Concentration inequalities

Ben Barber

February 11, 2019

Assume all random variables have well-defined means, variances etc. when mentioned. Results may not be stated in greatest possible generality.

1 Basic inequalties

Theorem 1 (Markov)**.** *Let* $X \geq 0$*. For* $t > 0$, $\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}(X)}{t}$.

Proof. Observe that $t\mathbf{1}_{X\geq t} \leq X$, then take expectations of each side. □

Theorem 2 (Chebyshev)**.** $\mathbb{P}(|X - \mathbb{E}(X)| \geq t) \leq \frac{\mathrm{Var}(X)}{t^2}$.

Proof. Apply Markov to $(X - \mathbb{E}(X))^2$. □

'The second moment method' = 'we computed a variance and applied Chebyshev'. Computing variances can be difficult. Quadratic error probability can be problematic if you want many events to hold simultaneously.

Could look at higher moments (e.g. $\mathbb{E}(X)/t^4$ on rhs) if so inclined.

2 Exponential concentration

Theorem 3 (Chernoff)**.** *Let* $X = \sum_{i=1}^n X_i$ *with* $X_i = \pm 1$ *independently with probablity* $1/2$*. Then*

$$\mathbb{P}(X \geq t) \leq e^{-t^2/2n}.$$

Proof. Apply Markov to e^{hX} for some $h > 0$.

$$\begin{aligned}\mathbb{P}(X \geq t) &= \mathbb{P}(e^{hX} \geq e^{ht}) \\ &\leq \frac{\mathbb{E}(e^{hX})}{e^{ht}} = \frac{\mathbb{E}(\prod_{i=1}^n e^{hX_i})}{e^{ht}} = \frac{\prod_{i=1}^n \mathbb{E}(e^{hX_i})}{e^{ht}},\end{aligned}$$

since the X_i are independent. Now

$$\mathbb{E}(e^{hX_i}) = \frac{e^h + e^{-h}}{2} = \cosh h \leq e^{h^2/2},$$

by comparing Taylor series, so

$$\mathbb{P}(X \geq t) \leq e^{\frac{nh^2}{2} - ht} = e^{\frac{nh}{2}(h - \frac{2t}{n})} = e^{-t^2/2n}$$

for $h = t/n$. □

Theorem 4 (Hoeffding). *Let $X = \sum_{i=1}^n X_i$ be a sum of independent random variables with $X_i \in [-a_i, b_i]$ and $\mathbb{E}(X_i) = 0$. Then*

$$\mathbb{P}(X \geq t) \leq e^{-2t^2/\sum_{i=1}^n (a_i+b_i)^2}.$$

The means being 0 is just a techinical condition to make the proof easier to write. It can always be ensured by subtracting off the means of the random variables that you started with; you then get a bound on the probability of X exceeding its mean that depends on the lengths $c_i = a_i + b_i$ of the intervals that the X_i take values in.

Proof. As before, for $h > 0$,

$$\begin{aligned}\mathbb{P}(X \geq t) &= \mathbb{P}(e^{hX} \geq e^{ht})\\ &\leq \frac{\mathbb{E}(e^{hX})}{e^{ht}} = \frac{\mathbb{E}(\prod_{i=1}^n e^{hX_i})}{e^{ht}} = \frac{\prod_{i=1}^n \mathbb{E}(e^{hX_i})}{e^{ht}},\end{aligned}$$

since the X_i are independent. For every $x \in [-a, b]$ we have $x = \frac{b-x}{a+b}(-a) + \frac{a+x}{a+b}b$, so by convexity

$$e^{hx} \leq \frac{b-x}{a+b}e^{-ha} + \frac{a+x}{a+b}e^{hb}.$$

Hence

$$\mathbb{E}(e^{hX_i}) \leq \mathbb{E}\left(\frac{b_i - X_i}{a_i+b_i}e^{-ha_i} + \frac{a_i + X_i}{a_i + b_i}e^{hb_i}\right) = \frac{b_i e^{-ha_i} + a_i e^{hb_i}}{a_i + b_i} = e^{-ha_i}\frac{b_i + a_i e^{h(a_i+b_i)}}{a_i + b_i}.$$

