Asymptotic Tail Bound and Applications

IT University of Copenhagen

Thomas D. Ahle

December 15 2017

Abstract

In the field of Computer Science, the Chernoff bound is an extremely useful found bounding the error probabilities of various algorithms. Chernoff gives an exponentially decaying upper bound on the probability that a sum of independent random variables is many standard deviations away from its expectation.

However sometimes we need more than an upper bound, and we need bounds that are tight within a constant factor or better. (There is a fairly standard tail lower bound, but it deviates from Chernoff by a factor of $\sqrt{n}$.)

Questions

In this project I will explore and derive multiple upper and lower bounds for Chernoff using methods ranging from geometric series and generating functions to saddlepoint approximations and laplace approximation. All these results will be known, but they don't seem have a good exposition in Computer Science.

The reason also to consider more simple methods is to help an intuition for deriving similar bounds for different problems. In particular I will derive a tight bound for the size of the intersection between two hamming balls, which does not seem to exist in the literature.

I will use the formulas to answer whether algorithms exists for the following problems: Locality Sensitive Filters in hamming space with limited use of random bits (as opposed to current methods, requiring gaussian samples), LSF with improved performance for low dimensional spaces. (dimension $< 2\log n$) Linear space bit sampling LSH, which I have previously partially analyzed, but only in a different context, in which a full analysis was not necessary.

Contents

Chapter 1

Introduction

Large Deviation Theory, including what is commonly known as Tail Inequalities, is the field of probability theory concerned with the parts of probability distributions, where the Central Limit Theorem, CLT, 'law of large numbers' isn't useful. Usually the distributions can be interpreted as some sort of sum of random variables, for which we would normally seek to apply a 'normal approximation' using the CLT. It is not that the CLT is not correct for these applications. It is simply that it converges way to slowly, and thus gives error terms that by far drown the signal.

Large Deviation Theory is also useful when considering geometry in high dimensional spaces. This is due to probabilities of many independent, or somewhat independent, variables living in a high dimensional product space of these individual variables' behavior. There are thus direct parallels between, say, the tail of the binomial distribution and the volume of spherical caps or small balls in hamming space.

In Computer Science, the most common way to deal with tail probabilities or 'rare events' is the Chernoff bound. There are also other bounds by Hoeffding and many more advanced bounds for variables with degrees of dependence. The main thing to notice is, however, that these are merely 'bounds' and not necessarily asymptotics. They are usually used when people just want to show that something is exponentially unlikely, but don't really care about how exponentially. That is, the constants in the exponents are not claimed to be sharp, and the lower bounds in the tail, when the exist at all, tend to leave a huge gap up to the upper bound.

One area where this is not sufficient is the modern field of High Dimensional Computational Geometry. The seminal paper of Indyk and Motwani studied a data structure for the fundamental Approximate Near Neighbor problem (ANN). This is the classical data structure problem of finding the nearest stored point to a query point, but with the adaptation that there is a certain distance 'gap' separating the 'near' points, which we can return, from the 'far' points, which should be ignored. In this approach to near neighbor data structures, each data point is stored, possibly with duplication, in each region covering it. To answer a query, one simply computes the regions covering the query point and computes the distance to the points stored in these regions. This tends to be much faster than classical approaches, such as CD-trees, when the dimension is large, $\succeq \log n$.

The connection to Large Deviation Theory is, that often these regions are chosen to be the isoperimetric shape of the space, usually some kind of ball. Since, again, the volume of high dimensional balls tend to be related to the tails of probability distributions, we end up needing good bounds on these tails. Often the volume of the shapes are taken to be around $1/n$ of the space, so central limit approximations, with errors of the size $\Omega(1/\sqrt{d})$ are not useful. What's worse, the exponent used for making the balls small, tend to also go into the exponent n^ρ of the final query time of the algorithm. Hence, if one uses a bound that's off by a factor two in the exponent, an $n^{2/3}$ algorithm may turn into a $n^{4/3}$.

Some books on coding theory get close to what we want. Niederreiter's book [**?**], even considers volumes of intersections, which is something we will eventually need for our analysis.

However even when they do that, they don't care about the polynomial dependence on d, which we also want to get right. This will be the difference between a dn^ρ algorithm and a d^5n^ρ algorithm.

However, while modern sources seem to ignore the issue, older sources do not. In this report we will review classical results and discover many new ones. The following section gives the historical context. Then follows a chapter with our mathematical results, and finally a chapter with various applications and new algorithms.

1.1 History

Since De Moivre, most of the focus on binomial approximation has been on the central case. This has given results as resent as the Camp Paulsen approximation [?], which gives an approximation with error uniformly bounded by $0.007/\sqrt{np(1-p)}$. Unfortunately results in this area never seems to give tight bounds when the probability we are interested in is asymptotically smaller than $1/\sqrt{n}$.

Some probabilists have however ventured into the realm of asymptotically sharp 'tail approximations'. The original attack on the binomial distribution was Cramér 1938[?], but unfortunately I haven't been able to recover the article. Cramér also did a lot of more general work on the so called 'rate' of probability distributions, part of which is known today as Cramér's Theorem. It should be noted that Cramér's theorem is not very related to what we discuss in this article.

Later Bahadur and Rao[?] and Littlewood[?] have given approximations of increasing accuracy and complexity. Littlewood examined the binomial distribution in 1969, when he was 84. The proof is 28 pages and is known as his 'last hard paper'[?]. Apparently he was interested in the binomial distribution, because he was surprised by his own ability to 'guess cards' [?]. In 1989 McKay fixed some typos [?], and it is his version we state. The Littlewood bound has error term $(1 + O(1/n))$, which is even better than the Cramér bound, though more complicated.

The exact statement by Littlewood/McKay is the following:

Theorem 1. *Let p, $0 < p < 1$ be fixed. Let $t = t(n)$ be such that $pn + t$ is an integer and $0 \leq \frac{3}{4}qn$. Define $x = t/\sigma$ and $\rho = q - t/n$. Then*

$$\begin{aligned} B(pn+t;n,p) &= Q(x)\exp(A_1 + A_2/\sqrt{\rho(1-\rho)n} + O(1/n)) \\ A_1 &= \frac{t^2}{2pqn} - (pn+t-1/2)\log\left(1+\frac{t}{pn}\right) - (qn-t+1/2)\log\left(1-\frac{t}{qn}\right) \\ A_2 &= \tfrac{1}{6}(1-2\rho)((1-x^2)/Y(x)+x^3) + \tfrac{1}{2}(1/Y(x)-x) \end{aligned}$$

where $B(pn+t;n,p) = \sum_{j=pn+t}^{n}\binom{n}{j}p^jq^{n-j}$ and $\phi(x) = \exp(-x^2/2)/\sqrt{2\pi}$, $Q(x) = \int_x^\infty \phi(y)dy$, $Y(x) = \Phi(x)/\phi(x)$.

In the first chapter we will state this theorem in a (hopefully) slightly more readable version, which compares the result to that of Cramér.

In 1965 Petrov[?, ?] gave generalizations of Cramér for non-identical random variables. Raič, Martin [?] looked at non-independent variables. Short, Michael [?] looked practical approximations for computer calculations. Hwang, Hsien-Kuei[?] give a nice review of various results, including Cramér type bounds for the Poisson Distribution.

Chapter 2

Bounds

2.1 Preliminaries

In this chapter we will use a number of classical results from probability theory and in particular from asymptotic analysis. These are included here for reference, but most readers can likely skip the section.

2.1.1 Entropy

It will often be useful to write our formulas in terms of the following quantities:

$$\mathrm{H}(x) = x\log\frac{1}{x} + (1-x)\log\frac{1}{1-x} \tag{2.1}$$

$$\mathrm{D}(a \parallel b) = a\log\frac{a}{b} + (1-a)\log\frac{1-a}{1-b}. \tag{2.2}$$

As is common, we define $\mathrm{H}(0) = \mathrm{H}(1) = 0$, and $\mathrm{D}(a \parallel b)$ is similarly completed at $a, b = 0$ by its limits.

These are the Binary entropy function and Binary KL-divergence respectively. Intuitively these appear because of the Asymptotic Equipartition Property, which states that on a probability space of n independent variables, can be roughly thought of as consisting of $\exp(n\,\mathrm{H}(X))$ equally likely values. See e.g. [**?**].

2.1.2 Classical binomial bounds

The classical way to bound the binomial distribution, is by the Central Limit Theorem. Berry Esseen gives the quantified version of Moivre-Laplace:

Theorem 2 (Berry Esseen [**?**]). *Let X have zero mean, unit variance, and finite third moment. Let $S_n = (X_1 + \cdots + X_n)/\sqrt{n}$, where $X_1, \ldots, X_n$ are iid copies of X. Then we have*

$$\Pr[S_n \le m] = \Pr[G \le m] + O(\mathrm{E}\,|X|^3/\sqrt{n}) \tag{2.3}$$

uniformly for all $m \in \mathbb{R}$, where $G \sim N(0,1)$.

We will often use the symbols $\phi(x) = \frac{1}{\sqrt{2\pi}}\exp(-y^2/2)$ and $\Phi(x) = \Pr[G \le x] = \int_{-\infty}^{x}\phi(y)dy$ for respectively the pdf and cdf of the normal distribution.

For $X \sim B(n,p)$ binomially distributed, we have mean $\mu = np$ and variance $\sigma^2 = np(1-p)$. As $X = X_1 + \cdots + X_n$ where X_i has Bernoulli distribution, we can write $X' = (X-\mu)\sigma$ and $E|(X_1 - p)/\sqrt{p(1-p)}|^3 = (1-2(1-p)p)/\sqrt{p(1-p)})$ This allows us to apply Berry Esseen:

$$\Pr[X \le \mu - x\sigma] = \Phi(-x) + O(1/\sqrt{p(1-p)n}) = \Phi(-x) + O(1/\sigma). \tag{2.4}$$

Mostly we will consider p bounded away from 0 and 1, in which case we can simply write $\Pr[X \le \mu - x\sigma] = \Phi(-x) + O(1/\sqrt{n})$.

The Berry Esseen bound is nice, when $\Phi(-x)$ is in the order of $1/\sqrt{n}$, however in many useful cases, this is not the case. The most well known bound used outside of this range, in the tails, is the Chernoff bound:

Theorem 3 (Chernoff bound)**.** *Let $X \sim B(n,p)$ be Binomially distributed, and let $m \le n/2$, then*

$$\frac{1}{3\sqrt{m}} \exp(-n\,\mathrm{D}(m/n\|p)) \le \Pr[X \le m] \le \exp(-n\,\mathrm{D}(m/n\|p)). \tag{2.5}$$

Here the upper bound is due to Bernstein [**?**], while the lower bound is simply a bound on the largest term of the sum, namely $\binom{n}{m}p^m(1-p)^{n-m}$. Using the following precise Stirling bound, (see e.g. [**?**]),

$$\sqrt{2\pi}\; n^{n+\frac{1}{2}}e^{-n} \le n! \le e\; n^{n+\frac{1}{2}}e^{-n} \tag{2.6}$$

we get $\binom{n}{m}\exp(-n\,\mathrm{H}(m/n)) \ge \frac{\sqrt{2\pi n}}{e\sqrt{m}e\sqrt{n-m}} \ge \frac{1}{3\sqrt{m}}$.

While the Chernoff bound can give us exponentially small probabilities, it has the issue of not being sharp. For $m \sim n$, it is lose by a factor $\sqrt{n}$, which will be a problem for our eventual applications.

A simple, but very tight bound, which will also be useful eventually, is the following bound tail bound very close to the mean:

Theorem 4 (Uhlmann [**?**] or [**?**])**.** *For $n \ge 2$, let $X_1 \sim B(n, \frac{k-1}{n-1})$ and $X_2 \sim B(n, \frac{k}{n+1})$. Then for $k < \frac{n+1}{2}$*

$$\Pr[X_1 \ge k] < 1/2 < \Pr[X_2 \ge k] \tag{2.7}$$

For $k > \frac{n+1}{2}$ the reverse inequalities hold, and for $k = \frac{n+1}{2}$ (n odd) all three terms are equal.

