Expected Errors for Least-Squares with Gaussian Matrices

Nicholas Richardson & Summer Sheng

April 13, 2023

1 Introduction

The main question we wish to answer comes from solving the least-squares problem when our measurements y are some (known) random linear combination of an unknown vector x, possibly with noise. More precisely, this problem is to recover a vector $x \in \mathbb{R}^n$ given a random matrix $A \in \mathbb{R}^{m\times n}$, $A_{i,j} \sim \mathcal{N}(0,1)$, and a measurement vector $y = Ax + z \in \mathbb{R}^m$. Here, the vector $z \in \mathbb{R}^m$ is some unknown noise vector. One simple approach to estimate x is to solve the least-squares problem to obtain $\hat{x} = \arg\min_{x'} \|Ax' - y\|_2$ where we expect $\hat{x} \approx x$. When $m > n$ and the columns of A are independent, this problem has the closed form solution $\hat{x} = A^\dagger y = (A^\top A)^{-1} A^\top y$.

It is here we ask the natural question about the recovered $\hat{x}$: how close is it to the true x? There are a number of ways of assessing closeness, so we formalize this question as finding $\mathbb{E}\,\|\hat{x} - x\|_2^2$. We state the main, and to our knowledge, novel theorems of the report in Section 2, provide numerical experiments in Section 3, and conclude with final remarks and research directions in Section 4.

2 Theoretical Results

Theorem 1. *Let $A \in \mathbb{R}^{m\times n}$, $m > n$, $A_{i,j} \sim \mathcal{N}(0,1)$, $x \in \mathbb{R}^n$, $z \in \mathbb{R}^m$, and $y = Ax + z \in \mathbb{R}^m$. If $\hat{x} = A^\dagger y$, then,*

$$\mathbb{E}\,\|\hat{x} - x\|_2^2 = \frac{1}{m}\left(\mathbb{E}\left\|A^\dagger\right\|_F^2\right)\|z\|_2^2 \asymp \frac{n}{m(m-n)}\|z\|_2^2 . \tag{1}$$

Theorem 2. *Let $A \in \mathbb{R}^{m\times n}$, A have independent columns, $x \in \mathbb{R}^n$, $z \in \mathbb{R}^m$, $z_i \sim \mathcal{N}(0,1)$, and $y = Ax + z \in \mathbb{R}^m$. If $\hat{x} = A^\dagger y$, then,*

$$\mathbb{E}\,\|\hat{x} - x\|_2^2 = \left\|A^\dagger\right\|_F^2 . \tag{2}$$

Note. If we wanted uniform upper bounds, we would be left with:

$$\mathbb{E}\sup_{\|z\|\le\nu}\|\hat{x} - x\|_2^2 = \left(\mathbb{E}\left\|A^\dagger\right\|_2^2\right)\nu^2 \le \frac{1}{(\sqrt{m}-\sqrt{n})^2}\nu^2 = \frac{m}{n}\frac{\sqrt{\frac{m}{n}}+1}{\sqrt{\frac{m}{n}}-1}\frac{n}{m(m-n)}\nu^2 \tag{3}$$

in place of Equation 1, and

$$\mathbb{E}_{z'}\sup_{\|z\|\le\|z'\|}\|\hat{x} - x\|_2^2 = m\left\|A^\dagger\right\|_2^2 \le m\left\|A^\dagger\right\|_F^2 \tag{4}$$

in place of Equation 2. The uniform bound in Equation 3 could possibly be made tighter, but is given to illustrate the coefficient in front of the noise in Theorem 1 is asymptotically a factor of m/n smaller than what a uniform bound would give. Furthermore, the random variable $\left\|A^\dagger\right\|_F^2$ concentrates which Equation 1 is be more representative of the error in practice which is shown in Section 3.

In Equation 4, the notation is shorthand for "take the supremum over all noise vectors z that have norm less than or equal to a random Gaussian vector z'". This lets the noise pick an adversarial direction without increasing its norm parallel to singular vector with the smallest singular value. Theorem 2 gives an error that is lower up to a factor of m. See Appendix A for justification of these uniform bounds.