Let

$$f(h) = \log(\mathbb{E}(e^{hX_i})) = -ha_i + \log(b_i + a_i e^{h(a_i+b_i)}) - \log(a_i + b_i).$$

Then

$$f'(h) = -a_i + \frac{a_i e^{h(a_i+b_i)}}{b_i + a_i e^{h(a_i+b_i)}}(a_i + b_i)$$

and

$$f''(h) = \frac{a_i b_i e^{h(a_i+b_i)}}{(b_i + a_i e^{h(a_i+b_i)})^2}(a_i + b_i)^2.$$

Now $f(0) = f'(0) = 0$ and $f''(h) \leq (a_i + b_i)^2/4$ by the arithmetic mean-geometric mean inequality, so, by Taylor's theorem, $f(h) \leq h^2(a_i + b_i)^2/8$ and

$$\mathbb{E}(e^{hX_i}) \leq e^{h^2(a_i+b_i)^2/8}.$$

Thus

$$\mathbb{P}(X \geq t) \leq e^{\frac{h^2}{8}\sum_{i=1}^n (a_i+b_i)^2 - ht} = e^{\frac{h}{8}\sum_{i=1}^n (a_i+b_i)^2 (h - \frac{8t}{\sum_{i=1}^n (a_i+b_i)^2})} = e^{-2t^2/\sum_{i=1}^n (a_i+b_i)^2},$$

for $h = \frac{4t}{\sum_{i=1}^n (a_i+b_i)^2}$. □

3 Martingale concentration

A σ-algebra is an object that tells you what events are detectable/measurable/have probabilities. There's some general definition, but over finite probability spaces they're very easy to describe. You have a partition of the probability space into smallest measurable events, and then every set you can make by taking a union of these *atoms* is also a measurable event. So finite σ-algebras essentially look like power sets, except that the smallest events might not be singletons. For example, suppose that the underlying probability space is $\Omega = \{0,1\}^n$, modelling n flips of a coin. The following are all examples of σ-algebras.

- The power set $\mathcal{F} = \mathcal{P}(\Omega)$, which lets you ask about any conceivable property of the sequence of coin flips.
- The σ-algebra $\mathcal{G}$ whose atomic events are the sets A_i for $0 \leq i \leq n$, where A_i is the set of sequences of flips with exactly n heads. $\mathcal{G}$ knows about the number of heads, but not the result of any particular coin flip.
- The σ-algebra $\mathcal{F}_k$ whose atomic events have the form $\{x_1\} \times \cdots \times \{x_k\} \times \{0,1\}^{n-k}$. $\mathcal{F}_k$ knows about the first k flips, but nothing about what happens later.

The σ-algebra $\mathcal{A}$ *refines* the σ-algebra B if its atomic events are obtained by splitting the atoms of $\mathcal{B}$ into smaller pieces. That is, we look inside the atoms to distinguish outcomes that we couldn't distinguish before. The finest possible σ-algebra is $\mathcal{P}(\Omega)$; the coarsest is $\{\emptyset, \Omega\}$. A *filtration* is a sequence $(\mathcal{A}_i)$ of σ-algebras such that $\mathcal{A}_{i+1}$ is a refinement of $\mathcal{A}_i$ for each i. Thus $(\mathcal{F}_k)_{k=0}^n$ is a filtration where at each stage we learn about the outcome of the next coin flip.

