Markov, Chebychev and Hoeffding Inequalities

Robert L. Wolpert
Department of Statistical Science
Duke University, Durham, NC, USA

For each constant $c > 0$, any non-negative integrable random variable Y satisfies the inequalities

$$\begin{aligned} 0 \le Y &\le c\,\mathbf{1}_{\{Y \ge c\}} \quad \text{for every } \omega \in \Omega, \text{ so} \\ \mathsf{E}Y &\le \mathsf{E}\{c\,\mathbf{1}_{\{Y \ge c\}}\} = c\,\mathsf{P}[Y \ge c] \\ \mathsf{P}[Y \ge c] &\le \mathsf{E}Y/c, \end{aligned} \tag{1}$$

a result known as **Markov's Inequality**. A special case of this arises for $Y = |X - \mu|^2$, for any L_2 random variable X with mean μ and variance σ^2— by Markov's inequality,

$$\begin{aligned} \mathsf{P}\big[|X - \mu| \ge c\big] &= \mathsf{P}\big[|X - \mu|^2 \ge c^2\big] \\ &\le \mathsf{E}|X - \mu|^2/c^2 \\ &= \sigma^2/c^2, \end{aligned} \tag{2}$$

the well-known **Chebychev Inequality**. The one-sided bound

$$\mathsf{P}[X - \mu \ge c] \le \sigma^2/c^2 \tag{3}$$

follows immediately, but one can do better— for any t,

$$\begin{aligned} \mathsf{P}[X - \mu \ge c] &= \mathsf{P}[(X - \mu + t) \ge (c + t)] \\ &\le \frac{\sigma^2 + t^2}{(c + t)^2}, \end{aligned}$$

and the optimal $t = \sigma^2/c$ (found by setting the derivative of the logarithm to zero) isn't quite 0. With the optimal t,

$$\mathsf{P}[X - \mu \ge c] \le \frac{\sigma^2}{\sigma^2 + c^2}, \tag{4}$$

a slight improvement on Equation (3).

The technique of applying Markov's inequality with a free parameter (here t) and choosing it optimally can be very powerful; one of the best applications of this idea leads to what are variously called the Chernoff Bounds or (UNC's own Wassily) Hoeffding's Inequality. Here's a sketch of how they work.

First, an aside on Kullback-Leibler Divergence. Let F and G be two probability distributions on the same set $\mathcal{X}$ with $F \ll G$; then the "Kullback-Leibler divergence from F to G" is defined by

$$\mathcal{K}[F : G] := \mathsf{E}_F\left[\log \frac{F(dx)}{G(dx)}\right] = \int_{\mathcal{X}} \log \frac{F(dx)}{G(dx)}\, F(dx),$$

where as usual $F(dx)/G(dx)$ denotes the Radon-Nikodym derivative. The K-L divergence may always be computed using densities with respect to any dominating measure $m(dx)$ as

$$= \int_{\mathcal{X}} \log \frac{f(x)}{g(x)}\, f(x)\, m(dx),$$

a quantity that depends only on the distributions and not on the choice of dominating measure $m(dx)$. The quantity $\mathcal{K}[F : G]$ is always non-negative, vanishes only if F and G coincide, and is easy to compute, so it is often used as a measure of discrepancy between F and G, even though it is not a true "distance" meaasure because it is not symmetric in F and G, and does not satisfy the triangle inequality. The topology (i.e. measure of "closeness" and hence definition of convergence) induced by $\mathcal{K}[F_{\theta_0} : F_{\theta_1}]$ on the parameter space Θ for a distributional family $\{F_\theta\}$ is the same as the "information metric" topology generated by the Riemannian distance metric[1] $d_I(\cdot, \cdot)$ based on the Fisher Information I; one can show that

$$\mathcal{K}[F_{\theta_0} : F_{\theta_1}] \asymp \frac{1}{2} d_I(\theta_0, \theta_1)^2$$

as $d_I(\theta_0, \theta_1) \to 0$. The K-L divergence from one Bernoulli distribution to another is

$$\begin{aligned} K(p : q) &:= \mathcal{K}[\mathsf{Bi}(1, p) : \mathsf{Bi}(1, q)] \\ &= p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q} \\ &= p \log \frac{p}{q} + \bar{p} \log \frac{\bar{p}}{\bar{q}} \end{aligned}$$

where we denote $\bar{p} \equiv 1 - p$ and $\bar{q} \equiv 1 - q$. This vanishes for $q = p$ and, for $0 < q < p < 1$, is

$$\begin{aligned} K(p : q) &= \int_q^p \left\{ \frac{p}{x} - \frac{1 - p}{1 - x} \right\} dx = \int_q^p \left\{ \frac{p - x}{x(1 - x)} \right\} dx \\ &\geq 4 \int_q^p (p - x)\, dx = 2(p - q)^2, \end{aligned}$$

since $x(1 - x) \leq 1/4$ for all $x \in \mathbb{R}$.

