Lecture 4. Random Projections and Johnson-Lindenstrauss Lemma

Yuan Yao

Hong Kong University of Science and Technology

March 11, 2020

Outline

PCA and MDS

- Data matrix: $X = [x_1, \ldots, x_n] \in \mathbb{R}^{p\times n}$
  - Centering: $Y = XH$, where $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$

- Singular Value Decomposition $Y = USV^T$, $S = \mathbf{diag}(\sigma_j)$, $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{\min(n,p)}$
  - PCA is given by top-k SVD (S_k, U_k): $U_k = (u_1, \ldots, u_k) \in \mathbb{R}^{p\times k}$, with embedding coordinates $U_k S_k$
  - MDS is given by top-k SVD (S_k, V_k): $V_k = (v_1, \ldots, v_k) \in \mathbb{R}^{n\times k}$, with embedding coordinates $V_k S_k$
  - Kernel PCA (MDS): for $K \succeq 0$, $K_c = HKH^T$, $K_c = U\Lambda U^T$ gives MDS embedding $U_k \Lambda_k^{1/2} \in \mathbb{R}^{n\times k}$

Computational Concerns: Big Data and High Dimensionality

- Big Data: n is large
  - Downsample for approximate PCA:

$$\widehat{\Sigma}_{n'} = \frac{1}{n'} \sum_{i=1}^{n'} (x_i - \widehat{\mu}_{n'})(x_i - \widehat{\mu}_{n'})^T, \qquad \widehat{\Sigma}_{n'} = U \Lambda U^T$$

  - Nyström Approximation for MDS: $V_k = (v_1, \ldots, v_k) \in \mathbb{R}^{n \times k}$ (we'll come to this in Manifold Learning - ISOMAP)
- High Dimensionality: p is large
  - Random Projections for PCA: $RXH = \tilde{U}\tilde{S}\tilde{V}^T$ with random matrix $R^{d \times p}$ (today): $\tilde{U}_k = (\tilde{u}_1, \ldots, \tilde{u}_k) \in \mathbb{R}^{d \times k}$
  - Perturbation of MDS: $\tilde{V}_k = (\tilde{v}_1, \ldots, \tilde{v}_k) \in \mathbb{R}^{n \times k}$

Outline

Random Projections: Examples

- $R = [r_1, \cdots, r_k]$, $r_i \sim U(S^{d-1})$, e.g. $r_i = (a_1^i, \cdots, a_d^i) / \parallel a^i \parallel$ $a_k^i \sim N(0,1)$

- $R = A/\sqrt{k} \quad A_{ij} \sim N(0,1)$

- $R = A/\sqrt{k} \quad A_{ij} = \begin{cases} 1 & p = 1/2 \\ -1 & p = 1/2 \end{cases}$

- $R = A/\sqrt{k/s} \quad A_{ij} = \begin{cases} 1 & p = 1/(2s) \\ 0 & p = 1 - 1/s \\ -1 & p = 1/(2s) \end{cases}$

  where $s = 1, 2, \sqrt{D}, D/\log D$, etc.

Example: Human Genomics Diversity Project

- Now consider a SNPs (Single Nucleid Polymorphisms) dataset in Human Genome Diversity Project (HGDP),

  http://www.cephb.fr/en/hgdp_panel.php

  – Data matrix of n-by-p for $n = 1,064$ individuals around the world and $p = 644,258$ SNPs.

  – Each entry in the matrix has 0, 1, 2, and 9, representing "AA", "AC", "CC", and "missing value", respectively.

  – After removing 21 rows with all missing values, we are left with a matrix X of size $1,043 \times 644,258$.

Original MDS (PCA)

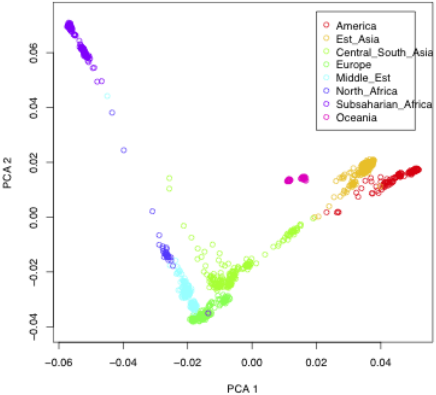
- Projection of 1,043 persons on the top-2 MDS (PCA) coordinates.
  - Define

$$K = HXX^T H = U\Lambda U^T, \qquad H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$$

    which is a positive semi-define matrix as centered Gram matrix whose eigenvalue decomposition is given by $U\Lambda U^T$.
  - Take the first two eigenvectors $\sqrt{\lambda_i}u_i$ $(i = 1, \ldots, 2)$ as the projections of n individuals.

Figure: Original MDS (PCA)

Projection of 1,043 individuals on the top-2 MDS principal components, shows a continuous trajectory of human migration in history: human origins from Africa, then migrates to the Middle East, followed by one branch to Europe and another branch to Asia, finally spreading into America and Oceania.



Random Projection MDS (PCA)

- To reduce the computational cost due to the high dimensionality $p = 644,258$, we randomly select (without replacement) $\{n_i, i = 1, \ldots, k\}$ from $1, \ldots, p$ with equal probability. Let $R \in \mathbb{R}^{k \times p}$ is a Bernoulli random matrix satisfying:

$$R_{ij} = \begin{cases} 1/k & j = n_i, \\ 0 & otherwise. \end{cases}$$

Now define

$$\widetilde{K} = H(XR^T)(RX^T)H$$

whose eigenvectors leads to new principal components of MDS.

Figure: Comparisons of Random Projected MDS with Original One

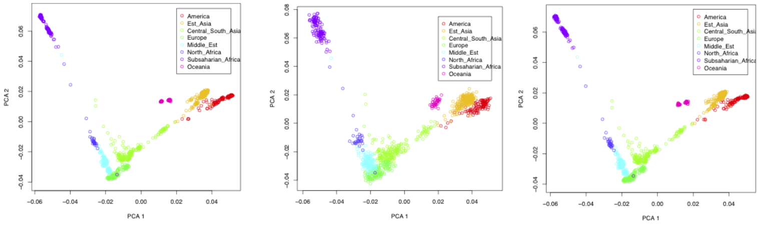


Figure: (Left) Projection of 1043 individuals on the top 2 MDS principal components. (Middle) MDS computed from 5,000 random columns. (Right) MDS computed from 100,000 random columns. Pictures are due to Qing Wang.

