Appendix B

Measure Concentration

Let $Z_1, \ldots, Z_m$ be an i.i.d. sequence of random variables and let μ be their mean. The strong law of large numbers states that when m tends to infinity, the empirical average, $\frac{1}{m}\sum_{i=1}^{m} Z_i$, converges to the expected value μ, with probability 1. Measure concentration inequalities quantify the deviation of the empirical average from the expectation when m is finite.

B.1 MARKOV'S INEQUALITY

We start with an inequality which is called Markov's inequality. Let Z be a nonnegative random variable. The expectation of Z can be written as follows:

$$\mathbb{E}[Z] = \int_{x=0}^{\infty} \mathbb{P}[Z \geq x] dx. \tag{B.1}$$

Since $\mathbb{P}[Z \geq x]$ is monotonically nonincreasing we obtain

$$\forall a \geq 0, \quad \mathbb{E}[Z] \geq \int_{x=0}^{a} \mathbb{P}[Z \geq x] dx \geq \int_{x=0}^{a} \mathbb{P}[Z \geq a] dx = a\, \mathbb{P}[Z \geq a]. \tag{B.2}$$

Rearranging the inequality yields Markov's inequality:

$$\forall a \geq 0, \quad \mathbb{P}[Z \geq a] \leq \frac{\mathbb{E}[Z]}{a}. \tag{B.3}$$

For random variables that take value in $[0, 1]$, we can derive from Markov's inequality the following.

Lemma B.1. *Let Z be a random variable that takes values in $[0, 1]$. Assume that $\mathbb{E}[Z] = \mu$. Then, for any $a \in (0, 1)$,*

$$\mathbb{P}[Z > 1 - a] \geq \frac{\mu - (1 - a)}{a}.$$

This also implies that for every $a \in (0, 1)$,

$$\mathbb{P}[Z > a] \geq \frac{\mu - a}{1 - a} \geq \mu - a.$$

Proof. Let $Y = 1 - Z$. Then Y is a nonnegative random variable with $\mathbb{E}[Y] = 1 - \mathbb{E}[Z] = 1 - \mu$. Applying Markov's inequality on Y we obtain

$$\mathbb{P}[Z \leq 1 - a] = \mathbb{P}[1 - Z \geq a] = \mathbb{P}[Y \geq a] \leq \frac{\mathbb{E}[Y]}{a} = \frac{1 - \mu}{a}.$$

Therefore,

$$\mathbb{P}[Z > 1 - a] \geq 1 - \frac{1 - \mu}{a} = \frac{a + \mu - 1}{a}.$$

□

B.2 CHEBYSHEV'S INEQUALITY

Applying Markov's inequality on the random variable $(Z - \mathbb{E}[Z])^2$ we obtain Chebyshev's inequality:

$$\forall a > 0, \quad \mathbb{P}[|Z - \mathbb{E}[Z]| \geq a] = \mathbb{P}[(Z - \mathbb{E}[Z])^2 \geq a^2] \leq \frac{\mathrm{Var}[Z]}{a^2}, \tag{B.4}$$

where $\mathrm{Var}[Z] = \mathbb{E}[(Z - \mathbb{E}[Z])^2]$ is the variance of Z.

Consider the random variable $\frac{1}{m}\sum_{i=1}^{m} Z_i$. Since $Z_1, \ldots, Z_m$ are i.i.d. it is easy to verify that

$$\mathrm{Var}\left[\frac{1}{m}\sum_{i=1}^{m} Z_i\right] = \frac{\mathrm{Var}[Z_1]}{m}.$$

Applying Chebyshev's inequality, we obtain the following:

Lemma B.2. *Let $Z_1, \ldots, Z_m$ be a sequence of i.i.d. random variables and assume that $\mathbb{E}[Z_1] = \mu$ and $\mathrm{Var}[Z_1] \leq 1$. Then, for any $\delta \in (0, 1)$, with probability of at least $1 - \delta$ we have*

$$\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\right| \leq \sqrt{\frac{1}{\delta m}}.$$

Proof. Applying Chebyshev's inequality we obtain that for all $a > 0$

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\right| > a\right] \leq \frac{\mathrm{Var}[Z_1]}{m a^2} \leq \frac{1}{m a^2}.$$

The proof follows by denoting the right-hand side δ and solving for a. □

The deviation between the empirical average and the mean given previously decreases polynomially with m. It is possible to obtain a significantly faster decrease. In the sections that follow we derive bounds that decrease exponentially fast.

