Using random projections to estimate condition numbers and solve linear systems
CMPUT 501 Project

Bernardo Ávila Pires

December 14, 2012

1 Introduction

In this project I explore the use random projections for a topic of interest for the course: solving linear systems. This work is of theoretical nature, and it is not concerned with methods for *computing* solutions to linear systems using random projections, but rather with the quality of these solutions, and the asymptotic cost of computing them.

(Linear) random projections are functions $f : \mathbb{R}^d \to \mathbb{R}^k$ of the form $f(x) = Hx$, where $H \in \mathbb{R}^{k \times d}$ is a random projection matrix. Definitions 1.1, 1.2 and 1.3 contain three examples of random projection matrices.

Definition 1.1 (Rademacher random projection matrix, [Achlioptas, 2003])**.** *Let $H_{\mathcal{R}} \in \mathbb{R}^{k \times d}$ be such that (s.t.) each $(H_{\mathcal{R}})_{i,j}$ is an i.i.d. observation of a Rademacher random variable $\frac{1}{\sqrt{k}}R$, which is s.t. $\mathbb{P}(R = 1) = \mathbb{P}(R = -1) = \frac{1}{2}$.*

Definition 1.2 ("Sparse" random projection matrix, [Achlioptas, 2003])**.** *Let $H_{\mathcal{S}} \in \mathbb{R}^{k \times d}$ be s.t. each $(H_{\mathcal{S}})_{i,j}$ is an i.i.d. observation of random variable $\frac{1}{\sqrt{k}}S$, which is s.t. $\mathbb{P}\left(S = -\sqrt{3}\right) = 4\mathbb{P}(S = 0) = \mathbb{P}\left(S = -\sqrt{3}\right)\frac{1}{6}$.*

Definition 1.3 (Gaussian projection matrix, [Dasgupta and Gupta, 2003])**.** *Let $H_{\mathcal{N}} \in \mathbb{R}^{k \times d}$ be such that its rows are (transposed) i.i.d. observations of $\frac{1}{\sqrt{d}}Z$, where $Z \sim \mathcal{N}(\mathbf{0}_d, I_d)$.*

The useful property of random projections is that distances between vectors are nearly preserved after projection, as it is made explicit in the Johnson-Lindenstrauss Theorem (JL-Theorem), which will be the cornerstone of all results in this text and is given as the following lemma.

Lemma 1.4 (Theorem 1.1 in [Achlioptas, 2003]). *Fix a set V of n points in $\mathbb{R}^d$. For any $0 < \delta, \varepsilon < 1$, let k be a positive integer such that*

$$k \geq 2\left(\frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3}\right)^{-1}\left(2\ln n + \ln\frac{1}{\delta}\right).$$

Then with probability (w.p.) at least $1 - \delta$, for all $u, v \in V$

$$(1-\varepsilon)\|u-v\|_2^2 \leq \|H(u-v)\|_2^2 \leq (1+\varepsilon)\|u-v\|_2^2,$$

where H is $H_{\mathcal{R}}$ or $H_{\mathcal{S}}$.

Dasgupta and Gupta [2003] show a similar result to Lemma 1.4 when the random projection matrix is generated using normal random variables, as in Definition 1.3.

Why is the result of Lemma 1.4 useful to us? Since we know from [Heath, 2002] that the quality of solutions $\hat{x}$ of linear systems of the form $Ax = b$ can suffer significantly from ill-conditioning of A, there are two ways we can use random projections to our benefit:

- to bound the condition number of $A \in \mathbb{R}^{d\times d}$ in terms of the condition number of HA;
- to solve the linear system $HAH^\top x = Hb$.

Of course, the scenario of interest is when d is too large for us to be able to perform computations that take much more than $O(d^2)$ time (e.g. a LU decomposition of A, which would take $O(d^3)$). The idea is that if we can pick k small enough without having ε too large, we may be able to meaningfully upper-bound the condition number of A in terms of the condition number of HA. Similarly, we may be able to find approximate solutions to $Ax = b$ by solving $HAH^\top x = Hb$. Doing so will be advantageous if the computational gain offsets the error introduced by using random projections instead of the actual matrices in computing the condition number or solving the linear system.

