Lecture 11

February 25, 2025

Instructor: Sepehr Assadi *Scribe: Parth Mittal*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

Topics of this Lecture

1 String Similarity **1**

1.1 Attempt 1: Random indices 1

2 Johnson-Lindenstrauss Lemma (JLL) **2**

2.1 Attempt 2: Gaussians 2

1 String Similarity

In the **string similarity** problem, we are given n strings $x_1, \ldots, x_n \in \{0,1\}^d$, and are interested in answering queries on normalized hamming distance between pairs of them. For any pairs of strings $x, y \in \{0,1\}^d$, define

$$\overline{\Delta}(x,y) = \frac{\Delta(x,y)}{d},$$

where $\Delta(x,y)$ is the hamming distance between x and y, i.e., the number of indices where they differ.

Our goal in the string similarity problem is to compress the data such that given a query (i,j), we can output whether $\overline{\Delta}x_i, x_j > 0.1$, say, or not. Of course, this problem is easy if we store the x_i's as is; in this lecture, we will see how to solve this problem approximately while storing only a roughly $(\log n)$-dimensional representation of each x_i.

1.1 Attempt 1: Random indices

The most natural thing to try is to pick t random indices from $[d]$ independently and uniformly (i.e. with replacement). Let $\mathbf{S}$ denote the random variable containing all the indices we chose, and for $j \in [t]$, let $\mathbf{S}_j$ denote the j-th element of $\mathbf{S}$. For $i \in [n]$, let $\mathbf{y}_i$ denote the projection of x_i to the coordinates in $\mathbf{S}$. Then, we have the following claim.

Claim 1. *For $t = \frac{10\ln(2n)}{\varepsilon^2}$, $\overline{\Delta}(\mathbf{y}_i, \mathbf{y}_j) \in [\overline{\Delta}(x_i, x_j) \pm \varepsilon]$ for all $i, j \in [n]$ with probability $\geqslant 1 - 1/n^2$.*

Note that the guarantee of the claim is additive — this approach cannot give multiplicative guarantees: e.g., if $\Delta(x_i, x_j) = O(1)$, then the indices where they differ will w.h.p not appear in $\mathbf{S}$.

Proof. Let $\phi : \{0,1\}^d \to \{0,1\}^t$ be the map that projects x down to $x_{\mathbf{S}}$. First, we will show that $\overline{\Delta}(x,y)$ is preserved with high probability for any pair of strings x, y.

Fix $x, y \in \{0,1\}^d$; for $i \in [t]$, let $\mathbf{Z}_i = 1$ iff $x_{\mathbf{S}_i} \neq y_{\mathbf{S}_i}$ and let $\mathbf{Z} = \sum_{i=1}^t \mathbf{Z}_i$. Observe that $\mathbf{Z} = \Delta(\phi(x), \phi(y))$. Then by the additive Chernoff bound, we have that

$$\Pr\left[\left|\mathbf{Z} - \mathbb{E}[\mathbf{Z}]\right| \geqslant \varepsilon t\right] \leqslant 2\exp\left(-\frac{\varepsilon^2 t^2}{2t}\right) = 2\exp\left(-\frac{\varepsilon^2 t}{2}\right) \leqslant 2\exp(-5\ln(2n)) = \frac{2}{(2n)^5} \leqslant \frac{1}{n^4}.$$

On the other hand, since each $\mathbf{Z}_i$ is an unbiased estimator for $\overline{\Delta}(x,y)$ we know that $\mathbb{E}[\mathbf{Z}] = t \cdot \overline{\Delta}(x,y)$, and so we have that

$$\left|\frac{\Delta(\phi(x), \phi(y))}{t} - \overline{\Delta}(x,y)\right| < \varepsilon$$

with probability $\geqslant 1 - 1/n^4$.