Question

How does the Random Projection work?

General MDS

- Given pairwise distances d_{ij} between n sample points, MDS aims to find $Y := [y_i]_{i=1}^n \in \mathbb{R}^{k \times n}$ such that the following sum of square is minimized,

$$\begin{aligned} \min_{Y=[y_1,\ldots,y_n]} \quad & \sum_{i,j} (\|y_i - y_j\|^2 - d_{ij}^2)^2 \\ \text{subject to} \quad & \sum_{i=1}^n y_i = 0 \end{aligned} \tag{1}$$

i.e. the total distortion of distances is minimized.

Metric MDS

- When $d_{ij} = \|x_i - x_j\|$ is exactly given by the distances of points in Euclidean space $x_i \in \mathbb{R}^p$, classical (metric) MDS defines a positive semidefinite kernel matrix $K = -\frac{1}{2}HDH$ where $D = (d_{ij}^2)$ and $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$. Then, the minimization (1) is equivalent to

$$\min_{Y \in \mathbb{R}^{k \times n}} \quad \|Y^T Y - K\|_F^2 \tag{2}$$

  i.e. the total distortion of distances is minimized by setting the column vectors of Y as the eigenvectors corresponding to k largest eigenvalues of K.

MDS toward Minimal Total Distortion

- The main features of MDS are the following.
  - MDS looks for Euclidean embedding of data whose *total* or *average* metric distortion are minimized.
  - MDS embedding basis is *adaptive* to the data, *e.g.* as a function of data via spectral decomposition.
- Can we have a tighter control on metric distortions, *e.g.* uniform distortion control?

Uniformly Almost-Isometry?

- What if a *uniform* control on metric distortion: there exists a $\epsilon \in (0,1)$, such that for every (i,j) pair,

$$(1-\epsilon) \leq \frac{\|y_i - y_j\|^2}{d_{ij}^2} \leq (1+\epsilon)?$$

  It is a uniformly almost isometric embedding or a Lipschitz mapping from metric space $\mathcal{X}$ to $\mathcal{Y}$.

- An beautiful answer is given by Johnson-Lindenstrauss Lemma, if $\mathcal{X}$ is an Euclidean space (or more generally Hilbert space), that $\mathcal{Y}$ can be a subspace of dimension $k = O(\log n/\epsilon^2)$ via random projections to obtain an almost isometry with high probability.

Johnson-Lindenstrauss Lemma

Theorem (Johnson-Lindenstrauss Lemma)

For any $0 < \epsilon < 1$ and any integer n, let k be a positive integer such that

$$k \geq (4 + 2\alpha)(\epsilon^2/2 - \epsilon^3/3)^{-1} \ln n, \quad \alpha > 0.$$

Then for any set V of n points in $\mathbb{R}^p$, there is a map $f : \mathbb{R}^p \to \mathbb{R}^k$ such that for all $u, v \in V$

$$(1-\epsilon) \parallel u - v \parallel^2 \leq \parallel f(u) - f(v) \parallel^2 \leq (1+\epsilon) \parallel u - v \parallel^2 \qquad (3)$$

Such a f in fact can be found in randomized polynomial time. In fact, inequalities (3) holds with probability at least $1 - 1/n^\alpha$.

Remark

- Almost isometry is achieved with a **uniform** metric distortion bound (*Bi-Lipschitz* bound), with high probability, rather than average metric distortion control;
- The mapping is **universal**, rather than being adaptive to the data.
- The theoretical basis of this method was given as a lemma by Johnson and Lindenstrauss (1984) in the study of a Lipschitz extension problem in Banach space.
- In 2001, Sanjoy Dasgupta and Anupam Gupta, gave a simple proof of this theorem using elementary probabilistic techniques in a four-page paper. Below we are going to present a brief proof of Johnson-Lindenstrauss Lemma based on the work of Sanjoy Dasgupta, Anupam Gupta, and Dimitris Achlioptas.

Note

- The distributions of the following two events are identical:

$$\begin{array}{rl} & \text{unit vector was randomly projected to } k\text{-subspace} \\ \Longleftrightarrow & \text{random vector on } S^{d-1} \text{ fixed top-}k \text{ coordinates.} \end{array}$$

  Based on this observation, we change our target from random k-dimensional projection to random vector on sphere S^{d-1}.

  - Let $x_i \sim N(0,1)$ $(i = 1, \cdots, p)$, and $X = (x_1, \cdots, x_p)$, then $Y = X/\|x\| \in S^{p-1}$ is uniformly distributed.
  - Fixing top-k coordinates, we get $z = (x_1, \cdots, x_k, 0, \cdots, 0)^T/\|x\| \in \mathbb{R}^p$. Let $L = \|z\|^2$ and $\mu := k/p$. Note that $\mathbf{E}\,\|(x_1, \cdots, x_k, 0, \cdots, 0)\|^2 = k = \mu \cdot \mathbf{E}\,\|x\|^2$.
  - The following lemma shows that L is concentrated around μ.

Key Lemma

Lemma

For any $k < p$, there hold

(a) *if $\beta < 1$ then*

$$\mathbf{Prob}[L \leq \beta\mu] \leq \beta^{k/2}\left(1 + \frac{(1-\beta)k}{p-k}\right)^{(p-k)/2} \leq \exp\left(\frac{k}{2}(1 - \beta + \ln\beta)\right)$$

(b) *if $\beta > 1$ then*

$$\mathbf{Prob}[L \geq \beta\mu] \leq \beta^{k/2}\left(1 + \frac{(1-\beta)k}{p-k}\right)^{(p-k)/2} \leq \exp\left(\frac{k}{2}(1 - \beta + \ln\beta)\right)$$

Here $\mu = k/p$.

