Random Methods for Linear Algebra

Alex Gittens
gittens@acm.caltech.edu

Applied and Computational Mathematics
California Institue of Technology

October 2, 2009

Outline

The Johnson-Lindenstrauss lemma says that a collection of n points in any Euclidean space can be mapped into a Euclidean space of dimension $k = \mathrm{O}(\log n)$ with little distortion:

Theorem

Let $\epsilon \in (0, \frac{1}{2})$, and let $P = \{p_1, \ldots, p_n\}$ be a set of n points in $\mathbb{R}^n$. Let k be an integer with $k \geq \mathrm{C}\epsilon^{-2} \log n$, where C is a sufficiently large absolute constant. Then there exists a mapping $f : \mathrm{R}^n \to \mathrm{R}^k$ such that

$$(1 - \epsilon)\|p_i - p_j\| \leq \|f(p_i) - f(p_j)\| \leq (1 + \epsilon)\|p_i - p_j\|$$

for all $i, j = 1, 2, \ldots, n$.

- The JL-lemma allows us to compress (an approximation of) the representation of a set of points from $\mathrm{O}(n^2)$ space to $\mathrm{O}(n \log n)$
- Better, the JL transforms can be done by using multiplication with a random matrix. Projection of all the points takes $\mathrm{O}(n^2 \log n)$ time.
- Even better, you can use particularly nice random matrices, and obtain a Fast JL Transform. Projection of all the points takes $\mathrm{O}(n \log n + \mathrm{polylog}(1/\epsilon, \log n))$.

Ailon and Chazelle introduced the FJLT

- earlier work showed you need $k = \mathrm{O}(\log n)$, so to decrease computation, considered using sparse projection matrices
- problem is that sparse projection matrices distort sparse vectors
- key idea: use FFT to increase the support of sparse vectors. use randomized FFT to avoid sparsification of dense vectors
- form of the random projection doesn't depend on the data
- with probability $2/3$, one application of the FJLT succeeds

Structure of the FJLT

$$\Phi = PHD$$

- P is a $k \times n$ matrix such that $P_{ij} \sim N(0, q^{-1})$ w.p. q and $P_{ij} = 0$ w.p. $1 - q$, where q depends on n, ϵ.
- H is an $n \times n$ normalized Hadamard matrix (real analog of the Fourier matrix)
- D is a $n \times n$ diagonal matrix with independent diagonal entries uniformly drawn from $1, -1$.

The magic lies in HD. Fix $p \in \mathbb{R}^n$ with unit length, then examine

$$(HDp)_1 = \sum_{j=1}^{n} H_{1j} d_{jj} p_j$$

Apply Hoeffding's inequality and the fact $|H_{ij}| = n^{-1/2}$ (gives an upper bound on the probability for a sum of bounded random variables to deviate from its mean):

$$\mathrm{P}(|(HDp)_1| \geq \sqrt{n} nt) \leq 2 \exp\left(-\frac{n(\sqrt{n}t)^2}{2}\right)$$

so with very high probability $|(HDp)_1| = \mathrm{O}(n^{-1/2})$.

H and D are isometries, and the $(HDp)_i$ are i.i.d, so it follows that they are all on the order of $n^{-1/2}$.

- we have densification of sparse vectors without sparsification of dense vectors.
- P is tailored appropriately to get a JLT.
- Randomized FFTs have become very popular (SRFTs)

The Low Rank Approximation Problem

Given *large* $A \in \mathbb{R}^{m \times n}$, efficiently approximate the solution to

$$\min_{\text{rank}(B) \leq k} \|A - B\|_\xi, \quad \xi \in \{F, 2\}.$$

- Cost of dense SVD (entire SVD): $\mathrm{O}(\min\{mn^2, m^2 n\})$
- Cost of sparse SVD (top k singular vectors): $\mathrm{O}(kmn + k^2 m)$

Ways to use randomization:

- Reduce the size of the matrix, e.g. by biased sampling of its rows or columns, take the SVD of this matrix, show this is close to the SVD of the original. The 'Monte-Carlo' approach.
- Approximate the matrix with a random sparse matrix, show that the SVD of this is close to the SVD of the original.
- Find a rank-k basis Q which minimizes $\|A - QQ^\star A\|$ and take $B = QQ^\star A$.

Frieze, Kannan, Vempala's work on Monte Carlo method is seminal

- key observation is that a good low rank approximation to A (in Frobenius norm) lies in the span of a small subset of its rows.
- algorithm samples rows of A to form X, biased according to their share of $\|A\|_F$. After scaling, SVD of $X^T X$ approximates that of $A^T A$.

Rudelson and Vershynin showed that if numerical rank of A is r, then by sampling $\mathrm{O}(r \log r)$ rows of A, can construct a matrix P_k so

$$\|A - AP_k\|_2 \leq \sigma_{k+1}(A) + \epsilon \|A\|_2.$$

Randomized Sparsification Schemes

- Achlioptas and McSherry considered replacing A with a sparse X using the scheme

$$X_{jk} = \begin{cases} a_{jk}/p_{jk}, & \text{w. p. } p_{jk} \\ 0, & \text{otherwise} \end{cases}$$

  where $p_{jk} = p(a_{ij}/b)^2$ (more or less). b is the largest absolute value in A.
- Key idea: sparsification similar to adding gaussian noise, which has weak spectral features.