To finish the proof we will union bound over ($\binom{n}{2}$-many) pairs x_i, x_j in our input; using the bound we showed above, one can see that

$$\left|\frac{\Delta(\phi(x_i), \phi(x_j))}{t} - \overline{\Delta}(x_i, y_i)\right| < \varepsilon$$

for all $i, j \in [n]$ with probability $\geqslant 1 - 1/n^2$. ☐

Remark. Notice that ϕ is a linear map — it has a $t \times d$ matrix where the (i,j)-th entry is 1 iff $\mathbf{S}_i = j$. This means that $\phi(x+y) = \phi(x) + \phi(y)$, and hence we can easily update the representation of any x_i should only a few bits of x_i change, without having to recompute the entire map from the beginning.

We will now see a second idea that can get multiplicative error bounds, even for the vector analogue of our string similarity problem.

2 Johnson-Lindenstrauss Lemma (JLL)

We begin by defining the **vector similarity** problem; here we are given vectors $x_1, \ldots, x_n \in \mathbb{R}^d$, and want to store low dimension representations $y_1, \ldots, y_n \in \mathbb{R}^t$ that preserve the ℓ_2-norm. In particular, we want[1]

$$\|y_i - y_j\|_2 \approx_\varepsilon \|x_i - x_j\|_2$$

for all $i, j \in [n]$.

2.1 Attempt 2: Gaussians

Recall that $\mathcal{N}(\mu, \sigma^2)$ is the gaussian random variable with mean μ and variance σ^2, whose PDF is:

$$p(x) := \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

See Figure 1 for the familiar "bell curve" shape of this distribution with different parameters.

We are now ready to state the main lemma of this lecture:

Lemma 2 (Johnson-Lindenstrauss Lemma [JL84])**.** *For vectors $x_1, \ldots, x_n \in \mathbb{R}^d$, define $\mathbf{y}_1, \ldots, \mathbf{y}_n \in \mathbb{R}^t$ such that $\mathbf{y}_i = \mathbf{S}x_i/\sqrt{t}$, where $\mathbf{S}$ is a $t \times d$ matrix of independent $\mathcal{N}(0,1)$ variables, and $t = 100(\ln n)/\varepsilon^2$. Then with high probability (over the choice of $\mathbf{S}$), $\|\mathbf{y}_i - \mathbf{y}_j\| \approx_\varepsilon \|x_i - x_j\|$ for all $i, j \in [n]$.*

To prove the lemma, we first claim that $\mathbf{S}$ preserves the norm of a fixed unit vector.

[1] Here and throughout this note, we will use $a \approx_\varepsilon b$ to mean $(1-\varepsilon) \cdot b \leqslant a \leqslant (1+\varepsilon) \cdot b$.