B.3 CHERNOFF'S BOUNDS

Let $Z_1, \ldots, Z_m$ be independent Bernoulli variables where for every i, $\mathbb{P}[Z_i = 1] = p_i$ and $\mathbb{P}[Z_i = 0] = 1 - p_i$. Let $p = \sum_{i=1}^{m} p_i$ and let $Z = \sum_{i=1}^{m} Z_i$. Using the monotonicity

of the exponent function and Markov's inequality, we have that for every $t > 0$

$$\mathbb{P}[Z > (1+\delta)p] = \mathbb{P}[e^{tZ} > e^{t(1+\delta)p}] \leq \frac{\mathbb{E}[e^{tZ}]}{e^{(1+\delta)tp}}. \tag{B.5}$$

Next,

$$\begin{aligned}
\mathbb{E}[e^{tZ}] = \mathbb{E}[e^{t\sum_i Z_i}] &= \mathbb{E}\left[\prod_i e^{tZ_i}\right] \\
&= \prod_i \mathbb{E}[e^{tZ_i}] && \text{by independence} \\
&= \prod_i \left(p_i e^t + (1-p_i)e^0\right) \\
&= \prod_i \left(1 + p_i(e^t - 1)\right) \\
&\leq \prod_i e^{p_i(e^t-1)} && \text{using } 1+x \leq e^x \\
&= e^{\sum_i p_i(e^t-1)} \\
&= e^{(e^t-1)p}.
\end{aligned}$$

Combining the equation with Equation (B.5) and choosing $t = \log(1+\delta)$ we obtain

Lemma B.3. *Let $Z_1, \ldots, Z_m$ be independent Bernoulli variables where for every i, $\mathbb{P}[Z_i = 1] = p_i$ and $\mathbb{P}[Z_i = 0] = 1 - p_i$. Let $p = \sum_{i=1}^m p_i$ and let $Z = \sum_{i=1}^m Z_i$. Then, for any $\delta > 0$,*

$$\mathbb{P}[Z > (1+\delta)p] \;\leq\; e^{-h(\delta)\,p},$$

where

$$h(\delta) = (1+\delta)\log(1+\delta) - \delta.$$

Using the inequality $h(a) \geq a^2/(2 + 2a/3)$ we obtain

Lemma B.4. *Using the notation of Lemma B.3 we also have*

$$\mathbb{P}[Z > (1+\delta)p] \;\leq\; e^{-p\frac{\delta^2}{2+2\delta/3}}.$$

For the other direction, we apply similar calculations:

$$\mathbb{P}[Z < (1-\delta)p] = \mathbb{P}[-Z > -(1-\delta)p] = \mathbb{P}[e^{-tZ} > e^{-t(1-\delta)p}] \leq \frac{\mathbb{E}[e^{-tZ}]}{e^{-(1-\delta)tp}}, \tag{B.6}$$

and

$$\begin{aligned}
\mathbb{E}[e^{-tZ}] = \mathbb{E}[e^{-t\sum_i Z_i}] &= \mathbb{E}\left[\prod_i e^{-tZ_i}\right] \\
&= \prod_i \mathbb{E}[e^{-tZ_i}] && \text{by independence} \\
&= \prod_i \left(1 + p_i(e^{-t} - 1)\right) \\
&\le \prod_i e^{p_i(e^{-t}-1)} && \text{using } 1 + x \le e^x \\
&= e^{(e^{-t}-1)p}.
\end{aligned}$$

Setting $t = -\log(1-\delta)$ yields

$$\mathbb{P}[Z < (1-\delta)p] \;\le\; \frac{e^{-\delta p}}{e^{(1-\delta)\log(1-\delta)\,p}} = e^{-ph(-\delta)}.$$

It is easy to verify that $h(-\delta) \ge h(\delta)$ and hence

Lemma B.5. *Using the notation of Lemma B.3 we also have*

$$\mathbb{P}[Z < (1-\delta)p] \;\le\; e^{-ph(-\delta)} \le e^{-ph(\delta)} \le e^{-p\frac{\delta^2}{2+2\delta/3}}.$$

B.4 HOEFFDING'S INEQUALITY

Lemma B.6 (Hoeffding's Inequality). *Let $Z_1, \ldots, Z_m$ be a sequence of i.i.d. random variables and let $\bar{Z} = \frac{1}{m}\sum_{i=1}^m Z_i$. Assume that $\mathbb{E}[\bar{Z}] = \mu$ and $\mathbb{P}[a \le Z_i \le b] = 1$ for every i. Then, for any $\epsilon > 0$*

