CS731 Spring 2011 Advanced Artificial Intelligence

Random Projection

Lecturer: Xiaojin Zhu *jerryzhu@cs.wisc.edu*

Random projection is a powerful technique behind compressive sensing and matrix completion.

1 The Johnson-Lindenstrauss Lemma

When n points in some high dimensional space are *randomly* projected down to $O(\frac{\log n}{\epsilon^2})$ dimensions, with large probability the pairwise squared distances between the points change by a factor of no more than $1 \pm \epsilon$. Note the original dimensionality does not matter. This is a statistical property based on concentration of measure. It is useful as an efficient dimension reduction tool.

The following theorem considers projecting a random vector onto a fixed subspace, which is equivalent to projecting a fixed vector onto a random subspace.

Lemma 1 *Let $Y \in \mathbb{R}^d$ be chosen uniformly from the surface of the d-dimensional sphere. Let $Z = (Y_1, Y_2, \ldots, Y_k)$ be the projection onto the first k coordinates, where $k < d$. Then for any $\alpha < 1$ and $\beta > 1$,*

$$Pr\left(\frac{d}{k}\|Z\|^2 \le \alpha\right) \quad \le \quad \exp\left(\frac{k}{2}(1-\alpha+\log\alpha)\right) \tag{1}$$

$$Pr\left(\frac{d}{k}\|Z\|^2 \ge \beta\right) \quad \le \quad \exp\left(\frac{k}{2}(1-\beta+\log\beta)\right). \tag{2}$$

With this, one can prove the Johnson-Lindenstrauss Lemma.

Theorem 1 (Johnson-Lindenstrauss) *Let $x_1, \ldots, x_n \in \mathbb{R}^d$, and let $\epsilon \in (0,1)$. Let k be a positive integer satisfying*

$$k \ge \frac{8\delta \log n}{\epsilon^2 - 2\epsilon^3/3} \tag{3}$$

where $\delta \ge 1$. Then a random projection $\Pi_k : \mathbb{R}^d \mapsto \mathbb{R}^k$ satisfies

$$Pr\left((1-\epsilon)\frac{k}{d}\|x_i - x_j\|^2 \le \|\Pi_k(x_i) - \Pi_k(x_j)\|^2 \le (1+\epsilon)\frac{k}{d}\|x_i - x_j\|^2, \forall i \ne j\right) \ge 1 - \frac{n(n-1)}{n^{2\delta}}. \tag{4}$$

If Π_k is a good projection, then the scaled mapping $f_k(x) = \sqrt{\frac{d}{k}}\Pi_k(x)$ satisfies

$$(1-\epsilon)\|x_i - x_j\|^2 \le \|f_k(x_i) - f_k(x_j)\|^2 \le (1+\epsilon)\|x_i - x_j\|^2, \forall i \ne j. \tag{5}$$

That is, f_k approximately preserves distance.

In addition to preserving pairwise distances, random projection also approximately preserves inner products.

Theorem 2 *Let $x, y \in R^d$ with $\|x\|_2, \|y\|_2 \le 1$. Assume that Φ is a $k \times d$ random matrix with independent $N(0, 1/d)$ entries. Then for all $\epsilon > 0$,*

$$Pr\left(\left|\frac{d}{k}(\Phi x)^\top(\Phi y) - x^\top y\right| \ge \epsilon\right) \le 2\exp\left(\frac{-k\epsilon^2}{C_1 + C_2\epsilon}\right) \tag{6}$$

where $C_1 = 4e/\sqrt{6\pi} \approx 2.5$ and $C_2 = \sqrt{8e} \approx 7.7$.

2 Compressive Sensing

Consider signal $f \in \mathbb{R}^n$, e.g., an image with n pixels. Assuming there is some orthonormal basis $\Psi_{n\times n} = [\psi_1 \dots \psi_n]$, e.g. wavelets, that

$$f(t) = \sum_{i=1}^{n} \psi_i(t) x_i. \tag{7}$$

The intuition is that the coefficients $x = [x_1 \dots x_n]$ is sparse (having many zeros) or nearly so for many real signals under an appropriate basis. You may not know which coefficients are significant, though (i.e., x may not be sorted in any way). Say you don't see f or x. Instead, you can take a few measurements. A measurement is

$$y_j = \phi_j^\top f + \epsilon_j = \phi_j^\top \Psi x + z, \tag{8}$$

where $\phi_j \in \mathbb{R}^n$ is a sensing vector that you choose, and z is noise. Your noisy measurement is y_j. How many measurements do you need in order to recover f? Clearly, if it is noiseless, n measurements with $\phi_j = e_j$ (the canonical basis, or in fact any basis) is sufficient to recover f. Can you do better?

Say x is S-sparse, i.e., having S nonzero elements. If you know the location of those nonzero elements, you only need S measurements with $\phi_j = \psi_k$ where k is a nonzero location in x. What if you do not know the nonzero locations? What if you do not even know Ψ before you measure the signal? Is there a way to take advantage of the knowledge that x is S-sparse?

Compressive sensing offers a surprising solution: you only need $O(S \log(n/S))$ *random* measurements, and there is a very efficient way to recover x (or f). Let us consider the $m \times n$ sensing matrix

$$A = \Phi\Psi \tag{9}$$

where $\Phi = [\phi_1 \dots \phi_m]^\top$ and $m \le n$. We have

$$y = Ax + z, \tag{10}$$

where y is the vector of m measurements.