2.1 Proof of Theorems 1 & 2

Lemma 1. *When A has independent columns, then $\hat{x} - x = A^\dagger z$. Note if $A \in \mathbb{R}^{m\times n}$ has Gaussian entries, and $m > n$, then A has n, length-m independent column vectors with probability 1.*

Proof. $\hat{x} - x = A^\dagger y - x = A^\dagger(Ax + z) - x = (A^\dagger A - I_n)x + A^\dagger z = ((A^\top A)^{-1}A^\top A - I_n)x + A^\dagger z = A^\dagger z$ □

Lemma 2. *If either A is a Gaussian matrix or z is Gaussian vector, then $\left\|A^\dagger z\right\|_2 \overset{d}{=} \left\|\Sigma^\dagger s\right\|_2 \|z\|_2$ where $A = U\Sigma V^\top$ is the singular value decomposition (SVD) of A and $s \sim \mathcal{U}(S^{m-1})$ is uniform on the sphere.*

Proof. **Case 1: z is Gaussian.** Since a unitary matrix V preserves the 2-norm, we have $\left\|A^\dagger z\right\|_2 = \left\|V\Sigma^\dagger U^\top z\right\|_2 = \left\|\Sigma^\dagger U^\top z\right\|_2$. Because z is Gaussian and hense isotropic, $U^\top z \overset{d}{=} z \overset{d}{=} s\|z\|_2$.

Case 2: A is Gaussian. Let B be an $m \times m$ unitary matrix chosen uniformly at random. For a Gaussian vector g, $Bg \overset{d}{=} g$, so for our matrix A with Gaussian column vectors, we have $BA \overset{d}{=} A \implies A^\dagger B^{-1} \overset{d}{=} A^\dagger \implies A^\dagger z \overset{d}{=} A^\dagger B^{-1} z \implies \left\|A^\dagger z\right\|_2 \overset{d}{=} \left\|A^\dagger B^{-1} z\right\|_2 \overset{d}{=} \left\|(V\Sigma^\dagger U^\top)s\right\|_2 \|z\|_2 \overset{d}{=} \left\|\Sigma^\dagger s\right\|_2 \|z\|_2$ by noting $B^{-1}z \overset{d}{=} Bz \overset{d}{=} s\|z\|_2$ because B is a uniform unitary matrix, and $U^\top s \overset{d}{=} s$ because s is isotropic. □

Lemma 3. *Let $D \in \mathbb{R}^{n\times m}$ be diagonal and independent of $s \sim \mathcal{U}(S^{m-1})$, then $\mathbb{E}\|Ds\|_2^2 = \frac{1}{m}\mathbb{E}\|D\|_F^2$.*

Proof. By linearity of expectation, and symmetry of entries of a uniform vector on a sphere,

$$\mathbb{E}\|Ds\|_2^2 = \mathbb{E}\sum_{i=1}^{\min(m,n)} D_{ii}^2 s_i^2 = \sum_{i=1}^{\min(m,n)} \mathbb{E}D_{ii}^2 \mathbb{E}s_i^2 = (\mathbb{E}s_1^2)\mathbb{E}\sum_{i=1}^{\min(m,n)} D_{ii}^2 = \mathbb{E}s_1^2 \mathbb{E}\|D\|_F^2.$$