Proof of Johnstone-Lindenstrauss Lemma

- If $p \leq k$,the theorem is trivial.
- Otherwise take a random k-dimensional subspace S, and let v_i' be the projection of point $v_i \in V$ into S, then setting $L = \|v_i' - v_j'\|^2$ and $\mu = (k/p)\|v_i - v_j\|^2$ and applying Lemma 1(a), we get that

$$
\begin{aligned}
\mathbf{Prob}[L \leq (1-\epsilon)\mu] &\leq \exp\left(\frac{k}{2}(1-(1-\epsilon)+\ln(1-\epsilon))\right) \\
&\leq \exp\left(\frac{k}{2}(\epsilon - (\epsilon + \frac{\epsilon^2}{2}))\right), \\
&\qquad \text{by } \ln(1-x) \leq -x - x^2/2 \text{ for } 0 \leq x < 1 \\
&= \exp\left(-\frac{k\epsilon^2}{4}\right) \leq \exp(-(2+\alpha)\ln n), \\
&\qquad \text{for } k \geq 4(1+\alpha/2)(\epsilon^2/2)^{-1}\ln n \\
&= \frac{1}{n^{2+\alpha}}
\end{aligned}
$$

Proof of Johnstone-Lindenstrauss Lemma (continued)

- Similarly, we can apply Lemma 1(b) to get

$$\begin{aligned}
\mathbf{Prob}[L \geq (1+\epsilon)\mu] &\leq \exp\left(\frac{k}{2}(1-(1+\epsilon)+\ln(1+\epsilon))\right) \\
&\leq \exp\left(\frac{k}{2}(-\epsilon+(\epsilon-\frac{\epsilon^2}{2}+\frac{\epsilon^3}{3}))\right), \\
&\qquad \text{by } \ln(1+x) \leq x - x^2/2 + x^3/3 \text{ for } x \geq 0 \\
&= \exp\left(-\frac{k}{2}(\epsilon^2/2-\epsilon^3/3)\right) \leq \exp(-(2+\alpha)\ln n), \\
&\qquad \text{for } k \geq 4(1+\alpha/2)(\epsilon^2/2-\epsilon^3/3)^{-1}\ln n \\
&= \frac{1}{n^{2+\alpha}}
\end{aligned}$$

-

Proof of Johnstone-Lindenstrauss Lemma (continued)

- Now set the map $f(x) = \sqrt{\frac{d}{k}} x' = \sqrt{\frac{d}{k}} (x_1, \ldots, x_k, 0, \ldots, 0)$. By the above calculations, for some fixed pair i, j, the probability that the distortion
$$\frac{\|f(v_i) - f(v_j)\|^2}{\|v_i - v_j\|^2}$$
does not lie in the range $[(1-\epsilon), (1+\epsilon)]$ is at most $\frac{2}{n^{(2+\alpha)}}$. Using the trivial union bound with $\binom{n}{2}$ pairs, the chance that some pair of points suffers a large distortion is at most:
$$\binom{n}{2} \frac{2}{n^{(2+\alpha)}} = \frac{1}{n^\alpha}\left(1 - \frac{1}{n}\right) \leq \frac{1}{n^\alpha}.$$
Hence f has the desired properties with probability at least $1 - \frac{1}{n^\alpha}$. This gives us a randomized polynomial time algorithm. $\square$

Proof of Lemma 1

- For Lemma 1(a),

$$\begin{aligned}
\mathbf{Prob}(L \le \beta\mu) &= \mathbf{Prob}\left(\sum_{i=1}^{k} x_i^2 \le \beta\mu(\sum_{i=1}^{p} x_i^2)\right) \\
&= \mathbf{Prob}\left(\beta\mu \sum_{i=1}^{p} x_i^2 - \sum_{i=1}^{k} x_i^2 \ge 0\right) \\
&= \mathbf{Prob}\left[\exp\left(t\beta\mu \sum_{i=1}^{p} x_i^2 - t\sum_{i=1}^{k} x_i^2\right) \ge 1\right], \quad (t > 0) \\
&\le \mathbf{E}\left[\exp\left(t\beta\mu \sum_{i=1}^{p} x_i^2 - t\sum_{i=1}^{k} x_i^2\right)\right]
\end{aligned}$$

(by Markov's inequality)

Proof of Lemma 1 (continued)

$$
\begin{aligned}
r.h.s. =& \Pi_{i=1}^{k} \mathbf{E} \exp(t(\beta\mu - 1)x_i^2) \Pi_{i=k+1}^{p} \mathbf{E} \exp(t\beta\mu x_i^2) \\
=& (\mathbf{E} \exp(t(\beta\mu - 1)x^2))^k (\mathbf{E} \exp(t\beta\mu x^2))^{p-k} \\
=& (1 - 2t(\beta\mu - 1))^{-k/2} (1 - 2t\beta\mu)^{-(p-k)/2} =: g(t)
\end{aligned}
$$

where the last equation uses the fact that if $X \sim \mathcal{N}(0,1)$, then

$$
\mathbf{E}[e^{sX^2}] = \frac{1}{\sqrt{(1-2s)}},
$$

for $-\infty < s < 1/2$.

Proof of Lemma 1 (continued)

- Now we will refer to last expression as $g(t)$.
  - The last line of derivation gives us the additional constraints that $t\beta\mu \leq 1/2$ and $t(\beta\mu - 1) \leq 1/2$, and so we have $0 < t < 1/(2\beta\mu)$.
  - Now to minimize $g(t)$, which is equivalent to maximize
  $$h(t) = 1/g(t) = (1 - 2t(\beta\mu - 1))^{k/2}(1 - 2t\beta\mu)^{(p-k)/2}$$
  in the interval $0 < t < 1/(2\beta\mu)$. Setting the derivative $h'(t) = 0$, we get the maximum is achieved at
  $$t_0 = \frac{1-\beta}{2\beta(p - \beta k)}$$
  Hence we have
  $$h(t_0) = \left(\frac{p-k}{p-k\beta}\right)^{(p-k)/2}\left(\frac{1}{\beta}\right)^{k/2},$$
  and this is exactly what we need.
- Similar derivation is for the proof of Lemma 1 (b). □

Outline

Locality Sensitive Hashing (LSH)

- (M.S. Charikar 2002) A **locality sensitive hashing** scheme is a distribution on a family $\mathcal{F}$ of hash functions operating on a collection of objects, such that for two objects x, y

$$\mathop{\mathbf{Prob}}_{h \in \mathcal{F}}[h(x) = h(y)] = \mathrm{sim}(x, y)$$

  where $\mathrm{sim}(x, y) \in [0, 1]$ is some similarity function defined on the collection of objects.

