Chernoff Bounds

Wolfgang Mulzer

1 The General Bound

Let $P = (p_1, \ldots, p_m)$ and $Q = (q_1, \ldots, q_m)$ be two distributions on m elements, i.e., $p_i, q_i \geq 0$, for $i = 1, \ldots, m$, and $\sum_{i=1}^m p_i = \sum_{i=1}^m q_i = 1$. The *Kullback-Leibler divergence* or *relative entropy* of P and Q is defined as

$$D_{\mathrm{KL}}(P \| Q) := \sum_{i=1}^m p_i \ln \frac{p_i}{q_i}.$$

If $m = 2$, i.e., $P = (p, 1-p)$ and $Q = (q, 1-q)$, we also write $D_{\mathrm{KL}}(p \| q)$. The Kullback-Leibler divergence provides a measure of distance between the distributions P and Q: it represents the expected loss of efficiency if we encode an m-letter alphabet with distribution P with a code that is optimal for distribution Q. We can now state the general form of the Chernoff Bound:

Theorem 1.1. *Let $X_1, \ldots, X_n$ be independent random variables with $X_i \in \{0, 1\}$ and $\Pr[X_i = 1] = p$, for $i = 1, \ldots n$. Set $X := \sum_{i=1}^n X_i$. Then, for any $t \in [0, 1-p]$, we have*

$$\Pr[X \geq (p+t)n] \leq e^{-D_{\mathrm{KL}}(p+t \| p)n}.$$

2 Four Proofs

2.1 The Moment Method

The usual proof of Theorem 1.1 uses the exponential function exp and Markov's inequality. It is called *moment method* because exp simultaneously encodes all *moments* of X, i.e., X, X^2, X^3, etc. The proof technique is very general and can be used to obtain several variants of Theorem 1.1. Let $\lambda > 0$ be a parameter to be determined later. We have

$$\Pr[X \geq (p+t)n] = \Pr[\lambda X \geq \lambda(p+t)n] = \Pr\left[e^{\lambda X} \geq e^{\lambda(p+t)n}\right].$$

From Markov's inequality, we obtain

$$\Pr\left[e^{\lambda X} \geq e^{\lambda(p+t)n}\right] \leq \frac{\mathbf{E}[e^{\lambda X}]}{e^{\lambda(p+t)n}}.$$

Now, the independence of the X_i yields

$$\mathbf{E}[e^{\lambda X}] = \mathbf{E}\left[e^{\lambda \sum_{i=1}^n X_i}\right] = \mathbf{E}\left[\prod_{i=1}^n e^{\lambda X_i}\right] = \prod_{i=1}^n \mathbf{E}\left[e^{\lambda X_i}\right] = \left(pe^\lambda + 1 - p\right)^n.$$

Thus,

$$\Pr[X > (p+t)n] \le \Big(\frac{pe^{\lambda} + 1 - p}{e^{\lambda(p+t)}}\Big)^n, \tag{1}$$

for every $\lambda > 0$. Optimizing for λ using calculus, we get that the right hand side is minimized if

$$e^{\lambda} = \frac{(1-p)(p+t)}{p(1-p-t)}.$$

Plugging this into (1), we get

$$\Pr[X > (p+t)n] \le \left[\Big(\frac{p}{p+t}\Big)^{p+t}\Big(\frac{1-p}{1-p-t}\Big)^{1-p-t}\right]^n = e^{-D_{\mathrm{KL}}(p+t\|p)n},$$

as desired.

