18.409 An Algorithmist's Toolkit Nov. 10, 2009

Lecture 17

Lecturer: Jonathan Kelner

1 Johnson-Lindenstrauss Theorem

1.1 Recap

We first recap a theorem (isoperimetric inequality) and a lemma (concentration) from last time:

Theorem 1 (Measure concentration on the sphere) *Let $\mathbb{S}^{n-1}$ be the unit sphere in $\mathbb{R}^n$ and $A \in \mathbb{S}^{n-1}$ be a measurable set with vol$(A) \geq 1/2$, and let A_ε denote the set of points of $\mathbb{S}^{n-1}$ with distance at most ε from A. Then vol$(A_\varepsilon) \geq 1 - e^{-n\varepsilon^2/2}$.*

This theorem basically says that: When we get a set A which is greater or equal to half of the sphere, if we further incorporate points at most ε away from A, we *almost* have the whole sphere.

Definition 2 (c-Lipschitz) *A function $f : A \to B$ is c-Lipschitz if, for any $u, v \in A$, we have $\|f(u) - f(v)\| \leq c \cdot \|u - v\|$*

For a unit vector $x \in \mathbb{S}^{n-1}$, the projection of the first k dimension is a 1-Lipschitz function,:

$$f(x) = \sqrt{x_1^2 + x_2^2 + \cdots + x_k^2}$$

Lemma 3 *For a unit vector $x \in \mathbb{S}^{n-1}$, and $f(x) = \sqrt{x_1^2 + x_2^2 + \cdots + x_k^2}$. Let x be a vector randomly chosen with uniform distribution from $\mathbb{S}^{n-1}$ and M be the median of $f(x)$. Then $f(x)$ is sharply concentrated with:*

$$Pr[|f(x) - M| \geq t] \leq 2e^{-t^2 n/2}$$

1.2 Metric Embedding

Definition 4 (D-embedding) *Suppose that $X = \{x_1, x_2, \cdots x_n\}$ is a finite set, d is a metric on X, and $f : X \to \mathbb{R}^k$ is 1-Lipschitz, with $\|f(x_i) - f(x_j)\| \leq d(x_i, x_j)$. The "distortion" of f is the minimum D for which*

$$\|f(x_i) - f(x_j)\| \leq d(x_i, x_j) \leq D\|f(x_i) - f(x_j)\|$$

for some positive constant α. We refer to f as a D-embedding of X.

Claim of Johnson-Lindenstrauss Theorem: The Euclidean metric on any finite set X (a bunch of high dimensional points) can be embedded with distortion $D = 1 + \varepsilon$ in $\mathbb{R}^k$ for $k = O(\varepsilon^{-2} \log n)$.

If we lose ε ($\varepsilon = 0$), it becomes almost impossible to do better than that in $\mathbb{R}^n$. Nevertheless, it is not hard to construct a counter example to this: a simplex of $n + 1$ points. The Johnson-Lindenstrauss theorem gives us an interesting result: if we project x to a random subspace, the projection y give us an approximate length of x for some fixed multiplication factor c, i.e. $\|x\| \sim c \cdot \|y\|$. And $c \cdot y$ is embedded with distortion $D = 1 + \varepsilon$.

1.3 Proof of the Theorem

Next, we provide a more precise statement about Johnson-Lindenstrauss Theorem:

Theorem 5 (Johnson-Lindenstrauss) *Let $X = \{x_1, x_2, \cdots x_n\} \in \mathbb{R}^m$ (for any m) and let $k = O(\varepsilon^{-2} \log n)$. For:*

- *$\mathfrak{L} \subseteq \mathbb{R}^m$ be a uniform random k dimensional subspace.*
- *$\{y_1, y_2, \cdots y_n\}$ be projections of x_i on $\mathfrak{L}$.*
- *$y_i' = cy_i$ for some fixed constant c, and $c = \Theta(\frac{k}{m})$*

Then, with high probability $\mathfrak{L}$ is a $(1+\varepsilon)$-embedding of X into $\mathbb{R}^k$, i.e. for $x_i, x_j \in X$

$$\|x_i - x_j\| \leq \|y_i' - y_j'\| \leq (1+\varepsilon)\|x_i - x_j\|$$

Proof Let $\Pi_{\mathfrak{L}} : \mathbb{R}^m \to \mathfrak{L}$ be the orthogonal projection of $\mathbb{R}^m$ vector into subspace $\mathfrak{L}$. For $x_i, x_j \in X$, we let x be the normalized unit vector of $x_i - x_j$, and we need to prove that

$$(1-\phi) \cdot M\|x\| \leq \|\Pi_{\mathfrak{L}}(x)\| \leq (1+\phi) \cdot M\|x\|$$

holds with high probability, where M is the median of the of the function $f = \sqrt{x_1^2 + \cdots + x_m^2}$.