- Such a scheme leads to efficient (sub-linear) algorithms for approximate nearest neighbor search and clustering.

LSH via Random Projections

- (Goemans and Williamson (1995); Charikar (2002)) Given a collection of vectors in R^d, we consider the family of hash functions defined as follows: We choose a random vector $\vec{r}$ from the d-dimensional Gaussian distribution (i.e. each coordinate is drawn the 1-dimensional Gaussian distribution). Corresponding to this vector $\vec{r}$, we define a hash function $h_{\vec{r}}$ as follows:

$$h_{\vec{r}}(\vec{u}) = \mathbf{sign}(\vec{r} \cdot \vec{u}) = \begin{cases} 1 & \text{if } \vec{r} \cdot \vec{u} \geq 0 \\ -1 & \text{if } \vec{r} \cdot \vec{u} < 0 \end{cases}$$

Then for vectors $\vec{u}$ and $\vec{v}$

$$\mathbf{Pr}\left[h_{\vec{r}}(\vec{u}) = h_{\vec{r}}(\vec{v})\right] = 1 - \frac{\theta(\vec{u}, \vec{v})}{\pi}$$

Compressed Sensing

- Compressive sensing can be traced back to 1950s in signal processing in geography. Its modern version appeared in LASSO (Tibshirani, 1996) and Basis Pursuit (Chen-Donoho-Saunders, 1998), and achieved a highly noticeable status after 2005 due to the work by Candes and Tao et al.

- The basic problem of compressive sensing can be expressed by the following under-determined linear algebra problem. Assume that a signal $x^* \in \mathbb{R}^p$ is sparse with respect to some basis (measurement matrix) $A \in \mathbb{R}^{n \times p}$ or $A \in \mathbb{R}^{n \times p}$ where $n < p$, given measurement $b = Ax^* = Ax^* \in \mathbb{R}^n$, how can one recover x^* by solving the linear equation system
$$Ax = b? \tag{4}$$

Sparsity

- As $n < p$, it is an under-determined problem, whence without further constraint, the problem does not have an unique solution. To overcome this issue, one popular assumption is that the signal x^* is sparse, namely the number of nonzero components $\|x^*\|_0 := \#\{x_i^* \neq 0 : 1 \leq i \leq p\}$ is small compared to the total dimensionality p. Figure below gives an illustration of such sparse linear equation problem.

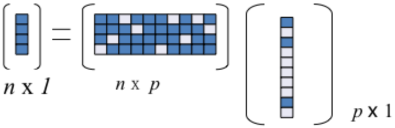


Figure: Illustration of Compressive Sensing (CS). A is a rectangular matrix with more columns than rows. The dark elements represent nonzero elements while the light ones are zeroes. The signal vector x^*, although high dimensional, is sparse.

P_0

Without loss of generality, we assume each column of design matrix $A = [A_1, \ldots, A_p]$ has being standardized, that is, $\|A_j\|_2 = 1$, $j = 1, ..., p$.

- With such a sparse assumption above, a simple idea is to find the sparsest solution satisfying the measurement equation:

$$\begin{aligned} (P_0) \quad \min \quad & \|x\|_0 \\ s.t. \quad & Ax = b. \end{aligned} \tag{5}$$

- This is an NP-hard combinatorial optimization problem.

A Greedy Algorithm: Orthogonal Matching Pursuit

Input A, b.

Output x.

initialization: $r_0 = b$, $x_0 = 0$, $S_0 = \emptyset$.

repeat if $\|r_t\|_2 > \varepsilon$,

1. $j_t = \arg\max_{1 \le j \le p} | \langle A_j, r_{t-1} \rangle |$.
2. $S_t = S_{t-1} \cup j_t$.
3. $x_t = \arg\min_{x \in \mathbb{R}^p} \|b - A_{S_t} x\|$.
4. $r_t = b - A x_t$.

return x^t.

- Stephane Mallat and Zhifeng Zhang (1993), choose the column of maximal correlation with residue, as the steepest descent in residue.
- Joel Tropp (2004) shows that OMP recovers x^* under the Incoherence condition; Tony Cai and Lie Wang (2011) extended it to noisy cases.

Basis Pursuit (BP): P_1

- A convex relaxation of (5) is called *Basis Pursuit* (Chen-Donoho-Saunders, 1998),

$$(P_1) \quad \min \quad \|x\|_1 := \sum |x_i| \qquad (6)$$
$$s.t. \quad Ax = b.$$

  This is a tractable linear programming problem.

- Now a natural problem arises, under what conditions the linear programming problem (P_1) has the solution exactly solves (P_0), *i.e.* exactly recovers the sparse signal x^* ?
  - Donoho and Huo (2001) proposed Incoherence condition; Joel Tropp (2004) shows that BP recovers x^* under the Incoherence condition.

Illustration

Figure shows different projections of a sparse vector x^* under l_0, l_1 and l_2, from which one can see in some cases the convex relaxation (6) does recover the sparse signal solution in (5).

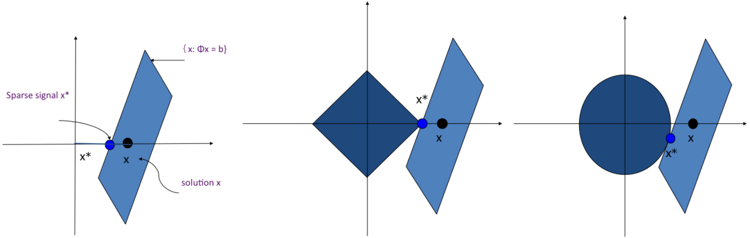


Figure: Comparison between different projections. Left: projection of x^* under $\|\cdot\|_0$; middle: projection under $\|\cdot\|_1$ which favors sparse solution; right: projection under Euclidean distance.