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^m Z_i - \mu\right| > \epsilon\right] \;\le\; 2\exp\left(-2m\epsilon^2/(b-a)^2\right).$$

Proof. Denote $X_i = Z_i - \mathbb{E}[Z_i]$ and $\bar{X} = \frac{1}{m}\sum_i X_i$. Using the monotonicity of the exponent function and Markov's inequality, we have that for every $\lambda > 0$ and $\epsilon > 0$,

$$\mathbb{P}[\bar{X} \ge \epsilon] = \mathbb{P}[e^{\lambda\bar{X}} \ge e^{\lambda\epsilon}] \le e^{-\lambda\epsilon}\,\mathbb{E}[e^{\lambda\bar{X}}].$$

Using the independence assumption we also have

$$\mathbb{E}[e^{\lambda\bar{X}}] = \mathbb{E}\left[\prod_i e^{\lambda X_i/m}\right] = \prod_i \mathbb{E}[e^{\lambda X_i/m}].$$

By Hoeffding's lemma (Lemma B.7 later), for every i we have

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

Therefore,

$$\mathbb{P}[\bar{X} \ge \epsilon] \le e^{-\lambda\epsilon}\prod_i e^{\frac{\lambda^2(b-a)^2}{8m^2}} = e^{-\lambda\epsilon + \frac{\lambda^2(b-a)^2}{8m}}.$$

Setting $\lambda = 4m\epsilon/(b-a)^2$ we obtain

$$\mathbb{P}[\bar{X} \geq \epsilon] \leq e^{-\frac{2m\epsilon^2}{(b-a)^2}}.$$

Applying the same arguments on the variable $-\bar{X}$ we obtain that $\mathbb{P}[\bar{X} \leq -\epsilon] \leq e^{-\frac{2m\epsilon^2}{(b-a)^2}}$. The theorem follows by applying the union bound on the two cases. □

Lemma B.7 (Hoeffding's Lemma). *Let X be a random variable that takes values in the interval $[a,b]$ and such that $\mathbb{E}[X] = 0$. Then, for every $\lambda > 0$,*

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

Proof. Since $f(x) = e^{\lambda x}$ is a convex function, we have that for every $\alpha \in (0,1)$, and $x \in [a,b]$,

$$f(x) \leq \alpha f(a) + (1-\alpha) f(b).$$

Setting $\alpha = \frac{b-x}{b-a} \in [0,1]$ yields

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

Taking the expectation, we obtain that

$$\mathbb{E}[e^{\lambda X}] \leq \frac{b-\mathbb{E}[X]}{b-a} e^{\lambda a} + \frac{\mathbb{E}[x]-a}{b-a} e^{\lambda b} = \frac{b}{b-a} e^{\lambda a} - \frac{a}{b-a} e^{\lambda b},$$

where we used the fact that $\mathbb{E}[X] = 0$. Denote $h = \lambda(b-a)$, $p = \frac{-a}{b-a}$, and $L(h) = -hp + \log(1 - p + pe^h)$. Then, the expression on the right-hand side of the equation can be rewritten as $e^{L(h)}$. Therefore, to conclude our proof it suffices to show that $L(h) \leq \frac{h^2}{8}$. This follows from Taylor's theorem using the facts $L(0) = L'(0) = 0$ and $L''(h) \leq 1/4$ for all h. □

B.5 BENNET'S AND BERNSTEIN'S INEQUALITIES

Bennet's and Bernsein's inequalities are similar to Chernoff's bounds, but they hold for any sequence of independent random variables. We state the inequalities without proof, which can be found, for example, in Cesa-Bianchi and Lugosi (2006).

Lemma B.8 (Bennet's Inequality). *Let $Z_1, \ldots, Z_m$ be independent random variables with zero mean, and assume that $Z_i \leq 1$ with probability 1. Let*

$$\sigma^2 \geq \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}[Z_i^2].$$

Then for all $\epsilon > 0$,

$$\mathbb{P}\left[\sum_{i=1}^{m} Z_i > \epsilon\right] \leq e^{-m\sigma^2 h\left(\frac{\epsilon}{m\sigma^2}\right)}.$$

where

$$h(a) = (1+a)\log(1+a) - a.$$

By using the inequality $h(a) \geq a^2/(2+2a/3)$ it is possible to derive the following:

Lemma B.9 (Bernstein's Inequality). *Let $Z_1, \ldots, Z_m$ be i.i.d. random variables with a zero mean. If for all i, $\mathbb{P}(|Z_i| < M) = 1$, then for all $t > 0$:*

$$\mathbb{P}\left[\sum_{i=1}^{m} Z_i > t\right] \leq \exp\left(-\frac{t^2/2}{\sum \mathbb{E} Z_j^2 + Mt/3}\right).$$

B.5.1 Application

Bernstein's inequality can be used to interpolate between the rate $1/\epsilon$ we derived for PAC learning in the realizable case (in Chapter 2) and the rate $1/\epsilon^2$ we derived for the unrealizable case (in Chapter 4).

