1 A few useful probabilistic facts

Let X be a random variable (r.v.) taking real values: for instance, in $\mathbb{R}$ or $\mathbb{N}$. We assume X has an expectation and a variance.[1] A few useful things:

Fact 1.1. *If X takes values in $\mathbb{N} = \{0, 1, 2, \dots, \}$,*

$$\mathbb{E}[X] = \sum_{n=0}^{\infty} n \Pr[\, X = n \,] = \sum_{n=1}^{\infty} \Pr[\, X \geq n \,]$$

To remember whether the sum in the last expression starts at $n = 0$ or $n = 1$: either reprove it (a bit time-consuming), or take X to be the "useless" random variable equal to 0 with probability 1. Then $\mathbb{E}[X] = 0$, but $\sum_{n=0}^{\infty} \Pr[\, X \geq n \,] = \Pr[\, X \geq 0 \,] = 1$. So we shouldn't have the term $n = 0$.

Fact 1.2. *If X takes values in $\mathbb{N} = \{0, 1, 2, \dots, \}$,*

$$\mathrm{Var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$$

As a direct consequence, $\mathrm{Var}[X] \leq \mathbb{E}\left[X^2\right]$ (sometimes useful).

Lemma 1.3 (Jensen's Inequality)**.** *If $f : \mathbb{R} \to \mathbb{R}$ is convex (and $\mathbb{E}[f(X)]$ is well-defined)*

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)] \,.$$

For f concave, the inequality is reversed.

To remember the direction: check with $f(x) = x^2$ (convex). The variance is non-negative, so $0 \leq \mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$.

Fact 1.4 (Linearity of Expectation)**.** *For any X, Y and $a, b \in \mathbb{R}$,*

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$$

(We do not *need X, Y to be independent!)*

This extends to more random variables: for instance, $\mathbb{E}[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} \mathbb{E}[X_i]$. (No independence needed!)

Fact 1.5 (Variance)**.** *For any X and $a \in \mathbb{R}$,*

$$\mathrm{Var}[aX] = a^2 \,\mathrm{Var}[X] \,.$$

Moreover, if X, Y are independent,

$$\mathrm{Var}[aX + bY] = a^2 \,\mathrm{Var}[X] + b^2 \,\mathrm{Var}[Y] \,.$$

More generally, **if** $X_1, \dots, X_n$ are mutually independent (or, weaker condition, *pairwise independent*: any two X_i, X_j with $i \neq j$ are independent, but $X_1, \dots, X_n$ as a whole might not be mutually independent.), then

$$\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathrm{Var}[X_i]. \tag{1}$$

[1] This is not necessarily always true! Some random variables do not even have a well-defined expectation. For instance, the random variable defined on $\mathbb{Z}$ by $\Pr[\, X = k \,] = \frac{1}{C} \cdot \frac{1}{1+k^2}$ with $C = 1 + \pi \coth \pi$ (so that the probabilities sum to 1) is well-defined, but does not have an expectation since $\sum_{k \in \mathbb{Z}} k \cdot \Pr[\, X = k \,]$ is not defined (does not converge).

The proof is not too hard: basically, since $\mathrm{Var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$, consider $\mathbb{E}\left[\left(\sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right)^2\right]$ and expand the square, then use linearity of expectation:

$$\begin{aligned}\mathrm{Var}\left[\sum_{i=1}^n X_i\right] &= \mathbb{E}\left[\sum_{i=1}^n \sum_{j=1}^n (X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])\right] = \mathbb{E}\left[\sum_{i=1}^n (X_i - \mathbb{E}[X_i])^2\right] + \mathbb{E}\left[\sum_{i \neq j} (X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])\right] \\ &= \sum_{i=1}^n \mathbb{E}\left[(X_i - \mathbb{E}[X_i])^2\right] + \sum_{i \neq j} \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])]\end{aligned}$$

The first term is exactly $\sum_{i=1}^n \mathrm{Var}[X_i]$; the second, by pairwise independence, is 0, since $\mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])] = \mathbb{E}[(X_i - \mathbb{E}[X_i])]\mathbb{E}[(X_j - \mathbb{E}[X_j])] = 0 \cdot 0$.