Theorem 1.5. *Consider a full-rank $A \in \mathbb{R}^{d\times d}$. Then for $0 < \delta < 1$ and any k and any $0 < \varepsilon < 1$ s.t.*

$$\varepsilon \geq \sqrt{\frac{24}{k} \ln \frac{d}{\delta}},$$

and s.t. H is $H_{\mathcal{R}}$ or $H_{\mathcal{S}}$, we have, for any $x \in \mathbb{R}^d$, w.p. at least $1 - \delta$,

$$(1 - \varepsilon)\|Ax\|_2^2 \leq \|HAx\|_2^2 \leq (1 + \varepsilon)\|Ax\|_2^2 \tag{1}$$

Proof of Theorem 1.5. It suffices to choose a set of d orthonormal vectors spanning $\mathbb{R}^d$ and apply Lemma 1.4, by noting that if k, ε satisfy the condition in the theorem, *i.e.*, if they are s.t.

$$\varepsilon^2 \left(\frac{1}{2} - \frac{1}{3}\right) \geq \frac{4}{k} \ln \frac{d}{\delta}$$

then they satisfy the condition in Lemma 1.4, *i.e.*,

$$\left(\frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3}\right) \geq \frac{2}{k}\left(2 \ln d + \frac{1}{\delta}\right)$$

□

The remainder of this text is as follows. In the next section, I will investigate how to use random projections to estimate, for a full-rank matrix $A \in \mathbb{R}^{d\times d}$, its condition number $\kappa(A)$. In Section 3, I investigate the errors in solutions of linear systems of the form $HAH^\top x = Hb$, and in Section 4 I discuss computational aspects of using random projections to estimate bounds on the condition number of A and to solve linear systems. Finally, in Sections 4.1 and 5, I point out some related work and some conclusions for the project.

2 Condition number bounds

This section contains bounds on the condition number of a full-rank matrix $A \in \mathbb{R}^{d\times d}$ in terms of the condition number of a matrix in $\mathbb{R}^{k\times d}$. These bounds are derived from the JL-theorem.

Proposition 2.1. *For any $H \in \mathbb{R}^{k\times d}$ and full-rank $A \in \mathbb{R}^{d\times d}$ s.t. Inequalities (3) hold for all $x \in \mathbb{R}^d$, we have $\kappa(A) \leq \sqrt{\frac{1+\varepsilon}{1-\varepsilon}}\kappa(HA)$.*

Proof of Proposition 2.1. If Inequalities (3) hold for any $x \in \mathbb{R}^d$, it hold for the first and last right-eigenvectors of A, and so there exist $u, v \in \mathbb{R}^d$ s.t.

$$\begin{aligned}(1+\varepsilon)\lambda_{\min}^2(A) &= (1+\varepsilon)\|Av\|_2^2 \\ &\geq \|HAv\|_2^2 \\ &\geq \lambda_{\min}^2(HA),\end{aligned}$$

and

$$\begin{aligned}(1-\varepsilon)\lambda_{\max}^2(A) &= (1-\varepsilon)\|Au\|_2^2 \\ &\leq \|HAu\|_2^2 \\ &\leq \lambda_{\max}^2(HA),\end{aligned}$$

so that the conditioning number κ of A is bounded by

$$\kappa(A) \leq \sqrt{\frac{1+\varepsilon}{1-\varepsilon}}\kappa(HA).$$

□

For example, if we take $\varepsilon = \frac{1}{2}$ and apply Theorem 1.5 with Proposition 2.1, we have $k = 96 \ln \frac{d}{\delta}$ and $\kappa(A) \leq \sqrt{3}\kappa(HA)$ with high probability (w.h.p.). This is a large constant, but as we will see in Section 4, the worst problem if this approach is the cost of computing HA, more than moderately large constant factors on the size of the small space.

3 Solving linear systems

Proposition 3.1. *Consider a full-rank $A \in \mathbb{R}^{d\times d}$, and $b \in \mathbb{R}^d$. For any $H \in \mathbb{R}^{k\times d}$, k and $0 < \delta, \varepsilon < 1$ satisfying Inequalities (3), we have*

$$\frac{\|H^\top \hat{x} - x^*\|_2^2}{\|x^*\|_2^2} \leq \frac{1+\varepsilon}{1-\varepsilon}\kappa^2(A)\frac{\|HAH^\top \hat{x} - Hb\|_2^2}{\|Hb\|_2^2}.$$

Proof of Proposition 3.1. We have

$$\frac{\|H^\top \hat{x} - x^*\|_2^2}{\|x^*\|_2^2} \leq \frac{\lambda_{\max}^2(A)}{\lambda_{\min}^2(A)} \frac{\|A(H^\top \hat{x} - x^*)\|_2^2}{\|Ax^*\|_2^2}.$$

