda \in \left(0, \frac{3}{b}\right)$, and so $\left(1 - \frac{b\lambda}{3}\right) \in (0, 1)$.

For $Z > 0$ we use Taylor's expansion of the exponent, $e^x = 1 + x + \frac{x^2}{2} + \sum_{i=3}^{\infty} \frac{1}{i!} x^i$, which gives

$$\begin{aligned} e^{\lambda Z} &= 1 + \lambda Z + \frac{\lambda^2 Z^2}{2} + \sum_{i=3}^{\infty} \frac{1}{i!} (\lambda Z)^i \\ &\le 1 + \lambda Z + \frac{\lambda^2 Z^2}{2} + \frac{\lambda^2 Z^2}{2} \sum_{i=3}^{\infty} \left(\frac{1}{3}\lambda Z\right)^{i-2} \\ &= 1 + \lambda Z + \frac{\lambda^2 Z^2}{2} \sum_{i=0}^{\infty} \left(\frac{1}{3}\lambda Z\right)^{i} \\ &\le 1 + \lambda Z + \frac{\lambda^2 Z^2}{2} \sum_{i=0}^{\infty} \left(\frac{1}{3}\lambda b\right)^{i} \\ &= 1 + \lambda Z + \frac{\lambda^2 Z^2}{2\left(1 - \frac{b\lambda}{3}\right)}. \end{aligned}$$

By combining this with the earlier inequality, we obtain that for all Z

$$e^{\lambda Z} \le 1 + \lambda Z + \frac{\lambda^2 Z^2}{2\left(1 - \frac{b\lambda}{3}\right)}.$$

And, therefore,

$$\mathbb{E}\left[e^{\lambda Z}\right] \le 1 + \lambda \mathbb{E}[Z] + \frac{\lambda^2 \mathbb{E}\left[Z^2\right]}{2\left(1 - \frac{b\lambda}{3}\right)} \le 1 + \frac{\lambda^2 \nu}{2\left(1 - \frac{b\lambda}{3}\right)} \le \exp\left(\frac{\lambda^2 \nu}{2\left(1 - \frac{b\lambda}{3}\right)}\right),$$

where in the second step we used the facts that $\mathbb{E}[Z] = 0$ and $\mathbb{E}\left[Z^2\right] \le \nu$, and the last step is based on the inequality $1 + x \le e^x$ that holds for all x. □

Now we need a couple of technical results, which we leave as an exercise.

Lemma 3.36. *For $x > 0$ let $f(x) = 1 + x - \sqrt{1+2x}$, then for any $\varepsilon > 0$*

$$\sup_{\lambda \in \left(0, \frac{1}{c}\right)} \left(\varepsilon\lambda - \frac{\lambda^2 \nu}{2(1-c\lambda)} \right) = \frac{\nu}{c^2} f\left(\frac{c\varepsilon}{\nu}\right).$$

Lemma 3.37. *For $x > 0$ let $f(x) = 1 + x - \sqrt{1+2x}$, then $f^{-1}(x) = x + \sqrt{2x}$.*

And now we are ready to present Bernstein's inequality.

Theorem 3.38 (Bernstein's Inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables, such that for all i we have $\mathbb{E}[X_i] - X_i \le b$ and $\mathbb{V}[X_i] \le \nu$. Then*

$$\mathbb{P}\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n X_i\right] \ge \frac{1}{n}\sum_{i=1}^n X_i + \sqrt{\frac{2\nu \ln\frac{1}{\delta}}{n}} + \frac{b \ln\frac{1}{\delta}}{3n} \right) \le \delta.$$

Note that if ν is close to zero, Bernstein's inequality provides "fast convergence rate", meaning that $\frac{1}{n}\sum_{i=1}^n X_i$ converges to $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n X_i\right]$ at the rate of $\frac{1}{n}$ rather than at the rate of $\frac{1}{\sqrt{n}}$.

