CS271 Randomness & Computation **Fall 2022**

Lecture 13: October 6

Instructor: Alistair Sinclair

So far we have talked about the expectation of random variables, and moved on to the variance or the second moment. However, neither of these methods is strong enough to produce the tight bounds needed for some applications. We have already stated that if all we are given is the variance of a r.v., then Chebyshev's Inequality is essentially the best we can do. Thus, more information is needed to acheive substantially tighter bounds. We begin with the case in which the r.v. is the sum of a sequence of *independent* r.v.'s; this case arises very frequently in applications.

13.1 Chernoff/Hoeffding Bounds

These methods are usually referred to as Chernoff bounds, but should probably be referred to as Hoeffding bounds. Chernoff established the bounds for the domain of independent coin flips [C52], but it was Hoeffding who extended them to the general case [H63]. The method of proof used here is due to Bernstein.

Let $X_1 \dots X_n$ be i.i.d. random variables. $\mathbf{E}[X_i] = \mu_i$, and $\mathbf{Var}[X_i] = \sigma_i^2$, with all μ_i and all σ_i equal.

Let $X = \sum_{i=1}^{n} X_i$. What can we say about the distribution of X?

For illustration, suppose each X_i is a flip of the same (possibly biased) coin. Identifying Heads with 1 and Tails with 0, we are interested in the distribution of the total number of heads, X, in n flips.

Let

$$\mu = \mathbf{E}[X] = n\mu_i$$

by linearity of expectation. Let

$$\sigma^2 = \mathbf{Var}[X] = n\sigma_i^2$$

by linearity of the variance of independent random variables.

The Central Limit Theorem states that as $n \to \infty$, $\frac{X-\mu}{\sigma}$ approaches a standard normal distribution $N(0,1)$. Thus, as $n \to \infty$, for any fixed $\beta > 0$ we have

$$\mathbf{Pr}[|X - \mu| > \beta\sigma] \to \frac{1}{\sqrt{2\pi}} \int_{\beta}^{\infty} e^{-\frac{t^2}{2}}\, dt \approx \frac{1}{\sqrt{2\pi}\beta} e^{-\frac{\beta^2}{2}}.$$

Note that the probability of a β-deviation from the mean (in units of the standard deviation σ) decays exponentially with β. Unfortunately, there are two deficiencies in this bound for typical combinatorial and computational applications:

- The above result is asymptotic only, and says nothing about the rate of convergence (i.e., the behavior for finite n).
- The above result applies only to deviations on the order of the standard deviation, while we would like to be able to talk about arbitrary deviations.

Chernoff/Hoeffiding bounds deal with both of these deficiencies. Here is the most general (though not often used) form of the bounds:

Theorem 13.1 *Let $X_1 \ldots X_n$ be independent 0-1 r.v.s with $\mathbf{E}[X_i] = p_i$ (not necessarily equal). Let $X = \sum_{i=1}^n X_i$, $\mu = \mathbf{E}[X] = \sum_{i=1}^n p_i$, and $p = \frac{\mu}{n}$ (the average of the p_i).*

1. $\mathbf{Pr}[X \geq \mu + \lambda] \leq \exp\{-nH_p(p + \frac{\lambda}{n})\}$ *for* $0 < \lambda < n - \mu$;
2. $\mathbf{Pr}[X \leq \mu - \lambda] \leq \exp\{-nH_{1-p}(1 - p + \frac{\lambda}{n})\}$ *for* $0 < \lambda < \mu$,

where $H_p(x) = x \ln(\frac{x}{p}) + (1 - x) \ln(\frac{1-x}{1-p})$ is the relative entropy of x with respect to p.

The theorem gives bounds on the probabilities of deviating by λ above or below the mean. Note that the relative entropy is always positive (and is zero iff $x = p$). The relative entropy $H_p(x)$ can be viewed as a measure of distance between the two-point distributions $(p, 1-p)$ and $(x, 1-x)$. Thus the quantity appearing in the exponent of the tail probability in the theorem can be interpreted as the distance between the "true" two-point distribution $(p, 1 - p)$ and the distribution $(p + \frac{\lambda}{n}, 1 - p - \frac{\lambda}{n})$, in which the mean has been shifted to the tail value $p + \frac{\lambda}{n}$.

Proof: The bounds in 1 and 2 are symmetrical (replace x by $n - x$ to get 2 from 1). Therefore, we need only prove 1.

