Chernoff-Hoeffding Inequality

When dealing with modern big data sets, a very common theme is reducing the set through a random process. These generally work by making "many simple estimates" of the full data set, and then judging them as a whole. Perhaps magically, these "many simple estimates" can provide a very accurate and small representation of the large data set. The key tool in showing how many of these simple estimates are needed for a fixed accuracy trade-off is the *Chernoff-Hoeffding* inequality [2, 5]. This document provides a simple form of this bound, and two examples of its use.

2.6 Chernoff-Hoeffding Inequality

We consider a two specific form of the Chernoff-Hoeffding bound. It is not the strongest form of the bound, but is for many applications asymptotically equivalent, and it also fairly straight-forward to use. It is more similar to the form of Azuma's inequality which deals with Martingales that have more complicated dependence structure.

Theorem 2.6.1. *Consider a set of* r independent *random variables* $\{X_1, \ldots, X_r\}$. *If we know* $a_i \leq X_i \leq b_i$, *then let* $\Delta_i = b_i - a_i$. *Let* $M = \sum_{i=1}^{r} X_i$. *Then for any* $\alpha \in (0, 1/2)$

$$\mathbf{Pr}[|M - \mathbf{E}[M]| > \alpha] \leq 2\exp\left(\frac{-2\alpha^2}{\sum_{i=1}^{r} \Delta_i^2}\right).$$

Theorem 2.6.2. *Consider a set of* r *independent identically distributed (iid) random variables* $\{X_1, \ldots, X_r\}$ *such that* $-\Delta \leq X_i \leq \Delta$ *and* $\mathbf{E}[X_i] = 0$ *for each* $i \in [r]$. *Let* $M = \sum_{i=1}^{r} X_i$ *(a sum of* X_i*s). Then for any* $\alpha \in (0, 1/2)$

$$\mathbf{Pr}[|M| > \alpha] \leq 2\exp\left(\frac{-\alpha^2}{2r\Delta^2}\right).$$

We follow by stating a bound that depends only on the variance, but has an unfortunately quite strong requirement on α. It is an open question (as far as I know) if this variance-only form can be shown with an improved dependence on α.

Theorem 2.6.3. *Consider a set of* r independent *random variables* $\{X_1, \ldots, X_r\}$. *Let* $M = \sum_{i=1}^{r} X_i$. *Then for* $\alpha \in (0, 2\mathbf{Var}[M]/(\max_i |X_i - \mathbf{E}[X_i]|))$

$$\mathbf{Pr}[|M - \mathbf{E}[M]| > \alpha] \leq 2\exp\left(\frac{-\alpha^2}{4\sum_{i=1}^{r} \mathbf{Var}[X_i]}\right).$$

2.6.1 The Union Bound

The *Robin* to Chernoff-Hoeffding's *Batman* is the *union bound*. It shows how to apply this single bound to many problems at once. It may appear crude, but can usually only be significantly improved if special structure is available in the class of problems.

Theorem 2.6.4. *Consider* t *possibly* dependent *random events* $X_1, \ldots, X_t$. *The probability that all events occur is at least*

$$1 - \sum_{i=1}^{t} (1 - \mathbf{Pr}[X_i]).$$

That is, all events are true if no event is not true.

2.7 Johnson-Lindenstrauss Lemma

The first example use is the Johnson-Lindenstrauss Lemma [8]. It describes, in the worst case, how well are distances preserved under random projections. A random projection $\phi : \mathbb{R}^d \to \mathbb{R}^k$ can be defined by the k independent (not necessarily orthogonal) coordinates, each expressed separately $\phi_i : \mathbb{R}^d \to \mathbb{R}^1$ for $i \in [k]$. Specifically, ϕ_i is associated with an independent random vector $u_i \in \mathbb{S}^{d-1}$, that is a random unit vector in $\mathbb{R}^d$. Then $\phi_i(p) = \langle p, u_i \rangle$, the inner (aka dot) product between p and the random vector u_i.