Following definition 4, this shows that the mapping $\Pi_{\mathfrak{L}}$ is a D-embedding of X into $\mathbb{R}^k$ with $D = \frac{1+\phi}{1-\phi}$. We let $\phi = \frac{\varepsilon}{3}$ so that $D = \frac{1+\varepsilon/3}{1-\varepsilon/3} \leq 1+\varepsilon$. Since $\|x\| = 1$, it is equivalent to showing that the following inequality holds with high probability

$$||\|\Pi_{\mathfrak{L}}(x)\| - M| < \frac{\varepsilon}{3} M \tag{1}$$

Lemma 3 describes the case when we have a random unit vector and project it onto a fixed subspace. It is actually identical to fixing a vector and projecting it onto a *random subspace* (we will describe how this random subspace is generated in the next subsection). We use Lemma 3 and plug in $t = \frac{\varepsilon}{3} M$; the probability inequality (1) *does not* hold is bounded by

$$\begin{aligned} Pr\left[||\|\Pi_{\mathfrak{L}}(x)\| - M| \geq \frac{\varepsilon}{3} M\right] &\leq 4e^{-t^2 m/2} \\ &= 4e^{-\varepsilon^2 M^2 m/18} \\ &\leq 4e^{-\varepsilon^2 k/72} \\ &\leq 1/m^2 \end{aligned}$$

Line 4 holds since $k = O(\varepsilon^{-2} \log n)$ (for further details, please see [1]). Line 3 holds since $M = \Omega(\sqrt{\frac{k}{m}})$, based on the following reasoning: We have that

$$1 = \mathbb{E}[\|X\|^2] = \sum \mathbb{E}[x_i^2],$$

which implies that $\mathbb{E}[x_i^2] = \frac{1}{m}$. Consequently,

$$\frac{k}{m} = \mathbb{E}[f^2] \leq \Pr[f \leq M+t](M+t)^2 + \Pr[f > M+t]\max(f^2) \leq (M+t)^2 + 2e^{-t^2 m/2},$$

where we used the fact that $f^2 = \sum_{i=1}^k x_i^2$. Taking $t = \Theta(\sqrt{\frac{k}{m}})$, we have that $M = \Omega(\sqrt{\frac{k}{m}})$. ■

1.4 Random Subspace

Here we describe how a random subspace is generated. We first provide a quick review about Gaussians, a multivariate Gaussian has PDF:

$$p_x(x_1, x_2, \cdots, x_N) = \frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} \exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$$

where Σ is a nonsingular covariance matrix and vector μ is the mean of x.

Gaussians have several nice properties. The following operations on Gaussian variables also yield Gaussian variables:

- Project onto a lower dimensional subspace.
- Restrict to a lower dimensional subspace, i.e. conditional probability.
- Any linear operations.

In addition, we can generate a vector with multi-dimensional Gaussian distribution by picking *each coordinate* according to a 1-dimensional Gaussian distribution.

How do we generate a random vector from a sphere? The idea here is to pick a point from a multi-dimensional Gaussian distribution (generate each coordinate with mean = 0 and variance = 1, $N(0, 1)$) so most n-dimensional vectors have norm $\sqrt{n}$. As the shape of an independent Gaussian distribution's PDF is *symmetric*, this procedure does indeed generate a point randomly and uniformly from a sphere (after normalizing it). Generating a random vector from a uniform distribution does not work, since it is *not* sampling uniformly from a sphere after normalization.