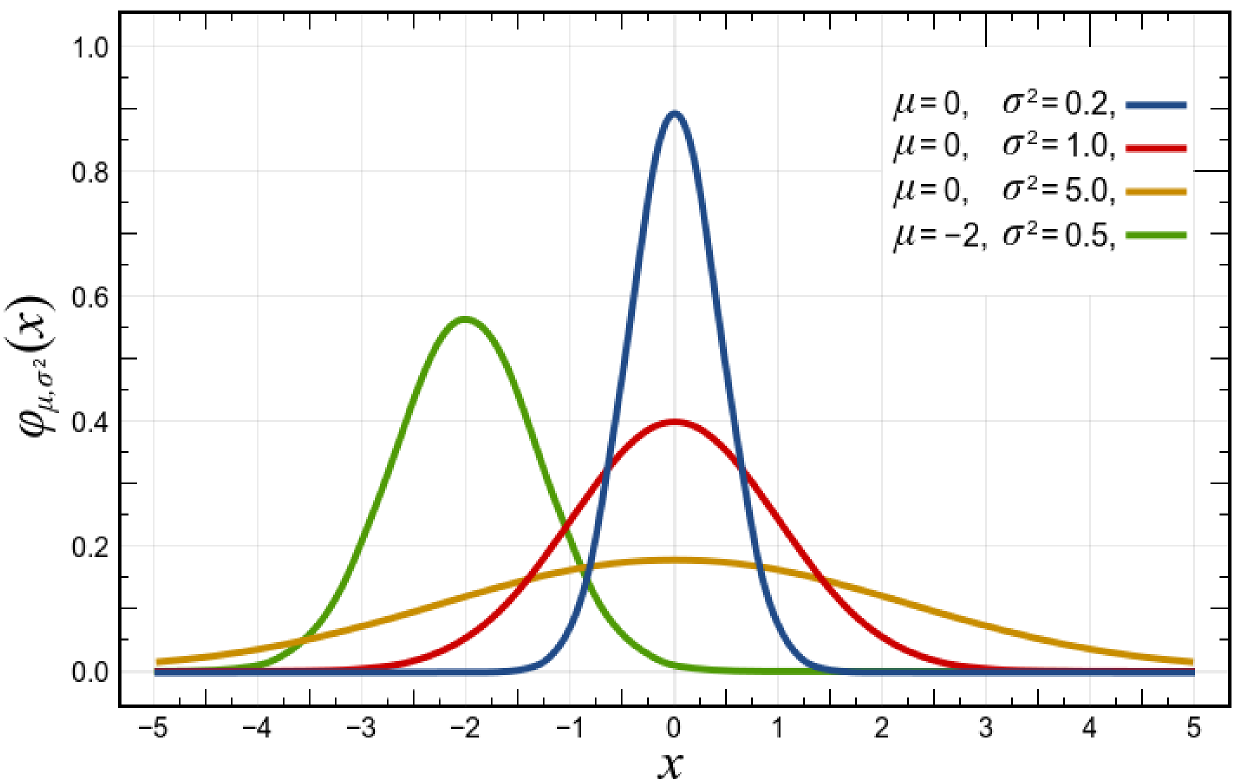

Figure 1: A selection of Normal Distribution Probability Density Functions (PDFs). Both the mean, μ, and variance, σ^2, are varied. The key is given on the graph.
Source: By Inductiveload - Own work (Original text: self-made, Mathematica, Inkscape), Public Domain, https://commons.wikimedia.org/w/index.php?curid=3817954.

Claim 3. *For a vector $v \in \mathbb{R}^d$ such that $\|v\| = 1$ and a matrix $\mathbf{S}$ sampled as in Lemma 2 with dimension $t = 10\ln(1/\delta)/\varepsilon^2$,*

$$\Pr_{\mathbf{S}}[\frac{\|\mathbf{S}v\|}{\sqrt{t}} \approx_\varepsilon 1] \geqslant 1 - 2\delta.$$

Before proving the claim, we see how it implies the lemma.

Proof of Lemma 2. For $i, j \in [n]$ define $v_{ij} = (x_i - x_j)/\|x_i - x_j\|$. Since $t = 100(\ln n)/\varepsilon^2$, we can apply Claim 3 on all v_{ij}'s with $\delta = 1/n^{10}$, to get that for any $i, j \in [n]$, $\Pr_{\mathbf{S}}[\|\mathbf{S}v_{ij}\|/\sqrt{t} \not\approx_\varepsilon 1] \leqslant 2/n^{10}$. Union-bounding over i, j, we obtain that $\|\mathbf{S}v_{ij}\|/\sqrt{t} \approx_\varepsilon 1$ for all $i \neq j \in [n]$ with probability $\geqslant 1 - 1/n^8$.

