

The Johnson-Lindenstrauss Lemma

Sariel Har-Peled

November 18, 2005

598shp - Randomized Algorithms

1 The Johnson-Lindenstrauss lemma

1.1 Some Probability

Definition 1.1 Let $N(0, 1)$ denote the one dimensional *normal distribution*. This distribution has density $n(x) = e^{-x^2/2}/\sqrt{2\pi}$.

Let $N^d(0, 1)$ denote the d -dimensional *Gaussian distribution*, induced by picking each coordinate independently from the standard normal distribution $N(0, 1)$.

Let $\text{Exp}(\lambda)$ denote the *exponential distribution*, with parameter λ . The density function of the exponential distribution is $f(x) = \lambda \exp(-\lambda x)$.

Let $\Gamma_{\lambda,k}$ denote the *gamma distribution*, with parameters λ and k . The density function of this distribution is $g_{\lambda,k}(x) = \lambda \frac{(\lambda x)^{k-1}}{(k-1)!} \exp(-\lambda x)$. The cumulative distribution function of $\Gamma_{\lambda,k}$ is $\Gamma_{\lambda,k}(x) = 1 - \exp(-\lambda x) \left(1 + \frac{\lambda x}{1!} + \dots + \frac{(\lambda x)^i}{i!} + \dots + \frac{(\lambda x)^{k-1}}{(k-1)!} \right)$. As we prove below, gamma distribution is how much time one has to wait till k experiments succeed, where an experiment duration distributes according to the exponential distribution.

A random variable X has the *Poisson distribution*, with parameter $\eta > 0$ (which is a discrete distribution) if $\Pr[X = i] = \frac{\eta^i}{i!} e^{-\eta}$.

Lemma 1.2 If $X \sim \text{Exp}(\lambda)$ then $\mathbf{E}[X] = \frac{1}{\lambda}$.

Proof:
$$\int_{x=0}^{\infty} x \cdot \lambda e^{-\lambda x} dx = \left[-\frac{1}{\lambda} e^{-\lambda x} - x e^{-\lambda x} \right]_{x=0}^{\infty} = \frac{1}{\lambda}. \quad \blacksquare$$

Lemma 1.3 The following properties hold for the d dimensional Gaussian distribution $N^d(0, 1)$:

- (i) The distribution $N^d(0, 1)$ is centrally symmetric around the origin.
- (ii) If $X \sim N^d(0, 1)$ and u is a unit vector, then $X \cdot u \sim N(0, 1)$.
- (iii) If $X, Y \sim N(0, 1)$ are two independent variables, then $Z = X^2 + Y^2$ follows the exponential distribution with parameter $\lambda = \frac{1}{2}$.
- (iv) Given k independent variables X_1, \dots, X_k distributed according to the exponential distribution with parameter λ , then $Y = X_1 + \dots + X_k$ is distributed according to the Gamma distribution $\Gamma_{\lambda,k}(x)$.

Proof: (i) Let $x = (x_1, \dots, x_d)$ be a point picked from the Gaussian distribution. The density $\phi_d(x) = \phi(x_1)\phi(x_2) \cdot \phi(x_d)$, where $\phi(x_i)$ is the normal distribution density function, which is $\phi(x_i) = \exp(-x_i^2/2)/\sqrt{2\pi}$. Thus $\phi_d(x) = (2\pi)^{-n/2} \exp(-(x_1^2 \cdots + x_d^2)/2)$. Consider any two points $x, y \in \mathbb{R}^n$, such that $r = \|x\| = \|y\|$. Clearly, $\phi_d(x) = \phi_d(y)$. Namely, any two points of the same distance from the origin, have the same density (i.e., “probability”). As such, the distribution $N^d(0, 1)$ is centrally symmetric around the origin.

(ii) Consider $e_1 = (1, 0, \dots, 0) \in \mathbb{R}^n$. Clearly, $x \cdot e_1 = x_1$, which is distributed $N(0, 1)$. Now, by the symmetry of $N^d(0, 1)$, this implies that $x \cdot u$ is distributed $N(0, 1)$. Formally, let R be a rotation matrix that maps u to e_1 . We know that Rx is distributed $N^d(0, 1)$ (since $N^d(0, 1)$ is centrally symmetric). Thus $x \cdot u$ has the same distribute as $Rx \cdot Ru$, which has the same distribution as $x \cdot e_1$, which is $N(0, 1)$.