Corollary 1. *If $X \sim B(n, \frac{k}{n})$, then $\Pr[X \ge k] \ge 1/2$ for all k.*

This follows because $\frac{k}{n} \ge \frac{k-1}{n-1}$ and $\frac{k}{n} \ge \frac{k}{n+1}$, and that the binomial cdf is monotone in p. See also Neuman's bound [**?**] or [**?**] for an asymptotic expansion of this bound.

2.1.3 Asymptotics

We will often not be able to find exact values for the expressions we consider. Hence we'll apply asymptotic analysis. The following expansions will all be useful:

$$e^x = 1 + x + \frac{x^2}{2} + O(x^3) \tag{2.8}$$

$$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} + O(x^4) \tag{2.9}$$

$$\frac{1}{1-x} = 1 + x + x^2 + O(x^3) \tag{2.10}$$

$$(1+x)^\alpha = 1 + \alpha x + \binom{\alpha}{2} x^2 + O(x^3) \tag{2.11}$$

$$\mathrm{H}(x+\epsilon) = \mathrm{H}(x) + \epsilon \log \frac{1-x}{x} - \frac{\epsilon^2}{2x(1-x)} + O(\epsilon^3) \tag{2.12}$$

$$\mathrm{D}(x \parallel x+\epsilon) = \frac{\epsilon^2}{2x(1-x)} + \frac{1}{3}\left(\frac{1}{(1-x)^2} - \frac{1}{x^2}\right)\epsilon^3 + O(\epsilon^4) \tag{2.13}$$

$$n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \left(1 + \frac{1}{12n} + O\left(\frac{1}{n^2}\right)\right) \tag{2.14}$$

$$\Phi(-x)/\phi(x) = \frac{1}{x} - \frac{1}{x^3} + O\left(\frac{1}{x^5}\right) \tag{2.15}$$

$$\Phi(-\epsilon) = \frac{1}{2} + \frac{\epsilon}{\sqrt{2\pi}} - \frac{\epsilon^3}{6\sqrt{2\pi}} + O(\epsilon^5) \tag{2.16}$$

$$\binom{2n}{n} = \frac{2^{2n}}{\sqrt{\pi n}}\left(1 - \frac{1}{8n} + \frac{1}{128n^2} + O\left(\frac{1}{n^3}\right)\right) \tag{2.17}$$

of which most follows from Taylor's Theorem. See e.g. [**?**]. In particular (2.15) is known as the Mill's Ratio [**?**].

Sometimes we will use the $\succ$ operator instead of the more familiar $o(n)$, $\omega(n)$ notation:

$$f(n) \prec g(n) \iff f(n) = o(g(n)) \iff \lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$$
$$f(n) \succ g(n) \iff g(n) \prec f(n) \iff f(n) = \omega(g(n)).$$

This is useful because it allows us to write $1 \prec f(n) \prec \sqrt{n}$ rather than the less readable $f(n) = \omega(1) \cap o(\sqrt{n})$.

In most cases we won't specify the moving variable for our asymptotics. If nothing is specified it will be either d or n going to ∞.

2.1.4 Stirling's approximation

It's well known (see e.g. [**?**]) that we can expand the factorial function as:

$$\log n! = n \log n - n + \frac{1}{2}\log(2\pi n) + \frac{1}{12n} - O(1/n^3) \tag{2.18}$$

We can use that to make a bound for binomial coefficients:

$$\begin{aligned}\log\binom{n}{m} &= \log n! - \log m! - \log(n-m)! \\ &= n\,\mathrm{H}\left(\frac{m}{n}\right) + \frac{1}{2}\log\frac{n}{2\pi m(n-m)} - \frac{1}{12}\left(\frac{1}{m} + \frac{1}{n-m} - \frac{1}{n}\right) + O\left(\frac{1}{m^3} + \frac{1}{(n-m)^3}\right)\end{aligned}$$

We can bound $\frac{1}{n} - \frac{1}{m} - \frac{1}{n-m} \leq -\frac{3}{n}$. However, if we don't know anything about the size of m, the second term may be constant. Hence the most general approximation (we only assume $1 \leq m \leq n/2$) of the binomial is

$$\begin{aligned}\binom{n}{m} &= \sqrt{\frac{n}{2\pi m(n-m)}} \exp(nH(m/n)) \exp\left(-\frac{1}{4n} + O\left(\frac{1}{m^3} + \frac{1}{(n-m)^3}\right)\right) \\ &= \sqrt{\frac{n}{2\pi m(n-m)}} \exp(nH(m/n)) \left(1 + O\left(\frac{1}{n} + \frac{1}{m^3} + \frac{1}{(n-m)^3}\right)\right)\end{aligned}$$

Though with the assumption general in this note, that $m = \Omega(n)$ or x is not $\sqrt{np/q} - \epsilon$, it should be safe to write

$$\begin{aligned}\binom{n}{m} &= \sqrt{\frac{n}{2\pi m(n-m)}} \exp(nH(m/n)) \left(1 - O\left(\frac{1}{n}\right)\right) \\ &= \frac{\exp(nH(m/n))}{\sqrt{2\pi}} \left(\sqrt{\frac{n}{m(n-m)}} - O\left(\frac{1}{n^{3/2}}\right)\right) \\ &= \frac{\exp(nH(m/n))}{\sqrt{2\pi}} \left(\sqrt{\frac{n}{n^2/4 - (1-2p)n^{3/2}x/2 - x^2n/4}} - O\left(\frac{1}{n^{3/2}}\right)\right)\end{aligned}$$

And if further $x = o(\sqrt{n})$:

$$\begin{aligned}\binom{n}{m} &= \frac{\exp(nH(m/n))}{\sqrt{2\pi}} \left(\frac{2}{\sqrt{n}} \sqrt{1 + O\left(\frac{x}{\sqrt{n}}\right)} - O\left(\frac{1}{n^{3/2}}\right)\right) \\ &= \frac{\exp(nH(m/n))}{\sqrt{2\pi}} \left(\frac{2}{\sqrt{n}} + O\left(\frac{x}{n}\right)\right)\end{aligned}$$

2.2 Binomial Distribution

In the preliminaries we reviewed two useful results, which can be used to bound the Cumulative Distribution Function of the binomial distribution $\Pr[X \leq \mu - x\sigma]$: The Berry Esseen bound, for central approximation, and the Chernoff bound, for tail approximation. The first, (2.4), was only sharp when $\Phi(-x) \approx 1/\sqrt{n}$, that is $x \leq \sqrt{\log n}$. The second, (2.5) was always off by around a factor $\sqrt{n}$.

In this section, we review important asymptotic approximations to the tail of the Binomial Distribution, going back to Cramér and Littlewood. We derive various simple, memorable forms, which give $1 + o(1)$ multiplicative approximations to the tail probability in the most important cases. Finally we give a new proof of Cramér's original bound, which shows how to apply the Berry Esseen approximation in the tail of the Binomial and is an order of magnitude shorter than previous proofs.

Without further ado,

Theorem 5 ([?]). *Let X $B(n,p)$ be binomially distributed, let p be bounded away from 0 and 1, and let $m = np - x\sqrt{npq}$ where $x \gg 1$, then*

$$\Pr[X \leq m] = \frac{\Phi(-x)}{\phi(x)} \frac{1}{\sqrt{2\pi}} \exp\left(-n\,\mathrm{D}\left(m/n \| p\right)\right) \left(1 + O\left(\frac{x}{\sqrt{n}}\right)\right). \tag{2.19}$$

The Cramér bound is nearly simple enough to be rememberable: Take the $\exp(-n\,\mathrm{D})$ part from the Chernoff bound and multiply it with the Mills ratio of the normal distribution.

As promised in the introduction, we present a rewriting of Littlewoods theorem 1 in a notation comparable to Cramér:

Theorem 6. *Let p, $0 < p < 1$ be fixed. Let $m = np - x\sqrt{npq}$ be an integer, and $0 \leq \frac{3}{4}(1-p)n$. Then*

$$\Pr[X \geq \mu + x\sigma] = \frac{\Phi(-x)}{\phi(x)}\sqrt{\frac{r}{2\pi}}\exp\left(-n\,\mathrm{D}(m/n\|p)\right)\exp\left(A_2/\sqrt{\tau n} + O(1/n)\right)$$
$$r = \frac{(1-p)m/n}{p(1-m/n)}, \quad \tau = (m/n)(1-m/n)$$
$$A_2 = \frac{2m/n-1}{6}\left((1-x^2)\frac{\phi(x)}{\Phi(-x)} + x^3\right) + \frac{1}{2}\left(\frac{\phi(x)}{\Phi(-x)} - x\right)$$

For large x we can take $A_2/\tau = O(1/(x\sqrt{n}))$ and $\sqrt{r} = 1 + O(x/\sqrt{n})$. This this recovers the Cramér bound.

In many cases however, we are interested in a more gradual trade-off between sharpness and simplicity. In particular we can derive the following simpler bounds:

Corollary 2. *Again, let X $B(n,p)$ be binomially distributed, let p be bounded away from 0 and 1, and let $m = np - x\sqrt{npq}$ where $x \gg 1$, then*

$$\Pr[X \leq m] = \frac{1}{\sqrt{2\pi}x}\exp\left(-n\,\mathrm{D}\left(m/n\|p\right)\right) \quad \left(1 + O\left(\frac{1}{x^2} + \frac{x}{\sqrt{n}}\right)\right) \quad \textit{for } 1 \ll x \ll \sqrt{n} \tag{2.20}$$

$$= \frac{1}{\sqrt{2\pi}x}\exp\left(-x^2/2\right) \quad \left(1 + O\left(\frac{1}{x^2} + \frac{x^3}{\sqrt{n}}\right)\right) \quad \textit{for } 1 \ll x \ll n^{1/6} \tag{2.21}$$

$$= \frac{1}{\sqrt{2\pi}x}\exp\left(-x^2/2\right) \quad \left(1 + O\left(\frac{1}{x^2} + \frac{x^4}{n}\right)\right) \quad \textit{for } 1 \ll x \ll n^{1/4} \textit{ and } p = 1/2 \tag{2.22}$$

The corollary is proven simply by expanding the $\frac{\Phi(-x)}{e^{-x^2/2}} = \frac{1}{\sqrt{2\pi}x}(1 + O(1/x^2))$ ratio, as by (2.15). We have also expanded $\mathrm{D}(p - \epsilon \parallel p) = -\epsilon^2/(2pq) + \epsilon^3(1-2p)/(6p^2q^2) + O(\epsilon)^4$, where notably the ϵ^3 term disappears when $p = 1/2$.

The bounds are fairly easy to remember. For $x = o(n^{1/6})$ they are basically equal to the normal distribution cdf: $Q(x) = \frac{1}{\sqrt{2\pi}x}\exp(-x^2/2)(1 + O(1/x^2))$. These bounds also allow us to see what side of the Chernoff bound (2.5) is sharp: For $x \sim \sqrt{n}$, the lower bound is sharp, while for $x \sim 1$, the upper bound is sharp.

For the sake of completion, we also provide the equivalent bounds for the volume of a hamming ball. Mostly these follow directly from $\mathrm{D}(x \parallel 1/2) = \log 2 - \mathrm{H}(x)$, where H is the binary entropy function:

Corollary 3.

$$\sum_{k=0}^{m}\binom{n}{k} = 2^n P[X \leq m]$$

$$= \frac{\Phi(-x)}{e^{-x^2/2}}\exp\left(n\,\mathrm{H}(m/n)\right) \quad \left(1 + O\left(\frac{x}{\sqrt{n}}\right)\right) \quad \textit{for } 1 \ll x \ll \sqrt{n} \tag{2.23}$$

$$= \frac{1}{\sqrt{2\pi}x}\exp\left(n\,\mathrm{H}(m/n)\right) \quad \left(1 + O\left(\frac{1}{x^2} + \frac{x}{\sqrt{n}}\right)\right) \quad \textit{for } 1 \ll x \ll \sqrt{n} \tag{2.24}$$

$$= \frac{1}{\sqrt{2\pi}x}2^n\exp\left(-x^2/2\right) \quad \left(1 + O\left(\frac{1}{x^2} + \frac{x^4}{n}\right)\right) \quad \textit{for } 1 \ll x \ll n^{1/4} \tag{2.25}$$

It is interesting to note how similarly these bounds are to normal approximations, such as Berry Esseen. However their range is much larger. While we noted that Berry Esseen is only sharp for $x \ll \sqrt{\log n}$, these bounds give sharp result for $x = n^{O(1)}$!