To finish, we expand the definition of v_{ij} and use the linearity of $\mathbf{S}$:

$$\frac{\|\mathbf{S}v_{ij}\|}{\sqrt{t}} \approx_\varepsilon 1 \iff \frac{\|\mathbf{S}(x_i - x_j)\|}{\sqrt{t} \cdot \|x_i - x_j\|} \approx_\varepsilon 1 \iff \left\| \frac{\mathbf{S}x_i}{\sqrt{t}} - \frac{\mathbf{S}x_j}{\sqrt{t}} \right\| \approx_\varepsilon \|x_i - x_j\| \iff \|\mathbf{y}_i - \mathbf{y}_j\| \approx_\varepsilon \|x_i - x_j\|,$$

which concludes the proof. □

So it "only" remains to show Claim 3. Emulating the proof of Claim 1, we will first argue that each row of $\mathbf{S}$ gives an unbiased estimator for $\|v\|^2$. Let $\mathbf{g} = (\mathbf{g}_1, \ldots, \mathbf{g}_d) \sim \mathcal{N}(0,1)^d$ be a vector of d independent $\mathcal{N}(0,1)$'s, and look at the random variable $\langle \mathbf{g}, v \rangle$. Because the $\mathbf{g}_i$'s are mean-0, the expectation of $\langle \mathbf{g}, v \rangle$ is also 0, and gives us no information. The quantity we should really care about (because we are computing

$\|\mathbf{S}v\|$, which sums the squares of each entry of $\mathbf{S}v$) is the expectation of its square:

$$\mathbb{E}\big[\langle \mathbf{g}\,,\, v\rangle^2\big] = \mathbb{E}\left[\left(\sum_{i=1}^{t} \mathbf{g}_i v_i\right)^2\right] = \mathbb{E}\left[\sum_{i=1}^{t} (\mathbf{g}_i v_i)^2 + \sum_{i \neq j} \mathbf{g}_i v_i \mathbf{g}_j v_j\right] = \sum_{i=1}^{t} \mathbb{E}\big[(\mathbf{g}_i v_i)^2\big] + \sum_{i \neq j} \mathbb{E}[\mathbf{g}_i v_i \mathbf{g}_j v_j],$$

where the last inequality is by linearity of expectation. Since $\mathbf{g}_i$ and $\mathbf{g}_j$ are independent when $i \neq j$ the second sum is 0, whereas the i-th term of the first is equal to:

$$v_i^2 \cdot \mathbb{E}[\mathbf{g}_i]^2 = v_i^2 \cdot (\mathrm{Var}[\mathbf{g}_i] - \mathbb{E}[\mathbf{g}_i]^2) = v_i^2.$$

Hence $\mathbb{E}[\langle \mathbf{g}\,,\, v\rangle^2] = \|v\|^2$, and $\mathbb{E}[\|\mathbf{S}v\|/\sqrt{t}] = \|v\| = 1$. We note that thus far we only used the fact that each entry of $\mathbf{g}$ is independent, has mean 0 and variance 1.

To finish the proof, we need to show a concentration result on $\|\mathbf{S}v\|$. We have

$$\Pr\left[\frac{\|\mathbf{S}v\|}{\sqrt{t}} \not\approx_\varepsilon 1\right] = \Pr\left[\|\mathbf{S}v\|^2 \notin \big[(1-\varepsilon)^2 \cdot t, (1+\varepsilon)^2 \cdot t\big]\right] \leqslant \Pr\big[\|\mathbf{S}v\|^2 \not\approx_\varepsilon t\big],$$

where the inequality holds because for $0 < \varepsilon < 1$, $[(1-\varepsilon), (1+\varepsilon)] \subseteq [(1-\varepsilon)^2, (1+\varepsilon)^2]$ and $t > 0$. Let $\mathbf{S}_1, \ldots, \mathbf{S}_t$ denote the rows of $\mathbf{S}$, and define the random variables $\mathbf{X}_i := \langle \mathbf{S}_i\,,\, v\rangle$ and $\mathbf{X} = \sum_i \mathbf{X}_i^2$. Since we showed above that $\mathbb{E}[\mathbf{X}] = \mathbb{E}[\|\mathbf{S}v\|^2] = t$, all we need is a bound on the probability $\Pr[|\mathbf{X} - \mathbb{E}[\mathbf{X}] \geqslant \varepsilon t|]$. This is precisely a concentration inequality and it additionally has the familiar form that $\mathbf{X}$ is sum of *independent* random variables. However, we cannot readily use Chernoff-like bounds on the $\mathbf{X}$ directly since the variables $\mathbf{X}_i^2$ used in the sum-definition of $\mathbf{X}$ are not bounded.