(iii) If $X, Y \sim N(0, 1)$, and consider the density function $g(x, y) = \frac{1}{2\pi} \exp\left(-\frac{x^2+y^2}{2}\right)$ and the associated integral $\int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} g(x, y) dx dy$. We would like to change the integration variables to $x(r, \alpha) = \sqrt{r} \sin \alpha$ and $y(r, \alpha) = \sqrt{r} \cos \alpha$. The Jacobian of this change of variables is

$$I(r, \alpha) = \begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \alpha} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \alpha} \end{vmatrix} = \begin{vmatrix} \frac{\sin \alpha}{2\sqrt{r}} & \sqrt{r} \cos \alpha \\ \frac{\cos \alpha}{2\sqrt{r}} & -\sqrt{r} \sin \alpha \end{vmatrix} = -\frac{1}{2}(\sin^2 \alpha + \cos^2 \alpha) = -\frac{1}{2}.$$

As such, we have

$$\begin{aligned} \Pr[Z = z] &= \int_{x^2+y^2=z} \frac{1}{2\pi} \exp\left(-\frac{x^2+y^2}{2}\right) dx dy \\ &= \int_{\alpha=0}^{2\pi} \frac{1}{2\pi} \exp\left(-\frac{x(\sqrt{z}, \alpha)^2 + y(\sqrt{z}, \alpha)^2}{2}\right) \cdot |I(z, \alpha)| d\alpha \\ &= \frac{1}{2\pi} \cdot \frac{1}{2} \cdot \int_{\alpha=0}^{2\pi} \exp\left(-\frac{z}{2}\right) = \frac{1}{2} \exp\left(-\frac{z}{2}\right). \end{aligned}$$

As such, Z has an exponential distribution with $\lambda = 1/2$.

(iv) For $k = 1$ the claim is trivial. Otherwise, let $g_{k-1}(x) = \lambda \frac{(\lambda x)^{k-2}}{(k-2)!} \exp(-\lambda x)$. Observe that

$$\begin{aligned} g_k(t) &= \int_0^t g_{k-1}(t-x)g_1(x) dx = \int_0^t \left(\lambda \frac{(\lambda(t-x))^{k-2}}{(k-2)!} \exp(-\lambda(t-x)) \right) (\lambda \exp(-\lambda x)) dx \\ &= \int_0^t \lambda^2 \frac{(\lambda(t-x))^{k-2}}{(k-2)!} \exp(-\lambda t) dx \\ &= \lambda \exp(-\lambda t) \int_0^t \lambda \frac{(\lambda x)^{k-2}}{(k-2)!} dx = \lambda \exp(-\lambda t) \frac{(\lambda t)^{k-1}}{(k-1)!} = g_k(x). \quad \blacksquare \end{aligned}$$

1.2 Proof of the Johnson-Lindenstrauss Lemma

Lemma 1.4 *Let u be a unit vector in \mathbb{R}^d . For any even positive integer k , let U_1, \dots, U_k be random vectors chosen independently from the d -dimensional Gaussian distribution $N^d(0, 1)$. For $X_i = u \cdot U_i$, define $W = W(u) = (X_1, \dots, X_k)$ and $L = L(u) = \|W\|^2$. Then, for any $\beta > 1$, we have:*

1. $\mathbf{E}[L] = k$.
2. $\Pr[L \geq \beta k] \leq \frac{k+3}{2} \exp\left(-\frac{k}{2}(\beta - (1 + \ln \beta))\right)$.

3. $\Pr[L \leq k/\beta] < O(k) \times \exp(-\frac{k}{2}(\beta^{-1} - (1 - \ln \beta)))$.

Proof: By Lemma 1.3 (ii) each X_i is distributed as $N(0, 1)$, and X_1, \dots, X_k are independent. Define $Y_i = X_{2i-1}^2 + X_{2i}^2$, for $i = 1, \dots, \tau$, where $\tau = k/2$. By Lemma 1.3 (iii) Y_i follows the exponential distribution with parameter $\lambda = 1/2$. Let $L = \sum_{i=1}^{\tau} Y_i$. By Lemma 1.3 (iv), the variable L follows the Gamma distribution $(k/2, 1/2)$, and its expectation is $\mathbf{E}[L] = \sum_{i=1}^{k/2} \mathbf{E}[Y_i] = \tau \times 2 = k$, since $\mathbf{E}[Y_i] = 2$ by Lemma 1.2.