Proof. The proof is based on Chernoff's bounding technique and follows the same strategy as earlier proofs of Hoeffding's and kl inequalities, just now using Bernstein's lemma instead of Hoeffding's or kl lemma. Let $Z_i = \mathbb{E}[X_i] - X_i$, then $\mathbb{E}[Z_i] = 0$, $\mathbb{E}\left[Z_i^2\right] = \mathbb{V}[X_i] \le \nu$, and $Z_i \le b$. For any $\lambda \in \left(0, \frac{b}{3}\right)$ we have

$$\begin{aligned}
\mathbb{P}\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n X_i\right] \ge \frac{1}{n}\sum_{i=1}^n X_i + \varepsilon \right) &= \mathbb{P}\left(\sum_{i=1}^n Z_i \ge n\varepsilon \right) \\
&= \mathbb{P}\left(e^{\lambda \sum_{i=1}^n Z_i} \ge e^{\lambda n \varepsilon} \right) \\
&\le e^{-\lambda n\varepsilon} \mathbb{E}\left[e^{\lambda \sum_{i=1}^n Z_i} \right] \\
&= e^{-\lambda n\varepsilon} \prod_{i=1}^n \mathbb{E}\left[e^{\lambda Z_i} \right] \\
&\le e^{-\lambda n\varepsilon} \prod_{i=1}^n \exp\left(\frac{\lambda^2 \nu}{2\left(1 - \frac{b\lambda}{3}\right)} \right) \\
&= \exp\left(-n \left(\lambda\varepsilon - \frac{\lambda^2 \nu}{2\left(1 - \frac{b\lambda}{3}\right)} \right) \right),
\end{aligned}$$

where the first inequality is by Markov's inequality and the second inequality is by Bernstein's lemma.

Since the bound holds for any $\lambda \in \left(0, \frac{b}{3}\right)$, we have

$$\mathbb{P}\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n X_i\right] \ge \frac{1}{n}\sum_{i=1}^n X_i + \varepsilon \right) \le \exp\left(-n \sup_{\lambda \in \left(0, \frac{3}{b}\right)} \left(\lambda\varepsilon - \frac{\lambda^2 \nu}{2\left(1 - \frac{b\lambda}{3}\right)} \right) \right) = \exp\left(-n a f(u) \right),$$

where $a = \frac{9\nu}{b^2}$, $u = \frac{b\varepsilon}{3\nu}$, $f(u) = 1 + u - \sqrt{1+2u}$, and the inequality follows by Theorem 3.36. Note that the right hand side of the inequality above is deterministic (independent of the random variable $\frac{1}{n}\sum_{i=1}^n X_i$), meaning that the optimal λ can be selected deterministically before observing the sample.

Finally, taking $\exp(-naf(u)) = \delta$ and using Theorem 3.37 to express ε in terms of δ, we obtain the statement in the theorem. □

3.9 Empirical Bernstein's Inequality

Bernstein's inequality (Theorem 3.38) assumes access to an upper bound ν on the variance. Empirical Bernstein's inequality presented in this section constructs a high-probability upper bound on the variance based on the sample, and then applies it within Bernstein's inequality, i.e., replaces ν with a high-probability upper bound on $\mathbb{V}[X]$.

Theorem 3.39 (Empirical Bernstein's Inequality (Maurer and Pontil, 2009)). *Let $X_1, \dots, X_n$ be independent identically distributed random variables taking values in $[0,1]$. Let $\hat{\nu}_n = \frac{1}{n(n-1)} \sum_{1 \le i < j \le n} (X_i - X_j)^2$. Then for any $\delta \in (0,1]$:*

$$\mathbb{P}\left(\mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} X_i \right] \ge \frac{1}{n} \sum_{i=1}^{n} X_i + \sqrt{\frac{2\hat{\nu}_n \ln \frac{2}{\delta}}{n}} + \frac{7 \ln \frac{2}{\delta}}{3(n-1)} \right) \le \delta.$$

The $\ln \frac{2}{\delta}$ term in the bound comes from a union bound over empirical bound on the variance and a bound on expectation of X. Note that the overall cost of replacing the true variance with its estimate is relatively small. Gao and Zhou (2013) offer a slight improvement of the bound, showing that for $n \ge 5$ the denominator in the last term can be increased from $n-1$ to n. We omit a proof of the theorem, but in Exercise 3.12 you are given a chance to prove a slightly weaker version of the theorem yourself.