Inequalities ensure that

$$\|A(H^\top \hat{x} - x^*)\|_2^2 \leq (1+\varepsilon)\|HAH^\top \hat{x} - Hb\|_2^2,$$
$$\|Hb\|_2^2 \geq (1-\varepsilon)\|Hb\|_2^2,$$

and the result follows. □

Consider a computed solution $\hat{x}'$ of the larger system $Ax = b$. Then $\frac{\|\hat{x}' - x^*\|_2^2}{\|x^*\|_2^2} \leq \kappa^2(A)\frac{\|A\hat{x}' - b\|_2^2}{\|b\|_2^2}$, and we see from Proposition 3.1 that the squared backward error of solving the smaller system is larger than that of solving the original system by a factor of only $(1+\varepsilon)\frac{\|HAH^\top \hat{x} - Hb\|_2^2}{\|A\hat{x}' - b\|_2^2}$.[1]

We might expect $\frac{\|HAH^\top \hat{x} - Hb\|_2^2}{\|A\hat{x}' - b\|_2^2}$ to be small when the solution to the system lies mostly in a k-dimensional subspace of A that is well-conditioned, and if $HAH^\top$ can be shown to also lie in this subspace w.h.p.. However, if the random projection discards meaningful "directions" of A, then the residual will grow and the quality of the solution, evidently, deteriorates. We might also expect the random projection to provide poor solutions when A is well-conditioned, since we will lose a lot by ignoring/discarding meaningful subspaces of A with the projection.

4 Computational issues

In this section, I will discuss the computational costs of the operations needed in order to use random projections for condition estimation and for solving linear systems. Unless otherwise noted, the computational complexities mentioned here can be easily derived from the results in [Heath, 2002].

The choice of H plays a really important role in the complexity of performing the operations we are interested in. Suppose that H has m nonzero elements. Then the cost of performing HA is $O(dm)$. This is $O(d^2k)$

[1] Although I have no proof that the residual $\|HAH^\top \hat{x} - Hb\|_2^2$ is often smaller than $\|A\hat{x}' - b\|_2^2$, intuitively this seems to be the case, so I conjecture that the factor introduced by the random projections is not big.

for $H_{\mathcal{N}}$, $H_{\mathcal{R}}$ and $H_{\mathcal{S}}$, but because $H_{\mathcal{S}}$ is $\frac{1}{3}$ sparse, there is some gain of a constant factor. The cost of this multiplication is prohibitive for using random projections to estimate an upper-bound $\kappa(A)$, since one can compute the largest and smallest eigenvalues of A using the power iteration method [Heath, 2002].[2] Since the number of iterations required for this method to approximately converge is logarithmic on the ratio between the largest and second-largest eigenvalues, I would expect the power iteration method on A to take $\tilde{O}(d^2)$ time to run.

Is this cost also prohibitive for solving linear systems? The cost of computing $HAH^\top$ is again $O(d^2k)$, but now the cost of factorizing $HAH^\top$ is only $O(k^3)$, whereas that of factorizing A is $O(d^3)$. After factorization, the cost of solving the small linear system is $O(k^2 + kd)$ (since Hb must be computed), which is much lower than the $O(d^2)$-cost of solving the larger system. Therefore, the use of random projections can be advantageous for solving linear systems, but the cost of projecting A still dominates the costs of the other operations. Can we reduce it?

From the way the JL-Theorem works and uses concentration inequalities, to find H that has sparsity of higher order than a factor of dk is not a trivial task. Ailon and Liberty [2008] have investigated computing random projections in $O(d^2 \ln k)$ time, which allows choosing $k < d$ even so that $O(\ln k)$ is in $O(\ln d)$. This means that we can choose k in $O\left(d^{\frac{2}{3}}\right)$, enjoy tighter theoretical guarantees about the quality of the solution to the system, and perform $O(d^2 \ln d)$ computation, which asymptotically as much as we might need to perform using $H_{\mathcal{N}}$, $H_{\mathcal{R}}$ or $H_{\mathcal{S}}$.

The other question that may arise is whether we can improve the computational cost of solving linear systems (or estimating the condition number of A) when A is sparse. Dasgupta et al. [2010] have developed random projections based on hashing that exploit sparsity in the projected vectors in order to reduce the computational cost of the projection. They show the following theorem.