2.2.1 Proofs

As mentioned in the introduction, we won't restate Cramér's original proof. Instead we will derive our own, which is hopefully more intuitive, and certainly shorter, if more specific. After the full proof of theorem 5 we investigate weaker proofs, which will get increasingly close to the multiplicative $(1 + o(1))$ bound we want. The reason for this is to gain a better understanding for which tools are needed, and how the binomials behave, which will help us in the next section on binomial intersections.

The idea of the proof is to use a trick from [?] to 'shift' the mean of the Binomial from np to m. This way we are suddenly close to the center and can apply variations of the central limit theorem. In [?] they approximate the CTL tail as a constant, giving a reasonable (constant) approximation. However we will show how to apply Berry Esseen in the tail, giving the full $(1 + o(1)$ power of the Cramér bound.

Proof. Let $Y \sim B(n, m/n)$ such that $EY = m$.

$$\begin{aligned}
\frac{\Pr[X = m - i]}{\Pr[Y = m - i]} &= \frac{\binom{n}{m-i} p^{m-i}(1-p)^{n-m+i}}{\binom{n}{m-i}(m/n)^{m-i}(1-m/n)^{n-m+i}} \\
&= \left(\frac{p}{m/n}\right)^m \left(\frac{1-p}{1-m/n}\right)^{n-m} \left(\frac{(1-p)m/n}{p(1-m/n)}\right)^i \\
&= \exp\left(-n\,\mathrm{D}(m/n\|p)\right) r^i
\end{aligned}$$

Where r is the 'ratio of odds'. In [?] this was considered a constant, however for us $m \asymp np$, or more precisely for the range of x in which we are interested, we can expand r as $1 - \frac{x}{\sigma}(1 + \frac{x\sigma}{nq})^{-1} = 1 - \frac{x}{\sigma} + O(\frac{x^2}{n})$.

The trick is now to consider

$$\begin{aligned}
\Pr[X \le m] &= \sum_{i=0}^{\infty} \Pr[X = m - i] \\
&= \left[\sum_{i=0}^{\infty} \Pr[Y = m - i] r^i\right] \exp\left(-n\,\mathrm{D}(m/n\|p)\right) \\
&= \left[\sum_{i=0}^{\infty} \left(\frac{1}{\sqrt{2\pi}\tau} \exp\left(\frac{-i^2}{2\tau^2}\right) + S_i\right) r^i\right] \exp\left(-n\,\mathrm{D}(m/n\|p)\right)
\end{aligned}$$

Here $\sum_{i=0}^{\infty} S_i = O(1/\sqrt{n})$. That is, we have managed to 'shift' our probability distribution, such that we can now apply Berry Esseen around the mean, rather than in the tail!

A slight annoyance is that we had to introduce $\tau = \sqrt{VY} = \sqrt{n\frac{m}{n}(1 - \frac{m}{n})}$, the standard deviation of Y. This turns out to be easy to deal with, however, since $\tau^2 = \sigma^2 + (1 - 2p)\sigma x + (\sigma x)^2/n$ or simply $\tau = \sigma(1 + O(x/\sqrt{n}))$.

We also still have the $r^i = \exp[-i(\frac{x}{\sigma} + O(\frac{x^2}{n}))]$ factor to worry about. However we can simply complete the square $\frac{-i^2}{2\tau^2} - i(\frac{x}{\sigma} + c\frac{x^2}{\sigma}) = \frac{-(i/\tau + x\xi)^2}{2} + \frac{(x\xi)^2}{2}$, where $\xi = (1 - c\sigma x/n)\tau/\sigma$ and $c = c(n) = O(1)$. This shifts the normal approximation slightly, but not more than keeping us in the ‘CLT area’. We continue as

$$\begin{aligned}
\frac{\Pr[X \le m]}{\exp\left(-n\,\mathrm{D}(m/n\|p)\right)} &= \int_0^\infty \frac{1}{\sqrt{2\pi}\tau} \exp\left[\frac{-i^2}{2\tau^2}\right] r^i di + O\left(\frac{1}{\sqrt{2\pi}\tau}\right) + \sum_{i=0}^{\infty} S_i r^i \\
&= \int_0^\infty \frac{1}{\sqrt{2\pi}\tau} \exp\left[\frac{-(i/\tau + x\xi)^2}{2}\right] di \exp\left(\frac{(x\xi)^2}{2}\right) + O(1/\sqrt{n}) \\
&= \int_{x\xi}^\infty \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-i^2}{2}\right] di \exp\left(\frac{(x\xi)^2}{2}\right) + O(1/\sqrt{n}) \\
&= \Phi(x\xi) \exp\left(\frac{(x\xi)^2}{2}\right) + O(1/\sqrt{n})
\end{aligned}$$

In the first line we have used that $\sum_{i=a}^b f(i) = \int_a^b f(i)di + O(f(a))$ for decreasing functions. We now expand $\xi = (1 + c'x/\sqrt{n})$, which completes the proof:

$$\begin{aligned}
\Phi(x\xi) \exp\left(\frac{(x\xi)^2}{2}\right) + O(1/\sqrt{n}) &= \frac{\Phi(x)}{\exp(-x^2/2)} \left(1 + \left[x^2 - \frac{\exp(-x^2/2)x}{\sqrt{2\pi}\Phi(x)}\right] O(x/\sqrt{n})\right) + O(1/\sqrt{n}) \\
&= \frac{\Phi(x)}{\exp(-x^2/2)} \left(1 + O(1)O(x/\sqrt{n}) + O(x/\sqrt{n})\right) \\
&= \frac{\Phi(x)}{\exp(-x^2/2)} \left(1 + O(x/\sqrt{n})\right)
\end{aligned}$$

■

Tada! □

Just around one page?

2.2.2 Weaker proofs

It turns out that we can get reasonably close using surprisingly simple techniques. In the remainder of this note, we’ll see how to get increasingly close to the Cramér bound, using only Stirling’s approximation to the binomial coefficient and the sum of powers $\sum_k x^k = (1-x)^{-1}$.

A well known method for upper bounding the binomial distribution is to use the generating function:

$$\begin{aligned}
\Pr[X \le m] &= \sum_{k=0}^{m} \binom{n}{k} p^k q^{n-k} \\
&\le \sum_{k} \binom{n}{k} p^k q^{n-k} z^{k-m} \quad \text{when} \quad z \le 1 \\
&= (q + pz)^n z^{-m} \\
&\le \left(q + p\frac{mq}{(n-m)p}\right)^n \left(\frac{mq}{(n-m)p}\right)^{-m} \\
&= \exp(-n\,\mathrm{D}(m/n\|p)) \qquad (2.26)
\end{aligned}$$

The magical $\frac{mq}{(n-m)p}$ is found by simply solving $\frac{d}{dz}(q+pz)^n z^{-m} = 0$ for z. Also notice, that this is exactly equal to the ‘Chernoff/Markov’ method, as $\Pr[X \le m] = \Pr[z^X \ge z^m] \le \mathrm{E}[z^X]z^{-m} = (q+pz)^n z^{-m}$.

A simple lower bound can be given from just considering the largest coefficient (since $m \leq n/2$), and expanding using Stirling's approximation (see the last section for details):

$$\begin{aligned}
\Pr[X \leq m] &= \sum_{k=0}^{m} \binom{n}{k} p^k q^{n-k} \\
&\geq \binom{n}{m} p^m q^{n-m} \\
&= \frac{\exp(n\,\mathrm{H}(m/n))}{\sqrt{2\pi}} \left(\sqrt{\frac{n}{(n-m)m}} + O(1/n^{3/2}) \right) p^m q^{n-m} \\
&= \frac{\exp(-n\,\mathrm{D}(m/n \| p))}{\sqrt{2\pi}} \left(\frac{1}{\sigma} + O(x/n) \right) \qquad (2.27)
\end{aligned}$$

These two bounds are enough to show that

$$\lim_{n\to\infty} \frac{\log \Pr[X \leq m]}{-n\,\mathrm{D}(m/n\|p)} = 1$$

That is, we get the exponent right. However there is still a multiplicative gap of size σ. This is inevitable, when we try to approximate the sum from just one term, since the total sum may be as large as 1, but the individual elements are never larger than $\sqrt{\frac{2}{\pi n}}$.

Concrete Mathematics [?] show us a way to get a better upper bound. The authors consider bounding $\sum_{k=0}^{m} \binom{n}{k}$ by extracting the largest element, and approximating the tail as a geometric sum. We can use the same technique for the binomial distribution:

$$\begin{aligned}
\Pr[X \leq m] &= \sum_{k=0}^{m} \binom{n}{k} p^k q^{n-k} \\
&\leq \binom{n}{m} p^m q^{n-m} \left(1 + \frac{m}{n-m+1}\frac{q}{p} + \frac{m}{n-m+1}\frac{m-1}{n-m+2}\left(\frac{q}{p}\right)^2 + \dots \right) \\
&\leq \binom{n}{m} p^m q^{n-m} \left(1 + \left(\frac{m}{n-m}\frac{q}{p}\right) + \left(\frac{m}{n-m}\frac{q}{p}\right)^2 + \dots \right) \\
&= \binom{n}{m} p^m q^{n-m} \frac{1}{1 - \frac{m}{n-m}\frac{q}{p}} \\
&= \frac{\exp(-n\,\mathrm{D}(m/n\|p))}{\sqrt{2\pi}} \left(\frac{1}{\sigma} + O(x/n) \right) \left(\frac{\sigma}{x} + p \right) \\
&= \frac{1}{\sqrt{2\pi}x} \exp(-n\,\mathrm{D}(m/n\|p)) \left(1 + O\left(\frac{x}{\sqrt{n}} \right) \right) \qquad (2.28)
\end{aligned}$$

The first thing we notice is that (2.28) is compatible with the Cramér approximation (2.19). We have tightened the upper bound (2.26) by a factor $1/x$.

What remains is to tighten the lower bound equally, such that we essentially have derived the Cramér approximation, up to constant factors.

A simple way to take advantage of more than one element of the distribution, is to take a partial tail, and lower bound it by its smallest element. The trick is to take a long enough tail remains of how many elements to include. To get a tight bound, we need a factor $l = \sigma/x$ over the single element lower bound. Luckily, including this many elements turns out to work very well.

Our previous lower bound (2.27) is already sharp enough when $x = \Omega(\sigma)$, and the Moivre-Laplace is sharp for $x = O(1)$, so in the below, we might as well assume that $x = o(\sigma) \cap \omega(1)$.

$$
\begin{aligned}
\Pr[X \le m] &= \sum_{k=0}^{m} \binom{n}{k} p^k q^{n-k} \\
&\ge l \binom{n}{m-l} p^{m-l} q^{n-m+l} \\
&\ge \binom{n}{m} p^m q^{n-m} l \left(\frac{m-l}{n-m+l}\right)^l \left(\frac{q}{p}\right)^l \\
&= \binom{n}{m} p^m q^{n-m} l \left(1 - \frac{x\sigma + l}{\sigma^2 + xp\sigma + lp}\right)^l \\
&= \binom{n}{m} p^m q^{n-m} l \exp\left(\log\left(1 - \frac{x}{\sigma} + O\left(\frac{1}{\sigma x}\right)\right) l\right) \\
&= \binom{n}{m} p^m q^{n-m} \frac{\sigma}{x} \exp\left(\left(-\frac{x}{\sigma} + O\left(\frac{1}{\sigma x} + \frac{x^2}{\sigma^2}\right)\right) \frac{\sigma}{x}\right) \\
&= \binom{n}{m} p^m q^{n-m} \frac{\sigma}{x} \exp\left(-1 + O\left(\frac{1}{x^2} + \frac{x}{\sigma}\right)\right) \\
&= \frac{\exp(-n\,\mathrm{D}(m/n \| p))}{\sqrt{2\pi}} \left(\frac{1}{\sigma} + O\left(\frac{x}{\sigma^2}\right)\right) \frac{\sigma}{x}\frac{1}{e} \left(1 - O\left(\frac{1}{x^2}\right) + O\left(\frac{x}{\sigma}\right)\right) \\
&= \frac{1}{\sqrt{2\pi} x e} \exp(-n\,\mathrm{D}(m/n \| p)) \left(1 + \left(\frac{1}{x^2}\right) + O\left(\frac{x}{\sqrt{n}}\right)\right)
\end{aligned} \tag{2.29}
$$

Notice that this is exactly equal to (2.20), except for a factor $\frac{1}{e}$. It doesn't appear possible to improve this simply by tweaking l in the above, since $\exp(-1/t)/t$ is largest at $t = 1$. It does appear that using a truncated geometric series can improve things a bit however. In any case, it is nice that we can get such a sharp bound without any fancy technology.