Let $m = \mu + \lambda > \mu$. We are trying to estimate $\mathbf{Pr}[X \geq m]$. The general steps we will use to obtain the bound (and which can be used in other, similar scenarios) are as follows:

- Exponentiate both sides of the inequality $X \geq m$, with a free parameter t. (This is like taking a Laplace transform.)
- Apply Markov's Inequality.
- Use the independence of the X_i to transform the expectation of a sum into a product of expectations.
- Plug in specific information about the distribution of each X_i to evaluate the expectations.
- Use concavity to bound the resulting product as a function of p rather than of the individual p_i.
- Minimize the final bound as a function of t, using basic calculus.

Here is the derivation:

$$\begin{aligned}
\mathbf{Pr}[X \geq m] &= \mathbf{Pr}[e^{tX} \geq e^{tm}] && \text{for any } t > 0 \text{ (to be chosen later)} \\
&\leq e^{-tm}\mathbf{E}[e^{tX}] && \text{by Markov's inequality} \\
&= e^{-tm}\prod_{i=1}^{n}\mathbf{E}[e^{tX_i}] && \text{by independence of the } X_i \\
&= e^{-tm}\prod_{i=1}^{n}(e^t p_i + 1 - p_i) && \text{from the distributions of the } X_i \qquad (13.1) \\
&\leq e^{-tm}(e^t p + 1 - p)^n && \text{by the Arithmetic Mean/Geometric Mean inequality}
\end{aligned}$$

More generally, we can view the last step as an application of concavity of the function $\gamma(z) = \ln((e^t - 1)z + 1)$.

Finally, we need to optimize this bound over t. Rewriting the final expression above as $\exp\{n\ln(pe^t + (1-p)) - tm\}$ and differentiating w.r.t. t, we find that the minimum is attained when $e^t = \frac{m(1-p)}{(n-m)p}$ (and note that this is indeed > 1, so $t > 0$ as required). Plugging this value in, we find that

$$\mathbf{Pr}[X \geq m] \leq \exp\left\{n\ln\left(\tfrac{m(1-p)}{(n-m)} + 1 - p\right) - m\ln\left(\tfrac{m(1-p)}{(n-m)p}\right)\right\}.$$

At this point, we need only massage this formula into the right form to get our result. It equals

$$\begin{aligned} & \exp\left\{n\left(\ln(\tfrac{1-p}{1-m/n}) - \tfrac{m}{n}\ln(\tfrac{(1-p)m/n}{(1-m/n)p})\right)\right\} \\ = \; & \exp\left\{n\left(\left(1-\tfrac{m}{n}\right)\ln\left(\tfrac{1-p}{1-m/n}\right) + \tfrac{m}{n}\ln\left(\tfrac{p}{m/n}\right)\right)\right\} \\ = \; & \exp\left\{-nH_p\left(p + \tfrac{\lambda}{n}\right)\right\} \end{aligned}$$

by the definition of entropy and the substitution $m = \mu + \lambda$, which is equivalent to $\frac{m}{n} = \frac{\mu+\lambda}{n} = p + \frac{\lambda}{n}$.

This proves part 1 of Theorem 13.1, and hence also part 2 by symmetry. ■

Note that the above theorem and proof applies equally to independent r.v.'s X_i taking any values in the interval $[0,1]$ (not just the values 0 and 1). The only step in the proof where we used the distributions of the X_i was step (13.1), and one can check that this still holds (with $=$ replaced by $\leq$) for any X_i on $[0,1]$ (i.e., 0-1 is the worst case, for given mean p_i). Similarly, an analogous bound holds when the X_i live on any bounded intervals $[a_i, b_i]$ (and then the quantities $|a_i - b_i|$ will appear in the bound). In fact, one can obtain similar bounds even when the X_i are unbounded, provided their distributions fall off quickly enough, as in the case (e.g.) of a geometric random variable. See the **exercises** following Corollary 13.3 below for a derivation of some of these claims.

13.2 Some Corollaries

We now give some more useful versions of the Chernoff bound, all obtained from the above basic form.