Theorem 2.7.1 ([8]). *Consider a point set $P \subset \mathbb{R}^d$ of size n. Let $Q = \phi(P)$ be a random linear projection of P to $\mathbb{R}^k$ where $k = (8/\varepsilon)^2 \ln(n/\delta)$. Then with probability at least $1 - \delta$ for* all *$p, p' \in P$, and with $\varepsilon \in (0, 1/2]$*

$$(1-\varepsilon)\left\|p - p'\right\| \leq \sqrt{\frac{d}{k}}\left\|\phi\left(p\right) - \phi\left(p'\right)\right\| \leq (1+\varepsilon)\left\|p - p'\right\|. \tag{2.1}$$

To prove this we first note that the squared version of $\|\phi(p) - \phi(p')\|$ can be decomposed as follows:

$$\|\phi(p) - \phi(p')\|^2 = \sum_{i=1}^{k} \|\phi_i(p) - \phi_i(p')\|^2.$$

Then since $(1-\varepsilon) > (1-\varepsilon)^2$ and $(1+\varepsilon) < (1+\varepsilon)^2$ for $\varepsilon \in (0, 1/2]$, it is sufficient and simpler to prove

$$(1-\varepsilon) \leq \frac{d}{k}\frac{\|\phi(p) - \phi(p')\|^2}{\|p - p'\|^2} \leq (1+\varepsilon). \tag{2.2}$$

Now we consider the random variable $M = (d/k)\|\phi(p) - \phi(p')\|^2/\|p - p'\|^2$ as the sum over k random events $X_i = (d/k)\|\phi_i(p) - \phi_i(p')\|^2/\|p - p'\|^2$. Now two simple observations follow:

- $\mathbf{E}[X_i] = 1/k$. To see this, for each u_i (independent of other $u_{i'}$, $i \neq i'$) consider a random rotation of the standard orthogonal basis, restricted only so that one axis is aligned to u_i (which itself was random). Then, in expectation each axis of this rotated basis contains $1/d$ of the squared norm of any vector, in particular $(p - p')$. So $\mathbf{E}[(\langle u_i, p - p' \rangle)^2] = (1/d)\|p - p'\|^2$. Then $\mathbf{E}[X_i] = 1/k$ follows from the linearity of ϕ.
- $\mathbf{Var}[X_i] \leq 1/k^2$. Since $\|\phi_i(p) - \phi(p')\|^2 \geq 0$, if the variance were larger than $1/k^2$, then the average distance from $E[X_i] = 1/k$ would be larger that $1/k$, and then the expected value would need to be larger than $1/k$.

Now plugging these terms into Theorem 2.6.3 yields (for some parameter γ)

$$\mathbf{Pr}[|M - \mathbf{E}[M]| > \alpha] \leq 2\exp\left(\frac{-\alpha^2}{4k(1/k^2)}\right) \leq \gamma,$$

and hence solving for k

$$k \geq 4\frac{1}{\alpha^2}\ln\left(\frac{2}{\gamma}\right).$$

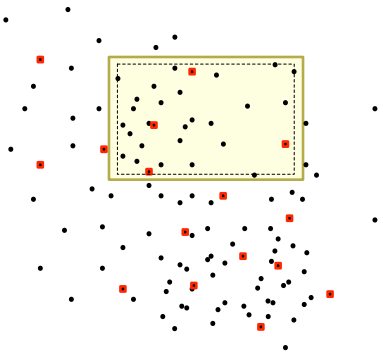
Set the middle term in (2.2) to M and note $\mathbf{E}[M] = 1$. Now by setting $\alpha = \varepsilon$[1], it follows (2.2) is satisfied with probability $1 - \gamma$ for any one pair $p, p' \in P$ when $k \geq (4/\varepsilon^2)\ln(2/\gamma)$. Since there are $\binom{n}{2} < n^2$ pairs in P, by the union bound, setting $\gamma = \delta/n^2$ reveals that for $k \geq (8/\varepsilon^2)\ln(n/\delta)$ ensures that *all pairs* $p, p' \in P$ satisfy (2.1) with probability at least $1 - \delta$. □

There are several other (often more general) proofs of this theorem [4, 6, 3, 1, 9, 7, 11].

[1] This is not the most general proof since Theorem 2.6.3 requires $\varepsilon = \alpha \leq 2(k/k^2)/(d/k) = 2/d$ which is typically much smaller than $1/2$.