2.3 The Volume of the Intersection of two Hamming Balls

A d-dimensional "Hamming ball" is a set of points in $\{0,1\}^d$, which contains all points within some radius r from some center point x under the hamming metric. Recall that the hamming distance, $|x - y|$, between $x, y \in \{0,1\}^d$ is the number of coordinates on which the two points differ.

Specifically we define the d-dimensional t-ball centered around x:

$$B_d(x,t) = \{p \in \{0,1\}^d : |x - p| \le t\}. \tag{2.30}$$

When we the dimension is implicit, we'll just write $B(x,t)$, and if only the volume is of interest, we may ignore the center and just write $B(r)$[1].

Hamming balls are interesting, among other reasons because of Harper's theorem [**?**], which says that for any set $S \subset \{0,1\}^d$ with $|S| \ge B(t)$ we have that $|S \cup \Gamma(S)| \ge B(t+1)$. Here $\Gamma(S) = \{x \in \{0,1\}^d : |x - y| = 1 \text{ for some } y \in S\}$ is the neighborhood of S. That is, the hamming balls have the smallest surface among subsets with the same volume.

In this section will study the size of the intersection between two hamming balls. This is of particular interest to, among other things, nearest neighbor data structures. As far as I am aware, this has not been presented previously in the literature.

Firstly, to connect this section with the previous one, we note that the size of hamming balls can be described by the binomial sums studied in the previous section. As in that section, we'll

[1] This is allowed, since the hamming distance is translation invariant, so t-balls centered around any coordinate have the same volume.

often write the radius as $t = d/2 - s\sqrt{d}/2$, such that we may consider s the number of 'standard deviations' from the 'mean'. At other times it will make sense to write t as $t = \delta d$, so that δ is the 'relative radius' in the space. This notation also let us be consistent with typical notation from Locality Sensitive Hashing: We'll let r denote the distance between two hamming balls, and t be their radius.

Now to calculate the volume of a t-ball, we consider the number of points at distance exactly r from 0. These the points vectors in $\{0,1\}^d$ with exactly r 1's and $d-r$ 0's. The number of such points are $\binom{d}{r}$, and so

$$B_d(t) = \sum_{s=0}^{t} \binom{d}{r} = \frac{1}{\sqrt{2\pi s}} \exp\left(d\,\mathrm{H}(\delta)\right)\left(1 + O\left(\frac{1}{s^2} + \frac{s}{\sqrt{d}}\right)\right)$$

when $1 \preceq s \preceq \sqrt{d}$. This follows directly from Cramérs theorem (2.23).

For intersections we won't be able to directly apply any theorems like Cramér. The quantities $|\,\mathrm{B}(x,t) \cap \mathrm{B}(y,t)|$ are studied in the theory of error correcting codes, but unfortunately with multiplicative $d^{O(1)}$ errors, which drown the signal when the volume is about $2^d d^{-\Omega(1)}$, as we'll often want it to be.

In particular we show the following bounds:

Theorem 7. *For $t = \frac{d}{2} - \frac{s\sqrt{d}}{2}$, $1 \le s \le \sqrt{d}/2$ and $\delta < 1/2$, let $I = |B_d(x,t) \cap B_d(y,t)|$ be the volume of the intersection between two t-balls at distance r. Then*

$$I \ge \Omega\left(\frac{1}{s^2}\exp\left[(d-r)\,\mathrm{H}\left(\frac{1}{2} - \frac{s/\sqrt{d}}{2(1-\delta)}\right) + r\log 2\right]\right) \tag{2.31}$$

$$I \le O\left(\exp\left[(d-r)\,\mathrm{H}\left(\frac{1}{2} - \frac{s/\sqrt{d}}{2(1-\delta)}\right) + r\log 2\right]\right) \tag{2.32}$$

and for $s = O(d^{1/4})$ we may write this simply as

$$\Omega\left(\frac{1}{s^2}\exp\left[-\frac{s^2}{2(1-\delta)}\right]\right) \le I \cdot 2^{-d} \le O\left(\exp\left[-\frac{s^2}{2(1-\delta)}\right]\right). \tag{2.33}$$

Note that the simplification for $s = O(d^{1/4})$ is analogous to that for the bounds on binomial distribution. This will have interesting consequences later, when we consider applications, as the range $d^{1/4} \prec s \prec d^{1/2}$ will appear in low dimensional, $d \preceq (\log n)^2$, data sets.

The rest of this section is devoted to the proof of theorem 7.

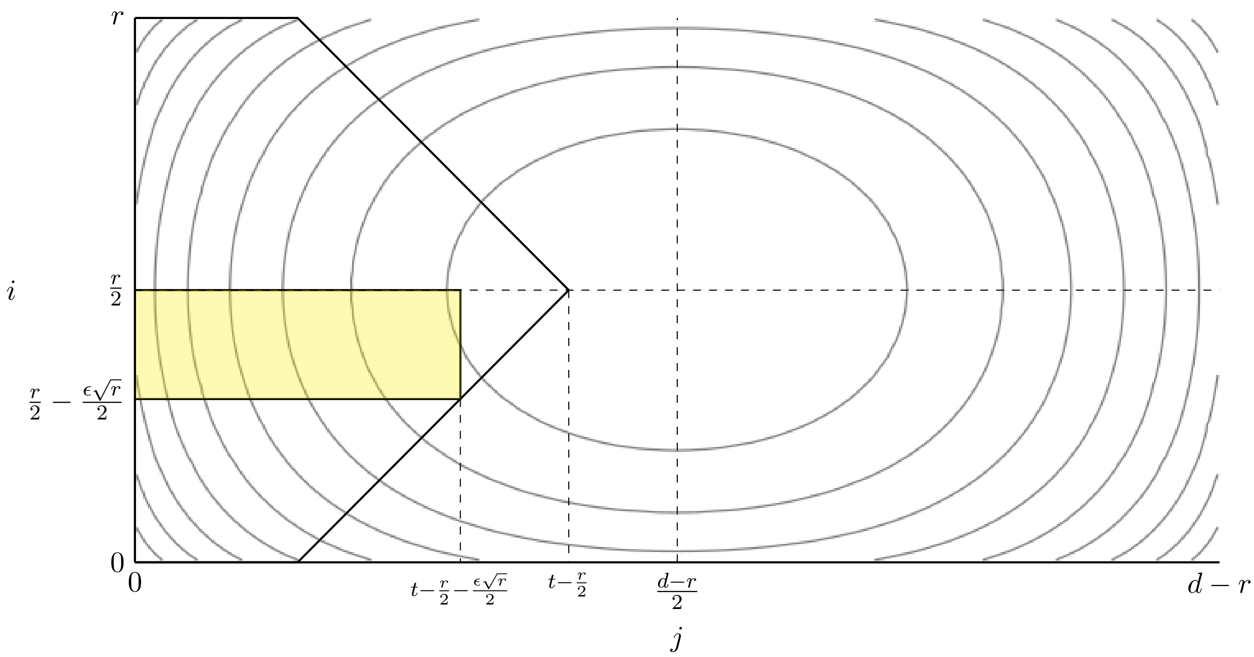
Proof. From figure 2.1 we observe that I may be computed as a two dimensional sum:

$$|\,\mathrm{B}(x,t) \cap \mathrm{B}(y,t)| = \sum_{\substack{i+j \le t \\ j-i \le t-r}} \binom{r}{i}\binom{d-r}{j}. \tag{2.34}$$

Here we are counting over groups of points with the same distance to x and y, namely $i+j$ and $j+r-i$. We can get a quick estimate of the sum, by considering the largest element $\binom{r}{r/2}\binom{d-r}{t-r/2}$ and using $\binom{r}{r/2} \approx 2^r$ and the normal approximation to $\binom{d-r}{t-r/2} = \binom{d-r}{\frac{d-r}{2} - \frac{s\sqrt{d-r}}{2\sqrt{1-\delta}}} = 2^{d-r}\exp(\frac{-s^2}{2(1-\delta)})$, giving combined $I \approx \exp(\frac{-s^2}{2(1-\delta)})$, which promisingly looks a lot like the theorem.

	$d-r$	r
x	0	0
y	0	1
z	j	i

Figure 2.1: To calculate how many points are within distance t from two points x and y, we consider without loss of generality $x = 0 \ldots 0$. For a point, z, lying in the desired region, we let i specify the number of 1's where x and y differ, and j the number of 1's where they are equal. With this notation we get $d(x, z) = i + j$ and $d(y, z) = j + r - i$.



r
i
$\frac{r}{2}$
$\frac{r}{2} - \frac{\epsilon\sqrt{r}}{2}$
0
0
$t - \frac{r}{2} - \frac{\epsilon\sqrt{r}}{2}$
$t - \frac{r}{2}$
$\frac{d-r}{2}$
$d-r$
j

Figure 2.2: A contour plot over the two dimensional binomial. The pentagon on the left marks the region over which we want to sum. For the upper bound we sum i from 0 to r and j from 0 to $t - r/2$. The small yellow rectangle marks the region used for the lower bound.

More rigorously, observe the following two inequalities:

$$\left(B_r\left(\frac{r}{2}\right) - B_r\left(\frac{r}{2} - \frac{\epsilon\sqrt{r}}{2}\right)\right) B_{d-r}\left(t - \frac{r}{2} - \frac{\epsilon\sqrt{r}}{2}\right) \le I \le B_r(r) B_{d-r}\left(t - \frac{r}{2}\right). \tag{2.35}$$

These are illustrated in figure 2.2 and correspond, respectively, to a particular rectangle inside the region specified by (2.34), and a rectangle bounding the region. The inspiration for this particular bound is (2.29), which lower bounded the binomial tail, also by a rectangle.