[1] In $d = 1$ dimension, this is $d_I(\theta_0, \theta_1) := \left| \int_{\theta_0}^{\theta_1} \sqrt{I(\theta)}\, d\theta \right|$, where $I(\theta)$ denotes the Fisher information; in higher dimensions it is $d_I(\theta_0, \theta_1) := \inf_\gamma \int_0^1 \sqrt{\dot{\gamma}'(s)\, I(\theta)\, \dot{\gamma}(s)}\, ds$, where the infimum is over all smooth paths $\gamma : [0, 1] \to \Theta$ with "velocity" $\dot{\gamma}(s) := d\gamma/ds$ connecting $\gamma(0) = \theta_0$ to $\gamma(1) = \theta_1$.

Theorem 1 (Hoeffding)**.** *Let $\{X_i\}$ be i.i.d. random variables taking values in the unit interval $[0,1]$, with mean μ. Then for all $c > 0$ with $p \equiv \mu + c \le 1$,*

$$P[\bar{X}_n - \mu \ge c] \le e^{-nK(p:\mu)} \le e^{-2\,n\,c^2}.$$

Proof. Set $S_n := \sum_{i \le n} X_i$ and let $t > 0$. Then by Markov's inequality applied to $Y = e^{tS_n}$,

$$\begin{aligned} \mathsf{P}[\bar{X}_n - \mu \ge c] &\le \mathsf{E}e^{tS_n} e^{-nt(\mu+c)} \\ &= [\mathsf{E}e^{tX_1}]^n e^{-nt\,p} \\ &\le [\bar{\mu} + \mu e^t]^n e^{-nt\,p} \quad \text{since } e^{t\,X} \le (1-X) + X\,e^t \text{ by convexity} \\ &= [\bar{\mu} e^{-t\,p} + \mu e^{t\,\bar{p}}]^n \end{aligned}$$

with $\bar{\mu} := 1 - \mu$. Using logarithms and derivatives again, one discovers that the minimum over all $t \ge 0$ is attained where $e^t = \frac{p\,\bar{\mu}}{\bar{p}\,\mu}$, so $[\bar{\mu} + \mu e^t] = \bar{\mu}/\bar{p}$ and:

$$\mathsf{P}[\bar{X}_n - \mu \ge c] \le \left[\frac{\mu^p\ \bar{\mu}^{\bar{p}}}{p^p\ \bar{p}^{\bar{p}}}\right]^n = e^{-np\log(p/\mu) - n\bar{p}\log(\bar{p}/\bar{\mu})} = e^{-nK(p:\mu)}.$$

∎

Applying the same result to $(1 - X_j)$ and summing gives the two-sided bound

$$\mathsf{P}\big[|\bar{X}_n - \mu| \ge c\big] \le 2e^{-nK(p:\mu)} \le 2e^{-2\,n\,c^2}. \tag{5}$$

This is much stronger than (for example) Chebychev's inequality

$$\mathsf{P}\big[|\bar{X}_n - \mu| \ge c\big] \le \frac{\sigma^2}{n\,c^2}$$

because it shrinks geometrically in n; by the Borel-Cantelli Lemma it leads immediately to a version of the strong law of large numbers (SLLN)

$$\mathsf{P}[\bar{X}_n \to \mu] = 1,$$

for example, while Chebychev can only get the weak LLN $\mathsf{P}\big[|\bar{X}_n - \mu| > \epsilon\big] \to 0$. Of course, Hoeffding's inequality requires $\{X_j\}$ to be bounded, while Chebychev doesn't, but the bounds don't have to be zero and one. If, say, $a \le Y_i \le b$ with probability one, then apply Hoeffding's inequality to $X_i := (Y_i - a)/(b - a)$ to find:

$$\mathsf{P}[\bar{Y}_n - \mu \ge c] = \mathsf{P}\left[\bar{X}_n - \frac{\mu - a}{b - a} \ge \frac{c}{b - a}\right] \le e^{-2\,n\,c^2/(b-a)^2} \tag{6}$$

(with a similar two-sided version). The $\{Y_i\}$ don't even have to be identically distributed— independence and boundedness are enough. If they satisfy $a_i \le Y_i \le b_i$ for each i, then

$$\mathsf{P}[\bar{Y}_n - \mu_n \ge c] \le e^{-2\,n^2\,c^2/\sum_{i \le n}(b_i - a_i)^2} \tag{7}$$

where $\mu_n = \mathsf{E}\hat{Y}_n$ is the average mean $\frac{1}{n}\sum_{i \le n} \mathsf{E}[Y_i]$.

Hoeffding proved this improvement on Chebychev's inequality (at UNC) in 1963. It is closely related to the earlier **Azuma**'s inequality (1967), **Chernoff** bounds (1952), and **Bernstein**'s inequality (1937). In modern probability theory the distribution measure for Y_n is said to be *concentrated* near μ, making this one of the first of the now popular "concentration inequalities."