1 Chernoff Bound

Let X be a real-valued random variable with distribution D:

$$\Pr[X \in S] = D(S), S \subseteq \mathbb{R}$$

Definition 1 *The* moment-generating function*, or characteristic function for X (or, more precisely but less commonly, for D) is defined for $t \in \mathbb{R}$ by*

$$g_D(t) = E\left[e^{tx}\right]$$

Note that, for $t \in \mathbb{C}$, this gives the characteristic function for t pure-real, and the Fourier transform for t pure-imaginary. For any D, $g_D(0) = E[1] = 1$.

Assume $E[X] = \theta$. We would like to find a large deviation bound. That is, if we sample $x_1, \ldots, x_n$ from D and take $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, we would like to know how the distribution of $\overline{X}$ is concentrated around θ. Last time we bounded the tails, in the form $\Pr\left[\left|\overline{X} - \theta\right| > c\right] \leq f(c)$, with a polynomial function, f, that dropped off as $\frac{1}{c^2}$. This polynomial bound is good in general for small c. However, further out on the tail we can get an exponential tail drop-off if D is tame enough (in particular, does not have a "heavy" tail). Without loss of generality, take $\theta = 0$.

Theorem 2 (Chernoff) *If the integral defining $g_D(t)$ converges unconditionally in a neighborhood of 0, and $g_D(t)$ is differentiable at 0, then*

$$\forall \epsilon > 0\ \exists c_\epsilon < 1 : \Pr\left[\overline{X} > \epsilon\right] < c_\epsilon^n$$

The idea is that the quality of the large deviation bound depends on how heavy the tails of D are, and that this is measured by the smoothness of g_D at the origin; a moment-generating function that is differentiable at the origin guarantees exponential tails.

Proof :

$$\begin{aligned}
\Pr\left[\overline{X} > \epsilon\right] &= \Pr\left[e^{\beta n \overline{X}} > e^{\beta n \epsilon}\right] && \text{for any } \beta > 0 \\
&< \frac{E\left[e^{\beta n \overline{x}}\right]}{e^{\beta n \epsilon}} && \text{Markoff bound} \\
&= e^{-\beta n \epsilon} E\left[e^{\beta \sum_i x_i}\right] \\
&= e^{-\beta n \epsilon}\left(E\left[e^{\beta X}\right]\right)^n && x_i \text{ are independent} \\
&= \left(e^{-\beta \epsilon} E\left[e^{\beta X}\right]\right)^n \\
&= \left(e^{-\beta \epsilon} g_D(\beta)\right)^n
\end{aligned}$$

We now need to show that there is a $\beta > 0$ such that $e^{-\beta\epsilon} g_D(\beta) < 1$. At $\beta = 0$, $e^0 g_D(0) = 1$, so let's find the derivative of $e^{-\beta\epsilon} g_D(\beta)$ at 0. Since g_D is differentiable at 0 we have:

$$\begin{aligned}
\left.\frac{\partial g_D(\beta)}{\partial \beta}\right|_0 &= \left.\frac{\partial E\left[e^{\beta X}\right]}{\partial \beta}\right|_0 \\
&= E\left[\frac{\partial e^{\beta X}}{\partial \beta}\right]\Bigg|_0 && \text{can switch order of derivative and integral by the unconditional convergence of } g_D \text{ around } 0 \\
&= E\left[X e^{\beta X}\right]\Big|_0 \\
&= E[X] = \theta = 0
\end{aligned}$$

So, the moment-generating function is flat at 0. Now we can differentiate the whole function:

$$\begin{aligned}
\left.\frac{\partial e^{-\beta\epsilon} g_D(\beta)}{\partial \beta}\right|_0 &= \left.\frac{\partial e^{-\beta\epsilon} g_D(\beta)}{\partial \beta}\right|_0 \\
&= e^{-\epsilon\beta} g'_D(\beta) - \epsilon e^{-\epsilon\beta} g_D(\beta)\big|_0 && \text{product rule} \\
&= e^{-\epsilon 0} \underbrace{g'_D(0)}_{0} - \epsilon e^{-\epsilon 0} \underbrace{g_D(0)}_{1} && \text{at } \beta = 0 \\
&= -\epsilon
\end{aligned}$$

We have determined that $\exists \beta > 0 : e^{-\beta\epsilon} g_D(\beta) < 1$, and thus there is a $c_\epsilon < 1$ as stated in the theorem. $\square$

This method also allows us, in some cases, to find the value of c_ϵ which gives the tightest Chernoff bound. (Of course in for general D and ϵ this can be a complicated task and we often settle for bounds on the best c_ϵ.)