2.8 Subset Samples for Density Approximation

Again, consider a set of n points $P \subset \mathbb{R}^d$. Also consider a set $\mathcal{R}$ of *queries* we can ask on these points. Herein let each $q \in \mathcal{R}$ corresponds to a d-dimensional axis-aligned rectangle $R_q = [a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_d, b_d]$, and asks for how many points in P are in R_q. That is $q(P) = |P \cap R_q|$. For example, if P represents customers of a store with d attributes (e.g. total number of purchases, average number of purchase each week, average purchase amount, ...) and R_q is a desired profile (e.g. has between 100 and 1000 purchases total, averaging between 2.5 and 10 a week, with an average total purchase between \$10 and \$20, ...). Then queries return the number of customers who fit that profile. This pair $(P, \mathcal{R})$ is called a *range space*.



We now present a weak version of a theorem by Vapnik and Chervonenkis [14] about randomly sampling and range spaces.

Theorem 2.8.1. *Let $S \subset P$ be a random sample from P of size $k = (d/\varepsilon^2)\log(2n/\delta)$. Then with probability at least $1 - \delta$, for all $q \in \mathcal{R}$*

$$\left| \frac{q(P)}{|P|} - \frac{q(S)}{|S|} \right| \leq \varepsilon. \tag{2.3}$$

They key to this theorem is again the Chernoff-Hoeffding bound. Fix some $q \in \mathcal{R}$, and for each point s_i in S, let X_i be a random event describing the effect on $q(S)$ of s_i. That is $X_i = 1$ if $s_i \in R_q$ and $X_i = 0$ if $s_i \notin R_q$, so $\Delta_i = 1$ for all $i \in [k]$. Let $M = \sum_i X_i = q(S)$, and note that $\mathbf{E}[M] = |S| \cdot q(P)/|P|$.

Multiplying M by $k = |S|$ we can now apply Theorem 2.6.1 to say

$$\mathbf{Pr}\left[\left|\frac{q(S)}{|S|} - \frac{q(P)}{|P|}\right| \geq \varepsilon\right] = \mathbf{Pr}\left[|M - \mathbf{E}[M]| \geq \varepsilon k\right] \leq 2\exp\left(\frac{-2(\varepsilon k)^2}{\sum_{i=1}^k \Delta_i^2}\right) = 2\exp(-2\varepsilon^2 k) \leq \gamma.$$

Solving for k yields that if $k \geq (1/2\varepsilon^2)\ln(2/\gamma)$, then (2.3) is true with probability at least $1 - \gamma$ for our fixed $q \in \mathcal{R}$.

To extend this to all possible choices of $q \in \mathcal{R}$ we need to apply the union bound on some bounded number of possible queries. We can show that there are no more than n^{2d} distinct subsets of P that any axis-aligned-based query in $\mathcal{R}$ can represent.

To see this, take any rectangle R that contains some subset of $T \subset P$ of the points in P. Shrink this rectangle along each coordinate until no interval can be made smaller without changing the subset of points it contains. At this point R will touch at most $2d$ points, two for each dimension (if one side happens to touch two points simultaneously, this only lowers the number of possible subsets). Any rectangle can thus be mapped to one of at most n^{2d} rectangles without changing which points it contains, where this *canonical* rectangle (and importantly its subset of points) is described by this subset of $2d$ points.

Since the application of the Chernoff-Hoeffding bound above does not change if the subset defined by R_q does not change, to prove Theorem 2.8.1 we need to show (2.3) holds for only n^{2d} different subsets. Setting $\delta = \gamma/n^{2d}$ and apply the union bound (Theorem 2.6.4) indicates that $k \geq (d/\varepsilon^2)\ln(2n/\delta)$ random samples is sufficient. □

Extensions:

- Amazingly, Vapnik and Chervonenkis [14] proved an ever stronger result that only $k = O((d/\varepsilon^2)\log(1/\varepsilon\delta))$ random samples are needed. Note, this has no dependence on n, the number of points! And moreover, Talagrand [13], as reported by Li, Long, and Srinivasan [10] improved this further to $k =$

$O((1/\varepsilon^2)(d + \log(1/\delta))$. So basically the number of samples needed to guarantee any *one* query has at most ε-error, is sufficient to guarantee the same result for *all* queries!