An *$\mathcal{A}$-measurable random variable* is a function which is constant on the atoms of $\mathcal{A}$. For example, a random variable is $\mathcal{F}_k$-measurable if it only depends on the outcome of the first k flips. For an $\mathcal{F}_n$-measurable random variable Y, the *conditional expectation of Y with respect to $\mathcal{F}_k$* is the $\mathcal{F}_k$-measurable random variable $\mathbb{E}(Y|\mathcal{F}_k)$ whose value on the atom A is $\mathbb{E}(Y\mathbf{1}_A)/\mathbb{P}(A)$; informally, the average value of Y conditioned on A having occurred.[1]

The sequence of random variables $Y_k = \mathbb{E}(Y|\mathcal{F}_k)$ is a *martingale* with respect to the filtration $(\mathcal{F}_k)_{k=0}^n$. The key property of a martingale is that

$$\mathbb{E}(Y_{k+1}|\mathcal{F}_k) = \mathbb{E}(\mathbb{E}(Y|\mathcal{F}_{k+1})|\mathcal{F}_k) = \mathbb{E}(Y|\mathcal{F}_k) = Y_k,$$

where the middle inequality is the grandly named Law of Total Expectation: the average of the average values of Y over the atoms of $\mathcal{F}_{k+1}$ contained in a given atom of $\mathcal{F}_k$ is the just average value of Y over that atom.

For a martingale $(Y_k)_{k=0}^n$, its *difference sequence* is defined by $X_0 = \mathbb{E}(Y_n)$ and $X_i = Y_i - Y_{i-1}$. Note that $\mathbb{E}(X_i|\mathcal{F}_{i-1}) = 0$.

Theorem 5 (Hoeffding–Azuma)**.** *Let $(Y_k)_{k=0}^n$ be a martingale with difference sequence $(X_k)_{k=0}^n$. Suppose that on each atom of $\mathcal{F}_{i-1}$, X_i takes values only in some interval of length at most c_i. Then*

$$\mathbb{P}(Y_n \geq \mathbb{E}(Y_n) + t) \leq e^{-2t^2/\sum_{i=1}^n c_i^2}.$$

[1] What we've actually done is define this notion. Note that if $\mathbb{P}(A) = 0$ we get nonsense but don't care, as it doesn't matter if random variables are only defined with probability 1 rather than everywhere.

Proof. Let $X = X_1 + \cdots + X_n = Y_n - \mathbb{E}(Y_n)$. Following the same strategy as before, we have

$$\begin{aligned}\mathbb{P}(Y_n \geq \mathbb{E}(Y_n) + t) &= \mathbb{P}(X \geq t)\\ &= \mathbb{P}(e^{hX} \geq e^{ht})\\ &\leq \frac{\mathbb{E}(e^{hX})}{e^{ht}} = \frac{\mathbb{E}(\prod_{i=1}^n e^{hX_i})}{e^{ht}}.\end{aligned}$$

Now

$$\mathbb{E}\left(\prod_{i=1}^n e^{hX_i}\right) = \mathbb{E}\left(\mathbb{E}\left(\prod_{i=1}^n e^{hX_i}|\mathcal{F}_{n-1}\right)\right) = \mathbb{E}\left(\prod_{i=1}^{n-1} e^{hX_i}\mathbb{E}\left(e^{hX_n}|\mathcal{F}_{n-1}\right)\right),$$

because $X_1, \ldots, X_{n-1}$ are each constant on atoms of $\mathcal{F}_{n-1}$.

On each atom of $\mathcal{F}_{n-1}$, $\mathbb{E}(e^{hX_n}|\mathcal{F}_{n-1})$ is an expression of the type we were previously able to bound by $e^{h^2c_i^2/8}$, so by induction we have

$$\mathbb{E}\left(\prod_{i=1}^n e^{hX_i}\right) \leq e^{h^2\sum_{i=1}^n c_i^2/8},$$

and we can complete the proof as before. □

Corollary 6. *Let $\Omega = \Omega_1 \times \cdots \times \Omega_n$ be a product space, and let Y be a random variable on Ω such that*