3.10 Unexpected Bernstein's Inequality

Empirical Bernstein's inequality (Theorem 3.39) proceeds by bounding the variance using empirical variance estimate, and then using Bernstein's inequality to bound the expectation using the variance estimate. Unexpected Bernstein's inequality proceeds by bounding the expectation directly via the first and the second empirical moments. It is based on the following inequality due to Fan et al. (2015, Equation (4.12)): for $z \le 1$ and $\lambda \in [0, 1)$

$$e^{-\lambda z + z^2(\lambda + \ln(1-\lambda))} \le 1 - \lambda z. \tag{3.19}$$

The inequality can be used to prove the Unexpected Bernstein's lemma.

Lemma 3.40 (Unexpected Bernstein's lemma (Fan et al., 2015)). *Let X be a random variable bounded from above by $b > 0$. Then for all $\lambda \in \left[0, \frac{1}{b}\right)$*

$$\mathbb{E}\left[e^{\lambda(\mathbb{E}[X] - X) + \frac{b\lambda + \ln(1-b\lambda)}{b^2} X^2} \right] \le 1.$$

A proof is left as Exercise 3.13. Given the Unexpected Bernstein's lemma, we can use already well-established pipeline to prove the Unexpected Bernstein's inequality.

Theorem 3.41 (Unexpected Bernstein's Inequality (Fan et al., 2015, Mhammedi et al., 2019, Wu and Seldin, 2022)). *Let $X_1, \dots, X_n$ be independent identically distributed random variables bounded from above by b for $b > 0$. Let $\mu = \mathbb{E}[X_1]$, $\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$, and $\hat{s}_n = \frac{1}{n} \sum_{i=1}^{n} X_i^2$. Let $\psi(u) = u - \ln(1+u)$. Then for any $\lambda \in [0, 1/b)$ and $\delta \in (0,1]$:*

$$\mathbb{P}\left(\mu \ge \hat{\mu}_n + \frac{\psi(-\lambda b)}{\lambda b^2} \hat{s}_n + \frac{\ln \frac{1}{\delta}}{\lambda n} \right) \le \delta.$$

A proof is left as Exercise 3.14 and follows exactly the same steps as the proofs of Hoeffding's, kl, and Bernstein's inequalities. As already mentioned, the Unexpected Bernstein's inequality goes in one step from empirical first and second moments, $\hat{\mu}_n$ and $\hat{\nu}_n$, to a bound on μ. Note that in contrast to Hoeffding's and Bernstein's inequalities, the value of λ that minimizes the bound in Theorem 3.41 depends on $\hat{s}_n$, which is a random variable. Therefore, we cannot plug it into the bound. Instead, we can take a grid of λ values, a union bound over the grid, and then pick the best λ from the grid. Mhammedi et al. (2019) proposed to use the grid $\Lambda = \left\{1/2b, \dots, 1/(2^k b)\right\}$ for $k = \left\lceil \log_2 \left(\sqrt{n / \ln(1/\delta)}/2 \right) \right\rceil$, which works reasonably well in practice (Wu and Seldin, 2022). Formally, the bound then becomes:

$$\mathbb{P}\left(\mu \ge \hat{\mu}_n + \min_{\lambda \in \Lambda} \left(\frac{\psi(-\lambda b)}{\lambda b^2} \hat{s}_n + \frac{\ln \frac{k}{\delta}}{\lambda n} \right) \right) \le \sum_{\lambda \in \Lambda} \mathbb{P}\left(\mu \ge \hat{\mu}_n + \frac{\psi(-\lambda b)}{\lambda b^2} \hat{s}_n + \frac{\ln \frac{k}{\delta}}{\lambda n} \right) \le \delta.$$

(Note that $\ln \frac{1}{\delta}$ is replaced by $\ln \frac{k}{\delta}$ due to the union bound.)