2.2 Chvátal's Method

Let $B(n,p)$ the random variable that gives the number of heads in n independent Bernoulli trials with success probability p. It is well known that

$$\Pr[B(n,p) = l] = \binom{n}{l} p^l (1-p)^{n-l},$$

for $l = 0, \ldots, n$. Thus, for any $\tau \ge 1$ and $k \ge pn$, we get

$$\begin{aligned}\Pr[B(n,p) \ge k] &= \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i} \\ &\le \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i} \underbrace{\tau^{i-k}}_{\ge 1} + \underbrace{\sum_{i=0}^{k-1} \binom{n}{i} p^i (1-p)^{n-i} \tau^{i-k}}_{\ge 0} = \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} \tau^{i-k}.\end{aligned}$$

Using the Binomial theorem, we obtain

$$\Pr[B(n,p) \ge k] \le \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} \tau^{i-k} = \tau^{-k} \sum_{i=0}^{n} \binom{n}{i} (p\tau)^i (1-p)^{n-i} = \frac{(p\tau + 1 - p)^n}{\tau^k}.$$

If we write $k = (p+t)n$ and $\tau = e^{\lambda}$, we can conclude

$$\Pr[B(n,p) \ge (p+t)n] \le \Big(\frac{pe^{\lambda} + 1 - p}{e^{\lambda(p+t)}}\Big)^n.$$

This is the same as (1), so we can complete the proof of Theorem 1.1 as in Section 2.1.

2.3 The Impagliazzo-Kabanets Method

Let $\lambda \in [0,1]$ be a parameter to be chosen later. Let $I \subseteq \{1,\dots,n\}$ be a random index set obtained by including each element $i \in \{1,\dots,n\}$ with probability λ. We estimate $\Pr\bigl[\prod_{i\in I} X_i = 1\bigr]$ in two different ways, where the probability is over the random choice of $X_1,\dots,X_n$ and I.

On the one hand, using the union bound and independence, we have

$$\Pr\Bigl[\prod_{i\in I} X_i = 1\Bigr] \le \sum_{S\subseteq\{1,\dots,n\}} \Pr\Bigl[I = S \wedge \prod_{i\in S} X_i = 1\Bigr] = \sum_{S\subseteq\{1,\dots,n\}} \Pr[I=S]\cdot \prod_{i\in S}\Pr[X_i=1]$$

$$= \sum_{S\subseteq\{1,\dots,n\}} \lambda^{|S|}(1-\lambda)^{n-|S|}\cdot p^{|S|} = \sum_{s=0}^{n}\binom{n}{s}(\lambda p)^s(1-\lambda)^{n-s} = (\lambda p + 1 - \lambda)^n, \quad (2)$$

by the Binomial theorem. On the other hand, by the law of total probability,

$$\Pr\Bigl[\prod_{i\in I} X_i = 1\Bigr] \ge \Pr\Bigl[\prod_{i\in I} X_i = 1 \mid X \ge (p+t)n\Bigr]\Pr[X \ge (p+t)n].$$

Now, fix $X_1,\dots,X_n$ with $X \ge (p+t)n$. For the fixed choice of $X_1 = x_1,\dots,X_n = x_n$, the probability $\Pr\bigl[\prod_{i\in I} x_i = 1\bigr]$ is exactly the probability that I avoids all the $n-X$ indices i where $x_i = 0$. Thus,

$$\Pr\Bigl[\prod_{i\in I} x_i = 1\Bigr] = (1-\lambda)^{n-X} \ge (1-\lambda)^{(1-p-t)n}.$$

Since the bound holds uniformly for every choice of $x_1,\dots,x_n$ with $X \ge (p+t)n$, we get

$$\Pr\Bigl[\prod_{i\in I} X_i = 1 \mid X \ge (p+t)n\Bigr] \ge (1-\lambda)^{(1-p-t)n},$$

so

$$\Pr\Bigl[\prod_{i\in I} X_i = 1\Bigr] \ge (1-\lambda)^{(1-p-t)n}\Pr[X \ge (p+t)n].$$

Combining with (2),

$$\Pr[X \ge (p+t)n] \le \left(\frac{\lambda p + 1 - \lambda}{(1-\lambda)^{(1-p-t)}}\right)^n. \quad (3)$$

Using calculus, we get that the right hand side is minimized for $\lambda = t/(1-p)(p+t)$ (note that $\lambda \le 1$ for $t \le 1-p$). Plugging this into (3),

$$\Pr[X > (p+t)n] \le \left[\Bigl(\frac{p}{p+t}\Bigr)^{p+t}\Bigl(\frac{1-p}{1-p-t}\Bigr)^{1-p-t}\right]^n = e^{-D_{\mathrm{KL}}(p+t\|p)n},$$

as desired.