We proceed to bound the different balls. On the upper bound side we get

$$\begin{aligned} B_r(r) &= 2^r, \text{ and} \\ B_{d-r}\left(t - \frac{r}{2}\right) &= B_{d-r}\left(\frac{d-r}{2} - \frac{s\sqrt{d-r}}{2\sqrt{1-\delta}}\right) \\ &= \frac{1}{\sqrt{2\pi x}} \exp\left[(d-r)\,\mathrm{H}\left(\frac{1}{2} - \frac{s}{2\sqrt{1-\delta}\sqrt{d-r}}\right)\right], \end{aligned}$$

which easily combines to prove (2.32). On the lower bound side we note by Uhlmanns theorem and Berry Esseen:

$$\begin{aligned} B_r(r/2) &\ge 2^{r-1}, \text{ and} \\ B_r\left(\frac{r}{2} - \frac{\epsilon\sqrt{r}}{2}\right) &= (\Phi(-\epsilon) + O(1/\sqrt{d})) \cdot 2^r \\ &= \left(\frac{1}{2} - \frac{\epsilon}{\sqrt{2\pi}} + O(\epsilon^2) + O(1/\sqrt{d})\right) \cdot 2^r. \end{aligned}$$

We will eventually take $\epsilon = 1/x = O(1)$, so this far everything is bounded quite well. The most difficult part is bounded the last term on the lower bound side. We expand as follows, using the rules from the table in the preliminaries:

$$\begin{aligned} B_{d-r}\left(t - \frac{r}{2} - \frac{\epsilon\sqrt{r}}{2}\right) &= \frac{1}{\sqrt{2\pi s}} \exp\left[(d-r)\,\mathrm{H}\left(\frac{1}{2} - \frac{s\sqrt{d-r} + \epsilon\sqrt{r}}{2\sqrt{1-\delta}(d-r)}\right)\right] \\ &= \frac{1}{\sqrt{2\pi s}} \exp\left[(d-r)\,\mathrm{H}\left(\frac{1}{2} - \frac{s}{2\sqrt{1-\delta}\sqrt{d-r}}\right) - \tau\right] \\ \text{where } \tau &= \frac{\epsilon\sqrt{r}}{2\sqrt{1-\delta}} \log\frac{1-\xi}{\xi} + O\left(\frac{\epsilon^2 r}{(d-r)\xi(1-\xi)}\right) \\ \text{and } \xi &= \frac{1}{2} - \frac{s}{2\sqrt{1-\delta}\sqrt{d-r}}. \end{aligned}$$

Here ξ is bounded away from 0, since $s \le \sqrt{d}/2 \le (1-\delta)\sqrt{d}$ by the assumption in the theorem. Since ξ is also bounded away from 1, and δ is bounded away from 0 and 1, the error term on τ reduces to $O(\epsilon)$. Meanwhile $\log\frac{1-\xi}{\xi} = \frac{4s}{2\sqrt{1-\delta}\sqrt{d-r}} + O(\frac{s^2}{d})$, so $\tau = O(\epsilon\sqrt{d})O(s/\sqrt{d}) = O(\epsilon \cdot s)$. Now taking $\epsilon = 1/s$, gives us our intended bound:

$$\begin{aligned} &\left(B_r\left(\frac{r}{2}\right) - B_r\left(\frac{r}{2} - \frac{\epsilon\sqrt{r}}{2}\right)\right) B_{d-r}\left(t - \frac{r}{2} - \frac{\epsilon\sqrt{r}}{2}\right) \\ &\ge 2^r\left(\frac{1}{2} - \left(\frac{1}{2} - O(\epsilon)\right)\right) \frac{1}{\sqrt{2\pi s}} \exp\left[(d-r)\,\mathrm{H}\left(\frac{1}{2} - \frac{s}{2\sqrt{1-\delta}\sqrt{d-r}}\right) - O(\epsilon s)\right] \\ &\ge \Omega\left(\frac{1}{s^2} \exp\left[(d-r)\,\mathrm{H}\left(\frac{1}{2} - \frac{s}{2\sqrt{1-\delta}\sqrt{d-r}}\right)\right] 2^r\right) \end{aligned}$$

which is what we wanted for (2.31).

The simple expression for $s = O(d^{1/4})$ now follows from the expansion of H around 1/2:

$$(d-r)\,\mathrm{H}\left(\frac{1}{2} - \frac{s}{2\sqrt{1-\delta}\sqrt{d-r}}\right) = (d-r)\left(\mathrm{H}\left(\frac{1}{2}\right) - \left(\frac{s}{2\sqrt{1-\delta}\sqrt{d-r}}\right)^2 / \left(2(\tfrac{1}{2})^2\right)\right)$$
$$= (d-r)\log 2 - \frac{s^2}{2(1-\delta)}.$$

□

2.3.1 Charikar regime, $t = d/2$

There is a particular data structure by Moses Charikar [?], which is really an adaptation of earlier algorithms for LPs and SDPs. The algorithm considers the volume of the intersection between two d-dimensional Euclidean half-spheres.

The data structure can also be applied in hamming space (though not with as good guarantees as the algorithm we'll derive in the applications). This requires a bound on hamming ball intersection when the radius is $t = d/2$. To my knowledge, this has not been done before.

The previous bounds don't give us anything useful for t this large. For one thing, they were made with the assumption that $s = \sqrt{d}(1 - 2t/d) = \Omega(1)$, but even if we follow the proofs without that assumption, we just get $I \cdot 2^{-d} = \Theta(1)$, which doesn't really tell us anything useful.

Instead we will show the following

Theorem 8. *For $t = \frac{d}{2}$, let $I = |B_d(x,t) \cap B_d(y,t)|$ be the volume of the intersection between two t-balls at distance r. Then*

$$2^{d-1}\left(\frac{1}{2} + \sqrt{\frac{2}{\pi d}}\right) \le I \le 2^{d-1}.$$

Note that this is tight, at least up to a factor of two. We can compare this to the Euclidean case, in which we get $I \approx 2^{d-1}(1 - \arccos(1-2\delta)/\pi) \sim 2^{d-1}(1 - 2\sqrt{\delta}/\pi)$.

It is possible, that we could use a nice argument to show that the hamming volume must be close to that of the euclidean, or we could use a central limit property, but in this section we simply see how close we can get by the methods we've used so far.

Proof. We split up the sum in three regions. (See figure 2.2 for the intuition.)

$$\sum_{\substack{i+j \le t \\ j+r-i \le t}} \binom{r}{i}\binom{d-r}{j} = 2(D + A + C)$$

where

$$D = \sum_{\substack{i \\ j \le d/2-r}} \binom{r}{i}\binom{d-r}{j}, \qquad A = \sum_{\substack{d/2-r<j\le(d-r)/2 \\ r/2-i<(d-r)/2-j}} \binom{r}{i}\binom{d-r}{j}$$
$$C = \sum_{\substack{d/2-r<j\le(d-r)/2 \\ i=(d-r)/2-j}} \binom{r}{i}\binom{d-r}{j}, \qquad B = \sum_{\substack{d/2-r<j\le(d-r)/2 \\ i>(d-r)/2-j}} \binom{r}{i}\binom{d-r}{j}.$$

Here B is the final part of the lower left quarter of the entire space, such that $A+B+C+D = 2^{d-2}$. We will show that $2C = \binom{d}{d/2}$ and $A \ge B$. This allows us to conclude $2A \ge 2^{d-2} - C - D$, and so

$$
\begin{aligned}
2(D+A+C) &\geq 2^{d-2}+C+D \geq 2^{d-2}+\binom{d}{d/2}/2 \\
&= 2^{d-1}\left(\frac{1}{2}+\sqrt{\frac{2}{\pi d}}-O(d)\right)
\end{aligned}
$$

which is the theorem.

For the first part, notice that C is just the border line splitting regions A and B. Since the line ends in the very center of the space, we can use symmetry and continue it on the other side. That gives us

$$
\begin{aligned}
2C &= \sum_{i=0}^{r}\binom{r}{i}\binom{d-r}{d/2-r+i} \\
&= \sum_{i=0}^{r}\binom{r}{i}\binom{d-r}{d/2-i}
\end{aligned}
$$

which is just a Vandermonde convolution, summing to $\binom{d}{d/2}$ as intended.

To show $A \geq B$ we set up a correspondence between points mirrored in the line represented by C. We'd like to show that for every i, j in the region, we have $\binom{r}{i}\binom{d-r}{j} \geq \binom{r}{r+j-d/2}\binom{d-r}{d/2-r+i}$. But this follows from lemma 1, setting $y = i - r/2$ and $x = j - (d-r)/2$, which completes the proof. □

Lemma 1. *for all $s \geq r$ and $y \geq x$, we have $\binom{2s}{s+y}\binom{2r}{r+x} \geq \binom{2s}{s+x}\binom{2r}{r+y}$.*

Proof. It is sufficient to show

$$
\binom{2s}{s+y}\bigg/\binom{2s}{s+x} = \binom{s-x}{y-x}\bigg/\binom{s+y}{y-x} \geq \binom{r-x}{y-x}\bigg/\binom{r+y}{y-x} = \binom{2r}{r+y}\bigg/\binom{2r}{r+x}
$$

where we have used trinomial revision of the binomial coefficients and monotonicity in the upper value. □

2.4 Conclusion

In this chapter we descried and proven a number of interesting results regarding the binomial distribution and spheres in high dimensions. The bounds have been stated with focus in usability and at different levels of trade-off between tightness and ease of use. Similarly, some of the results have been derived in different ways, hopefully illustrating what level of techniques are required to obtain different levels of tightness. We made use of this knowledge ourselves, when we derived the sphere intersection theorem by inscribing a rectangle, and got an s approximation, similarly to our last lower bound(2.29) of Cramér's theorem.

In particular we showed:

- The tail probabilities of the binomial distribution can be stated very succinctly using the entropy functions and the Mill's ratio of the Normal Distribution. The form we found was a near perfect match with the Chernoff bound, which should make it easy to remember for future uses.

- A proof of Cramérs theorem using only standard calculus and probability theory. All former approaches, including that of Littlewood, seem to use more advanced tools of complex analysis, which is usually lacking in Computer Science curriculum's.

- Upper and lower bounds for the volume of the intersection between two hamming balls. The bounds are tight up to polynomial factors in d, and for large balls, even up to constant factors. The bounds use a combination of combinatorial and geometric insight, and make crucial use of the Cramérs bound from the previous section.

As we will show in the next chapter, the bounds all have strong uses in the field of algorithm analysis, to the extent that useful and simple algorithms have not before been analyzed, because the bounds were lacking.

Open problems from this chapter include:

- Finding similarly tight bounds for other probability distributions, such as Poisson and the Hyper Geometric.
- Finding even tighter bounds for hamming ball intersection, perhaps using a transformation to the central limit theorem, as for Cramér.

Chapter 3

Applications

3.1 Preliminaries

In this report we will be working within the locality sensitive framework of creating high dimensional data structures. We make enough changes to the common data structures from these fields, that there is no big need for stating their theorems here.

We should however state the definition of the problem that these are all solving:

Definition 1 (Approximate Near Neighbour, ANN)**.** *Given a set of points $P \subseteq \{0,1\}^d$ of size $|P| = n$, let r and cr be distances under the hamming distance $\mathrm{d}(x,y) = |x - y|$, defined by the number of positions at which x and y differ.*

A solution to the (r, cr)-ANN problem is a data structure that supports the following query operation: on input $q \in \{0,1\}^d$, for which there exists a point $x \in P$ with $\mathrm{d}(q,x) \le r$, return some $x' \in P$ with $\mathrm{d}(x', q) \le cr$.

3.2 Time/Space Tradeoffs for LSH in Hamming Space

In this section we build a data structure for the Approximate Near Neighbor problem (ANN) fundamental problem in hamming space.

For some parameters to be specified: k, m integers and $x \in [0,1]$, our data structure uses m hash functions $h_i : \{0,1\}^d \to \{0,1\}^k$. This is done by, for each i, sampling k coordinates independently and uniformly at random. The exact building and querying procedures are as follows:

Building the data structure Given a set of points $P \subseteq \{0,1\}^d$ of size $|P| = n$, we create m hash tables T_i. For each point $p \in P$, we store p in $T_i[h_i(p)]$ for all $1 \le i \le m$. Note that we allow multiple points to be stored in the same hash-table bucket.

Querying the data structure Given a point $q \in \{0,1\}^d$, we consider all points in buckets $T_i[a]$ for $1 \le i \le m$ and $|a - h_i(p)| \le xk$. For each such point p we calculate $|q - p|$ and return the point if this distance is less than r.

See figure 3.1 for a picture of the data structure.

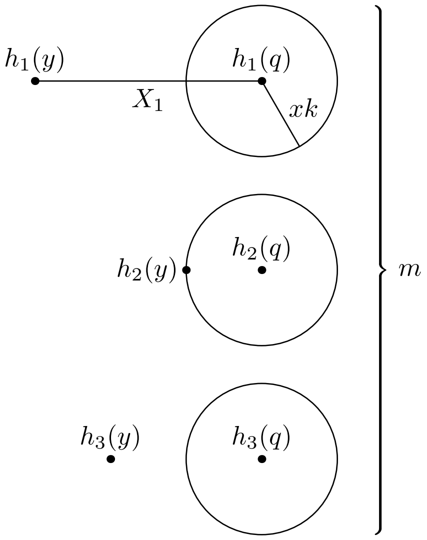


Figure 3.1: The m repetitions have independent projected distances $X_i = |h_i(y) - h_i(q)| \sim \mathrm{Binomial}(k, \delta)$.