Now, let $\eta = \lambda\beta k = \beta k/2 = \beta\tau$, we have

$$\Pr[L \geq \beta k] = 1 - \Pr[L \leq \beta k] = 1 - \Gamma_{1/2, \tau}(\beta k) = \sum_{i=0}^{\tau-1} e^{-\eta} \frac{\eta^i}{i!} \leq (\tau + 1) e^{-\eta} \frac{\eta^{\tau}}{\tau!},$$

since $\eta = \beta\tau > \tau$, as $\beta > 1$ and $\Gamma_{\lambda, k}(x) = 1 - \exp(-\lambda x) \left(1 + \frac{\lambda x}{1!} + \dots + \frac{(\lambda x)^i}{i!} + \dots + \frac{(\lambda x)^{k-1}}{(k-1)!}\right)$. Now, since $\tau! \geq (\tau/e)^{\tau}$, and thus

$$\begin{aligned} \Pr[L \geq \beta k] &\leq (\tau + 1) e^{-\eta} \frac{\eta^{\tau}}{\tau^{\tau}/e^{\tau}} = (\tau + 1) e^{-\eta} \left(\frac{e\eta}{\tau}\right)^{\tau} = (\tau + 1) e^{-\beta\tau} \left(\frac{e\beta\tau}{\tau}\right)^{\tau} \\ &= (\tau + 1) e^{-\beta\tau} \cdot \exp(\tau \ln(e\beta)) = (\tau + 1) \exp(-\tau(\beta - (1 + \ln \beta))) \\ &\leq \frac{k+3}{2} \exp\left(-\frac{k}{2}(\beta - (1 + \ln \beta))\right). \end{aligned}$$

Arguing in a similar fashion, we have, for a large constant $\rho \gg 1$

$$\begin{aligned} \Pr[L \leq k/\beta] &= \Gamma_{1/2, \tau}(k/\beta) = 1 - \sum_{i=0}^{\tau-1} e^{-\tau/\beta} \frac{(\tau/\beta)^i}{i!} = e^{-\tau/\beta} \sum_{i=0}^{\infty} \frac{(\tau/\beta)^i}{i!} - \sum_{i=0}^{\tau-1} e^{-\tau/\beta} \frac{(\tau/\beta)^i}{i!} \\ &= \sum_{i=\tau}^{\infty} e^{-\tau/\beta} \frac{(\tau/\beta)^i}{i!} \leq e^{-\tau/\beta} \sum_{i=\tau}^{\infty} \left(\frac{e\tau}{i\beta}\right)^i \\ &= e^{-\tau/\beta} \left[\sum_{i=\tau}^{\rho e\tau/\beta} \left(\frac{e\tau}{i\beta}\right)^i + \sum_{i=\rho e\tau/\beta+1}^{\infty} \left(\frac{e\tau}{i\beta}\right)^i \right] \end{aligned}$$

The second sum is very small for $\rho \gg 1$ and we bound only the first one. As the sequence $(\frac{e\tau}{i\beta})^i$ is decreasing for $i \geq \tau/\beta$, we can bound the first sum by

$$\frac{\rho e\tau}{\beta} \cdot e^{-\tau/\beta} \left(\frac{e}{\beta}\right)^{\tau} = O(\tau) \exp(-\tau(\beta^{-1} - (1 - \ln \beta))).$$

Since $\tau = k/2$, we obtain the desired result. \blacksquare

Next, we show how to interpret the above inequalities in a somewhat more intuitive way. Let $\beta = 1 + \varepsilon$, $\varepsilon > 0$. From Taylor expansion we know that $\ln \beta \leq \varepsilon - \varepsilon^2/2 + \varepsilon^3/3$. By plugging it into the upper bound for $\Pr[L \geq \beta k]$ we get

$$\begin{aligned} \Pr[L \geq \beta k] &\leq O(k) \times \exp\left(-\frac{k}{2}(1 + \varepsilon - 1 - \varepsilon + \varepsilon^2/2 - \varepsilon^3/3)\right) \\ &\leq O(k) \times \exp\left(-\frac{k}{2}(\varepsilon^2/2 - \varepsilon^3/3)\right) \end{aligned}$$