For integer S, define the isometry constant δ_S of a matrix A to be the smallest number such that

$$(1 - \delta_S)\|x\|^2 \le \|Ax\|^2 \le (1 + \delta_S)\|x\|^2 \tag{11}$$

for all S-sparse x. Roughly speaking, the matrix A has the *restricted isometry property* (RIP) of order S if δ_S is not close to one. If our goal is to recover S-sparse signal x from y and A is RIP of order $2S$, then any difference between two S-sparse targets $x_i - x_j$ (which is at most $2S$-sparse) is approximately preserved in the measurements y_i and y_j:

$$(1 - \delta_{2S})\|x_i - x_j\|^2 \le \|y_i - y_j\|^2 = \|A(x_i - x_j)\|^2 \le (1 + \delta_{2S})\|x_i - x_j\|^2. \tag{12}$$

Conceptually, this allows us to "enumerate" all S-sparse x' and compare its measurement $y' = Ax'$ to the actual observed measurement y. The closest x' is the solution. As we see below, there is a much more elegant algorithm.

Theorem 3 (Noiseless Case) *Assume $\delta_{2S} < \sqrt{2} - 1$. Given measurement $y = Ax$, the solution x^* to*

$$\min_{x' \in \mathbb{R}^n} \quad \|x'\|_1 \tag{13}$$

$$s.t. \quad Ax' = y \tag{14}$$

obeys

$$\|x^* - x\|_2 \le C_0/\sqrt{S}\|x - x_S\|_1 \tag{15}$$

$$\|x^* - x\|_1 \le C_0\|x - x_S\|_1, \tag{16}$$

where x_S is the vector x with all but the largest S components set to 0.

Note if x is already S-sparse, this indicates perfect recovery. Also note that this involves a tractable ℓ_1 minimization problem.

Theorem 4 (Noisy Case) *Assume $\delta_{2S} < \sqrt{2} - 1$. Given noisy measurement $y = Ax + z$, the solution x^* to the LASSO problem*

$$\min_{x' \in \mathbb{R}^n} \quad \|x'\|_1 \tag{17}$$

$$s.t. \quad \|Ax' - y\| \leq \epsilon \tag{18}$$

obeys

$$\|x^* - x\|_2 \quad \leq \quad C_0/\sqrt{S}\|x - x_S\|_1 + C_1\epsilon. \tag{19}$$

These theorems assume that we have A with the RIP property. Recall our $A = \Phi\Psi$ where Ψ is a fixed orthonormal basis. It turns out that one can let the entries of Φ be

1. sampling n column-vectors uniformly at random on the unit sphere in $\mathbb{R}^m$; or
2. iid samples from $N(0, 1/m)$; or
3. iid samples from Bernoulli$(0.5, 0.5)$ on $\phi_{ij} = \pm 1/\sqrt{m}$.

When

$$m \geq CS \log(n/S), \tag{20}$$

with overwhelming probability, the resulting A obeys the RIP. Also note that these designs of the sensing matrix Φ is independent of Ψ. This means that sensing is "universal" and can be done without knowing what is the sparse basis Ψ of the signal (of course, one needs to know Ψ during recovery).

3 Matrix Completion

Let M be an $n_1 \times n_2$ matrix of rank r. Suppose we observe m entries of M. How large does m have to be to recover M? We will show that it is a small number. However, there are a few conditions.

Note the observed entries cannot be adversarially placed – if we miss a whole row when M is rank-1 outer product, there is no way to recover M. Therefore, one assumes that the locations are sampled uniformly at random.

It is not enough for M to be low rank. Consider $M = e_1 e_1^\top$. It is very difficult to hit the 1 by chance. Instead, we consider the following family of M's.

Definition 1 *Let U be a subspace of $\mathbb{R}^n$ of dimension r, and P_U be the orthogonal projection onto U. Then the coherence of U is defined as*

$$\mu(U) = \frac{n}{r} \max_{1 \leq i \leq n} \|P_u e_i\|^2. \tag{21}$$

We are interested in low coherence subspaces. Let the SVD of M be

$$M = \sum_{k=1}^{r} \sigma_k u_k v_k^\top \tag{22}$$

with column and row spaces be U and V, respectively. The M we consider has two properties:

1. The coherence $\max(\mu(U), \mu(V)) \leq \mu_0$ for some positive μ_0;
2. The $n_1 \times n_2$ matrix $\sum_{k=1}^{r} u_k v_k^\top$ has a maximum entry bounded by $\mu_1\sqrt{r/(n_1 n_2)}$ in absolute value for some positive μ_1.

For such M, we have the following theorem.

Theorem 5 *Let M be an $n_1 \times n_2$ matrix of rank r satisfying the above two conditions. Suppose we observe m entries with locations sampled uniformly at random. Then there exist constants C, c such that if*

$$m \geq C \max(\mu_1^2, \mu_0^{1/2}\mu_1, \mu_0 n^{1/4}) nr(\beta \log n) \tag{23}$$

for some $\beta > 2$, then the minimizer to the nuclear norm minimization problem

$$\min_{X \in \mathbb{R}^{n_1 \times n_2}} \quad \|X\|_* \tag{24}$$

$$s.t. \quad X_{ij} = M_{ij} \text{ for observed locations } (i, j) \tag{25}$$

is unique and equal to M with probability at least $1 - cn^{-\beta}$. For $r \leq \mu_0^{-1} n^{1/5}$ the bound can be improved to

$$m \geq C\mu_0 n^{6/5} r(\beta \log n) \tag{26}$$

with the same probability of success.

Here, the nuclear norm $\|X\|_* = \sum_{k=1}^{r} \sigma_k$ is the sum of singular values of X. It is a convex approximation to the rank of X, i.e., the number of nonzero singular values. When X is symmetric and positive semi-definite, its nuclear norm is the same as its trace.