Theorem 4.1 (Theorem 1 in [Dasgupta et al., 2010]). *Let* $k = \lceil \frac{12}{\varepsilon^2} \ln \frac{4n}{\delta} \rceil$ *and* $c = \lceil \frac{16}{\varepsilon} \ln \frac{4n}{\delta} \ln^2 \frac{4nk}{\delta} \rceil$. *Let* $f : \{1, \ldots, cd\} \to \{1, \ldots, k\}$ *be a hashing function chosen uniformly at random and let* $Q \in \{0, \pm 1\}^{k \times cd}$ *be s.t.* $Q_{ij} = \mathbb{I}_{\{i=f(j)\}} R_j$,

[2] To compute the smallest eigenvalue, simply shift A by minus its trace, and then the method converges to largest eigenvalue in absolute value, which is the smallest eigenvalue of A.

where each R_j is an independent, identically distributed (i.i.d.) Rademacher random variable (r.v.). Let $P \in \{0, \pm 1\}^{cd \times d}$ be s.t. $P_{ij} = \frac{1}{\sqrt{c}} \mathbb{I}_{\{(j-1)c+1 \leq i \leq jc\}}$, and let $H_{\mathcal{D}} = QP$. Fix a set V of n vectors in $\mathbb{R}^d$. Then for any $v \in V$, w.p. at least $1 - \delta$,

$$(1 - \varepsilon)\|v\|_2^2 \leq \|H_{\mathcal{D}} v\|_2^2 \leq (1 + \varepsilon)\|v\|_2^2,$$

and $H_{\mathcal{D}} v$ can be computed in $O\left(c\|v\|_0\right)$ time.

In Theorem 4.1, $\|x\|_0$ denotes the number of non-zero elements of x, and $\mathbb{I}$ is the indicator function. Computing $H_{\mathcal{D}} A$ when A that has m non-zero elements can be done in $O\left(m\sqrt{k}\ln d\right)$ time. This is sub-optimal for dense matrices, but quite good an alternative for matrices of density in $O\left(d^{\frac{5}{3}}\right)$ as we can choose $k \in O\left(d^{\frac{2}{3}}\right)$ as in the case of the projections used by [Ailon and Liberty, 2008]. This will allow error bounds of the same order as if we were using the results of Ailon and Liberty [2008], and as A gets sparser (*e.g.*, $m \in O\left(d\right)$) the guarantees get significantly tighter.

4.1 Related work

Random projections have been widely studied. Some of the works, *e.g.* [Achlioptas, 2003, Ailon and Liberty, 2008, Andoni and Indyk, 2006, Dasgupta et al., 2010], are concerned with comparing points in the projected space, so their aim is to construct fast random projections. In our case it is also important to have a low computational cost for the projection, as it can often dominate the cost of calculating a bound on the condition number or the cost of solving the linear system in the smaller space.

One can wonder how the conditioning of the covariance of a set of points in $\mathbb{R}^d$ will behave after projection into $\mathbb{R}^k$. Dasgupta [1999, 2000] present the following result.

Lemma 4.2 ([Dasgupta, 1999, 2000])**.** *Consider any Gaussian in $\mathbb{R}^d$ with covariance Σ. Suppose this Gaussian is projected into a randomly chosen subspace of dimension k. There is a universal constant C s.t. for any $0 < \delta, \varepsilon < 1$, if the original dimension satisfies $d > C\frac{\kappa(\Sigma)^2}{\varepsilon^2}\left(\ln\frac{1}{\delta} + k\ln\frac{k}{\varepsilon}\right)$, then w.p. at least $1 - \delta$ over the choice of the random projection, the condition number of the covariance matrix of the projected points will be at most $1 + \varepsilon$. In particular, if $\kappa(\Sigma)$ is at*

most $n^{\frac{1}{2}}C^{-\frac{1}{2}}\left(\ln\frac{1}{\delta}+k\ln k\right)^{-\frac{1}{2}}$*, then w.p. at least* $1-\delta$ *the projected Gaussian will have condition number at most two.*

The bound in Lemma 4.2 says that if the original data are not too far from spherical, then the projected data will be close to spherical. This is a nice evidence toward the idea that the $HAH^\top$ should have better conditioning than A w.h.p..

The solutions of linear systems using random projections (or, more generally, linear least-squares) has been studied, *e.g.*, by Maillard and Munos [2012]. If we take a step further and start to look for the solution of overdetermined systems, we see that much work has been done in finding sparse solutions to these systems. For example, Candes and Tao [2007] show how to nearly recover sparse solutions from certain overdetermined systems. Curiously, the matrices A for which it is possible recover k-sparse solutions (provided that these sparse solutions exist) are those that satisfy, for all $c \in \mathbb{R}^k$ and all k-sized subsets T of its columns,[3]

$$(1-\varepsilon)\|c\|_2^2 \leq \|A_T c\|_2^2 \leq (1+\varepsilon)\|c\|_2^2$$

for some $0 < \varepsilon < 1$, and the quality of the recovery depends on who small ε is. This condition can be seen as having groups of columns of A behaving as random projection matrices, and in fact because $H_{\mathcal{N}}$ is a projection matrix, it can be shown that "Gaussian data" allows recovering sparse solutions to linear systems (again, if they exist, and if they are not too sparse: recall that k has to be $\Omega(\ln d)$ for Lemma 1.4 to hold).