Now, a few very trivial-looking (but useful!) facts. Suppose X takes values in $\{0, 1\}$, with $\Pr[\, X = 1\,] = p$ (this is a Bernoulli random variable). Then

- $X^2 = X$ (of course!), so $\mathbb{E}\left[X^2\right] = \mathbb{E}[X] = \Pr[\, X = 1\,] = p$
- That implies $\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = p - p^2 = p(1-p)$, which is at most $1/4$ (check it! $x(1-x) \leq 1/4$ for $x \in [0, 1]$, and the maximum is at $x = 1/2$).
- That implies that for a Binomial $X \sim \binom{n}{p}$, which is just the sum of n *independent, and identically distributed* (i.i.d.) Bernoullis with parameter p,

$$\mathbb{E}[X] = np, \qquad \mathrm{Var}[X] = np(1-p)\,.$$

Finally, an *indicator* random variable (for some "event" E) is just a Bernoulli random variable which is equal to 1 if the event occurs, and 0 otherwise (so, Bernoulli with parameter $\Pr(E)$). Usually denoted $\mathbb{1}_E$.

2 Concentration inequalities

We only mention here a few good bounds that we found to be useful, and sufficient in many or most settings. There are, of course, many others, and many refinements or variants of the bounds we present here. We refer the reader to, *e.g.*, [Ver18, Chapter 2] or [BLM13] for a much more comprehensive and insightful coverage.

We start with the mother of all concentration inequalities, Markov's inequality:

Theorem 2.1 (Markov's inequality). *Let X be a non-negative random variable with $\mathbb{E}[X] < \infty$. For any $t > 0$, we have*

$$\Pr[X \geq t] \leq \frac{\mathbb{E}[X]}{t}$$

Applying this to $(X - \mathbb{E}[X])^2$, we get

Theorem 2.2 (Chebyshev's inequality). *Let X be a random variable with $\mathbb{E}\left[X^2\right] < \infty$. For any $t > 0$, we have*

$$\Pr[|X - \mathbb{E}[X]| \geq t] \leq \frac{\mathrm{Var}[X]}{t^2}$$

By applying Markov's inequality to the moment-generating function (MGF) of $\sum_{i=1}^n X_i$ in various ways, one can also obtain the following statements:

Theorem 2.3 (Hoeffding bound). *Let $X_1, \ldots, X_n$ be independent random variables, where X_i takes values in $[a_i, b_i]$. For any $t \geq 0$, we have*

$$\Pr\left[\sum_{i=1}^n X_i > \sum_{i=1}^n \mathbb{E}[X_i] + t\right] \leq \exp(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}) \tag{2}$$

$$\Pr\left[\sum_{i=1}^n X_i < \sum_{i=1}^n \mathbb{E}[X_i] - t\right] \leq \exp(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}) \tag{3}$$

Corollary 2.4 (Hoeffding bound). *Let $X_1, \ldots, X_n$ be i.i.d. random variables taking value in $[0, 1]$, with mean μ. For any $\gamma \in (0, 1]$ we have*

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right| > \gamma\right] \leq 2\exp(-2\gamma^2 n) \tag{4}$$

Theorem 2.5 (Chernoff bound). *Let $X_1, \ldots, X_n$ be independent random variables taking value in $[0, 1]$, and let $P := \sum_{i=1}^n \mathbb{E}[X_i]$ For any $\gamma \in (0, 1]$ we have*

$$\Pr\left[\sum_{i=1}^n X_i > (1+\gamma)P\right] < \exp(-\gamma^2 P/3) \tag{5}$$

$$\Pr\left[\sum_{i=1}^n X_i < (1-\gamma)P\right] < \exp(-\gamma^2 P/2) \tag{6}$$

In particular, if $X_1, \ldots, X_n$ are i.i.d. with mean μ, then for any $\gamma \in (0, 1]$ we have