To find the second moment of a spherical coordinate, we calculate

$$\begin{aligned}
\mathbb{E}s_1^2 &= \frac{1}{S_{m-1}}\int_{S^{m-1}} s_1^2 dS \\
&= \frac{\int_0^\pi (\cos(\phi_1))^2 \sin^{m-2}(\phi_1)d\phi_1 \int_0^\pi \sin^{m-3}(\phi_2)d\phi_2 \cdots \int_0^\pi \sin(\phi_{n-2})d\phi_{n-2}\int_0^{2\pi} d\phi_{n-1}}{\int_0^\pi \sin^{m-2}(\phi_1)d\phi_1 \int_0^\pi \sin^{m-3}(\phi_2)d\phi_2 \cdots \int_0^\pi \sin(\phi_{n-2})d\phi_{n-2}\int_0^{2\pi} d\phi_{n-1}} \\
&\quad \text{by spherical parametrization with } s_1 = \cos(\phi_1) \text{ and } dS = \sin^{m-2}(\phi_1)\ldots\sin(\phi_{m-2})d\phi_1\ldots d\phi_{m-1} \\
&= \frac{\int_0^\pi (1-\sin^2(\phi_1))\sin^{m-2}(\phi_1)d\phi_1}{\int_0^\pi \sin^{m-2}(\phi_1)d\phi_1} \quad \text{all integrals over each coordinate cancel except for } \phi_1 \\
&= \frac{\int_0^\pi \sin^{m-2}(\phi_1)d\phi_1 - \frac{m-1}{m}\int_0^\pi \sin^{m-2}(\phi_1)d\phi_1}{\int_0^\pi \sin^{m-2}(\phi_1)d\phi_1} \text{ linearity and IBP with } u = \sin(\phi_1) \text{ and } dv = \sin^{m-1}(\phi_1) \\
&= 1 - \frac{m-1}{m} = \frac{1}{m}.
\end{aligned}$$

□

Lemma 4. *Let $A \in \mathbb{R}^{m\times n}$, $m > n$, have standard Gaussian entries, then $\mathbb{E}\left\|A^\dagger\right\|_F^2 \asymp \frac{n}{m-n}$.*

Proof. This lemma relies on a result from Wei [1] that says the singular values of a standard Gaussian matrix, when $m > n$, are $ci/\sqrt{m} \le \sigma_{m+1-i} \le di/\sqrt{m}$ for all $m+1-n \le i \le m$ with high probability $1 - \exp(-Ci)$ for some positive constants c, d, C. So we calculate the following:

$$\begin{aligned}
\left\|A^\dagger\right\|_F^2 = \left\|\Sigma^\dagger\right\|_F^2 = \sum_{j=1}^n 1/\sigma_j^2 = \sum_{i=m+1-n}^m 1/\sigma_i^2 &\le \frac{m}{c^2}\sum_{i=m+1-n}^m 1/i^2 \le \frac{m}{c^2}\int_{m-n}^m 1/x^2 dx = \frac{n}{c^2(m-n)} \\
\left\|A^\dagger\right\|_F^2 \ge \frac{m}{d^2}\sum_{i=m+1-n}^m 1/i^2 &\ge \frac{m}{d^2}\int_{m+1-n}^{m+1} 1/x^2 dx = \frac{n}{d^2(m-n+1)}\frac{m}{m+1}, \text{ w.h.p.}
\end{aligned}$$

Combining the inequalities, we obtain the asymptotic result in expectation desired. □

3 Numerical Experiments

This means that in practice, we do observe $\mathbb{E}\,\|\hat{x} - x\|_2^2 = \frac{n}{m(m-n)}$ rather than equality up to constants. Before we show how expectation of recovery error aligns with $\frac{n}{m(m-n)}$, we firstly show model's recovery capability using MNIST. MNIST is a hand-written digits dataset, with $784 = 28 \times 28$ features. Our sensing matrix is a $\mathbb{R}^{m \times 784}$, where m ranges from 784 to 1000. After we sensed the signal, we add standard gaussian noise to each entry. Then we recover the signal using the pseudo-inverse mentioned above, namely $\hat{x} = A^{\dagger} y$. With more measurements, our recovery gets closer to the real signal and with 800 measurements, it basically recovers the true signal. Our results shown in Figure.1 suggests that our theory align with real world experiments.