2.4 The Coding Theoretic Argument

The next proof, due to Luc Devroye, Gábor Lugosi, and Pat Morin, is inspired by coding theory. Let $\{0,1\}^n$ be the set of all bit strings of length n, and let $w : \{0,1\}^n \to [0,1]$ be a *weight function*. We call w *valid* if $\sum_{x \in \{0,1\}^n} w(x) \le 1$. The following lemma says that for any probability distribution p_x on $\{0,1\}^n$, a valid weight function is unlikely to be substantially larger than p_x.

Lemma 2.1. *Let $\mathcal{D}$ be a probability distribution on $\{0,1\}^n$ that assigns to each $x \in \{0,1\}^n$ a probability p_x, and let w be a valid weight function. For any $s \ge 1$, we have*

$$\Pr_{x \sim \mathcal{D}}[w(x) \ge sp_x] \le 1/s.$$

Proof. Let $Z_s = \{x \in \{0,1\}^n \mid w(x) \ge sp_x\}$. We have

$$\Pr_{x \sim \mathcal{D}}[w(x) \ge sp_x] = \sum_{\substack{x \in Z_s \\ p_x > 0}} p_x \le \sum_{\substack{x \in Z_s \\ p_x > 0}} p_x \frac{w(x)}{sp_x} \le (1/s) \sum_{x \in Z_s} w(x) \le 1/s,$$

since $w(x)/sp_x \ge 1$ for $x \in Z_s$, $p_x > 0$, and since w is valid. □

We now show that Lemma 2.1 implies Theorem 1.1. For this, we interpret the sequence $X_1, \dots, X_n$ as a bit string of length n. This induces a probability distribution $\mathcal{D}$ that assigns to each $x \in \{0,1\}^n$ the probability $p_x = p^{k_x}(1-p)^{n-k_x}$, where k_x denotes the number of 1-bits in x. We define a weight function $w : \{0,1\}^n \to [0,1]$ by $w(x) = (p+t)^{k_x}(1-p-t)^{n-k_x}$, for $x \in \{0,1\}^n$. Then w is valid, since $w(x)$ is the probability that x is generated by setting each bit to 1 independently with probability $p+t$. For $x \in \{0,1\}^n$, we have

$$\frac{w(x)}{p_x} = \left(\frac{p+t}{p}\right)^{k_x} \left(\frac{1-p-t}{1-p}\right)^{n-k_x}.$$

Since $((p+t)/p)((1-p)/(1-p-t)) \ge 1$, it follows that $w(x)/p_x$ is an increasing function of k_x. Hence, if $k_x \ge (p+t)n$, we have

$$\frac{w(x)}{p_x} \ge \left[\left(\frac{p+t}{p}\right)^{p+t} \left(\frac{1-p-t}{1-p}\right)^{1-p-t}\right]^n = e^{D_{\mathrm{KL}}(p+t\|p)n}.$$

We now apply Lemma 2.1 to $\mathcal{D}$ and w to get

$$\Pr[X \ge (p+t)n] = \Pr_{x \sim \mathcal{D}}[k(x) \ge (p+t)n] \le \Pr_{x \sim \mathcal{D}}\left[w(x) \ge p_x e^{D_{\mathrm{KL}}(p+t\|p)n}\right] \le e^{-D_{\mathrm{KL}}(p+t\|p)n},$$

as claimed in Theorem 1.1.