- This generalizes naturally to other types of range queries, using the idea of VC-dimension ν; where the bound is then $k = O((1/\varepsilon^2)(\nu + \log(1/\delta))$. For axis-aligned rectangles $\nu = 2d$, for balls it is $\nu = d+1$, and for half spaces it is $\nu = d+1$. This last bound for half spaces is particularly important for understanding how many samples are needed for determining approximate (linear) classifiers for machine learning.
- These bounds hold if P is a continuous distribution (in some sense it has an infinite number of points).

2.9 Delayed Proofs

Here we prove Theorem 2.6.1, inspired by the proof of Theorem 12.4 in Mitzenmacher and Upfal [12]. Then Theorem 2.6.2 follows as a corollary.

Markov inequality. Consider a random variable X such that all possible values of X are non-negative, then

$$\mathbf{Pr}[X > \alpha] \leq \frac{\mathbf{E}[X]}{\alpha}.$$

To see this, consider if it was not true, and $\mathbf{Pr}[X > \alpha] > \mathbf{E}[X]/\alpha$. Let $\gamma = \mathbf{Pr}[X > \alpha]$. Then, since $X > 0$, we need to make sure the expected value of X does not get too large. So, let the instances of X from the probability distribution of its values which are less than $\mathbf{E}[X]/\alpha$ be as small as possible, namely 0. Then we can still reach a contradiction:

$$\mathbf{E}[X] \geq (1-\gamma)0 + (\gamma)\alpha = \gamma\alpha > \frac{\mathbf{E}[X]}{\alpha}\alpha = \mathbf{E}[X].$$

Exponential inequalities. We state a simple fact about natural exponentials $e^x = \exp(x)$ that follows from its Taylor expansion.

$$\frac{1}{2}(e^x + e^{-x}) \leq e^{x^2/2} \tag{2.4}$$

Proof. We will prove the one-sided condition below. The other side is symmetric, and the two-sided version follows from the union bound.

$$\mathbf{Pr}[M - \mathbf{E}[M] > \alpha] \leq \exp\left(\frac{-2\alpha^2}{\sum_{i=1}^r \Delta_i^2}\right). \tag{2.5}$$

We start by letting $Y_i = X_i - \mathbf{E}[X_i]$ and rewriting

$$\begin{aligned} Y_i &= \Delta_i \frac{1 + Y_i/\Delta_i}{2} - \Delta_i \frac{1 - Y_i/\Delta_i}{2} \\ &= (\Delta_i)t + (-\Delta_i)(1-t), \end{aligned}$$

where $t = (1/2)(1 + Y_i/\Delta_i)$; note that since $|Y_i| \leq \Delta_i$ then $t \in [0,1]$. Now since $e^{\lambda x}$ is convex in x (we will set $\lambda = \alpha/\sum_{i=1}^r \Delta_i^2$ later), it follows that

$$\begin{aligned} e^{\lambda Y_i} &\leq e^{\lambda\Delta_i}\frac{1 + Y_i/\Delta_i}{2} + e^{-\lambda\Delta_i}\frac{1 - Y_i/\Delta_i}{2} \\ &= \frac{e^{\lambda\Delta_i} + e^{-\lambda\Delta_i}}{2} + \frac{Y_i}{2\Delta_i}(e^{\lambda\Delta_i} + e^{-\lambda\Delta_i}). \end{aligned}$$

Now since $\mathbf{E}[Y_i] = 0$ and equation (2.4) we have

$$\mathbf{E}\left[e^{\lambda Y_i}\right] \leq \mathbf{E}\left[\frac{e^{\lambda\Delta_i} + e^{-\lambda\Delta_i}}{2} + \frac{Y_i}{2\Delta_i}(e^{\lambda\Delta_i} + e^{-\lambda\Delta_i})\right] = \frac{e^{\lambda\Delta_i} + e^{-\lambda\Delta_i}}{2} \leq \exp\left(\frac{\lambda^2\Delta_i^2}{2}\right). \tag{2.6}$$