$$|Y(\omega) - Y(\omega')| \leq c_i$$

whenever ω and ω' differ only on the ith coordinate. Then

$$\mathbb{P}(Y_n \geq \mathbb{E}(Y_n) + t) \leq e^{-2t^2/\sum_{i=1}^n c_i^2}.$$

Proof. Let $(Y_k)_{k=0}^n$ be the martingale corresponding to the filtration $(\mathcal{F}_k)_{k=0}^n$ which looks at coordinates of the random sample ω one at a time. We claim that Theorem 5 applies. To see this, observe that

$$Y_i(\omega) = \sum_{\eta_{i+1}\in\Omega_{i+1}} \cdots \sum_{\eta_n\in\Omega_n} Y(\omega_1, \ldots, \omega_i, \eta_{i+1}, \eta_{i+2} \ldots, \eta_n)\mathbb{P}_{i+1}(\eta_{i+1}) \cdots \mathbb{P}_n(\eta_n),$$

where $\mathbb{P}_j(\eta_j)$ is the probability that a random sample from Ω_j is η_j. Similarly,

$$\begin{aligned}Y_{i+1}(\omega) &= \sum_{\eta_{i+2}\in\Omega_{i+1}} \cdots \sum_{\eta_n\in\Omega_n} Y(\omega_1, \ldots, \omega_i, \omega_{i+1}, \eta_{i+2} \ldots, \eta_n)\mathbb{P}_{i+2}(\eta_{i+2}) \cdots \mathbb{P}_n(\eta_n)\\ &= \sum_{\eta_{i+1}\in\Omega_{i+1}} \cdots \sum_{\eta_n\in\Omega_n} Y(\omega_1, \ldots, \omega_i, \omega_{i+1}, \eta_{i+2} \ldots, \eta_n)\mathbb{P}_{i+1}(\eta_{i+1}) \cdots \mathbb{P}_n(\eta_n).\end{aligned}$$

Hence on the atom of $\mathcal{F}_i$ picked out by $\omega_1, \ldots, \omega_i$,

$$\begin{aligned}|Y_i(\omega) - Y_{i+1}(\omega)| \leq &\sum_{\eta_{i+1}\in\Omega_{i+1}} \cdots \sum_{\eta_n\in\Omega_n} \mathbb{P}_{i+2}(\eta_{i+1}) \cdots \mathbb{P}_n(\eta_n)\big|Y(\omega_1, \ldots, \omega_i, \eta_{i+1}, \eta_{i+2} \ldots, \eta_n)\\ &\qquad\qquad - Y(\omega_1, \ldots, \omega_i, \omega_{i+1}, \eta_{i+2} \ldots, \eta_n)\big|\\ \leq &\sum_{\eta_{i+1}\in\Omega_{i+1}} \cdots \sum_{\eta_n\in\Omega_n} \mathbb{P}_{i+2}(\eta_{i+1}) \cdots \mathbb{P}_n(\eta_n)c_i\\ = &\; c_i.\end{aligned}$$

□

A similar argument can be made for functions of random permutations, rather than product spaces.

4 What can go wrong

Martingale concentration is a very powerful and flexible tool, but it does have some downsides.

(i) Concentration will typically only be in an interval of length $\sqrt{n}$, even if the expectation of the random variable is much lower than that.

(ii) It looks at worst case changes.

Problem (i) can be addressed by McDiarmid's inequality.

Theorem 7 (McDiarmid)**.** *Let $X = \sum_{i=1}^{n} X_i$ be a sum of independent random variables with $X_i \leq c$. Then*

$$\mathbb{P}(X \geq \mathbb{E}(X) + t) \leq e^{-\frac{t^2}{2\operatorname{Var}(X)\left(1+\frac{ct}{3\operatorname{Var}(X)}\right)}}.$$

The proof has the same basic shape as that of Hoeffding's inequality. The variance term arises from taking a Taylor expansion of e^x to second order.