We provide some coding-theoretic background to explain the intuition behind the proof. A *code* for $\{0,1\}^n$ is an injective function $C : \{0,1\}^n \to \{0,1\}^*$. The images of C are called *codewords*. A code is called *prefix-free* if no codeword is the prefix of another codeword, i.e., for all $x, y \in \{0,1\}^n$ with $x \ne y$, we have that if $|x| \le |y|$, then x and y differ in at least one bit position. A prefix-free code has a natural representation as a rooted binary tree in which the leaves correspond to elements of $\{0,1\}^n$. Even though the codeword lengths in a prefix-free code may vary, this structure imposes a restriction on the allowed lengths. This is formalized in *Kraft's inequality*.

Lemma 2.2 (Kraft's inequality)**.** *Let $C : \{0,1\}^n \to \{0,1\}^*$ be a prefix-free code. Then,*

$$\sum_{x \in \{0,1\}^n} 2^{-|C(x)|} \leq 1.$$

Conversely, given a function $\ell : \{0,1\}^n \to \mathbb{N}$ with

$$\sum_{x \in \{0,1\}^n} 2^{-\ell(x)} \leq 1,$$

there exists a prefix-free code $C : \{0,1\}^n \to \{0,1\}^*$ with $|C(x)| = \ell(x)$ for all $x \in \{0,1\}^n$.

Proof. Let $m = \max_{x \in \{0,1\}^n} |C(x)|$, and let y be random element of $y \in \{0,1\}^m$. Then, for each $x \in \{0,1\}^n$, the probability that $C(x)$ is a prefix of y is exactly $2^{-|C(x)|}$. Furthermore, since C is prefix-free, these events are mutually exclusive. Thus,

$$\sum_{x \in \{0,1\}^n} 2^{-|C(x)|} \leq 1,$$

as claimed.

Next, we prove the second part. Let $m = \max_{x \in \{0,1\}^n} \ell(x)$ and let T be a complete binary tree of height m. We construct C according to the following algorithm: we set $X = \{0,1\}^n$, and we pick $x^* \in X$ with $\ell(x^*) = \min_{x \in X} \ell(x)$. Then we select a node $v \in T$ with depth $\ell(x^*)$. We assign to $C(x^*)$ the codeword of length ℓ that corresponds to v, and we remove v and all its descendants from T. This deletes exactly $2^{m-\ell(x^*)}$ leaves from T. Next, we remove x^* from X and we repeat this procedure until X is empty. While $X \neq \emptyset$, we have

$$\sum_{x \in \{0,1\}^n \setminus X} 2^{m-\ell(x)} < 2^m,$$

so T contains in each iteration at least one leaf and thus also at least one node of depth $\ell(x^*)$. Since we assign the nodes by increasing depth, and since all descendants of an assigned node are deleted from the tree, the resulting code is prefix-free. □

Kraft's inequality shows that a prefix-free code C induces a valid weight function $w(x) = 2^{-|C(x)|}$. Thus, Lemma 2.1 implies that for any probability distribution p_x on $\{0,1\}^n$ and for any prefix-free code, the probability mass of the strings x with codeword length $\log(1/p_x) - s$ is at most 2^{-s}. Now, if we set $\ell(x) = \lceil -k_x \log(p+t) - (n-k_x)\log(1-p-t) \rceil$ for $x \in \{0,1\}^n$, the converse of Kraft's inequality shows that there exists a prefix free code C' with $|C'(x)| = \ell(x)$. The calculation above shows that C' saves roughly $n(p+t)\log((p+t)/p) + n(1-p-t)\log((1-p-t)/(1-p))$ bits over $\log(1/p_x)$ for any x with $k_x \geq (p+t)n$, which almost gives the desired result. We generalize to arbitrary valid weight functions to avoid the slack introduced by the ceiling function.