Our idea: develop bounds for the error of sparsification schemes which satisfy $\mathbb{E}X = A$ and the entries of X are independent. Want a bound on the expected spectral norm of $Z = A - X$

$$\begin{aligned}\mathbb{E}\|Z\|_2 \quad \leq \quad & C\left[\max_j \left(\sum_k \mathbb{E}Z_{jk}^2\right)^{1/2} + \max_k \left(\sum_j \mathbb{E}Z_{jk}^2\right)^{1/2}\right. \\ & \left. + \quad \left(\sum_{jk} \mathbb{E}(X_{jk} - a_{jk})^4\right)^{1/4}\right]\end{aligned}$$

This bound is optimal.

Found similar bounds for other useful norms

- $$\mathbb{E}\|Z\|_{\infty\to 1} \leq 2\mathbb{E}\left(\|Z\|_{\text{col}} + \|Z^T\|_{\text{col}}\right)$$
- $$\mathbb{E}\|Z\|_{\infty\to 2} \leq 2\mathbb{E}\|Z\|_F + 2\min_D \mathbb{E}\|ZD^{-1}\|_{2\to\infty}$$
- These bounds are also asymptotically optimal.

Deviations from the Expected Error

Let g be a measurable function of n random variables and $F = g(X_1, \ldots, X_n)$. Let F_i denote the random variable obtained by replacing the ith argument of g with an independent copy: $F_i = g(X_1, \ldots, X_i', \ldots, X_n)$.

Theorem

Assume that there exists a positive constant C *such that, a.s.* $\sum_{i=1}^{n}(F - F_i)^2 \mathbb{1}_{F>F_i'} \leq \mathrm{C}$. *Then for all* $t > 0$,

$$\mathrm{P}(F > \mathbb{E}F + t) \leq \exp\left(-\frac{t^2}{4\mathrm{C}}\right)$$

Let $F = \|Z\|_\xi$ and F_i be obtained by replacing $Z_{ij} = a_{ij} - X_{ij}$ with $Z'_{ij} = a_{ij} - X'_{ij}$. It follows that if the entries of X are bounded by $D/2$, then

- $$\mathrm{P}(\|A - X\|_2 > (1+\delta)\mathbb{E}\|A - X\|_2) \leq e^{\left(-\frac{\delta^2 \mathbb{E}\|A-X\|_2^2}{4D^2}\right)}$$
- $$\mathrm{P}(\|A - X\|_{\infty\to 1} > (1+\delta)\mathbb{E}\|A - X\|_{\infty\to 1}) \leq e^{\left(-\frac{\delta^2 \mathbb{E}\|A-X\|_{\infty\to 1}^2}{4D^2 nm}\right)}$$
- A similar bound holds for the $(\infty, 2)$ error.

The combination of our bounds for the expected error and the error tail bounds work as well as Achlioptas and McSherry's scheme. Additionally

- Analyzing other schemes is straightforward.
- Our tail bounds are sharper, and apply over a wider range of matrix sizes.
- comparison to another sparsification scheme (Arora, Hazan, and Kale) is equally favorable.

Find rank-k matrix Q minimizing $\|A - QQ^\star A\|$.
Idea:

- if A were exactly rank-k, take $Y = A\Omega$ where random matrix Ω has k columns. Then $Y = QR$ gives Q with high probability.
- since A not exactly rank-k, do an oversampling: let Ω have ℓ columns
- need a bound on $\|A - QQ^\star A\|$ so we know how to choose ℓ to minimize this quantity

A nice bound by Tropp, et.al. Let

$$A = U \begin{pmatrix} \Sigma_1 & \\ & \Sigma_2 \end{pmatrix} \begin{pmatrix} V_1^\star \\ V_2^\star \end{pmatrix}$$

where Σ_1 contains the top k singular values and Σ_2 contains the bottom $n - k$.
Further, let $\Omega_1 = V_1^\star \Omega$ and $\Omega_2 = V_2^\star \Omega$ so

$$Y = A\Omega = U \begin{pmatrix} \Sigma_1 \Omega_1 \\ \Sigma_2 \Omega_2 \end{pmatrix}.$$

Theorem

Assuming that Ω_1 has full row rank, the approximation error satisfies

$$\|(I - P_Y)A\|_\xi^2 \leq \|\Sigma_2\|_\xi + \|\Sigma_2 \Omega_2 \Omega_1^\dagger\|_\xi^2,$$

where $\xi \in \{\mathrm{F}, 2\}$.

Known results

If Ω is a $n \times (k+p)$ Gaussian,

$$\mathbb{E}\|(I - P_Y)A\|_F \leq \left(1 + \frac{k}{p-1}\right)^2 \left(\sum_{j>k} \sigma_j^2\right)^{1/2}.$$

If $\Omega = \sqrt{n/l}\, RHD$ is a SRFT — H and D are as before, and R is a random $\ell \times n$ matrix that restricts to l uniformly randomly chosen coordinates:

$$\|(I - P_Y)A\| \leq \sqrt{1 + 18n/l}\,\sigma_{k+1}$$

except with probability $\mathrm{O}(k^{-1/26})$.

Another useful form for Ω

Experimentally, taking Ω to be a submatrix of a product of random Givens rotations seems to give good results:

$$G_1 G_2 \dots G_s P =: \begin{pmatrix} \Omega_{\text{samp}} & \Omega_{\text{err}} \end{pmatrix}$$

where P is a random permutation matrix and G_i is a Givens rotation about a uniformly randomly chosen angle in a uniformly randomly chosen plane.

I was able to show

$$\|\Omega^{\dagger}\|^2 = \frac{1}{1 - \| V_1^{\star}\Omega_{\text{err}}\Omega_{\text{err}}^{\star} V_1 \|^2}$$

so the problem reduces to showing that $\| V_1^{\star}\Omega_{\text{err}}\Omega_{\text{err}}^{\star} V_1 \|^2$ is bounded away from 1 with reasonable probability.
HARD: connection to convergence of Kac Walk on sphere.