Finally we can show equation (2.5) as follows

$$\begin{aligned}
\mathbf{Pr}[M - \mathbf{E}[M] > \alpha] &= \mathbf{Pr}\left[\sum_i (X_i - \mathbf{E}[X_i]) \geq \alpha\right] = \mathbf{Pr}\left[\sum_i Y_i \geq \alpha\right] \\
&= \mathbf{Pr}\left[\exp\left(\lambda \sum_i Y_i\right) > \exp(\lambda\alpha)\right] \\
&\leq \frac{1}{\exp(\lambda\alpha)}\mathbf{E}\left[\exp\left(\lambda \sum_i Y_i\right)\right] = \frac{1}{\exp(\lambda\alpha)}\mathbf{E}\left[\prod_i \exp(\lambda Y_i)\right] \\
&\leq \frac{1}{\exp(\lambda\alpha)}\left(\prod_i \exp(\lambda^2\Delta_i^2/2)\right) = \exp\left(\frac{\lambda^2}{2}\sum_i \Delta_i^2 - \lambda\alpha\right) \\
&= \exp\left(\frac{-\alpha^2}{2\sum_i \Delta_i^2}\right).
\end{aligned}$$

The first inequality is from Markov inequality, the second from the equation (2.6), and the last equality uses our choice of $\lambda = \alpha/\sum_i \Delta_i^2$. □

To see Theorem 2.6.2 from Theorem 2.6.1, set each $\Delta_i = 2\Delta$ and $\mathbf{E}[M] = 0$.

2.9.1 On Independence and the Union Bound

The proof of the union bound is an elementary observation. Here we state a perhaps amazing fact that this seemingly crude bound is fairly tight even if the events are independent. Let $\mathbf{Pr}[X_i] = 1 - \gamma$ for $i \in [t]$. The union bound says the probability all events occur is at least $1 - t\gamma$. So to achieve a total of at most δ probability of failure, we need $\gamma \leq \delta/t$.

On the other hand, by independence, we can state the probability of all events is $(1 - \gamma)^t$. By the approximation for large s that $(1 - x/s)^s \approx e^{-x}$ we can approximate $(1 - \gamma)^t \approx e^{-\gamma t}$. So to achieve a total of at most δ probability of failure, we need $1 - \delta \geq e^{-\gamma t}$, which after some algebraic manipulation reveals $\gamma \leq \ln(1/(1 - \delta))/t$.

So for δ small enough (say $\delta = 1/100$, then $\ln(1/(1 - \delta)) = 0.01005\ldots$) the terms δ/t and $\ln(1/(1 - \delta))/t$ are virtually the same. The only way to dramatically improve this is to show that the events are *strongly negatively dependent*, as for instance is done in the proofs by Vapnik and Chervonenkis [14] and Talagrand [13].

Bibliography

[1] Dimitris Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coints. *Journal of Computer and System Science*, 66:671–687, 2003.

[2] Herman Chernoff. A measure of asymptotic efficiency for tests of hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23:493–509, 1952.

[3] Sanjoy Dasgupta and Anupam Gupta. An elmentary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22:60–65, 2003.

[4] P. Frankl and H. Maehara. The Johnson-Lindenstrauss lemma and the spericity of some graphs. *Journal of Combinatorial Theory, Series A*, (355–362), 1987.

[5] Wassily Hoeffding. Probability inequalities for the sum of bounded random variables. *Journal of the American Statisitcal Association*, 58:13–30, 1963.

[6] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings 30th Annual ACM Symposium on Theory of Computing*, 1998.

[7] Piotr Indyk and Assaf Naor. Nearest neighbor preserving embeddings. *ACM Transactions on Algorithms*, 3, 2007.

[8] William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz maps into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

[9] B. Klartag and S. Mendelson. Emperical processes and random projections. *Journal of Functional Analysis*, 225:229–245, 2005.

[10] Yi Li, Philip M. Long, and Aravind Srinivasan. Improved bounds on the samples complexity of learning. *Journal of Computer and System Science*, 62:516–527, 2001.

[11] Jirři Matoušek. On variants of the Johnson-Lindenstrauss lemma. *Random Structures & Algorithms*, 33:142–156, 2008.

[12] Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilisitic Analysis*. Cambridge University Press, 2005.

[13] M. Talagrand. Sharper bounds for Gaussian and emperical processes. *Annals of Probability*, 22:76, 1994.

[14] Vladimir Vapnik and Alexey Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.