Basis Pursuit De-Noising (BPDN)

- When measurement noise exists, *i.e.* $b = Ax^* + \varepsilon$ with bound $\|\varepsilon\|_2$, the following Basis Pursuit De-Noising (BPDN) are used instead

$$
\begin{aligned}
(BPDN) \quad \min \quad & \|x\|_1 \\
s.t. \quad & \|Ax - b\|_2 \leq \epsilon.
\end{aligned} \tag{7}
$$

  It's a convex quadratic programming problem.

- Similarly, Jiang-Yao-Liu-Guibas (2012) considers ℓ_∞-noise:

$$
\begin{aligned}
\min \quad & \|x\|_1 \\
s.t. \quad & \|Ax - b\|_\infty \leq \epsilon.
\end{aligned}
$$

  This is a linear programming problem.

LASSO

Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996) solves the following problem for noisy measurement:

$$(LASSO) \quad \min_{x \in \mathbb{R}^p} \|Ax - b\|_2^2 + \lambda \|x\|_1 \tag{8}$$

- A convex quadratic programming problem.
- Yu-Zhao (2006), Lin-Yuan (2007), Wainwright (2009) show the model selection consistency (support recovery of x^*) of LASSO under the Irrepresentable condition.

Dantzig Selector

The Dantzig Selector (Candes and Tao (2007)) is proposed to deal with noisy measurement $b = Ax^* + \epsilon$:

$$\begin{aligned} \min \quad & \|x\|_1 \\ s.t. \quad & \|A^T(Ax - b)\|_\infty \leq \lambda \end{aligned} \tag{9}$$

- A linear programming problem, more scalable than convex quadratic programming (LASSO) for large scale problems.
- Bickel, Ritov, Tsybakov (2009) show that Dantzig Selector and LASSO share similar statistical properties.

Differential Inclusion: Inverse Scaled Spaces (ISS)

Differential inclusion:

$$\dot{\rho}_t = \frac{1}{n} A^T (b - A x_t), \tag{10a}$$

$$\rho_t \in \partial \|x_t\|_1. \tag{10b}$$

starting at $t = 0$ and $\rho_0 = \beta_0 = 0$.

- Replace $\frac{\rho}{t}$ in KKT condition of LASSO by $\frac{\mathrm{d}\rho}{\mathrm{d}t}$,

$$\frac{\rho_t}{t} = \frac{1}{n} A^T (b - A x_t), \qquad t = \frac{1}{\lambda}$$

  to achieve unbiased estimator $\hat{x}_t$ when it is sign-consistent.

Differential Inclusion: Inverse Scaled Spaces (ISS) (more)

- Burger-Gilboa-Osher-Xu (2006) (in image recovery it recovers the objects in an inverse-scale order as t increases (larger objects appear in x_t first))
- Osher-Ruan-Xiong-Yao-Yin (2016) shows that its solution is a debiasing regularization path, achieving model selection consistency under nearly the same conditions of LASSO.
  - Note: if $\hat{x}_\tau$ is sign consistent $\mathbf{sign}(\hat{x}_\tau) = \mathbf{sign}(x^*)$, then $\hat{x}_\tau = x^* + (A^T A)^{-1} A^T \varepsilon$ which is unbiased.
  - However for LASSO, if $\hat{x}_\lambda$ is sign consistent $\mathbf{sign}(\hat{x}_\lambda) = \mathbf{sign}(x^*)$, then $\hat{x}_\lambda = x^* + \lambda (A^T A)^{-1} \mathbf{sign}(x^*) + (A^T A)^{-1} A^T \varepsilon$ which is biased.

Example: Regularization Paths of LASSO vs. ISS

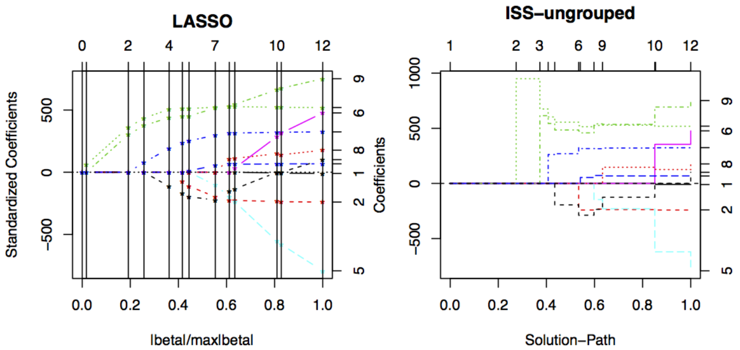


Figure: Diabetes data (Efron et al.'04) and regularization paths are different, yet bearing similarities on the order of parameters being nonzero

Linearized Bregman Iterations

A damped dynamics below has a continuous solution x_t that converges to the piecewise-constant solution of (10) as $\kappa \to \infty$.

$$\dot{\rho}_t + \frac{\dot{x}_t}{\kappa} = -\nabla_x \ell(x_t), \tag{11a}$$

$$\rho_t \in \partial\Omega(x_t), \tag{11b}$$

Its Euler forward discretization gives the *Linearized Bregman Iterations* (LBI, Osher-Burger-Goldfarb-Xu-Yin 2005) as

$$z_{k+1} = z_k - \alpha \nabla_x \ell(x_k), \tag{12a}$$

$$x_{k+1} = \kappa \cdot \mathrm{prox}_{\Omega}(z_{k+1}), \tag{12b}$$

where $z_{k+1} = \rho_{k+1} + \frac{x_{k+1}}{\kappa}$, the initial choice $z_0 = x_0 = 0$ (or small Gaussian), parameters $\kappa > 0,\ \alpha > 0,\ \nu > 0$, and the proximal map associated with a convex function Ω is defined by

$$\mathrm{prox}_{\Omega}(z) = \arg\min_x \frac{1}{2}\|z - x\|^2 + \Omega(x).$$

Uniform Recovery Conditions

- Under which conditions we can recover arbitrary k-sparse $x^* \in \mathbb{R}^p$ by those algorithms, for $k = |\mathbf{supp}(x^*)| \ll n < p$?