On the other hand, we also know that $\ln \beta \geq \varepsilon - \varepsilon^2/2$. Therefore

$$\begin{aligned} \Pr[L \leq k/\beta] &\leq O(k) \times \exp\left(-\frac{k}{2}(\beta^{-1} - 1 + \varepsilon - \varepsilon^2/2)\right) \\ &\leq O(k) \times \exp\left(-\frac{k}{2}\left(\frac{1}{1+\varepsilon} - 1 + \varepsilon - \varepsilon^2/2\right)\right) \\ &\leq O(k) \times \exp\left(-\frac{k}{2}\left(\frac{\varepsilon^2}{1+\varepsilon} - \varepsilon^2/2\right)\right) \\ &\leq O(k) \times \exp\left(-\frac{k}{2} \cdot \frac{\varepsilon^2 - \varepsilon^3}{2(1+\varepsilon)}\right) \end{aligned}$$

Thus, the probability that a given unit vector gets distorted by more than $(1 + \varepsilon)$ in any direction^① grows roughly as $\exp(-k\varepsilon^2/4)$, for small $\varepsilon > 0$. Therefore, if we are given a set P of n points in l_2 , we can set k to roughly $8\ln(n)/\varepsilon^2$ and make sure that with non-zero probability we obtain projection which does not distort distances^② between *any* two different points from P by more than $(1 + \varepsilon)$ in each direction.

Theorem 1.5 *Given a set P of n points in \mathbb{R}^d , and parameter ε , one can compute a random projection R into $k = 8\varepsilon^{-2} \ln n$ dimensions, such that the distances between points are roughly preserved. Formally, with constant probability, for any $p, q \in P$, we have*

$$(1 - \varepsilon)\|p - q\| \leq \|R(p) - R(q)\| \leq \|p - q\|.$$

The probability of success improves to high probability, if we use, say, $k = 10\varepsilon^{-2} \ln n$ dimensions.

2 Bibliographical notes

The probability review of Section 1.1 can be found in Feller [Fel71]. The proof of the Johnson-Lindenstrauss lemma of Section 1.2 is due to Indyk and Motwani [IM98]. The original proof of the Johnson-Lindenstrauss lemma is from [JL84].

It exposes the fact that the Johnson-Lindenstrauss lemma is no more than yet another instance of the concentration of mass phenomena (i.e., like the Chernoff inequality).

Interestingly, it is enough to pick each entry in the dimension reducing matrix randomly out of $-1, 0, 1$. This requires more involved proof [Ach01]. This is useful when one care about storing this dimension reduction transformation efficiently.

Magen [Mag01] observed that in fact the Johnson-Lindenstrauss lemma preserves angles, and in fact can be used to preserve any “ k dimensional angle”, by projecting down to dimension $O(k\varepsilon^{-2} \log n)$. In particular, Exercise 3.1 is taken from there.

Dimension reduction is crucial in learning, AI, databases, etc. One common technique that is being used in practice is to do PCA (i.e., principal component analysis) and take the first few main axes. Other techniques include independent component analysis, and MDS (multidimensional scaling). MDS tries to embed points from high dimensions into low dimension ($d = 2$ or 3), which preserving some properties. Theoretically, dimension reduction into really low dimensions is hopeless, as the distortion in the worst case is $\Omega(n^{1/(k-1)})$, if k is the target dimension [Mat90].

^①Note that this implies distortion $(1 + \varepsilon)^2$ if we require the mapping to be a contraction.

^②In fact, this statement holds even for the *square* of the distances.

3 Exercises

Exercise 3.1 [10 Points] Show that the Johnson-Lindenstrauss lemma also $(1 \pm \varepsilon)$ -preserves angles among triples of points of P (you might need to increase the target dimension however by a constant factor). [**Hint:** For every angle, construct a equilateral triangle that its edges are being preserved by the projection (add the vertices of those triangles [conceptually] to the point set being embedded). Argue, that this implies that the angle is being preserved.]

References

- [Ach01] D. Achlioptas. Database-friendly random projections. In *Proc. 20th ACM Sympos. Principles Database Syst.*, pages 274–281, 2001.
- [Fel71] W. Feller. *An Introduction to Probability Theory and its Applications*, volume II. John Wiley & Sons, NY, 1971.
- [IM98] P. Indyk and R. Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proc. 30th Annu. ACM Sympos. Theory Comput.*, pages 604–613, 1998.
- [JL84] W. B. Johnson and J. Lindenstrauss. Extensions of lipschitz mapping into hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.
- [Mag01] A. Magen. Dimensionality reductions that preserve volumes and distance to affine spaces, and its algorithmic applications. Submitted to STOC 2002, 2001.
- [Mat90] J. Matoušek. Bi-lipschitz embeddings into low-dimensional euclidean spaces. *Comment. Math. Univ. Carolinae*, 31:589–600, 1990.