On the Smallest Singular Value of Non-Centered Gaussian Designs

Stephen Tu
Google Brain Robotics
August 20, 2020

Abstract

Consider the stochastic process $x_i = \mu_i + w_i \in \mathbb{R}^d$, $i = 1, 2, ...$, where each $w_i$ is drawn independently across time from an isotropic Gaussian distribution, and $\mu_i$ is $(w_1, ..., w_{i-1})$-adapted. Let $X_N \in \mathbb{R}^{N \times d}$ be the design matrix after time $N$, where the $i$-th row of $X_N$ contains $x_i$. What is the behavior of the minimum singular value of $X_N$, denoted $\sigma_{\min}(X_N)$? In the most basic case where $\mu_i \equiv 0$, it is well-known that $\sigma_{\min}(X_N)$ scales as $\sqrt{N} - \sqrt{d}$ (we will only concern ourselves with the regime where $N > d$). In this note, we generalize this result to the setting where each $\mu_i$ is non-zero but also non-random. We show that a uniform lower bound on $\sigma_{\min}(X_N)$ scaling as $\sqrt{N} - \sqrt{d}$ also holds, irrespective of the magnitude of the $\mu_i$’s. Unfortunately, in the general setting where $\mu_i$ is allowed to adapt to the past history, we show that no such uniform lower bound on $\sigma_{\min}(X_N)$ is possible: for any fixed $N$, the minimum singular value of $X_N$ can be made arbitrarily small with constant probability.

1 Introduction

In this paper, we consider the following $\mathbb{R}^d$-valued stochastic process on $i = 1, 2, ...$ defined as:

$$x_i = \mu_i + w_i, \quad w_i \sim \mathcal{N}(0, I_d),$$

(1.1)

where each $\mu_i$ is $(w_1, ..., w_{i-1})$-measurable. Let $X_N \in \mathbb{R}^{N \times d}$ be the design matrix where the $i$-th row of $X_N$ contains $x_i$. We are interested in understanding how the bias terms $\mu_i$ affect the minimum singular value of $X_N$, denoted $\sigma_{\min}(X_N)$. Recall that:

$$\sigma_{\min}(X_N) = \sqrt{\inf_{\|v\|=1} \sum_{i=1}^{N} \langle x_i, v \rangle^2}.$$ 

Here, $\|\cdot\|$ and $\langle \cdot, \cdot \rangle$ denote the Euclidean norm and inner product on $\mathbb{R}^d$, respectively. We will restrict ourselves in this paper to the setting where $N > d$.

The most basic setting of (1.1) is when $\mu_i \equiv 0$, in which case $\sigma_{\min}(X_N)$ is characterized quite well by modern non-asymptotic random matrix theory. In particular, we have that (see e.g. [19, Section 7.3]),

$$\mathbb{E}\sigma_{\min}(X_N) \geq \sqrt{N} - \sqrt{d},$$

and furthermore for any $t > 0$, with probability at least $1 - e^{-\Theta/2}$,

$$\sigma_{\min}(X_N) \geq \sqrt{N} - \sqrt{d} - t.$$
On the other hand, the case when \( \mu_i = Ax_{i-1} \) for a fixed \( d \times d \) matrix \( A \) has received attention recently due to interest in non-asymptotic bounds for linear system identification [2, 3, 4, 5, 10, 13, 14, 15, 16, 18]. Most analyses of \( \sigma_{\min}(X_N) \) degrade as the \( \mu_i \)'s grow unbounded (equivalently when the spectral radius of \( A \) exceeds one). It is natural to wonder whether or not this degradation is fundamental, or a limitation of current proof techniques.

This note attempts to shed some light on this phenomenon. In this case where \( \mu_i = \beta_i \) and the \( \beta_i \)'s are fixed non-random biases, we show that a uniform lower bound on \( \sigma_{\min}(X_N) \) of \( \sqrt{N} - \sqrt{d} \) is indeed possible, irrespective of the size of the \( \beta_i \)'s. This gives an alternate proof, in the special case of Gaussian covariates, of a more general result from Oliveira [9] on lower tails of quadratic forms.

The situation changes, however, when the \( \mu_i \)'s are allowed to depend on the past history. We show that when \( d \geq 2 \), it is possible to drive \( \sigma_{\min}(X_N) \) arbitrarily close to zero with constant probability. This phenomenon is closely related to the inconsistency of ordinary least squares for unstable multivariate linear system identification and vector autoregression [11, 13]. For \( d = 1 \) uniform lower bounds are possible, and indeed this fact has already been used by Rantzer [12] in context of regret bounds for online learning of linear control systems.

2 Non-Centered Independent Design

The main result for this section is the following theorem.

**Theorem 2.1.** Let \( \{\beta_i\}_{i \geq 1} \) be a fixed sequence of vectors in \( \mathbb{R}^d \). Consider the process (1.1) with \( \mu_i = \beta_i \). Suppose that \( N - d \geq d \). We have that:

\[
\mathbb{E}\sigma_{\min}(X_N) \geq \sqrt{N - d} - \sqrt{d} - 1 .
\]  

Furthermore for any \( t > 0 \), with probability at least \( 1 - e^{-t^2/2} \),

\[
\sigma_{\min}(X_N) \geq \sqrt{N - d} - \sqrt{d} - 1 - t .
\]  

Before we prove Theorem 2.1, we note that it is not possible to obtain such a result using Mendelson’s small-ball method [6, 8], which provides a powerful and general framework for obtaining lower bounds on non-negative empirical processes. While it is true that the small-ball probability of \( \langle v, x_i \rangle \) can be lower bounded independently of \( \beta_i \) for any fixed unit vector \( v \), the Rademacher complexity \( \mathbb{E}\left\| \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i x_i \right\| \) clearly depends on the magnitude of the \( \beta_i \)'s.