- Now we turn to several conditions presented in literature, under which the algorithms above can recover x^*. Below A_S denotes the columns of A corresponding to the indices in $S = \mathbf{supp}(x^*)$; A^* denotes the conjugate of matrix A, which is A^T if A is real.

Uniform Recovery Conditions: a) Uniqueness

a) Uniqueness. The following condition ensures the uniqueness of k-sparse x^* satisfying $b = Ax^*$:

$$A_S^* A_S \geq rI, \qquad \text{for some } r > 0,$$

without which one may have more than one k-sparse solutions in solving $b = A_S x$.

Uniform Recovery Conditions: b) Incoherence

b) Incoherence. Donoho-Huo (2001) shows the following sufficient condition

$$\mu(A) := \max_{i \neq j} |\langle A_i, A_j \rangle| < \frac{1}{2k-1},$$

for sparse recovery by BP, which is later improved by Elad-Bruckstein (2001) to be

$$\mu(A) < \frac{\sqrt{2} - \frac{1}{2}}{k}.$$

This condition is numerically **verifiable**, so the simplest condition.

Uniform Recovery Conditions: c) Irrepresentable

c) Irrepresentable condition. It is also called the Exact Recovery Condition (ERC) by Joel Tropp (2004), which shows that under the following condition

$$M =: \|A_{S^c}^* A_S (A_S^* A_S)^{-1}\|_\infty < 1,$$

both OMP and BP recover x^*.

- This condition is **unverifiable** since the true support set S is unknown.
- "Irrepresentable" is due to Yu and Zhao (2006) for proving LASSO's model selection consistency under noise, based on the fact that the regression coefficients of $A_j \sim A_S \beta + \varepsilon$ for $j \in S^c$, are the row vectors of $A_{S^c}^* A_S (A_S^* A_S)^{-1}$, suggesting that columns of A_S can not be linearly represented by columns of A_{S^c}.

Incoherence vs. Irrepresentable

- Tropp (2004) also shows that Incoherence condition is strictly stronger than the Irrepresentable condition in the following sense:

$$\mu < \frac{1}{2k-1} \Rightarrow M \leq \frac{k\mu}{1-(k-1)\mu} < 1. \tag{13}$$

- On the other hand, Tony Cai et al. (2009, 2011) shows that the Irrepresentable and the Incoherence condition are **both tight** in the sense that if it fails, there exists data A, x^*, and b such that sparse recovery is not possible.

Uniform Recovery Conditions: d) Restricted Isometry Property

d) Restricted-Isometry-Property (RIP) For all k-sparse $x \in \mathbb{R}^p$, $\exists \delta_k \in (0, 1)$, s.t.

$$(1 - \delta_k)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta_k)\|x\|_2^2.$$

- This is the most popular condition by Candes-Romberg-Tao (2006).
- Although RIP is not easy to be verified, **Johnson-Lindestrauss Lemma** says some suitable random matrices will satisfy RIP with high probability.

Restricted Isometry Property for Uniform Exact Recovery

Candes (2008) shows that under RIP, uniqueness of P_0 and P_1 can be guaranteed for all k-sparse signals, often called *uniform exact recovery*.

Theorem

The following holds for all k-sparse x^* satisfying $Ax^* = b$.

- If $\delta_{2k} < 1$, then problem P_0 has a unique solution x^*;
- If $\delta_{2k} < \sqrt{2} - 1$, then the solution of P_1 (BP) has a unique solution x^*, *i.e.* recovers the original sparse signal x^*.

Restricted Isometry Property for Stable Noisy Recovery

Under noisy measurement $b = Ax^* + \varepsilon$, Candes (2008) also shows that RIP leads to stable recovery of the true sparse signal x^* using BPDN.

Theorem

Suppose that $\|\varepsilon\|_2 \leq \epsilon$. If $\delta_{2k} < \sqrt{2} - 1$, then

$$\|\hat{x} - x^*\|_2 \leq C_1 k^{-1/2} \sigma_k^1(x^*) + C_2 \epsilon,$$

where $\hat{x}$ is the solution of BPDN and

$$\sigma_k^1(x^*) = \min_{\mathbf{supp}(y) \leq k} \|x^* - y\|_1$$

is the best k-term approximation error in l_1 of x^*.

JL $\Rightarrow$ RIP

- Johnson-Lindenstrauss Lemma ensures RIP with high probability.
- Baraniuk, Davenport, DeVore, and Wakin (2008) show that in the proof of Johnson-Lindenstrauss Lemma, one essentially establishes that a random matrix $A \in \mathbb{R}^{n \times p}$ with each element i.i.d. sampled according to some distribution satisfying certain bounded moment conditions, has $\|Ax\|_2^2$ concentrated around its mean $\mathbf{E}\,\|Ax\|_2^2 = \|x\|_2^2$ (see Appendix), *i.e.*

$$\mathbf{Prob}\left(\left|\|Ax\|_2^2 - \|x\|_2^2\right| \geq \epsilon\|x\|_2^2\right) \leq 2e^{-nc_0(\epsilon)}. \qquad (14)$$

With this one can establish a bound on the action of A on k-sparse x by an union bound via covering numbers of k-sparse signals.

JL $\Rightarrow$ RIP: Key Lemma

Lemma

Let $A \in \mathbb{R}^{n \times p}$ be a random matrix satisfying the concentration inequality (14). Then for any $\delta \in (0, 1)$ and any set all T with $|T| = k < n$, the following holds

$$(1 - \delta)\|x\|_2 \leq \|Ax\|_2 \leq (1 + \delta)\|x\|_2 \tag{15}$$

for all x whose support is contained in T, with probability at least

$$1 - 2\left(\frac{12}{\delta}\right)^k e^{-c_0(\delta/2)n}. \tag{16}$$

Proof of Lemma

It suffices to prove the results when $\|x\|_2 = 1$ as A is linear.