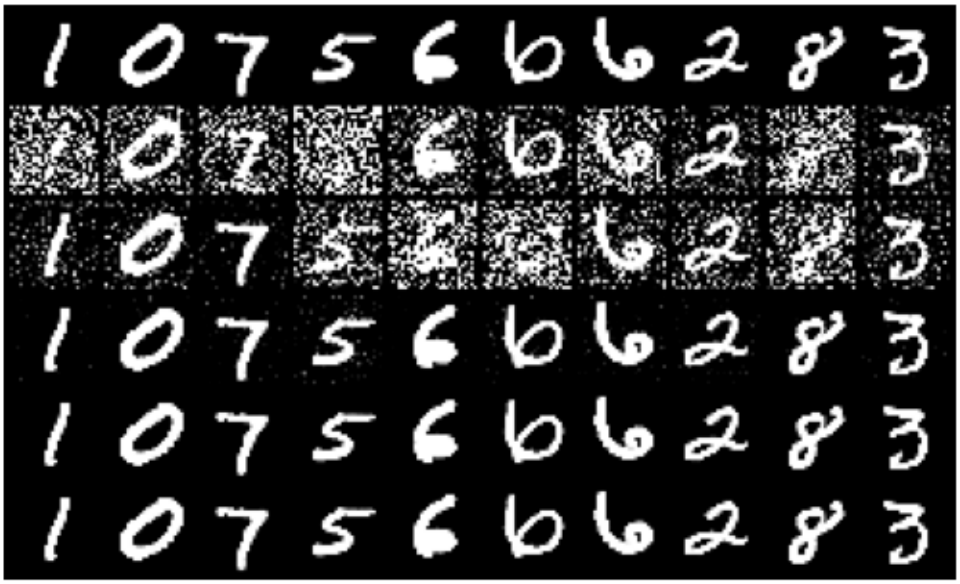

Figure 1: The first row is the original signal, followed by signal recovered using $784, 785, 800, 900, 1000$ measurements.

What's more, we also show empirical evidence of Lemma 4. We consider a simple case, where our signal is a mixture of cosine and sine waves, with 50 features. We recover the signal using the algorithm mentioned above. We compare the L_2 difference between two signals and run a few thousand times to get the empirical estimate of $E\|\hat{x} - x\|^2$. The blue line in the Figure 2 is $\frac{n}{m(m-n)}$, in this case $\frac{50}{m(m-50)}$. We see that the theoretical bound fit observation closely, especially with more measurements.