3 Useful Consequences

3.1 The Lower Tail

Corollary 3.1. *Let $X_1, \ldots, X_n$ be independent random variables with $X_i \in \{0,1\}$ and $\Pr[X_i = 1] = p$, for $i = 1, \ldots n$. Set $X := \sum_{i=1}^n X_i$. Then, for any $t \in [0,p]$, we have*

$$\Pr[X \le (p-t)n] \le e^{-D_{\mathrm{KL}}(p-t\|p)n}.$$

Proof.

$$\Pr[X \le (p-t)n] = \Pr[n - X \ge n - (p-t)n] = \Pr[X' \ge (1-p+t)n],$$

where $X' = \sum_{i=1}^n X_i'$ with independent random variables $X_i' \in \{0,1\}$ such that $\Pr[X_i' = 1] = 1-p$. The result follows from $D_{\mathrm{KL}}(1-p+t\|1-p) = D_{\mathrm{KL}}(p-t\|p)$. □

3.2 Motwani-Raghavan version

Corollary 3.2. *Let $X_1, \ldots, X_n$ be independent random variables with $X_i \in \{0,1\}$ and $\Pr[X_i = 1] = p$, for $i = 1, \ldots n$. Set $X := \sum_{i=1}^n X_i$ and $\mu = pn$. Then, for any $\delta \ge 0$, we have*

$$\Pr[X \ge (1+\delta)\mu] \le \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu}, \text{ and}$$
$$\Pr[X \le (1-\delta)\mu] \le \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu}.$$

Proof. Setting $t = \delta\mu/n$ in Theorem 1.1 yields

$$\begin{aligned}
\Pr[X \ge (1+\delta)\mu] &\le \exp\left(-n\left[p(1+\delta)\ln(1+\delta) + p\left(\frac{1-p}{p} - \delta\right)\ln\left(1 - \delta\frac{p}{1-p}\right)\right]\right) \\
&= \left(\frac{(1-\delta p/(1-p))^{\delta-(1-p)/p}}{(1+\delta)^{1+\delta}}\right)^{\mu} \\
&\le \left(\frac{e^{-\delta^2 p/(1-p)+\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu} \le \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu}.
\end{aligned}$$

Setting $t = \delta\mu/n$ in Corollary 3.1 yields

$$\begin{aligned}
\Pr[X \le (1-\delta)\mu] &\le \exp\left(-n\left[p(1-\delta)\ln(1-\delta) + p\left(\frac{1-p}{p} + \delta\right)\ln\left(1 + \delta\frac{p}{1-p}\right)\right]\right) \\
&= \left(\frac{(1+\delta p/(1-p))^{-\delta-(1-p)/p}}{(1-\delta)^{1-\delta}}\right)^{\mu} \\
&\le \left(\frac{e^{-\delta^2 p/(1-p)-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu} \le \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu}.
\end{aligned}$$

□

3.3 Handy Versions

Corollary 3.3. *Let $X_1, \ldots, X_n$ be independent random variables with $X_i \in \{0,1\}$ and $\Pr[X_i = 1] = p$, for $i = 1, \ldots n$. Set $X := \sum_{i=1}^n X_i$ and $\mu = pn$. Then, for any $\delta \in (0,1)$, we have*

$$\Pr[X \le (1-\delta)\mu] \le e^{-\delta^2\mu/2}.$$

Proof. By Corollary 3.2

$$\Pr[X \le (1-\delta)\mu] \le \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu}.$$

Using the power series expansion of $\ln(1-\delta)$, we get

$$(1-\delta)\ln(1-\delta) = -(1-\delta)\sum_{i=1}^{\infty}\frac{\delta^i}{i} = -\delta + \sum_{i=2}^{\infty}\frac{\delta^i}{(i-1)i} \ge -\delta + \delta^2/2.$$

Thus,

$$\Pr[X \le (1-\delta)\mu] \le e^{[-\delta+\delta-\delta^2/2]\mu} = e^{-\delta^2\mu/2},$$

as claimed. □

Corollary 3.4. *Let $X_1, \ldots, X_n$ be independent random variables with $X_i \in \{0,1\}$ and $\Pr[X_i = 1] = p$, for $i = 1, \ldots n$. Set $X := \sum_{i=1}^n X_i$ and $\mu = pn$. Then, for any $\delta \ge 0$, we have*