Lemma B.10. *Let $\ell : \mathcal{H} \times Z \to [0,1]$ be a loss function. Let $\mathcal{D}$ be an arbitrary distribution over Z. Fix some h. Then, for any $\delta \in (0,1)$ we have*

1. $$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\left[L_S(h) \geq L_D(h) + \sqrt{\frac{2L_D(h)\log(1/\delta)}{3m}} + \frac{2\log(1/\delta)}{m}\right] \leq \delta$$
2. $$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\left[L_D(h) \geq L_S(h) + \sqrt{\frac{2L_S(h)\log(1/\delta)}{m}} + \frac{4\log(1/\delta)}{m}\right] \leq \delta$$

Proof. Define random variables $\alpha_1, \ldots, \alpha_m$ s.t. $\alpha_i = \ell(h, z_i) - L_D(h)$. Note that $\mathbb{E}[\alpha_i] = 0$ and that

$$\begin{aligned}
\mathbb{E}[\alpha_i^2] &= \mathbb{E}[\ell(h,z_i)^2] - 2L_{\mathcal{D}}(h)\,\mathbb{E}[\ell(h,z_i)] + L_{\mathcal{D}}(h)^2 \\
&= \mathbb{E}[\ell(h,z_i)^2] - L_{\mathcal{D}}(h)^2 \\
&\leq \mathbb{E}[\ell(h,z_i)^2] \\
&\leq \mathbb{E}[\ell(h,z_i)] = L_{\mathcal{D}}(h),
\end{aligned}$$

where in the last inequality we used the fact that $\ell(h,z_i) \in [0,1]$ and thus $\ell(h,z_i)^2 \leq \ell(h,z_i)$. Applying Bernsein's inequality over the α_i's yields

$$\begin{aligned}
\mathbb{P}\left[\sum_{i=1}^{m} \alpha_i > t\right] &\leq \exp\left(-\frac{t^2/2}{\sum \mathbb{E}\alpha_j^2 + t/3}\right) \\
&\leq \exp\left(-\frac{t^2/2}{m\,L_{\mathcal{D}}(h) + t/3}\right) \stackrel{\text{def}}{=} \delta.
\end{aligned}$$

Solving for t yields

$$\begin{aligned}
&\frac{t^2/2}{m\,L_{\mathcal{D}}(h)+t/3} = \log(1/\delta) \\
&\Rightarrow\ t^2/2 - \frac{\log(1/\delta)}{3}t - \log(1/\delta)m\,L_{\mathcal{D}}(h) = 0 \\
&\Rightarrow\ t = \frac{\log(1/\delta)}{3} + \sqrt{\frac{\log^2(1/\delta)}{3^2} + 2\log(1/\delta)m\,L_{\mathcal{D}}(h)} \\
&\leq 2\frac{\log(1/\delta)}{3} + \sqrt{2\log(1/\delta)m\,L_{\mathcal{D}}(h)}
\end{aligned}$$

Since $\frac{1}{m}\sum_i \alpha_i = L_S(h) - L_{\mathcal{D}}(h)$, it follows that with probability of at least $1-\delta$,

$$L_S(h) - L_{\mathcal{D}}(h) \leq 2\frac{\log(1/\delta)}{3m} + \sqrt{\frac{2\log(1/\delta)\,L_{\mathcal{D}}(h)}{m}},$$

which proves the first inequality. The second part of the lemma follows in a similar way. □

B.6 SLUD'S INEQUALITY

Let X be a (m, p) binomial variable. That is, $X = \sum_{i=1}^{m} Z_i$, where each Z_i is 1 with probability p and 0 with probability $1-p$. Assume that $p = (1-\epsilon)/2$. Slud's inequality (Slud 1977) tells us that $\mathbb{P}[X \geq m/2]$ is lower bounded by the probability that a normal variable will be greater than or equal to $\sqrt{m\epsilon^2/(1-\epsilon^2)}$. The following lemma follows by standard tail bounds for the normal distribution.