Corollary 13.2

$$\left.\begin{array}{l}\mathbf{Pr}[X \leq \mu - \lambda] \\ \mathbf{Pr}[X \geq \mu + \lambda]\end{array}\right\} \leq \exp\left(-\frac{2\lambda^2}{n}\right).$$

Proof: (Sketch.) We consider only the upper tail; the lower tail follows by symmetry. Writing $z = \frac{\lambda}{n}$, and comparing the exponent in part 1 of Theorem 13.1 with our target value $-\frac{2\lambda^2}{n}$, we need to show that

$$f(z) \equiv (p+z)\ln\Big(\frac{p+z}{p}\Big) + (1-p-z)\ln\Big(\frac{1-p-z}{1-p}\Big) - 2z^2 \geq 0$$

on the interval $0 \leq z \leq 1-p$. To do this we note that $f(0) = 0$, so it is enough to show that $f(z)$ is non-decreasing on this interval. This can be verified by looking at the first and second derivatives of f. The details are left as an exercise. ■

We now prove the following alternative corollary, due to Angluin and Valiant [AV79], which is never much worse and is much sharper in cases where $\mu \ll n$ (e.g., $\mu = \Theta(\log n)$).

Corollary 13.3 *For $0 < \beta < 1$ we have*

$$\begin{aligned} \mathbf{Pr}[X \leq (1-\beta)\mu] \; & \leq \exp\{-\mu(\beta + (1-\beta)\ln(1-\beta))\} \\ & \leq \exp(-\tfrac{\beta^2\mu}{2}) \end{aligned}$$

And for $\beta > 0$ we have

$$\begin{aligned}\mathbf{Pr}[X \geq (1+\beta)\mu] &\leq \exp\{-\mu(-\beta + (1+\beta)\ln(1+\beta))\} \quad (*)\\ &\leq \begin{cases} \exp(-\frac{\beta^2\mu}{2+\beta}) & \beta > 0 \\ \exp(-\frac{\beta^2\mu}{3}) & 0 < \beta \leq 1 \end{cases}\end{aligned}$$

Proof: We first prove the bound on the lower tail. Plugging in $\lambda = \beta\mu = \beta np$ into part 2 of Theorem 13.1 gives

$$\begin{aligned}\mathbf{Pr}[X \leq (1-\beta)\mu] &\leq \exp\{-nH_{1-p}(1-p+\beta p)\} &(13.2)\\ &= \exp\Big\{n\Big((1-p+\beta p)\ln\Big(\frac{1-p}{1-p+\beta p}\Big) + (p-\beta p)\ln\Big(\frac{p}{p-\beta p}\Big)\Big)\Big\}\\ &\leq \exp\Big\{n\Big(-\beta p + p(1-\beta)\ln\Big(\frac{1}{1-\beta}\Big)\Big)\Big\} &(13.3)\\ &= \exp\{-\mu(\beta + (1-\beta)\ln(1-\beta))\}\\ &\leq \exp\Big\{-\mu\Big(\beta - \beta + \frac{\beta^2}{2}\Big)\Big\} &(13.4)\\ &= \exp\Big\{-\frac{\mu\beta^2}{2}\Big\}.\end{aligned}$$

The inequality in (13.2) comes directly from Theorem 13.1. Inequality (13.3) comes from the fact that

$$\ln\Big(\frac{1-p}{1-p+\beta p}\Big) = \ln\Big(1 - \frac{\beta p}{1-p+\beta p}\Big) \leq -\frac{\beta p}{1-p+\beta p}$$

using $\ln(1-x) < -x$. Finally, inequality (13.4) is due to the fact that $(1-x)\ln(1-x) \geq -x + x^2/2$ for $0 < x < 1$.

The proof for the upper tail is very similar. The first step is to plug in $\lambda = \beta\mu$ into part 1 of Theorem 13.1 and follow the same steps as above. The only difference is that this time we use the inequality $\ln(1+x) < x$ to get $(*)$, and then the inequality

$$\ln(1+\beta) > \frac{\beta}{1+\beta/2}$$

to get the final form. The details are left as an exercise. ■

Exercise: Is this bound ever worse than that of Corollary 13.2? If so, by how much?

Exercise: Verify that Theorem 13.1 (and therefore also the above Corollaries) hold when the X_i are not necessarily 0,1-valued but can take any values in the range $[0,1]$, subject only to their expectations being p_i. [Hint: Show that, if Y is any r.v. taking values in $[0,1]$, and Z is a (0,1)-valued r.v. with $EZ = EY$, then for any convex function f we have $E(f(Y)) \leq E(f(Z))$. Apply this fact to the function $f(x) = \mathrm{e}^{tx}$ at an appropriate point in the proof of Theorem 13.1.]