$$\Pr[X \ge (1+\delta)\mu] \le e^{-\min\{\delta^2,\delta\}\mu/4}.$$

Proof. We may assume that $(1+\delta)p \le 1$. Then Theorem 1.1 gives

$$\Pr[X \ge (1+\delta)pn] \le e^{-D_{\mathrm{KL}}((1+\delta)p\|p)n}.$$

Define $f(\delta) := D_{\mathrm{KL}}((1+\delta)p\|p)$. Then

$$f'(\delta) = p\ln(1+\delta) - p\ln(1-\delta p/(1-p))$$

and

$$f''(\delta) = \frac{p}{(1+\delta)(1-p-\delta p)} \ge \frac{p}{1+\delta}.$$

By Taylor's theorem, we have

$$f(\delta) = f(0) + \delta f'(0) + \frac{\delta^2}{2}f''(\xi),$$

for some $\xi \in [0,\delta]$. Since $f(0) = f'(0) = 0$, it follows that

$$f(\delta) = \frac{\delta^2}{2}f''(\xi) \ge \frac{\delta^2 p}{2(1+\xi)} \ge \frac{\delta^2 p}{2(1+\delta)}.$$

For $\delta \ge 1$, we have $\delta/(1+\delta) \ge 1/2$, for $\delta < 1$, we have $1/(\delta+1) \ge 1/2$. This gives for all $\delta \ge 0$

$$f(\delta) \ge \min\{\delta^2, \delta\}p/4,$$

and the claim follows. □

Corollary 3.5. *Let $X_1, \ldots, X_n$ be independent random variables with $X_i \in \{0,1\}$ and $\Pr[X_i = 1] = p$, for $i = 1, \ldots n$. Set $X := \sum_{i=1}^n X_i$ and $\mu = pn$. Then, for any $\delta > 0$, we have*

$$\Pr[|X - \mu| \geq \delta\mu] \leq 2e^{-\min\{\delta^2, \delta\}\mu/4}.$$

Proof. Combine Corollaries 3.3 and 3.4. □

Corollary 3.6. *Let $X_1, \ldots, X_n$ be independent random variables with $X_i \in \{0,1\}$ and $\Pr[X_i = 1] = p$, for $i = 1, \ldots n$. Set $X := \sum_{i=1}^n X_i$ and $\mu = pn$. For $t \geq 2e\mu$, we have*

$$\Pr[X \geq t] \leq 2^{-t}.$$

Proof. By Corollary 3.2

$$\Pr[X \geq (1+\delta)\mu] \leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu \leq \left(\frac{e}{1+\delta}\right)^{(1+\delta)\mu}.$$

For $\delta \geq 2e - 1$, the denominator in the right hand side is at least $2e$, and the claim follows. □

4 Generalizations

We mention a few generalizations of the proof techniques for Section 2. Since the consequences from Section 3 are based on simple algebraic manipulation of the bounds, the same consequences also hold for the generalized settings.

4.1 Hoeffding-Extension

Theorem 4.1. *Let $X_1, \ldots, X_n$ be independent random variables with $X_i \in [0,1]$ and $\mathbf{E}[X_i] = p_i$. Set $X := \sum_{i=1}^n X_i$ and $p := (1/n)\sum_{i=1}^n p_i$. Then, for any $t \in [0, 1-p]$, we have*

$$\Pr[X \geq (p+t)n] \leq e^{-D_{\mathrm{KL}}(p+t \| p)n}.$$

Proof. The proof generalizes the moment method. Let $\lambda > 0$ a parameter to be determined later. As before, Markov's inequality yields

$$\Pr\left[e^{\lambda X} \geq e^{\lambda(p+t)n}\right] \leq \frac{\mathbf{E}[e^{\lambda X}]}{e^{\lambda(p+t)n}}.$$

Using independence, we get

$$\mathbf{E}[e^{\lambda X}] = \mathbf{E}\left[e^{\lambda \sum_{i=1}^n X_i}\right] = \prod_{i=1}^n \mathbf{E}\left[e^{\lambda X_i}\right]. \tag{4}$$