	General expressioon	Small $c = 1 + \epsilon$	Large $c \to \infty$
Minimize space	$\rho_q = H(\delta)/(D(\delta\|\|c\delta) + H(\delta))$	$\rho_q \le 1 - 0.146\epsilon^2 + O(\epsilon^3)$	$\rho_q = O(\ln c/c)$
$(x = \delta)$	$\rho_u = 0$	$\rho_u = 0$	$\rho_u = 0$
Balance costs	$\rho_q = \ln(1-\delta)/\ln(1-c\delta)$	$\rho_q \le 1 - \epsilon + O(\epsilon^2)$	$\rho_q = 1/c$
$(x = 0)$	$\rho_u = \ln(1-\delta)/\ln(1-c\delta)$	$\rho_u \le 1 - \epsilon + O(\epsilon^2)$	$\rho_u = 1/c$
Minimize time*	$\rho_q = 0$	$\rho_q = 0$	$\rho_q = 0$
	$\rho_u = H(\delta)/D(\delta\|\|c\delta)$	$\rho_u \le 1/(0.146\epsilon^2) + O(1/\epsilon)$	$\rho_u = O(\ln c/c)$

Table 3.1: The extreme points of the trade-off. Answering a query takes time $O(n^{\rho_q})$, updates take time $O(n^{\rho_u})$ and space usage is $O(n^{1+\rho_u})$. The 0.146 constant becomes 0.721 for large delta, which is optimal for this type of lsh.

The idea of m and x are to balance the space and time used by the data structure. If v is the volume of a hamming ball of radius xk, the total memory usage is going to be mn and the time usage mv plus the distance computations between q and data points. If we use a small m, the space usage decreases, but so does the probability of the querying algorithm succeeding in finding an actual close point. This has to be balanced by increasing v, which in turn increases the time usage.

More precisely we get the following results:

Theorem 9. *For any $r = \delta d$, $c \ge 1$ and $x \in [0, \delta]$ chosen by the user, the algorithm described uses memory $O(n^{1+\rho_u})$, query time $O(n^{\rho_t})$ and solves the (r, cr) ANN problem with constant probability. For*

$$\rho_q = \frac{D(x||\delta) + H(x)}{D(x||c\delta) + H(x)} = \frac{x \ln \frac{1}{\delta} + (1-x)\ln\frac{1}{1-\delta}}{x \ln \frac{1}{c\delta} + (1-x)\ln\frac{1}{1-c\delta}}$$

$$\rho_u = \frac{D(x||\delta)}{D(x||c\delta) + H(x)} = \frac{x \ln \frac{x}{\delta} + (1-x)\ln\frac{1-x}{1-\delta}}{x \ln \frac{1}{c\delta} + (1-x)\ln\frac{1}{1-c\delta}}$$

The most interesting cases of theorem 9 are shown in table 3.1.

Note however that there is an issue with the theorem as stated: As δ goes to 0, $\rho_q = H(\delta)/(D(\delta||c\delta) + H(\delta))$ goes to 1. That is, as stated, the theorem doesn't even beat brute force, when δ is not bounded away from 0. After the proof of the theorem, we will show how to get fix this problem and get the result shown in the table.

Proof. We first show how to set the parameters, such that any point within distance r is found with constant probability. Then we show bound and optimize the space and time usage.

We will make heavy use of the hamming ball volume bound (2.24) (repeated below) and the corresponding bound for the tail of the binomial distribution.

$$V(xk) = \sum_{i=0}^{xk} \binom{k}{i} = \Theta(\exp(k\,\mathrm{H}(x))/s)$$

where $x \in [0, 1/2]$ and $s = (1-2x)\sqrt{k}$ is the number of standard deviations xk is from k.

Note that if s was asymptotically small, such as $O(k^{1/4})$, we would be able to use a simpler bound, such as (2.25). However in our case we let x take a wide range of values, and so we need the more general entropy bound. On the other hand, note that with a simple Chernoff bound, such as (2.5), we would get the $1/\sqrt{k}$ factor in the lower bound, but not the upper bound. This would eventually give us space and query times which were a factor $\sqrt{\log n}$ too large, and thus prevent us from ever getting down to $O(1)$ space or query time as in the theorem.

Correctness To show correctness of the algorithm, we need that for any point $p \in P$ with $|p - q| \leq r$, with constant probability, one of the buckets $T_i[a]$ visited in the query step contains p. The probability of this happening for a particular i is equal to the probability that $|h_i(p) - h_i(q)| \leq xk$. This is because we search all keys within radius xk of $h_i(q)$. Doing $m = 1/\Pr[|h_i(p) - h_i(q)| \leq xk]$ repetitions thus gives us correctness.

We can analyze this probability exactly: Let $X = |h_i(p) - h_i(q)|$, then X is the sum of k binary random variables $X_j \in \{0, 1\}$ where $X_j = 1$ exactly when h_i sampled coordinate j from those where p and q differ. Since p and q differ on $r = \delta d$ coordinates by assumption, this means X is binomially distributed $X \sim B(k, \delta)$. We can now plug in (2.20) to get

$$\Pr[X \leq xk] = \Theta(\exp(-k \operatorname{D}(x \| \delta))/s)$$

where $xk = EX - s\sqrt{VX} = \delta k - s\sqrt{\delta(1-\delta)k}$ or equivalently $s' = \sqrt{k}(\delta - x)/\sqrt{\delta(1-\delta)}$.

Space and usage The algorithm stores each point $p \in P$ once for each of the m hash-tables. Hence it uses space (mn). (We assume the actual points are stored elsewhere in $O(dn)$ space, such that we can simply refer them by unit space pointers.)

The expected time usage of the algorithm is $mV(xk)$ plus the number of distance calculations. In the worst case, every point in P has distance cr, which is minimal for points that should not be reported. By the same argument as for correctness above, we have to look at one of these points exactly when for some i, $|h_i(p) - h_i(q)| \leq xk$. Hence by linearity of expectation, the total time spent for a query is

$$mV(xk) + nm \Pr[|h_i(p) - h_i(q)| \leq xk].$$

This now suggests a way to choose the parameter k, let's set the two terms to be equal: $V(xk) = n \Pr[|h_i(p) - h_i(q)| \leq xk]$. The distance is binomially distributed $B(k, c\delta)$, so we roughly want

$$\exp(k \operatorname{H}(x)) = n \exp(-k \operatorname{D}(x \parallel c\delta)).$$

We see that setting $k = \frac{\log n}{\operatorname{H}(x)+\operatorname{D}(x \parallel c\delta)}$ suffices.

Plugging in our choices for m and k, we get space and time usage:

$$\begin{aligned}
\text{space} &= O(nm) \\
&= O(n \exp(k \operatorname{D}(x\|\delta))s') \\
&= O\left(n^{1+\frac{\operatorname{D}(x \parallel \delta)}{\operatorname{H}(x)+\operatorname{D}(x \parallel c\delta)}} \sqrt{\frac{\log n}{\operatorname{H}(x)+\operatorname{D}(x \parallel c\delta)}} \frac{\delta - x}{\sqrt{\delta(1-\delta)}}\right). \\
&= O\left(n^{1+\rho_u}\sqrt{\log n}\right). \\
\text{time} &\leq 2 \cdot O(mV(xk)) \\
&= O(\exp(k \operatorname{D}(x\|\delta))s' \exp(k \operatorname{H}(x))/s) \\
&= O\left(n^{\frac{\operatorname{D}(x \parallel \delta)+\operatorname{H}(x)}{\operatorname{H}(x)+\operatorname{D}(x \parallel c\delta)}} \frac{\delta - x}{\sqrt{\delta(1-\delta)}} \frac{1}{1-2x}\right) \\
&= O\left(n^{\rho_q}\right).
\end{aligned}$$

Finally, since the above bounds require x to be bounded away from 0 and δ, we note what happens in the extreme cases, when we set $xk = 1$ or $x = \delta$. In the first case we simply recover single probe LSH ala [**?**], which means we get space and time usage $O(n^{1+1/c})$ and $O(n^{1/c})$. In the other case we get a single hash-table data structure. Since the median of a binomial distribution is within a constant of its mean, searching a single ball of radius δk gives us constant probability of collision as wanted, so we can forget about the $\sqrt{\log n}$ which fell out of the analysis above.

□

As mentioned, theorem 9 has a problem for small value of δ. One solution to this problem is to take each of the k bits as the xor of t sampled 'sub bits'. This can be seen as similar to parity codes from coding theory. With this construction, a bit in $h(q)$ will differ from one in $h(y)$ exactly when we have sampled an odd number of different sub bits. Since the number of different sub bits is binomially distributed, we can use that a binomial variable is odd with probability $(1-(1-2p)^t)/2$. The distance's $X_i = |h_i(y) - h_i(q)|$ are then binomially distributed over this probability. That is, $X_i \sim \text{Binomial}(k, (1-(1-2\delta)^t)/2)$.

An alternative construction is to concatenate the t sub bits, and hash them with a 2-independent hash function. The probability for two bits being different is then $1 - ((1-\delta)^t + (1-(1-\delta)^t)(1/2)) = (1-(1-\delta)^t)/2$. This is nearly the same, but the first approach has the advantage of being able to recover our basic scheme for $t = 1$.

The optimal value of t turns out to be $t \approx \lceil 1/(4c\delta) \rceil$. In the case where kt is comparable to the dimension, we may want to make a different analysis.

In table 3.1, the constant 0.146 is derived using this distribution for the worst case $d \to 0$. If we instead use the best case $d = 1/(2c)$ we get the constant $1/(2\ln 2) \approx 0.721$. This constant may be recognized as what you get when you analyze classical bit sampling with $\delta = 1/(2c)$, in which case you get $\rho = \ln(1-1/(2c))/\ln(1/2) \approx 1/(2c\ln 2)$. Thijs Laarhoven gets (for hamming space) the constant 1. Certainly getting that would require using ball hashing. It is not clear what constant we should aim for at small δ. Maybe something around 0.146 is really optimal.

To optimize t for small c, let's first assume it doesn't go to infinity at $c = 1$. That is, we can assume t is constant with respect to c.

If we take $t = t(\delta, c)/(2\delta)$ then $\alpha = \frac{1-(1-2\delta)^{t/\delta}}{2} = \frac{1-e^{-2t}}{2} + O(\delta)$ and $\alpha_c = \frac{1-(1-2c\delta)^{t/\delta}}{2} = \alpha + \frac{t(1-2\alpha)}{1-2\delta}(c-1) + O(c-1)^2$.

$$\begin{aligned}
\frac{H(\alpha)}{D(\alpha||\alpha_c) + H(\alpha)} &= 1 - \frac{D(\alpha||\alpha_c)}{H(\alpha)} + O(D(\alpha, \alpha_c)^2) \\
&= 1 - \frac{D(\alpha||\alpha + \frac{t(1-2\alpha)}{1-2\delta}(c-1) + O(c-1)^2)}{H(\alpha)} + O(D(\alpha, \alpha_c)^2) \\
&= 1 - \frac{(t(1-2\alpha)(c-1))^2}{2(1-2\delta)\alpha(1-\alpha)H(\alpha)} + O(c-1)^4 \\
&= 1 - \frac{t^2(e^{-2t} + O(\delta))^2}{2(1-2\delta)(\frac{1-e^{-4t}}{4} + O(\delta))(H(\frac{1-e^{-2t}}{2}) + O(\delta))}(c-1)^2 + O(c-1)^4 \\
&= 1 - \left(\frac{2t^2 e^{-4t}}{(1-e^{-4t})H(\frac{1-e^{-2t}}{2})} + O(\delta) \right)(c-1)^2 + O(c-1)^4
\end{aligned}$$

Optimizing numerically, the best value turns out to be $0.146742\ldots$ at $t \approx 0.242246\ldots$.