Lemma B.11. *Let X be a (m, p) binomial variable and assume that $p = (1-\epsilon)/2$. Then,*

$$\mathbb{P}[X \geq m/2] \geq \frac{1}{2}\left(1 - \sqrt{1 - \exp(-m\epsilon^2/(1-\epsilon^2))}\right).$$

B.7 CONCENTRATION OF χ^2 VARIABLES

Let $X_1, \ldots, X_k$ be k independent normally distributed random variables. That is, for all i, $X_i \sim N(0,1)$. The distribution of the random variable X_i^2 is called χ^2 (chi square) and the distribution of the random variable $Z = X_1^2 + \cdots + X_k^2$ is called χ_k^2 (chi square with k degrees of freedom). Clearly, $\mathbb{E}[X_i^2] = 1$ and $\mathbb{E}[Z] = k$. The following lemma states that X_k^2 is concentrated around its mean.

Lemma B.12. *Let $Z \sim \chi_k^2$. Then, for all $\epsilon > 0$ we have*

$$\mathbb{P}[Z \leq (1-\epsilon)k] \leq e^{-\epsilon^2 k/6},$$

and for all $\epsilon \in (0,3)$ we have

$$\mathbb{P}[Z \geq (1+\epsilon)k] \leq e^{-\epsilon^2 k/6}.$$

Finally, for all $\epsilon \in (0,3)$,

$$\mathbb{P}[(1-\epsilon)k \leq Z \leq (1+\epsilon)k] \geq 1 - 2e^{-\epsilon^2 k/6}.$$

Proof. Let us write $Z = \sum_{i=1}^{k} X_i^2$ where $X_i \sim N(0,1)$. To prove both bounds we use Chernoff's bounding method. For the first inequality, we first bound $\mathbb{E}[e^{-\lambda X_1^2}]$, where $\lambda > 0$ will be specified later. Since $e^{-a} \leq 1 - a + \frac{a^2}{2}$ for all $a \geq 0$ we have that

$$\mathbb{E}[e^{-\lambda X_1^2}] \leq 1 - \lambda \mathbb{E}[X_1^2] + \frac{\lambda^2}{2} \mathbb{E}[X_1^4].$$

Using the well known equalities, $\mathbb{E}[X_1^2] = 1$ and $\mathbb{E}[X_1^4] = 3$, and the fact that $1 - a \leq e^{-a}$ we obtain that

$$\mathbb{E}[e^{-\lambda X_1^2}] \leq 1 - \lambda + \tfrac{3}{2}\lambda^2 \leq e^{-\lambda + \frac{3}{2}\lambda^2}.$$

Now, applying Chernoff's bounding method we get that

$$\begin{aligned}
\mathbb{P}[-Z \geq -(1-\epsilon)k] &= \mathbb{P}\left[e^{-\lambda Z} \geq e^{-(1-\epsilon)k\lambda}\right] \\
&\leq e^{(1-\epsilon)k\lambda}\, \mathbb{E}\left[e^{-\lambda Z}\right] \\
&= e^{(1-\epsilon)k\lambda} \left(\mathbb{E}\left[e^{-\lambda X_1^2}\right]\right)^k \\
&\leq e^{(1-\epsilon)k\lambda}\, e^{-\lambda k + \frac{3}{2}\lambda^2 k} \\
&= e^{-\epsilon k\lambda + \frac{3}{2}k\lambda^2}.
\end{aligned}$$

Choose $\lambda = \epsilon/3$ we obtain the first inequality stated in the lemma.

For the second inequality, we use a known closed form expression for the moment generating function of a χ_k^2 distributed random variable:

$$\forall \lambda < \tfrac{1}{2}, \quad \mathbb{E}\left[e^{\lambda Z^2}\right] = (1-2\lambda)^{-k/2}. \tag{B.7}$$

On the basis of the equation and using Chernoff's bounding method we have

$$\begin{aligned}
\mathbb{P}[Z \geq (1+\epsilon)k)] &= \mathbb{P}\left[e^{\lambda Z} \geq e^{(1+\epsilon)k\lambda}\right] \\
&\leq e^{-(1+\epsilon)k\lambda}\, \mathbb{E}\left[e^{\lambda Z}\right] \\
&= e^{-(1+\epsilon)k\lambda}\, (1-2\lambda)^{-k/2} \\
&\leq e^{-(1+\epsilon)k\lambda}\, e^{k\lambda} = e^{-\epsilon k\lambda},
\end{aligned}$$

where the last inequality occurs because $(1-a) \leq e^{-a}$. Setting $\lambda = \epsilon/6$ (which is in $(0, 1/2)$ by our assumption) we obtain the second inequality stated in the lemma.

Finally, the last inequality follows from the first two inequalities and the union bound. □