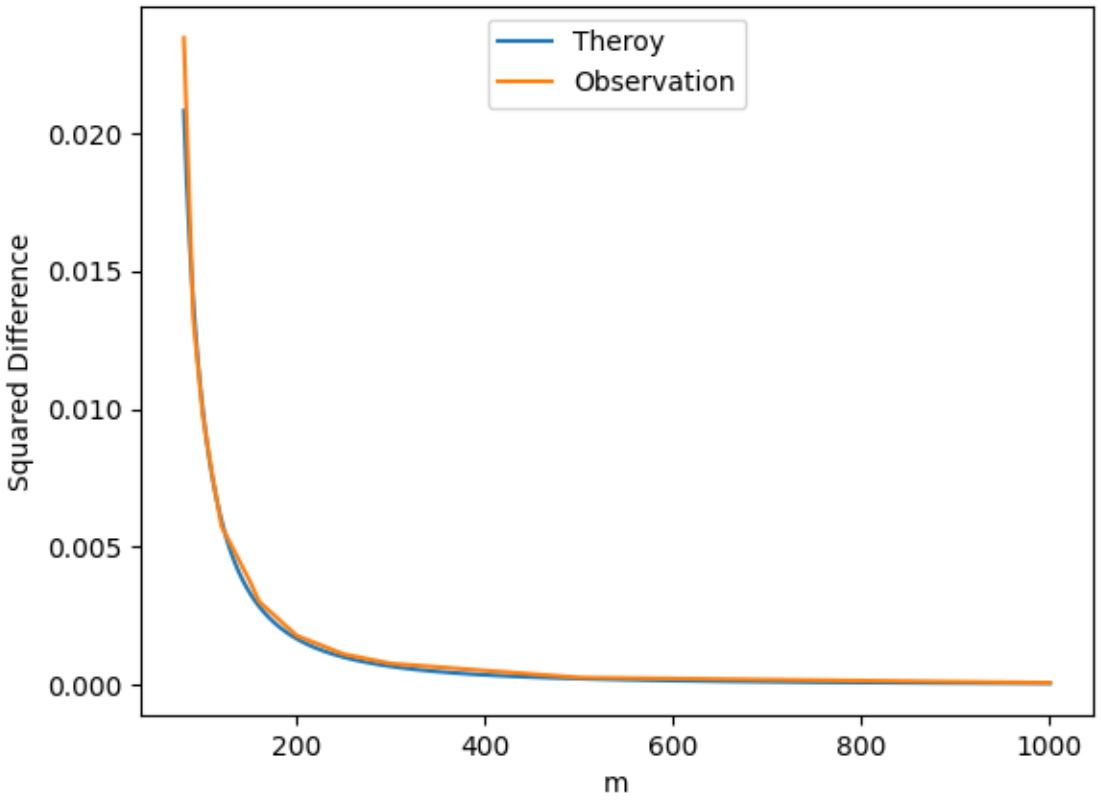

Figure 2: Formula matches with the simulation result

4 Conclusion & Further Study

What is great about the technique presented in Section 2 is that it can be generalized to many problems. The real heart of the proof lies in the combination of Lemmas 2 & 3 which says that the average case of the norm of an operator applied to a vector, if either the operator or vector acts randomly, is better measured by the Frobenius norm of that operator rather than the 2-norm which would instead always point to the worst case. This is well suited when there is concentration around your expected case so you can consider an average of all possible directions, rather than rely on a uniform bound given the worst direction always happens. In this case of concentration, the expectation on its own is more descriptive than a uniform bound. It would be great to extend this idea to problems where you are restricted to only looking at a subset of directions, like in the case of decent cones, or least squares on a constrained set, where this technique can be applied further.

5 References

[1] Feng Wei. "Upper bound for intermediate singular values of random matrices". In: *Journal of Mathematical Analysis and Applications* 445.2 (2017). A special issue of JMAA dedicated to Richard Aron, pp. 1530–1547. ISSN: 0022-247X. DOI: https://doi.org/10.1016/j.jmaa.2016.08.007.

[2] Roman Vershynin. *High-Dimensional Probability*. Cambridge Series in Statistical and Probabilistic Mathematics. University of California, Irvine: Cambridge University Press, Oct. 2018. ISBN: 978-1-108-24625-5.

[3] Mark Rudelson and Roman Vershynin. "Non-asymptotic Theory of Random Matrices: Extreme Singular Values". In: *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010)*, pp. 1576–1602. DOI: 10.1142/9789814324359_0111.

A Proof sketch of expected upper bounds

The upper bounds follow from the sub-multiplicative property of the 2-norm $\left\|A^\dagger z\right\|_2 \leq \left\|A^\dagger\right\|_2 \|z\|_2$, and noting equality can be achieved when z picks a direction in the singular vector associated with the largest singular value of $A^\dagger$. So $\sup_{\|z\|\leq\nu} \left\|A^\dagger z\right\|_2^2 = \left\|A^\dagger\right\|_2^2 \nu^2$.

Theorem 2.6 of [3] states that the expected value of the smallest singular value σ_n of a Gaussian matrix A is greater than $\sqrt{m}-\sqrt{n}$ which means the largest singular value of its psudo-inverse $1/\sigma_n$ is less than $\frac{1}{\sqrt{m}-\sqrt{n}}$. Furthermore, this expectation concentrates giving us

$$\mathbb{E} \sup_{\|z\|\leq\nu} \|\hat{x}-x\|_2^2 = \mathbb{E}\left\|A^\dagger\right\|_2^2 \nu^2 = \nu^2/(\sqrt{m}-\sqrt{n})^2.$$

The final equality in Equation 3 follows from standard algebra.