- Let $X_T := \{x : \mathbf{supp}(x) = T, \|x\|_2 = 1\}$. We first choose Q_T, a $\delta/4$-cover of X_T, such that for every $x \in X_T$ there exists $q \in Q_T$ satisfying $\|q - x\|_2 \leq \delta/4$. Since X_T has dimension at most k, it is well-known from covering numbers that the capacity $\#(Q_T) \leq (12/\delta)^k$.
- Now we are going to apply the union bound of (14) to the set Q_T with $\epsilon = \delta/2$. For each $q \in Q_T$, with probability at most $2e^{-c_0(\delta/2)n}$, $|Aq\|_2^2 - \|q\|_2^2 \geq \delta/2\|q\|_2^2$. Hence for all $q \in Q_T$, the same bound holds with probability at most

$$2\#(Q_T)e^{-c_0(\delta/2)n} \leq 2\left(\frac{12}{\delta}\right)^k e^{-c_0(\delta/2)n}.$$

Proof Lemma (continued)

- Now we define α to be the smallest constant such that

$$\|Ax\|_2 \leq (1+\alpha)\|x\|_2, \quad \text{for all } x \in X_T.$$

  We can show that $\alpha \leq \delta$ with the same probability.

- For this, pick up a $q \in Q_T$ such that $\|q - x\|_2 \leq \delta/4$, whence by the triangle inequality

$$\|Ax\|_2 \leq \|Aq\|_2 + \|A(x-q)\|_2 \leq 1 + \delta/2 + (1+\alpha)\delta/4.$$

  This implies that $\alpha \leq \delta/2 + (1+\alpha)\delta/4$, whence $\alpha \leq 3\delta/4/(1-\delta/4) \leq \delta$. This gives the upper bound. The lower bound also follows this since

$$\|Ax\|_2 \geq \|Aq\|_2 - \|A(x-q)\|_2 \geq 1 - \delta/2 - (1+\delta)\delta/4 \geq 1 - \delta,$$

  which completes the proof. □

RIP Theorem

- With this lemma, note that there are at most $\binom{p}{k}$ subspaces of k-sparse, an union bound leads to the following result for RIP.

Theorem

Let $A \in \mathbb{R}^{n \times p}$ be a random matrix satisfying the concentration inequality (14) and $\delta \in (0, 1)$. There exists $c_1, c_2 > 0$ such that if

$$k \leq c_1 \frac{n}{\log(p/k)}$$

the following RIP holds for all k-sparse x,

$$(1 - \delta_k)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta_k)\|x\|_2^2$$

with probability at least $1 - 2e^{-c_2 n}$.

Proof of RIP Theorem

Proof.

For each of k-sparse signal (X_T), RIP fails with probability at most

$$2\left(\frac{12}{\delta}\right)^k e^{-c_0(\delta/2)n}.$$

There are $\binom{p}{k} \leq (ep/k)^k$ such subspaces. Hence, RIP fails with probability at most

$$2\left(\frac{ep}{k}\right)^k \left(\frac{12}{\delta}\right)^2 e^{-c_0(\delta/2)n} = 2e^{-c_0(\delta/2)n+k[\log(ep/k)+\log(12/\delta)]}.$$

Thus for a fixed $c_1 > 0$, whenever $k \leq c_1 n/\log(p/k)$, the exponent above will be $\leq -c_2 n$ provided that

$$c_2 \leq c_0(\delta/2) - c_1(1 + (1 + \log(12/\delta))/\log(p/k)).$$

Note that one can always choose $c_2 > 0$ if $c_1 > 0$ is small enough. $\square$

Summary

The following results are about mean estimation under noise:

- Johnson-Lindenstrauss Lemma tells: random projections give a universal basis to achieve uniformly almost isometric embedding, using $O(\varepsilon^{-2}\log n)$ number of projections
- Various Applications
  - Dimensionality reduction: PCA or MDS
  - Locality Sensitive Hashing: clustering, nearest neighbor search, etc.
  - Compressed Sensing: random design satisfying Restricted Isometry Property with high probability

Outline

A Simple Version of Johnson-Lindenstrauss Lemma

Theorem (Simplified Johnson-Lindenstrauss Lemma)

Let $A = [A_{ij}]^{k \times d}$ where $A_{ij} \sim \mathcal{N}(0, 1)$ and $R = A/\sqrt{k}$. For any $0 < \epsilon < 1$ and any positive integer k, the following holds for all $0 \neq x \in \mathbb{R}^d$,

$$(1 - \epsilon) \leq \frac{\| Rx \|^2}{\|x\|^2} \leq (1 + \epsilon), \tag{17}$$

or for all $x \neq y \in \mathbb{R}^d$,

$$1 - \epsilon \leq \frac{\| Rx - Ry \|^2}{\|x - y\|^2} \leq 1 + \epsilon \tag{18}$$

with probability at least $1 - 2\exp\left(-\frac{k\varepsilon^2}{4}(1 - 2\varepsilon/3)\right)$.

Remark

- This version of JL-Lemma is essentially used in the derivation of RIP in compressed sensing.
- Extension to sub-Gaussian distributions with bounded moment conditions can be found in Joseph Salmon's lecture notes.
- Given n sample points $x_i \in V$. If we let

$$k \geq 4(1 + \alpha/2)(\epsilon^2/2 - \epsilon^3/3)^{-1} \ln n,$$

then

$$\mathbb{P}\left(\|Ru\|^2 \geq 1 + \varepsilon\right) \leq \exp(-(2 + \alpha) \log n) = \left(\frac{1}{n}\right)^{2+\alpha},$$

a union of $\binom{n}{2}$ probabilistic bounds gives the full JL-Lemma.

A Basic Lemma

Lemma

Let $X \sim \mathcal{N}(0, 1)$.

(a) For all $t \in (-\infty, 1/2)$,

$$\mathbf{E}(e^{tX^2}) = \frac{1}{1 - 2t}.$$

Proof.

(a) follows from Gaussian integral.