Finally, there are other approaches to bounding the condition number of large matrices. Tao and Vu [2009] explore one that is based on a perturbation analysis of the matrix norms. They study the condition number of a fixed matrix M that is perturbed with a random matrix N. The motivation is that in practice whereas we would start with an input matrix containing measurements of certain quantities, *e.g.*, 2.493, what we would in fact have would be a matrix containing slightly noisy numbers, *e.g.* 2.49293587 instead of 2.493. The technique they use is called smoothed analysis, and it was proposed so as to analyze the "conditioning" of problems whose pathological cases are so unlikely in the presence of noise that they are well-conditioned w.h.p. even though the conditioning for arbitrary input is bad.

[3] That is, for a set of indices T, A_T denotes the $d \times |T|$ matrix containing the columns of A indexed by T.

5 Future work and conclusion

Random projections are an interesting dimensionality reduction tool. As exposed in this text, they are not so interesting for estimating a bound on the condition number of large square matrices, but they can be useful for finding approximate solutions to linear systems.

There have been recent progresses in computing random projections that allow us to compute random projections fast, so that solving large linear systems via random projection is feasible even when solving the original system is not, at least from an asymptotic cost perspective. Because we wish to solve large linear systems, what we need to to is to look at the cost analysis in more detail, at least to verify if the constants are not too large that the bounds only hold when m, the number of non-zero elements in A, is also too large.

Another issue that needs to be investigated is the magnitude of $\|HAH^\top \hat{x} - Hb\|_2^2$ compared to $\|A\hat{x}' - b\|_2^2$, *i.e.*, verify the hypothesis that the latter is not much smaller than the former. A second hypothesis that should be empirically investigated is whether the random projection can project A onto a well-conditioned subspace of it (w.h.p.).

Bibliography

Dimitris Achlioptas. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66: 671–687, Jan 2003. URL http://linkinghub.elsevier.com/retrieve/pii/S0022000003000254.

Nir Ailon and Edo Liberty. Fast dimension reduction using rademacher series on dual bch codes. In *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, SODA '08, pages 1–9, Philadelphia, PA, USA, 2008. Society for Industrial and Applied Mathematics. URL http://dl.acm.org/citation.cfm?id=1347082.1347083.

A Andoni and P Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. Jan 2006. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4031381.

Emmanuel Candes and Terence Tao. The dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351, 2007.

Anirban Dasgupta, Ravi Kumar, and Tamás Sarlos. A sparse johnson: Lindenstrauss transform. In *Proceedings of the 42nd ACM symposium on Theory of computing*, STOC '10, pages 341–350, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0050-6. doi: 10.1145/1806689.1806737. URL http://doi.acm.org/10.1145/1806689.1806737.

Sanjoy Dasgupta. Learning mixtures of gaussians. *Foundations of Computer Science*, Jan 1999. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=814639.

Sanjoy Dasgupta. Experiments with random projection. *Uncertainty in Artificial Intelligence: Proceedings of the Sixteenth Conference (UAI-2000)*, pages 143–151, 2000.

Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Struct. Algorithms*, 22(1):60–65, January 2003. ISSN 1042-9832. doi: 10.1002/rsa.10073. URL http://dx.doi.org/10.1002/rsa.10073.

Michael T. Heath. *Scientific Computing, An Introductory Survey*. McGraw Hill, 2nd edition, 2002.

Odalric-Ambrym Maillard and Rémi Munos. Linear regression with random projections. *Journal of Machine Learning Research*, 13:2735–2772, Sep 2012.

Terence Tao and Van Vu. Smooth analysis of the condition number and the least singular value. In *Proceedings of the 12th International Workshop and 13th International Workshop on Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, APPROX '09 / RANDOM '09, pages 714–737, Berlin, Heidelberg, 2009. Springer-Verlag. ISBN 978-3-642-03684-2. doi: 10.1007/978-3-642-03685-9_53. URL http://dx.doi.org/10.1007/978-3-642-03685-9_53.