Now we need to estimate $\mathbf{E}\left[e^{\lambda X_i}\right]$. The function $z \mapsto e^{\lambda z}$ is convex, so $e^{\lambda z} \leq (1-z)e^{0\cdot\lambda} + ze^{1\cdot\lambda}$ for $z \in [0,1]$. Hence,

$$\mathbf{E}\left[e^{\lambda X_i}\right] \leq \mathbf{E}[1 - X_i + X_i e^\lambda] = 1 - p_i + p_i e^\lambda.$$

Going back to (4),

$$\mathbf{E}[e^{\lambda X}] \le \prod_{i=1}^{n}(1-p_i+p_ie^{\lambda}).$$

Using the arithmetic-geometric mean inequality $\prod_{i=1}^{n} x_i \le \left((1/n)\sum_{i=1}^{n} x_i\right)^n$, for $x_i \ge 0$, this is

$$\mathbf{E}[e^{\lambda X}] \le (1-p+pe^{\lambda})^n.$$

From here we continue as in Section 2.1. □

4.2 Hypergeometric Distribution

Chvátals proof generalizes to the *hypergeometric* distribution.

Theorem 4.2. *Suppose we have an urn with N balls, P of which are red. We randomly draw n balls from the urn without replacement. Let $H(N,P,n)$ denote the number of red balls in the sample. Set $p := P/N$. Then, for any $t \in [0, 1-p]$, we have*

$$\Pr[H(N,P,n) \ge (p+t)n] \le e^{-D_{\mathrm{KL}}(p+t\|p)n}.$$

Proof. It is well known that

$$\Pr[H(N,P,n)=l] = \binom{P}{l}\binom{N-p}{n-l}\binom{N}{l}^{-1},$$

for $l = 0, \dots, n$.

Claim 4.3. *For every $j \in \{0, \dots, n\}$, we have*

$$\binom{N}{n}^{-1}\sum_{i=j}^{n}\binom{P}{i}\binom{N-P}{n-i}\binom{i}{j} \le \binom{n}{j}p^j.$$

Proof. Consider the following random experiment: take a random permutation of the N balls in the urn. Let S be the sequence of the first n elements in the permutation. Let X be the number of j-subsets of S that contain only red balls. We compute $\mathbf{E}[X]$ in two different ways. On the one hand,

$$\mathbf{E}[X] = \sum_{i=j}^{n}\Pr[\text{S contains } i \text{ red balls}]\binom{i}{j} = \sum_{i=j}^{n}\binom{N}{n}^{-1}\binom{P}{i}\binom{N-P}{n-i}\binom{i}{j}. \tag{5}$$

On the other hand, let $I \subseteq \{1, \dots, n\}$ with $|I| = j$. Then the probability that all the balls in the positions indexed by I are red is

$$\frac{P}{N}\cdot\frac{P-1}{N-1}\cdot\dots\cdot\frac{P-j+1}{N-j+1} \le \left(\frac{P}{N}\right)^j = p^j.$$

Thus, by linearity of expectation $\mathbf{E}[X] \le \binom{n}{j}p^j$. Together with (5), the claim follows. □

Claim 4.4. *For every $\tau \geq 1$, we have*

$$\binom{N}{n}^{-1} \sum_{i=0}^{n} \binom{P}{i} \binom{N-P}{n-i} \tau^i \leq (1 + (\tau - 1)p)^n.$$

Proof. Using Claim 4.3 and the Binomial theorem (twice),

$$\begin{aligned}
\binom{N}{n}^{-1} \sum_{i=0}^{n} \binom{P}{i} \binom{N-P}{n-i} \tau^i &= \binom{N}{n}^{-1} \sum_{i=0}^{n} \binom{P}{i} \binom{N-P}{n-i} (1 - (\tau - 1))^i \\
&= \binom{N}{n}^{-1} \sum_{i=0}^{n} \binom{P}{i} \binom{N-P}{n-i} \sum_{j=0}^{i} \binom{i}{j} (\tau - 1)^j \\
&= \binom{N}{n}^{-1} \sum_{j=0}^{n} (\tau - 1)^j \sum_{i=j}^{n} \binom{P}{i} \binom{N-P}{n-i} \binom{i}{j} \\
&\leq \sum_{j=0}^{n} \binom{n}{j} ((\tau - 1)p)^j = (1 + (\tau - 1)p)^n,
\end{aligned}$$