Exercise: Suppose that the r.v.'s are independent and X_i takes values in the interval $[a_i, b_i]$. Prove the following generalization of Corollary 13.2.

$$\left.\begin{matrix}\mathbf{Pr}[X \leq \mu - \lambda] \\ \mathbf{Pr}[X \geq \mu + \lambda]\end{matrix}\right\} \leq \exp\left(-\frac{2\lambda^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

[Hint: You will need to go back and tweak the proof of Theorem 13.1. Is there a simpler "scaling" proof when all the intervals $[a_i, b_i]$ are the same?]

13.3 Simple Examples

Example 13.4 Fair coin flips.

Given a coin where each flip has a probability $p_i = p = 1/2$ of coming up heads, Corollary 13.2 gives the following bound on the probability that the number of heads X differs from the expected number $\mu = n/2$ by more than λ:

$$\mathbf{Pr}[|X - \mu| \geq \lambda] \leq 2e^{-\frac{2\lambda^2}{n}}.$$

Note that the standard deviation is $\sigma = \sqrt{np(1-p)} = \sqrt{n}/2$. So for $\lambda = \beta\sigma$ the deviation probability is $2e^{-\beta^2/2}$. This is similar to the asymptotic bound obtained from the CLT (but is valid for all n).

Example 13.5 Biased coin flips.

In this example, the bit flips still all have the same probability, but now the probability of coming up heads is $p_i = p = 3/4$. We can then ask for the value of

$$\mathbf{Pr}[\leq 1/2 \text{ of the flips come up heads }]$$

The expected number of heads is $\mu = 3n/4$. Thus, using Corollary 13.3 with $\beta = 1/3$ we get

$$\begin{aligned}
\mathbf{Pr}[X \leq n/2] &= \mathbf{Pr}[X \leq \mu - n/4] && (13.5)\\
&= \mathbf{Pr}[X \leq (1 - 1/3)\mu] && (13.6)\\
&\leq \exp(-\frac{(1/3)^2}{2} \cdot \frac{3n}{4}) && (13.7)\\
&= e^{-\frac{n}{24}}. && (13.8)
\end{aligned}$$

(A similar bound, with a slightly better constant in the exponent, is obtained by using Corollary 13.2.) Recall that we have already used a bound of this form twice in the course: in boosting the success probability of two-sided error randomized algorithms (BPP algorithms), and in the analysis of the "median trick" for fully polynomial randomized approximation schemes.

13.4 Randomized Routing

We will now see an example of a randomized algorithm whose analysis makes use of the above Chernoff bounds. The problem is defined as follows. Consider the network defined by the n-dimensional hypercube: i.e, the vertices of the network are the strings in $\{0,1\}^n$, and edges connect pairs of vertices that differ in exactly one bit. We shall think of each edge as consisting of two links, one in each direction. Let $N = 2^n$, the number of vertices. Now let π be any permutation on the vertices of the cube. The goal is to send one packet from each i to its corresponding $\pi(i)$, for all i simultaneously.

This problem can be seen as a building block for more realistic routing applications. A strategy for routing permutations on a graph can give useful inspiration for solving similar problems on real networks.

We will use a synchronous model, i.e., the routing occurs in discrete time steps, and in each time step, one packet is allowed to travel along each (directed) edge. If more than one packet wishes to traverse a given edge in the same time step, all but one of these packets are held in a queue at the edge. We assume any fair queueing discipline (e.g., FIFO).

The goal is to minimize the total time before all packets have reached their destinations. A priori, a packet only has to travel $O(n)$ steps (the diameter of the cube). However, due to the potential for congestion on the edges, it is possible that the process will take much longer than this as packets get delayed in the queues.

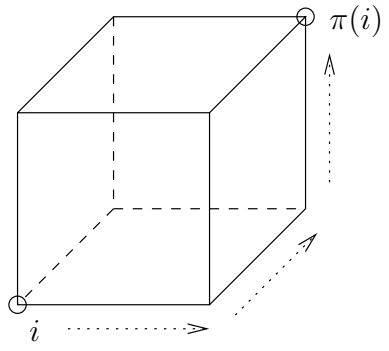

Figure 13.1: Randomized routing in a hypercube.