When working with larger c, we should also set t with respect to that. Something like $t \approx 1/(4c\delta)$ appears to work best, but a more careful analysis would have to be done.

Minimizing time If we want to minimize query time, at the cost of update time and memory usage, we can place extra points around the query point. The analysis is the same, except now the memory usage is mv and the time usage is mnp. We're mostly interested in the extreme case, so we set $m = 1$, $1 = np = n\exp(-kD(\delta||c\delta))$. Then $k = (\ln n)/D(\delta||c\delta)$ and $v = \exp(kH(\delta)) = n^{H(\delta)/D(\delta||c\delta)}$. We do however run into the same problem with small δ as in the case of minimal space, so we make sub-bits using the xor method. The results are in the table.

Comparison with previous work The most glaring issue in table 3.1 is the $\ln c$ in the denominator in the case of large c. This is better than Panigrahi's $2/\sqrt{c}$, but not better than

Kapralov's $4/(c+1)$. However interestingly, if we parametrize in terms of δ and $c\delta$, rather than δ and c. For a better comparison with Kapralov, see figure 3.2.

Conclusion It's not clear to me why this simple construction hasn't been tried before. It may be that others have not properly balanced the expected mismatches with the size of the hamming ball. Alternatively people haven't looked further into the parametrization, due to the nonoptimal $O(\ln c/c)$ exponent for large c.

Future work In this note, we have either added a ball around the query point or the data points. It is interesting to consider having balls around both points, however this analysis is a bit more involved, since now each far point may contribute a lot more than in the two simple cases. We may also wonder how the scheme can be improved to yield the correct asymptotic ρ_q for large c. It looks like we need a less symmetric hash function, so that we can look at only parts of the available space, ala Panigrahi. If we want to get rid of false negatives, we can use Lotto Designs. However I don't know any good Lotto constructions at present time. In particular we need a small $L(n,k,p,kp/n)$ design for the linear space case, and an $L(d,k,(1-\delta)d,(1-x)k) = L(d,d-k,r,r-xk)$ design of size around $\exp(kD(x||\delta))$ for the trade-off. Well, for the small design we can just use partitioning in n/k parts. Easy!

3.3 Locality Sensitive Filters in Hamming Space

In this section we build a data structure for the Approximate Near Neighbor problem (ANN) fundamental problem in hamming space. We consider the following construction:

Randomly, uniformly and independently, sample m points $F \subseteq \{0,1\}^d$. For parameters k and t to be determined, we'll be interested in F^k, which we'll call the 'filters'. A point $x \in \{0,1\}^d$ is said to be "caught by the filter f_i" if $f_i \in F^k$ and for all $a \in f_i$ we have $|x-a| \leq t$. That is, if x is in the intersection of all t balls centered at points in f_i.

Building the data structure Given a set of points $P \subseteq \{0,1\}^d$ of size $|P| = n$, we create a single hash table T with keys from F^k. For each point $p \in P$, we store p in $T[f_i]$ for all filters $f_i \subseteq F^k$ such that x is caught by f_i. Note that we allow multiple points to be stored in the same hash table bucket.

Querying the data structure Given a point $q \in \{0,1\}^d$, we compute the set of filters f_i catching q as in the building step. We then look at all points x in buckets $T[f_i]$ and compute their distances to q, $|x-q|$, returning if this value is less than or equal to r.

We will show that this simple data structure gives the following result:

Theorem 10. *For any $r = \delta d$ and $c \geq 1$ chosen by the user, the algorithm described uses memory $O(n^{1+\rho})$, query time $O(n^\rho)$ and solves the (r, cr) ANN problem with constant probability. For*

$$\rho = \frac{1-c\delta}{1-\delta}\frac{1}{c} + O(\tfrac{\log\log n}{\log n}) \leq \frac{1}{c} + O(\tfrac{\log\log n}{\log n})$$

In the case where far points are considered 'random noise', so $c \approx d/2$ and $\delta = 1/(2c)$, the theorem gives us $\rho = \frac{1}{2c-1} + o(1)$. Such a data structure has been created before, (e.g. [**?**] and [**?**]), however these algorithms used samples of gaussians requiring infinite random bits.

Proof. For the analysis, let x and y be arbitrary points in $\{0,1\}^d$ with $|x-y|=r=\delta d$, let f be an arbitrary filter in F^k and let a be an arbitrary point in F. The following three probability estimates will be of great importance in the analysis of the algorithm:

$$\begin{aligned} p^k &= \Pr[x \in T[f]] = (\Pr[|x-a| \le t])^k && = (|B(x,t)|\,2^{-d})^k \\ p_1^k &= \Pr[x \in T[f], y \in T[f]] = (\Pr[|x-a| \le t, |y-a| \le t])^k && = (|B(x,t) \cap B(y,t)|\,2^{-d})^k. \end{aligned}$$

We define p_2 similarly to p_1, but for two points at distance cr rather than r.

We first show that a close point within radius r is indeed found, if there is one. This happens exactly when at least k of the points in F land within distance t of the query point and data point, so both go in the same bucket of T. Because the points of F are independent, this is binomially distributed, so we require $\Pr[B(m,p_1) \ge k] \ge 1/2$. Taking $mp_1 = k$ is sufficient for this (as the median and mean coincide for binomial distributions with integer mean [**?**]).

To analyze performance, we have three main contributions: (1) Calculating the buckets to look in, (2) Looking in these (possibly empty) buckets, and (3) computing the distance to the points in these buckets. Note that (2) is always dominated by (1), and that (3) is only needed at query time, not when adding a point to the data structure. The space requirement will be equal to (2) times n.

For calculating the buckets, we can find all points in F within distance t in time m, and then take all subsets of this in constant time per bucket. By linearity of expectation, the number of buckets is $(mp)^k$, and the number of far points we have to look at is at most $n(mp_2)^k$. Balancing the later to terms, suggest setting $k = \frac{\log n}{\log p/p_2}$. Then the total expected time for (1) + (2) + (3) is bounded by

$$\begin{aligned} m + 2(mp)^k &= \frac{k}{p_1} + 2\left(\frac{kp}{p_1}\right)^k \\ &= n^{\frac{\log k/p_1}{\log n}} + n^{\frac{\log p/p_1}{\log p/p_2} + \frac{\log k}{\log p/p_2}}. \end{aligned} \tag{3.1}$$

We are now ready to apply our bounds from theorem 7. We'll write $t = d/2 - s\sqrt{d/4}$, where $1 \preceq s \preceq \sqrt{d^{1/4}}$:

$$\begin{aligned} \log 1/p &= \tfrac{s^2}{2} + O(\log s) \\ \log 1/p_1 &\le \tfrac{s^2}{2}\tfrac{1}{1-\delta} + O(\log s) \\ \log 1/p_2 &\ge \tfrac{s^2}{2}\tfrac{1}{1-c\delta} + \Omega(1). \end{aligned}$$

Plugging this into (3.1), and taking $\frac{\log 1/p_1}{\log n} = \frac{\log p/p_1}{\log p/p_2}$ to balance (1) with (2) and (3), gives us $s^2/2 = \frac{1-c\delta}{c}\log n$ and

$$(1) + (2) + (3) \le n^{\frac{1-c\delta}{1-\delta}\frac{1}{c} + O(\frac{\log\log n}{\log n})}$$

which is the theorem. □

To summarize we took $k = \frac{\log n}{\log p/p_2}$ and $m = n^\rho$. These parameters are of course integer, but the exact rounding doesn't really matter, since both quantities are $\omega(1)$, and so the rounding errors are lower order terms.

A very interesting thing to notice is, that in our the proof above, we chose $s = \Theta(\sqrt{\log n})$. However the bounds we used were only really valid for $1 \prec s \prec d^{1/4}$. We should also consider what happens when $s = \omega(d^{1/4})$ or equivalently $d = o(\log n)^2$. Intuitively this should help us, since we can take $m = |F|$ nearly exponential in d. With this many points in F, we can set the ball radius t to be substantially smaller than $d/2$. This is a nice, since if we could just take t to be $cr/2 = c\delta d/2$, we would be able to solve the problem with no approximation at all. We won't be able to set t that small though, but we can indeed get an improvement for the dense regime.

3.3.1 Low Dimensionality Regime

We'll write $d = \kappa \log n$ and $t = \tau d$ to symbolize this change in regime. We no longer need to worry about s and simply bound it by d. Using the general bounds from theorem 7 and corollary 3,

$$\begin{aligned}\log 1/p &= d \log 2 - d\,\mathrm{H}(\tau) + O(\log d)\\ \log 1/p_1 &\le d \log 2 - (d-r)\,\mathrm{H}\left(\frac{\tau - \delta/2}{1-\delta}\right) - r\log 2 + O(\log d)\\ \log 1/p_2 &\ge d \log 2 - (d-r)\,\mathrm{H}\left(\frac{\tau - c\delta/2}{1-c\delta}\right) - r\log 2 + \Omega(1).\end{aligned}$$

For simplicity we'll only consider the 'random' case $\delta = 1/(2c)$. The equation $\frac{\log 1/p_1}{\log n} = \frac{\log p/p_1}{\log p/p_2}$, which we use to set t, now becomes

$$\frac{\log 2 - (1-\frac{1}{2c})\,\mathrm{H}\left(\frac{4c\tau-1}{2(2c-1)}\right) - \frac{\log 2}{2c}}{1/\kappa} = \frac{(1-\frac{1}{2c})\,\mathrm{H}\left(\frac{4c\tau-1}{2(2c-1)}\right) + \frac{\log 2}{2c} - \mathrm{H}(\tau)}{\frac{1}{2}\,\mathrm{H}\left(\frac{4\tau-1}{2}\right) + \frac{1}{2}\log 2 - \mathrm{H}(\tau)}. \tag{3.2}$$

This equation doesn't allow any simple solutions for τ, but we can solve it numerically for different c and κ. See figure 3.3 and 3.4. The results are quite substantial, even for moderately large values of κ, and beats the dimension-oblivious $1/(2c-1)$ bound for all $\kappa = O(1)$.

For larger values of κ and c, we can consider the assignment $s^2/2 = \frac{1-c\delta}{c}\log n$ from before, which suggests taking $\tau = (1 - 1/\sqrt{c\kappa})/2$. If we do that, and expand for large c, we get approximately $\frac{1}{2c-1+\frac{4}{3k}}$, which is an improvement over $\frac{1}{2c-1}$ for all k, but only in the second term.

The best results are thus those we get for small values of c, close to 1. It turns out that our derived algorithm perfectly bridges the gap down to a certain kind of 'brute force' algorithm.

3.3.2 An Even Simpler Algorithm for Even Sparser Data

From looking at (3.2), one might wonder what happens in the case $t \le 1/4$, when the denominator on the right hand side becomes undefined. What is happening is that the balls becomes small enough that p_2 is 0. We should consider what interesting possibilities that gives us.

In particular in means that we have no reason in ever making τ smaller than $1/4$. For example $\tau = 1/3$ wouldn't help us at all in avoiding collisions with far points, compared to $\tau = 1/4$, but we would have to use more balls to capture the close points.

In this case, we can fall back to a simple algorithm: Choose m points for filters, and at query time report the first point you share a filter with. This takes time $1/p_1$, since we need enough filters to have a good chance of colliding with the near points. And

$$\begin{aligned}1/p_1 &= O\left(\exp\left[d(1-\frac{1}{2c})(\log 2 - \mathrm{H}\left(\frac{c-1}{2(2c-1)}\right))\right]\right)\\ &= O\left(n^{\kappa(1-\frac{1}{2c})(\log 2 - \mathrm{H}\left(\frac{c-1}{2(2c-1)}\right))}\right),\end{aligned}$$

which is of course just equal to the left side of (3.2). An interesting case of this algorithm, is when we want to solve the exact nearest neighbor problem, $c = 1$. Using the simple algorithm above, we get

Theorem 11. *For any $r = \delta d$ chosen by the user and dimension $d = \kappa \log n$, the 'simple algorithm' uses memory $O(n^{1+\rho})$, query time $O(n^\rho)$ and solves the exact r-NN problem with constant probability. For*

$$\rho = \kappa \log 2(1-\delta).$$

In particular for $r = d/2$, we get sub-linear query time and sub-quadratic memory whenever

$$\kappa < \tfrac{\log 2}{2} \approx 2.885.$$

Note that the simple brute force algorithm takes time $\approx 2^d = n^{\kappa \log 2}$ for the same problem and $r = d/2$. Thus, the brute force algorithm doesn't give sub-linear query time for any dimension $\geq \log_2 n$.