as claimed. □

Thus, for any $\tau \geq 1$ and $k \geq pn$, we get as before

$$\begin{aligned}
\Pr[H(N, P, n) \geq k] &= \binom{N}{n}^{-1} \sum_{i=k}^{n} \binom{P}{i} \binom{N-P}{n-i} \\
&\leq \binom{N}{n}^{-1} \sum_{i=0}^{n} \binom{P}{i} \binom{N-P}{n-i} \tau^{i-k} \leq \frac{(p\tau + 1 - p)^n}{\tau^k},
\end{aligned}$$

by Claim 4.4. From here the proof proceeds as in Section 2.2. □

4.3 General Impagliazzo-Kabanets

Theorem 4.5. *Let $X_1, \ldots, X_n$ be random variables with $X_i \in 0, 1$. Suppose there exist $p_i \in [0, 1]$, $i = 1, \ldots, n$, such that for every index set $I \subseteq \{1, \ldots, n\}$, we have $\Pr[\prod_{i \in I} X_i = 1] \leq \prod_{i \in I} p_i$. Set $X := \sum_{i=1}^{n} X_i$ and $p := (1/n) \sum_{i=1}^{n} p_i$. Then, for any $t \in [0, 1 - p]$, we have*

$$\Pr[X \geq (p + t)n] \leq e^{-D_{\mathrm{KL}}(p+t \| p)n}.$$

Proof. Let $\lambda \in [0, 1]$ be a parameter to be chosen later. Let $I \subseteq \{1, \ldots, n\}$ be a random index set obtained by including each element $i \in \{1, \ldots, n\}$ with probability λ. As before, we estimate the probability $\Pr\big[\prod_{i \in I} X_i = 1\big]$ in two different ways, where the probability is over the random choice of $X_1, \ldots, X_n$ and I. Similarly to before,

$$\begin{aligned}
\Pr\Big[\prod_{i \in I} X_i = 1\Big] &= \Pr\Big[\prod_{i \in I} X_i = 1\Big] \leq \sum_{S \subseteq \{1, \ldots, n\}} \Pr\Big[I = S \wedge \prod_{i \in S} X_i = 1\Big] \\
&\leq \sum_{S \subseteq \{1, \ldots, n\}} \Pr[I = S] \cdot \Pr\Big[\prod_{i \in S} X_i = 1\Big] \leq \sum_{S \subseteq \{1, \ldots, n\}} \lambda^{|S|} (1 - \lambda)^{n - |S|} \cdot \prod_{i \in S} p_i. \quad (6)
\end{aligned}$$

We define n independent random variables $Z_1, \ldots, Z_n$ as follows: for $i = 1, \ldots, n$, with probability $1 - \lambda$, we set $Z_i = 1$, and with probability λ, we set $Z_i = p_i$. By (6), and using independence and the arithmetic-geometric mean inequality.

$$\Pr\Big[\prod_{i \in I} X_i = 1\Big] = \mathbf{E}\Big[\prod_{i=1}^{n} Z_i\Big] = \prod_{i=1}^{n} \mathbf{E}[Z_i] = \prod_{i=1}^{n} (1 - \lambda + p_i \lambda) \leq (1 - \lambda + p\lambda)^n. \tag{7}$$

The proof of the lower bound remains unchanged and yields

$$\Pr\Big[\prod_{i \in I} X_i = 1\Big] \geq (1 - \lambda)^{(1-p-t)n} \Pr[X \geq (p+t)n],$$

as before. Combining with (7) and optimizing for λ finishes the proof, see Section 2.3. $\square$