Definition 13.6 *An* oblivious *strategy is one in which the route chosen for each packet does not depend on the routes of the other packets. That is, the path from i to $\pi(i)$ is a function of i and $\pi(i)$ only.*

An oblivious routing strategy is thus one in which there is no global synchronization, a realistic constraint if we are interested in real-world problems.

Theorem 13.7 *[KKT90] For any deterministic, oblivious routing strategy on the hypercube, there exists a permutation that requires $\Omega(\sqrt{N/n}) = \Omega(\sqrt{2^n/n})$ steps.*

This is quite an undesirable worst case. Fortunately, randomization can provide a dramatic improvement on this lower bound.

Theorem 13.8 *[VB81] There exists a randomized, oblivious routing strategy that terminates in $O(n)$ steps with very high probability.*

We now sketch this randomized strategy, which consists of two phases.

- In phase 1, each packet i is routed to $\delta(i)$, where the destination $\delta(i)$ is chosen u.a.r.
- In phase 2, each packet is routed from $\delta(i)$ to its desired final destination, $\pi(i)$.

In both phases, we use "bit-fixing" paths to route the packets. In the bit-fixing path from vertex x to vertex y of the cube, we flip each bit x_i to y_i (if necessary) in left-to-right order. For example, a bit-fixing path from $x = 0011001$ to $y = 1110001$ in the hypercube $\{0,1\}^7$ is $x = 0011001 \to 1011001 \to 1111001 \to 1110001 = y$. Bit-fixing paths are always shortest paths.

Note that δ is *not* required to be a permutation (i.e., different packets may share the same intermediate destination), so this strategy is oblivious.

This strategy breaks the symmetry in the problem by simply choosing a random intermediate destination for each packet. This makes it impossible for an adversary with knowledge of the strategy to engineer a bad permutation. In our analysis, we will see that each phase of this algorithm takes only $O(n)$ steps with high probability. In order to do this, we will take a union bound over all 2^n packets, so we will need an exponentially small probability of a single packet taking a long time. This is where Chernoff/Hoeffding bounds will be required.

Proof: Phase 1 starts from a fixed source i and routes to a random destination $\delta(i)$. Phase 2 starts from a random source $\delta(i)$ and routes to a fixed destination $\pi(i)$. By symmetry, it suffices to prove that phase 1 terminates in $O(n)$ steps w.h.p.

Let $D(i)$ be the delay suffered by packet i; then the total time taken is at most $n + \max_i D(i)$. We will soon prove:

$$\forall i \ \Pr[D(i) > cn] \leq e^{-2n}. \tag{13.9}$$

Taking a union bound, it follows that

$$\Pr[\exists i : D(i) > cn] \ \leq \ 2^n e^{-2n} \ < \ 2^{-n}.$$

■

It remains to prove (13.9). For a single packet traveling from $i \to \delta(i)$, let $P_i = (e_1, e_2, \ldots, e_k)$ be the sequence of edges on the route taken by packet i. Let $S_i = \{j \neq i \ : \ P_j \cap P_i \neq \emptyset\}$, the set of packets whose routes intersect P_i. The guts of the proof lie in the following claim.

Claim 13.9 $D(i) \ \leq \ |S(i)|$.

Proof: First note that, in the hypercube, when two "bit-fixing" paths diverge they will not come together again; i.e., routes which intersect will intersect only in one contiguous segment. With this observation, we can now charge each unit of delay for packet i to a *distinct* member of $S(i)$.

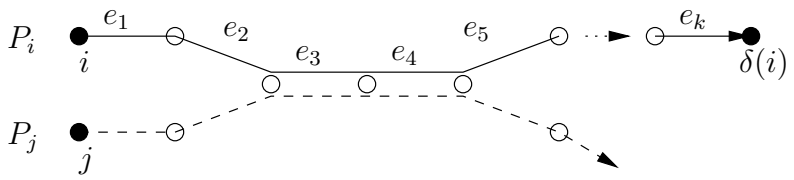

Figure 13.2: Routes P_i and P_j in the hypercube.

Definition 13.10 *Let j be a packet in $S(i) \cup \{i\}$. The* lag *of packet j at the start of time step t is $t - l$, where e_l is the next edge (on P_i) that packet j wants to traverse.*

Note that the lag of packet i proceeds through the sequence of values $0, 1, ..., D(i)$. Consider the time step t when the lag goes from L to $L + 1$; suppose this happens when i is waiting to traverse edge e_l. Then i must be held up in the queue at e_l (else its lag would not increase), so there exists at least one other packet at e_l with lag L, and this packet actually moves at step t.