3.4 Conclusion

In this chapter we have developed a number of simple algorithms, which all have one thing in common: Without the tight analysis permitted by chapter 1, we would not have been able to analyze them with enough precision to get interesting results. One might wonder how many people have conjured said algorithms, but scrapped them, because their lose bounds were unable to show the interesting behavior of the algorithms.

In particular we found:

- A complete space/time trade-off for the classical bit sampling LSH algorithm. Previous work has considered the edge cases: linear space or constant time queries, but perhaps because the lacked tight and easy binomial asymptotics, never for the full trade-off.
- A linear space LSH algorithm with sub-linear query time for any $c > 1$. Until this article, such algorithms were only known using the LSF framework, and often considered unsuitable for practical use.
- An optimal data independent LSF data structure for hamming space, using only a finite amount of randomness. Previous results all used Gaussian random vectors, which require an infinite number of random bits.
- An LSF data structure beating the $1/(2c-1)$ bound, for any dimension $d = \kappa \log n$, $\kappa = O(1)$. In particular we can solve even the exact nearest neighbor problem in time $\tilde{O}(2^{d-r})$, which is only possible in hamming space due to certain properties of hamming ball intersections.

Open problems from this chapter include:

- Explaining why the bit sampling LSH get $\rho = \Theta(\frac{\log c}{c})$ for large c, when used with linear space.
- Analyzing the hamming space LSF for space/time trade-offs.
- Finding nicer closed form expressions for the low dimensionality regime.

Figure 3.2: In the blue area we have $\rho_q \leq 4/(c+1)$. That is we beat that bound for 92% of the triangle $0 \leq \delta \leq c\delta \leq 1/2$. Interestingly, using our improved version for small δ doesn't appear to give us anything in this comparison.

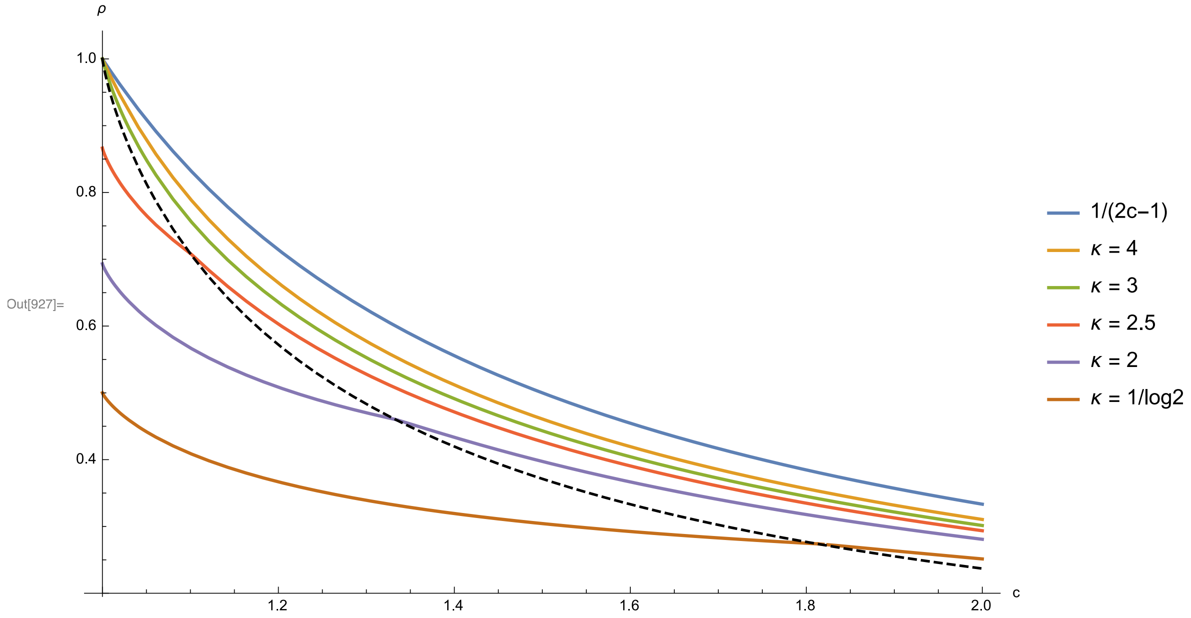


Figure 3.3: Query time exponent, ρ, as a function of c and κ for dimension $\kappa \log n$. We get an improvement for all κ, but as κ gets larger, we gradually gets closer to $1/(2c-1)$ as in the dimension oblivious case. Below the dashed line, the algorithm used is the simple scheme described, in which filters are small enough to never capture a far pair of points.

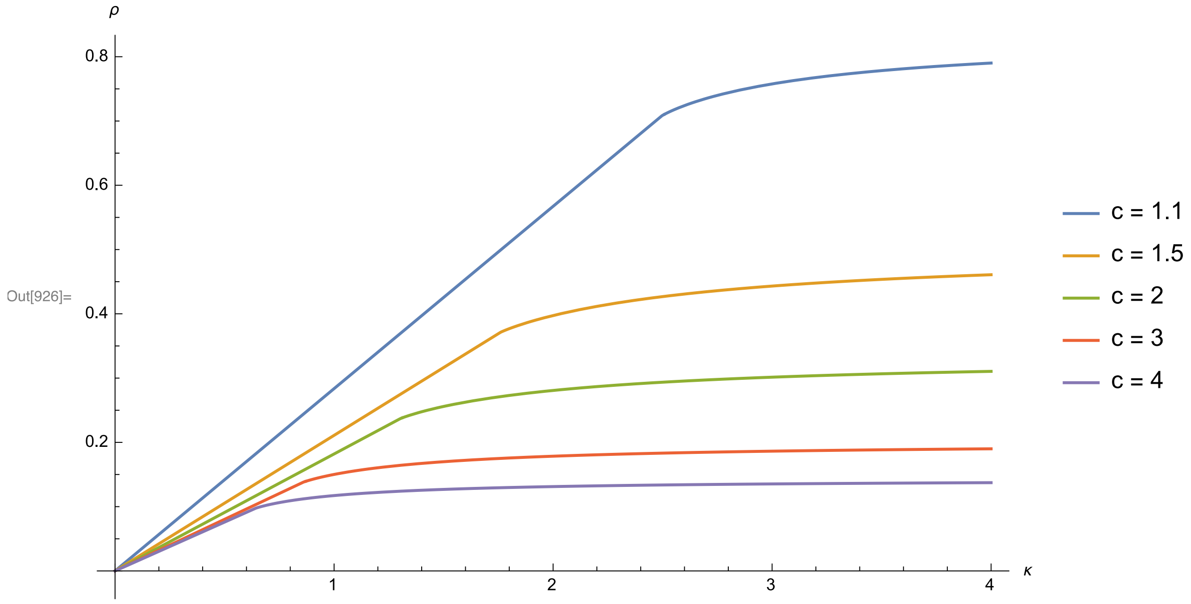


Figure 3.4: Query time exponent, ρ, as a function of κ and c for dimension $\kappa \log n$. Note that $\kappa < 1/\log 2 \approx 1.44$ doesn't make too much sense for hamming space, since the space is discrete and can only fit 2^d points. This doesn't mean the algorithm won't work in this area with the stated complexity though. Only that we might be able to do better by simply deduplicating the data.

Bibliography

[1] On a new théoreme-limit of théory of probabilityés.

[2] Alexandr Andoni, Thijs Laarhoven, Ilya Razenshteyn, and Erik Waingarten. Optimal hashing-based time-space trade-offs for approximate near neighbors. *arXiv preprint arXiv:1608.03580*, 2016.

[3] R Arratia and L Gordon. Tutorial on large deviations for the binomial distribution. *Bulletin of mathematical biology*, 51(1):125–131, 1989.

[4] Raghu Raj Bahadur, R Ranga Rao, et al. On deviations of the sample mean. *Ann. Math. Statist*, 31(4):1015–1027, 1960.

[5] SN Bernstein. Collected works, vol. 4. *Izdat. Akad. Nauk SSSR, Moscow*, 1964.

[6] Béla Bollobás. *Combinatorics: set systems, hypergraphs, families of vectors, and combinatorial probability.* Cambridge University Press, 1986.

[7] Béla Bollobás. *Modern graph theory*, volume 184. Springer Science & Business Media, 2013.

[8] Moses S Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 380–388. ACM, 2002.

[9] Tobias Christiani. A framework for similarity search with space-time tradeoffs using locality-sensitive filtering. *arXiv preprint arXiv:1605.02687*, 2016.

[10] Carl-Gustaf Esseen. *On the Liapounoff limit of error in the theory of probability.* Almqvist & Wiksell, 1942.

[11] Ronald L Graham, Donald E Knuth, and Oren Patashnik. Concrete mathematics: A foundation for computer science. 1994.

[12] Sariel Har-Peled, Piotr Indyk, and Rajeev Motwani. Approximate nearest neighbor: Towards removing the curse of dimensionality. *Theory of Computing*, 8(1):321–350, 2012.

[13] Hsien-Kuei Hwang. Asymptotic estimates of elementary probability distributions. *Studies in Applied Mathematics*, 99(4):393–417, 1997.

[14] Chris Impens. Stirling's series made easy. *The American mathematical monthly*, 110(8):730–735, 2003.

[15] Norman L Johnson, Adrienne W Kemp, and Samuel Kotz. *Univariate discrete distributions*, volume 444. John Wiley & Sons, 2005.

[16] Norman L Johnson, Samuel Kotz, and N Balakrishnan. Continuous univariate distributions , vol. 1john wiley & sons. *New York*, page 163, 1994.

[17] Rob Kaas and Jan M Buhrman. Mean, median and mode in binomial distributions. *Statistica Neerlandica*, 34(1):13–18, 1980.

[18] Alberto Leon-Garcia and Alberto. Leon-Garcia. *Probability, statistics, and random processes for electrical engineering.* Pearson/Prentice Hall Upper Saddle River, NJ, 2008.

[19] JE Littlewood. On the probability in the tail of a binomial distribution. *Advances in Applied Probability*, 1(1):43–72, 1969.

[20] John Edensor Littlewood and Béla Bollobás. *Littlewood's miscellany.* Cambridge University Press, 1986.

[21] Brendan D McKay. On littlewood's estimate for the binomial distribution. *Advances in Applied Probability*, pages 475–478, 1989.

[22] P Neumann. Über den mediafinder binomial-und poissonverteilung. *Wissenschaftliche Zeitschrift der Technischen Universität Dresden*, 15:223–226, 1966.

[23] Harald Niederreiter. *Coding theory and cryptology*, volume 1. World Scientific, 2002.

[24] Valentin Vladimirovich Petrov. On the probabilities of large deviations for sums of independent random variables. *Theory of Probability & Its Applications*, 10(2):287–298, 1965.

[25] VV Petrov. Sums of independent random variables. *Bull. Amer. Math. Soc. 83 (1977), 696-697 DOI: http://dx. doi. org/10.1090/S0002-9904-1977-14349-8 PII*, pages 0002–9904, 1977.

[26] Martin Raič. Clt-related large deviation bounds based on stein's method. *Advances in Applied Probability*, pages 731–752, 2007.

[27] Fraydoun Rezakhanlou. Optimal transport problem and contact structures. *preprint*, 2015.

[28] Michael Short. Improved inequalities for the poisson and binomial distribution and upper tail quantile functions. *ISRN Probability and Statistics*, 2013, 2013.

[29] Werner Uhlmann. Vergleich der hypergeometrischen mit der binomial-verteilung. *Metrika*, 10(1):145–158, 1966.