We will use the fact that a linear combination of independent Gaussians is still a Gaussian. In particular, the distribution of $\mathbf{X}_i$ is $\mathcal{N}(0, \|v\|) = \mathcal{N}(0,1)$. And so $\mathbf{X} = \sum_i \mathbf{X}_i^2$ has a χ-squared distribution, for which the following concentration bound is known:

Proposition 4 ([LM00]). *Suppose $\mathbf{X} = \sum_i \mathbf{X}_i^2$ where each $\mathbf{X}_i \sim \mathcal{N}(0,1)$ independently of the rest; then,*

$$\Pr[|\mathbf{X} - t| \geqslant \varepsilon t] \leqslant 2\exp\left(-\frac{\varepsilon^2 t}{8}\right).$$

Plugging this in, we have that

$$\Pr\left[\frac{\|\mathbf{S}v\|}{\sqrt{t}} \not\approx_\varepsilon 1\right] \leqslant 2\exp\left(-\frac{\varepsilon^2 t}{8}\right) = 2\exp\left(-\frac{\varepsilon^2 \cdot 10\ln(1/\delta)}{8\varepsilon^2}\right) \leqslant 2\exp(-\ln(1/\delta)) = 2/\delta.$$

This concludes the proof of Claim 3.

Detour: a "generic hack" for applying Chernoff to unbounded variables

Before concluding this lecture, let us mention a way of applying Chernoff bound itself to prove a weaker version of Claim 3, to show case a useful technique (although, in most cases, one should be able to replace this hack with a proper concentration inequality which is stronger than Chernoff bound).

Recall that the problem with applying Chernoff bound to $\mathbf{X} = \sum_{i=1}^{t} \mathbf{X}_i^2$ is that $\mathbf{X}_i^2$ variables are not bounded. We can get around this by defining the "clamping" variables $\mathbf{Y}_i := \min(\mathbf{X}_i^2, 8\log n)$, and show that with high probability, $\mathbf{Y}_i = \mathbf{X}_i^2$ for all i, because

$$\Pr\big[\mathbf{X}_i^2 > 8\log n\big] = \Pr\Big[|\mathbf{X}_i| > \sqrt{8\log n}\Big] \leqslant \frac{\exp(-4\log n)}{\sqrt{8\log n}} \leqslant \frac{1}{n^3},$$

where the first inequality is Mill's Inequality [Was04]. And since $\mathbf{Y} := \sum_i \mathbf{Y}_i$ is a sum of bounded, independent random variables, we can use Chernoff bound to finish the proof. Note that since $\mathbf{Y}_i \in [\pm 8\log n]$, to get a useful bound from Chernoff we will need $t = 1000(\log^2 n)/\varepsilon^2$ as opposed to $t = O(\log n/\varepsilon^2)$ of previous part. Nevertheless, this way we have,

$$\Pr(|\mathbf{X} - \mathbb{E}[\mathbf{X}]| \geqslant \varepsilon t) \leqslant \underbrace{\Pr(|\mathbf{Y} - \mathbb{E}[\mathbf{Y}]| \geqslant \varepsilon t)}_{\text{handled by Chernoff bound}} + \underbrace{\Pr(\mathbf{Y} \neq \mathbf{X})}_{\text{handled by Mill's inequality}},$$

and so we can use this technique to prove "some" concentration for $\mathbf{X}$ as well.

References

[JL84] William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in modern analysis and probability (New Haven, Conn., 1982)*, volume 26 of *Contemp. Math.*, pages 189–206. Amer. Math. Soc., Providence, RI, 1984. 2

[LM00] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338, 2000. 4

[Was04] Larry Wasserman. *All of statistics*. Springer Texts in Statistics. Springer-Verlag, New York, 2004. A concise course in statistical inference. 4