Now consider the last time at which there exists any packet with lag L: say this is time t'. Then some such packet must leave path P_i (or reach its destination) at time t', for at any step among all packets with a given lag at a given edge, at least one must move (and it would retain the same lag if it remained on path P_i). So we may charge this unit of delay to that packet. Each packet is charged at most once, because it is charged only when it leaves P_i (which, by the observation at the start of the proof, happens only once). ■

Proof: (of inequality (13.9))

Define

$$H_{ij} = \begin{cases} 1 & P_i \cap P_j \neq \emptyset \\ 0 & \text{otherwise} \end{cases}$$

By Claim 13.9, $D(i) \leq \sum_{j \neq i} H_{ij}$. And since the H_{ij} are independent we can use a Chernoff bound to bound the tail probability of this sum. First, we need a small detour to bound the mean, $\mu = \mathbf{E}[\sum_{j \neq i} H_{ij}]$ (since the expectations of the H_{ij} are tricky to get hold of).

Define

$$T(e_l) = \#\text{ of paths } P_j \text{ that pass through edge } e_l \text{ (on path } P_i).$$

Then,

$$\begin{aligned} \mathbf{E}[T(e_l)] &= \frac{\mathbf{E}[\text{total length of all paths}]}{\#\text{ directed edges}} \quad \text{(by symmetry)} \\ &= \frac{N\ (n/2)}{Nn} \ = \ 1/2, \end{aligned}$$

so the expected number of paths that intersect P_i is

$$\mathbf{E}[\sum_{j \neq i} H_{ij}] \leq k/2 \leq n/2. \tag{13.10}$$

Note that we used the fact that $\mathbf{E}[T(e_l)]$ does not depend on l, which follows by symmetry. (One might think this is false because we chose a particular bit ordering. However the ordering does not matter because for every source and every i with probability $1/2$ we choose a sink such that the bit-fixing algorithm flips the i^{th} bit.)

We can now apply the Chernoff bound (Angluin/Valiant version):

$$\Pr[D(i) \ \geq \ (1+\beta)\mu] \ \leq \ \exp\left(-\frac{\beta^2}{2+\beta}\mu\right),$$

where $\mu = \mathbf{E}[D(i)]$.

It is easy to check (**exercise!**) that, for any fixed value of the tail $(1+\beta)\mu$, the above bound is worst when μ is maximized. Thus we may plug in our upper bound $\mu \leq n/2$ and conclude that

$$\Pr[D(i) \geq (1+\beta)\tfrac{n}{2}] \leq \exp(-\tfrac{\beta^2}{2+\beta}\tfrac{n}{2}).$$

Taking $\beta = 6$,

$$\Pr[D(i) \geq \tfrac{7}{2}n] \leq \exp(-\tfrac{36}{16}n) = \exp(-\tfrac{9}{4}n) \leq \exp(-2n).$$

Thus, using a union bound as indicated earlier, we see that all packets reach their destinations in Phase 1 in at most $n + \frac{7}{2}n = \frac{9}{2}n$ steps w.h.p. The complete algorithm (phases 1 and 2) thus terminates in $\leq 9n$ steps. (To keep the phases separate, we assume that all packets wait until time $\frac{9}{2}n$ before beginning Phase 2.) This completes the proof that the randomized oblivious strategy routes any permutation in $O(n)$ steps w.h.p.

■

References

[AV79] D. Angluin and L.G. Valiant, "Fast probabilistic algorithms for Hamiltonian circuits and matchings", *Journal of Computer and System Sciences*, No. 19, 1979, pp. 155–193.

[C52] H. Chernoff, "A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the Sum of Observations," *Annals of Mathematic Statistics, 23*, 1952, pp. 493–507.

[H63] W. Hoeffding, "Probability Inequalities for Sums of Bounded Random Variables," *American Statistical Association Journal*, 1963, pp. 13–30.

[KKT90] C. Kaklamanis, D. Krizanc and A. Tsantilas, "Tight Bounds for Oblivious Routing in the Hypercube," *Proceedings of the Symposium on Parallel Algorithms and Architecture*, 1990, pp. 31–36.

[VB81] L. G. Valiant and G. J. Brebner, "Universal schemes for parallel communication," *Proceedings of the 13th annual ACM Symposium on Theory of Computing*, 1981, pp. 263–277.