$$\begin{aligned}
\frac{1}{(\sqrt{m}-\sqrt{n})^2} &= \frac{1}{\sqrt{m}-\sqrt{n}} \frac{1}{\sqrt{m}-\sqrt{n}} \frac{\sqrt{m}+\sqrt{n}}{\sqrt{m}+\sqrt{n}} \\
&= \frac{\sqrt{m}+\sqrt{n}}{\sqrt{m}-\sqrt{n}} \frac{1}{m-n} \\
&= \frac{\sqrt{\frac{m}{n}}+1}{\sqrt{\frac{m}{n}}-1} \frac{1}{m-n} \\
&= \frac{m}{n} \frac{\sqrt{\frac{m}{n}}+1}{\sqrt{\frac{m}{n}}-1} \frac{n}{m(m-n)}
\end{aligned}$$

This is a useful form not only to see the factor that makes this different than Equation 1, but also that this factor depends only on the ratio m/n i.e. how square the matrix is.[1] Finding the minimum around $m/n \approx 2.6$, you can show this factor $2.6\frac{\sqrt{2.6}+1}{\sqrt{2.6}-1}$ is at least bigger than 10.

For the uniform bound in Equation 4, Section 3.1 of [2] gives us $\mathbb{E}\|z\|^2 = m$.

What makes this a sketch is the need to rigorously justify $\mathbb{E}\sigma_n \geq \sqrt{m}-\sqrt{n} \implies \mathbb{E}\frac{1}{\sigma_n^2} \leq \frac{1}{(\sqrt{m}-\sqrt{n})^2}$ given the specific distribution of σ_n. These could be explored further to get more precise uniform bounds, but was not the focus of this project. These bounds are only meant to illustrate the expectation of $\|\hat{x}-x\|_2^2$ on its own is on the order of a factor of m smaller than what a uniform bound would give.

[1] Fun fact, the function $x\frac{\sqrt{x}+1}{\sqrt{x}-1}$, when $x>1$, has a global minimum at $x=\phi+1$, the golden ratio plus 1.

B Alternate intuitive proof sketch of Lemma 3

To show $\mathbb{E}\left\|Ds\right\|_2^2 = \mathbb{E}\left\|D\right\|_F^2/m$ alternatively, using $s \sim g/\left\|g\right\|$, one has

$$\left\|Ds\right\|_2^2 \sim \left\|Dg\right\|_2^2 / \left\|g\right\|_2^2 = \frac{1}{\left\|g\right\|_2^2} \sum_{i=1}^{\min(m,n)} D_{ii}^2 g_i^2.$$

So taking expectation, noting independence and the linearity of expectation, we have

$$\mathbb{E}\left\|Ds\right\|_2^2 = \mathbb{E}\left(\frac{1}{\left\|g\right\|_2^2}\right) \mathbb{E} \sum_{i=1}^{\min(m,n)} D_{ii}^2 g_i^2 = \frac{1}{m} \sum_{i=1}^{\min(m,n)} \mathbb{E} D_{ii}^2 \cdot 1 = \frac{1}{m} \mathbb{E}\left\|D\right\|_F^2.$$

Similar to Appendix A, this is only a sketch because you would need to rigorously show $\mathbb{E}\left\|g\right\|_2^2 = m \implies \mathbb{E}\frac{1}{\left\|g\right\|_2^2} = \frac{1}{m}$.

C Combining Lemmas 2 & 3

Using these two lemmas together, we actually have the following more general results.

Proposition 1. *If $A \in \mathbb{R}^{m\times n}$ is a Gaussian matrix or z is Gaussian vector, then $\mathbb{E}\left\|Az\right\|_2 = \frac{1}{n}\mathbb{E}\left\|A\right\|_F^2 \left\|z\right\|_2^2$.*

This offers an alternate way to prove the following Corollaries.

Corollary 1. *If $A \in \mathbb{R}^{m\times n}$ is a standard Gaussian matrix then $\mathbb{E}\left\|Az\right\|_2^2 = m\left\|z\right\|_2^2$.*

Corollary 2. *If $z \in \mathbb{R}^n$ is a standard Gaussian vector then $\mathbb{E}\left\|Az\right\|_2^2 = \left\|A\right\|_F^2$.*

Corollary 3. *If A and z are an independent standard Gaussian matrix and vector then $\mathbb{E}\left\|Az\right\|_2^2 = nm$.*