In typical applications t is of the order $\sqrt{\operatorname{Var}(X)}$ and $\operatorname{Var}(X)$ is tending to infinity. In this case the error probability looks like that for the Gaussian of the same variance, so cannot be improved. For large t the error probability is 'only' exponentially small, which is again the correct behaviour as can be seen by looking at binomial random variables.

There is a standard trick for addressing problem (ii), exemplified by Freedman's inequality.

Theorem 8 (Freedman)**.** *Let $(Y_k)_{k=0}^{n}$ be a martingale with difference sequence $(X_k)_{k=0}^{n}$ bounded above by c. Let $W = \sum_{k=1}^{n} \mathbb{E}(X_k^2 | \mathcal{F}_{k-1})$.*

Then

$$\mathbb{P}(Y \geq \mathbb{E}(Y) + t \text{ and } W \leq \sigma^2) \leq e^{-\frac{t^2}{2\sigma^2\left(1+\frac{ct}{3\sigma^2}\right)}}.$$

If the *predictable quadratic variation* W is unconditionally bounded, then Freedman's inequality is to McDiarmid's inequality as Hoeffding–Azuma is to Hoeffding. If there are possible but unlikely runs of the experiment where W is large, then we can introduce a new martingale $(Z_k)_{k=0}^{n}$ which, for each ω, agrees with $(Y_k)_{k=0}^{n}$ 'until W is about to become too large', after which point $j = j(\omega)$ we take $Z_j(\omega) = Z_{j+1}(\omega) = \cdots = Z_n(\omega)$. This is still a martingale and makes $W \leq \sigma^2$ by fiat at the cost of weakening applications to

$$\begin{aligned}
\mathbb{P}(Y \geq \mathbb{E}(Y) + t) &\leq \mathbb{P}(Z \geq \mathbb{E}(Z) + t \text{ or } Y \neq Z) \\
&= \mathbb{P}(Z \geq \mathbb{E}(Z) + t \text{ and } Y = Z) + \mathbb{P}(Y \neq Z) \\
&= \mathbb{P}(Y \geq \mathbb{E}(Y) + t \text{ and } W \leq \sigma^2) + \mathbb{P}(W \geq \sigma^2) \\
&\leq e^{-\frac{t^2}{2\sigma^2\left(1+\frac{ct}{3\sigma^2}\right)}} + \mathbb{P}(W \geq \sigma^2).
\end{aligned}$$

5 Talagrand

An alternative approach to problem (i) is offered by Talagrand's inequality, of which the following is a very special case.

Theorem 9 (Talagrand)**.** *Let $\Omega = \Omega_1 \times \cdots \times \Omega_n$ be a product space, and let Y be a random variable on Ω such that*

- *$|Y(\omega) - Y(\omega')| \leq c$ whenever ω and ω' differ only on the ith coordinate (that is, Y does not depend too much on any one coordinate);*
- *if $Y(\omega) \geq s$ then there is a set J of s coordinates such that $Y(\omega') \geq s$ whenever $\omega_j = \omega'_j$ for $j \in J$ (that is, if Y is large then we can explain that fact cheaply).*

Then, for every x, $\mathbb{P}(Y \leq x - t)\mathbb{P}(Y \geq x) \leq e^{-t^2/xc^2}$.

This inequality is almost always applied with $x = m$ or $x = m + t$ where m is a median of Y. Then one of the probabilities on the left-hand side is $1/2$, so

$$\begin{aligned} \mathbb{P}(Y \leq m - t) &\leq 2e^{-t^2/mc^2}, \\ \mathbb{P}(Y \geq m + t) &\leq 2e^{-t^2/(m+t)c^2}. \end{aligned}$$

The asymmetry here can be inconvenient, as can concentration near the median rather than the mean. However, it is frequently possible to use the concentration guaranteed by Talagrand and some crude bounds on Y to show that its mean and median cannot be too far apart.