How do we get a random projection? This is no more than sampling $n \times k$ times from a $N(0, 1)$ gaussian distributions. Each k samples are grouped to form a k-dimensional vector, so we have n total vectors: $v_1, v_2, \cdots v_n$. We can simply orthonormalize these vectors, denoted as $\hat{v}_i$, and form the random subspace $\mathfrak{L}$:

$$\begin{pmatrix} \vdots & \vdots & & \vdots \\ \hat{v}_1 & \hat{v}_2 & \cdots & \hat{v}_n \\ \vdots & \vdots & & \vdots \end{pmatrix}$$

1.5 Applications of Johnson-Lindenstrauss Theorem

The Johnson-Lindenstrauss Theorem is very useful in several application areas, since it can approximately solve many problems. Here we illustrate some of them:

- **Proximity Problems :** This is an immediate application of the J-L Theorem. This is the case when we get a set of points in a high dimensional space $\mathbb{R}^d$ and we want to compute any property defined in terms of distance between points. Using the J-L theorem, we can actually solve the problem in a lower dimensional space (up to a distortion factor). Example problems here include closest pair, furthest pair, minimum spanning tree, minimum cost matching, and various clustering problems.
- **On-line Problems :** The problems of this type involve answering queries in a high dimensional space. This is usually done through partitioning a high dimensional space according to some error (distance) measure. However, this operation tends to be exponentially dependent on the dimension of the space, e.g., $\left(\frac{1}{\varepsilon}\right)^d$ (referred to as the "curse of dimensionality"). Projecting points of higher dimensional space into lower dimensional space significantly helps with these types of problems.
- **Data Stream/Storage Problem :** We obtain data in a stream but we cannot store it all due to some storage space restriction. One way of dealing with it is to maintain a count for each data entry and then see how the counts are distributed. The idea is to provide "sketches" of such data based on the J-L Theorem. For further details, please refer to Piotr Indyk's course and his survey paper.

In summary, applications that are related to dimensionality reduction are very likely to be a good platform for the J-L Theorem.

2 Dvoretsky's Theorem

Dvoretsky's Theorem, proved by Aryeh Dvoretsky in his article "A Theorem on Convex Bodies and Applications to Banach Spaces" in 1959, tries to answer the following question:

- Let C be an origin-symmetric convex body in $\mathbb{R}^n$.
- $S \subseteq \mathbb{R}^n$ be a vector subspace.
- We would like to know: does $Y = C \cap S$ look like a sphere? Furthermore, for *how high a dimension* (we denote it as k) does there exist an S for which this occurs?

A formal statement of Y's similarity to a sphere can be characterized by whether Y has a small Banach-Mazur distance to the sphere, i.e. if there exists a linear transformation such that

$$\mathbb{S}^{k-1}(1) \leq Y \leq \mathbb{S}^{k-1}(1+\varepsilon)$$

where $\mathbb{S}^{k-1}(r)$ is denoted as a sphere with radius r.

It turns out that k varies with different types of convex bodies: for a ellipsoid $k = n$, for a cross-polytope $k = \Theta(n)$, and for a cube is $k = \log(n)$. It turns out that the cube case is the worst case scenario. Here is a formal statement of Dvoretsky's Theorem:

Theorem 6 (Dvoretsky) *There is a positive constant $c > 0$ such that, for all ε and n, every n-dimensional origin-symmetric convex body has a section within distance $1 + \varepsilon$ of the unit ball of dimension*

$$k \geq \frac{c\varepsilon^2}{\log(1+\varepsilon^{-1})} \log n$$

Instead of providing the whole proof, we give a sketch of the proof here:

1. When we are given an origin-symmetric convex body, denoted as C, it defines some norm with respect to the convex body: $C \to \|\cdot\|_C$.
2. We need a subspace S to be spherical. It is basically saying that when we take any vector θ on S, then $\|\theta\|_C$ is approximately *constant*.
3. This is similar to concentration of measures which we have shown before. It basically says that when we have a function defined as a norm $f : \theta \to \|\theta\|_C$, it is precisely concentrated for every θ on the sphere (i.e. every $\|\theta\|_C$ is close to median).
4. This is similar to Johnson-Lindenstrauss except that we need *every* vector in k-dimensional subspace satisfying point 2 (In the J-L theorem, we prove that *most* of the vectors (points) are close to a fixed constant, i.e. median).
5. What we do is to put a fine "mesh" on the k-dimensional subspace and show that every point on the grid is right. The number of points we need to check is approximately $O((\frac{4}{\delta})^k)$ where δ is the error. We can see that it is exponentially dependent on k and it looks similar to the dependency of k in the J-L theorem. For further details of the proof, please see [2].

References

1. Sariel Har-Peled, "*Geometric Approximation Algorithms*", http://valis.cs.uiuc.edu/~sariel/teach/notes/aprx
2. Aryeh Dvoretzky, "*Some results on convex bodies and Banach spaces*", Proceedings of the National Academy of Sciences, 1959.

18.409 Topics in Theoretical Computer Science: An Algorithmist's Toolkit
Fall 2009