$\square$

Proof of JL Lemma

Let us denote $x \in \mathbb{R}^d, u = \frac{x}{\|x\|}$ and Y_i the column values of the output, i.e $Y_i = (Ru)_i = \sum_{j=1}^{d} R_{i,j} u_j$. Then,

$$\mathbb{E}\left(Y_i\right) = \mathbb{E}\left(\sum_{j=1}^{d} R_{i,j} u_j\right) = \sum_{j=1}^{d} \mathbb{E}\left(R_{i,j} u_j\right) = \sum_{j=1}^{d} u_j \mathbb{E}\left(R_{i,j}\right) = 0$$

$$\begin{aligned} \operatorname{Var}\left(Y_i\right) &= \operatorname{Var}\left(\sum_{j=1}^{d} R_{i,j} u_j\right) = \mathbb{E}\left(\sum_{j=1}^{d} R_{i,j} u_j\right)^2 = \sum_{j=1}^{d} \operatorname{Var}\left(R_{i,j} u_j\right) \\ &= \sum_{j=1}^{d} u_j^2 \operatorname{Var}\left(R_{i,j}\right) = \frac{1}{k} \end{aligned}$$

Proof of JL Lemma (continued)

(Upper) . Defining $Z_i = \sqrt{k} Y_i \sim \mathcal{N}(0,1)$, one can state the following bound:

$$\begin{aligned}
\mathbb{P}\left(\|Ru\|^2 \geq 1+\varepsilon\right) &= \mathbb{P}(\sum_{i=1}^{k}((\sqrt{k}Y_i)^2 - 1) \geq \varepsilon k) \\
&= \mathbb{P}(\sum_{i=1}^{k}(Z_i^2 - 1) \geq \varepsilon k) \\
&\leq e^{-t\varepsilon k} \prod_{i=1}^{k} \mathbf{E} \exp\left(t(Z_i^2 - 1)\right), \quad \text{(Markov Ineq.)} \\
&= e^{-tk(1+\varepsilon)} [\mathbf{E}\, e^{tZ^2}]^k \\
&= e^{-tk(1+\varepsilon)} (1-2t)^{-k/2} =: g(t) \quad \text{(Lemma (a))}
\end{aligned}$$

Proof of JL Lemma (continued)

Let

$$h(t) := 1/g(t) = e^{tk(1+\varepsilon)}(1-2t)^{k/2}.$$

Hence $\min_t g(t)$ is equivalent to $\max_t h(t)$. Taking derivative of $h(t)$,

$$\begin{aligned}
0 = h'(t)|_{t^*} &= k(1+\varepsilon)e^{tk(1+\varepsilon)}(1-2t)^{k/2} - ke^{tk(1+\varepsilon)}(1-2t)^{k/2-1}\Big)\Big|_{t^*} \\
&= ke^{t^*k(1+\varepsilon)}(1-2t^*)^{k/2-1}\left[(1+\varepsilon)(1-2t^*)-1\right] \\
&\Rightarrow t^* = \frac{1}{2} - \frac{1}{2(1+\varepsilon)}
\end{aligned}$$

$$\begin{aligned}
\Rightarrow g(t^*) &= e^{-t^*k(1+\varepsilon)}(1-2t^*)^{-k/2} = e^{-k\varepsilon/2}(1+\varepsilon)^{k/2} \\
&= \exp\left(-\frac{k\varepsilon}{2} + \frac{k}{2}\ln(1+\varepsilon)\right) \\
&\leq \exp\left(-\frac{k\varepsilon}{2} + \frac{k}{2}(\varepsilon - \frac{\varepsilon^2}{2} + \frac{\varepsilon^3}{3})\right), \quad \text{using } \ln(1+x) \leq x - \tfrac{x^2}{2} + \tfrac{x^3}{3} \\
&= \exp\left(-\frac{k\varepsilon^2}{4} + \frac{k\varepsilon^3}{6}\right), \quad \varepsilon \in (0,1)
\end{aligned}$$

Proof of JL Lemma (continued)

(Lower) . Similarly

$$
\begin{aligned}
\mathbb{P}\left(\|Ru\|^2 \leq 1-\varepsilon\right) &= \mathbb{P}(\sum_{i=1}^{k}(1-(\sqrt{k}Y_i)^2) \geq \varepsilon k) \\
&= \mathbb{P}(\sum_{i=1}^{k}(1-Z_i^2) \geq \varepsilon k) \\
&\leq e^{-t\varepsilon k}\prod_{i=1}^{k}\mathbf{E}\exp\left(t(1-Z_i^2)\right), \quad \text{(Markov Ineq.)} \\
&= e^{tk(1-\varepsilon)}[\mathbf{E}\, e^{-tZ^2}]^k \\
&= e^{tk(1-\varepsilon)}(1+2t)^{-k/2} =: g(t) \quad \text{(Lemma (a))}
\end{aligned}
$$

Proof of JL Lemma (continued)

Let

$$h(t) := 1/g(t) = e^{tk(\varepsilon-1)}(1+2t)^{k/2}.$$

Taking derivative of $h(t)$,

$$0 = h'(t)|_{t^*} = k(\varepsilon-1)e^{tk(\varepsilon-1)}(1+2t)^{k/2} + ke^{tk(\varepsilon-1)}(1+2t)^{k/2-1}\Big)\Big|_{t^*}$$

$$= ke^{t^*k(\varepsilon-1)}(1+2t^*)^{k/2-1}\left[(\varepsilon-1)(1+2t^*)+1\right]$$

$$\Rightarrow t^* = \frac{1}{2(1-\varepsilon)} - \frac{1}{2}$$

$$\Rightarrow g(t^*) = e^{t^*k(1-\varepsilon)}(1+2t^*)^{-k/2} = e^{k\varepsilon/2}(1-\varepsilon)^{k/2}$$

$$= \exp\left(\frac{k\varepsilon}{2} + \frac{k}{2}\ln(1-\varepsilon)\right)$$

$$\leq \exp\left(\frac{k\varepsilon}{2} + \frac{k}{2}(-\varepsilon - \frac{\varepsilon^2}{2})\right), \quad \text{using } \ln(1-x) \leq -x - \frac{x^2}{2}$$

$$= \exp\left(-\frac{k\varepsilon^2}{4}\right), \quad \varepsilon \in (0,1) \qquad \square$$