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right| > \gamma\mu\right] \leq 2\exp(-\gamma^2 n\mu/3) \tag{7}$$

As a rule of thumb, the "multiplicative" (Chernoff) from Theorem 2.5 is preferable to the "additive" bound (Hoeffding) from Corollary 2.4 whenever $\mu := P/n \ll 1$. In case one only has an upper or lower bound on the quantity $P = \sum_{i=1}^n \mathbb{E}[X_i]$, the following version of the Chernoff bound can come in handy:

Theorem 2.6 (Chernoff bound (upper and lower bound version)). *In the setting of Theorem 2.5, suppose that $P_L \le P \le P_H$. Then for any $\gamma \in (0,1]$, we have*

$$\Pr\left[\sum_{i=1}^{n} X_i > (1+\gamma)P_H\right] < \exp(-\gamma^2 P_H/3) \tag{8}$$

$$\Pr\left[\sum_{i=1}^{n} X_i < (1-\gamma)P_L\right] < \exp(-\gamma^2 P_L/2) \tag{9}$$

Theorem 2.7 (Bernstein's inequality). *Let $X_1, \dots, X_n$ be independent random variables taking values in $[-a, a]$, and such that $\mathbb{E}\left[X_i^2\right] \le v_i$ for all i. Then, for every $t \ge 0$, we have*

$$\Pr\left[\left|\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mathbb{E}[X_i]\right| \ge t\right] \le \exp\left(-\frac{t^2}{2(\sum_{i=1}^{n} v_i + \frac{a}{3}t)}\right).$$

In particular, if $X_1, \dots, X_n$ are i.i.d. with mean μ and $\mathbb{E}\left[X_1^2\right] \le v$, then for any $\gamma \ge 0$ we have

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \ge \gamma\right] \le \exp\left(-\frac{\gamma^2 n}{2(v + \frac{a}{3}\gamma)}\right).$$

Observe that this tail bound exhibits both behaviours: it decays in a subgaussian fashion for small γ, before switching to a subexponential tail bound for large γ.

We conclude this section by providing a very convenient bound, specifically for Poisson random variables, which shares the same "two-tail" behaviour:

Theorem 2.8 (Poisson concentration). *Let X be a $\mathrm{Poisson}(\lambda)$ random variable, where $\lambda > 0$. Then, for any $t > 0$, we have*

$$\Pr[\, X \ge \lambda + t\,] \le e^{-\frac{t^2}{2\lambda}\psi\left(\frac{t}{\lambda}\right)} \le e^{-\frac{t^2}{2(\lambda+t)}} \tag{10}$$

and, for any $0 < t < \lambda$,

$$\Pr[\, X \le \lambda - t\,] \le e^{-\frac{t^2}{2\lambda}\psi\left(-\frac{t}{\lambda}\right)} \le e^{-\frac{t^2}{2(\lambda+t)}}, \tag{11}$$

where $\psi(u) := 2\frac{(1+u)\ln(1+u)-u}{u^2}$ for $u \ge -1$. In particular, for any $t \ge 0$,

$$\Pr[\, |X - \lambda| \ge t\,] \le 2e^{-\frac{t^2}{2(\lambda+t)}}. \tag{12}$$

Is the proof needed?

References

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities*. A nonasymptotic theory of independence, With a foreword by Michel Ledoux. Oxford University Press, Oxford, 2013, pp. x+481. ISBN: 978-0-19-953525-5. DOI: 10.1093/acprof:oso/9780199535255.001.0001. URL: https://doi.org/10.1093/acprof:oso/9780199535255.001.0001 (cit. on p. 3).

[Ver18] Roman Vershynin. *High-dimensional probability*. Vol. 47. Cambridge Series in Statistical and Probabilistic Mathematics. An introduction with applications in data science, With a foreword by Sara van de Geer. Cambridge University Press, Cambridge, 2018, pp. xiv+284. ISBN: 978-1-108-41519-4. DOI: 10.1017/9781108231596. URL: https://doi.org/10.1017/9781108231596 (cit. on p. 3).