Example 3 *Symmetric Random Walk*

Take D to be the probability with $\Pr[X = 1] = \Pr[X = -1] = \frac{1}{2}$. The moment-generating function is:

$$g_D(t) = \frac{1}{2}(e^t + e^{-t}) = \cosh t$$

Finding the optimal c_ϵ:

$$\begin{aligned}
c_\epsilon &= \inf_\beta c^{-\epsilon\beta} \cosh \beta \\
&= \cdots \text{insert calculus here} \cdots \\
&= (1-\epsilon)^{\frac{\epsilon-1}{2}} (1+\epsilon)^{-\frac{1+\epsilon}{2}} && \text{using } \beta = \frac{1}{2} \log \frac{1+\epsilon}{1-\epsilon}
\end{aligned}$$

Define:

$$\begin{aligned}
k_\epsilon &= -\log c_\epsilon \\
&= \frac{1-\epsilon}{2} \log(1-\epsilon) + \frac{(1+\epsilon)}{2} \log(1+\epsilon)
\end{aligned}$$

By the Chernoff bound we have:

$$\Pr[X > \epsilon] \leq e^{k_\epsilon n}$$

Consider two distributions: p, with probabilities $\left\{\frac{1}{2}, \frac{1}{2}\right\}$, the symmetric random walk from above, like a fair coin, and q, with probabilities $\left\{\frac{1-\epsilon}{2}, \frac{1+\epsilon}{2}\right\}$, like a biased coin. Let's rewrite k_ϵ:

$$\begin{aligned}
k_\epsilon &= \frac{1-\epsilon}{2} \log \frac{\frac{1-\epsilon}{2}}{\frac{1}{2}} + \frac{1+\epsilon}{2} \log \frac{\frac{1+\epsilon}{2}}{\frac{1}{2}} \\
&= \sum_x p(x) \log \frac{p(x)}{q(x)} && \text{defined as } D(p||q)
\end{aligned}$$

This value is the *Kullback-Leibler divergence* of p from q, also known as the information divergence or the relative entropy of p with respect to q. $D(p||q)$ is not a metric (it isn't symmetric and doesn't satisfy the triangle inequality). For example, if have a fair coing but we sample 90 heads out of 100 throws, $D(\{0.9, 0.1\}||\{0.5, 0.5\})$ quantifies how unlikely this event is. It isn't symmetric since, of course, the probability of getting 100 heads with a fair coin is not the same as the probability of getting 50 heads with a coin that has probability 1 of coming up heads. D is useful throughout information theory and statistics (and is closely related to the "Fisher information"); it's role in the Chernoff bound is one of the reasons for it's importance. For more information see the text by Cover and Thomas.

2 #DNF (Continued)

Recall, from last time, that we have an algorithm for estimating #DNF which runs in time poly$(n, \frac{1}{\epsilon}, \frac{1}{\delta})$ and that produces an unbiased estimator T of θ satisfying:

$$\Pr\left[(1-\epsilon)\theta \leq T \leq (1+\epsilon)\theta\right] \geq 1-\delta$$

Definition 4 *Algorithm A is a* FPRAS *(fully polynomial randomized approximation scheme) for quantity θ if:*

- *A is randomized,*
- *A runs in time poly$(n, \frac{1}{\epsilon})$, and*
- $\Pr\left[(1-\epsilon)\theta \leq T \leq (1+\epsilon)\theta\right] \geq \frac{2}{3}$.

Lemma 5 *Having a FPRAS implies that in time poly$(n, \frac{1}{\epsilon}, \log\frac{1}{\delta})$ we can produce T satisfying:*

$$\Pr\left[(1-\epsilon)\theta \leq T \leq (1+\epsilon)\theta\right] \geq 1-\delta$$

In our algorithm from last time, we started with an algorithm to approximate #DNF, and amplified it using the Chebyshev inequality to shrink the variance below ϵ, and then continued to shrink it below $\epsilon\delta$. The above lemma shows us that there is a way of avoiding going as far in the variance-reduction as we did last time, since we only need $\frac{3}{4}$ of the probability mass inside the $\theta(1\pm\epsilon)$ range to apply the lemma.

Proof : By assumption, we have a random variable X which we can produce in time poly$(n, \frac{1}{\epsilon})$ with $\frac{2}{3}$ of the probability mass inside the range $\theta(1\pm\epsilon)$. Collect $m = (\log\frac{1}{\delta})/D\left(\{\frac{1}{2}, \frac{1}{2}\}, \{\frac{2}{3}, \frac{1}{3}\}\right)$ samples $x_1, \ldots, x_m$, from this distribution. (Here, $D\left(\{\frac{1}{2}, \frac{1}{2}\}, \{\frac{2}{3}, \frac{1}{3}\}\right)$ is the divergence corresponding to an empirical "fair" distribution given a coin with probability 2/3 of coming up heads.) Select the median of $x_1, \ldots, x_m$ as the output. By assumption, $\mathrm{Var}(x_i) \leq \frac{\theta^2\epsilon^2}{3}$. Therefore, by the Chebyshev inequality, we have $\Pr\left[|x_i - \theta| > \theta\epsilon\right] < \frac{1}{3}$. Therefore, with probability $\frac{2}{3}$, each sample is in the $\theta(1\pm\epsilon)$ range, so:

$$\Pr\left[|\mathrm{median}(|x_i|) - \theta| > \theta\epsilon\right] \leq e^{D\left(\{\frac{1}{2}, \frac{1}{2}\}, \{\frac{2}{3}, \frac{1}{3}\}\right)m} = \delta$$

$\square$

Now our overall algorithm consists of m applications of a variance-reduction step, which averages the samples, and one median calculation on the m averages.

Next time we will discuss Karger's min-cut algorithm (as in CS 138), and put this together with the #DNF approximation algorithm, to solve the network reliability problem.