2.1 Proof of Theorem 2.1

The main tool will be a Gaussian min-max theorem which is attributed to Gordon. This allows us to generalize the standard proof when \( \mu_i \equiv 0 \). We state the version presented in Thrampoulidis et al. [17].

**Theorem 2.2** (Gaussian min-max theorem). Let \( A, \xi, g, h \) all have \( \mathcal{N}(0, 1) \) entries independent of each other. Let \( S_1, S_2 \) be two compact sets, and let \( \psi \) be a continuous function on \( S_1 \times S_2 \). Define:

\[
F(A, \xi) = \inf_{x \in S_1} \sup_{y \in S_2} y^T Ax + \xi \|x\|\|y\| + \psi(x, y) ,
\]

\[
G(g, h) = \inf_{x \in S_1} \sup_{y \in S_2} \|x\|g^T y + \|y\|h^T x + \psi(x, y) .
\]
Then for any \( t \in \mathbb{R} \) we have:
\[
P(F(A, \xi) \leq t) \leq P(G(g, h) \leq t)
\]

Let \( M \in \mathbb{R}^{N \times d} \) be the matrix where the \( i \)-th row contains \( \beta_i \). Let \( A \in \mathbb{R}^{N \times d} \) be a matrix where each entry is i.i.d. \( \mathcal{N}(0, 1) \). Then we have that \( X = A + M \). We write:
\[
\sigma_{\min}(X) = \inf_{\|x\|=1} \|Xx\| = \inf_{\|x\|=1} \sup_{\|y\|=1} y^T X x = \inf_{\|x\|=1} \sup_{\|y\|=1} y^T A x + y^T M x
\]
\[
= -\xi + \inf_{\|x\|=1} \sup_{\|y\|=1} y^T A x + \xi + y^T M x
\]
\[
= -\xi + F_s(A, \xi).
\]

Now define \( G_s(g, h) \) as:
\[
G_s(g, h) := \inf_{\|x\|=1} \sup_{\|y\|=1} g^T y + h^T x + y^T M x.
\]

By the Gaussian min-max theorem (Theorem 2.2), we have that \( P(F_s(A, \xi) > t) \geq P(G_s(g, h) > t) \) for all \( t \in \mathbb{R} \). We lower bound \( G_s(g, h) \) as follows. Write the SVD of \( M \) as \( M = U \Sigma V^T \), where \( U \in \mathbb{R}^{N \times d} \). Let \( U_\perp \in \mathbb{R}^{N \times N-d} \) denote the orthogonal complement of \( U \). We can then lower bound \( G_s \) by restricting the inner supremum over \( \{ y \in \mathbb{R}^N : \|y\| = 1 \} \) to \( \{ y \in \text{Span}(U_\perp) : \|y\| = 1 \} \). This latter set is equivalent to \( \{ U_\perp \alpha : \alpha \in \mathbb{R}^{N-d}, \|\alpha\| = 1 \} \). Hence
\[
G_s(g, h) \geq \inf_{\|x\|=1} \sup_{\|\alpha\|=1} g^T U_\perp \alpha + h^T x = \|U_\perp^T g\| - \|h\|.
\]

Next, note that \( U_\perp^T g \) has the same distribution as \( \tilde{g} \sim \mathcal{N}(0, I_{N-d}) \). Therefore we have \( P(F_s(A, \xi) > t) \geq P(\|\tilde{g}\| - \|h\| > t) \) for all \( t \in \mathbb{R} \), which implies:
\[
1 + \mathbb{E}\sigma_{\min}(X) = \mathbb{E}F_s(A, \xi) \geq \mathbb{E}\|\tilde{g}\| - \mathbb{E}\|h\| \geq \sqrt{N - d} - \sqrt{d}.
\]

The last inequality follows since the function \( f(n) = \mathbb{E}_{g \sim \mathcal{N}(0, I_n)} \|g\| - \sqrt{n} \) is increasing in \( n \) and we assumed \( N - d \geq d \). This proves (2.1). The tail inequality (2.2) follows since \( A \mapsto \sigma_{\min}(A + M) \) is a 1-Lipschitz function [19, Section 5.2.1].

3 The Non-Centered Adaptive Case

We now show that when \( d \geq 2 \), a universal lower bound of the type shown in Theorem 2.1 is not possible in the adapted case.

Theorem 3.1. Consider the process (1.1) where \( \mu_i = \rho x_{i-1} \) and where \( d = 2 \). Fix an \( N \geq N_0 \) for a universal \( N_0 \), and suppose that \( \rho \geq \rho(N) \), where \( \rho(N) \gg 1 \). With constant probability (say 9/10), we have:
\[
\sigma_{\min}(X_N) \leq O(\rho^{-1} \sqrt{N}).
\]
Note that Theorem 3.1 is similar to Lemma 2 of Phillips and Magdalinos [11] which states that for a fixed \( \rho > 1 \), the quantity \( \frac{1}{N} \sigma_{\min}(X_N)^2 \) converges to \( \frac{1}{\rho^2 - 1} \) in probability as \( N \to \infty \). Theorem 3.1 also provides a sharper characterization of \( \sigma_{\min}(X_N) \) compared to Proposition 19.1 of Sarkar and Rakhlin [13].

It is interesting to note that in the scalar case when \( d = 1 \), a universal lower bound is possible for arbitrary adapted \( \mu_t \)'s. In fact, it is an elementary calculation to show that \( \mathbb{E} \sigma_{\min}(X_N)^2 \) = \( \mathbb{E} \sum_{i=1}^{N} x_i^2 \) \( \geq N \). A uniform large deviation bound is given in the following theorem.

**Theorem 3.2.** Consider the process (1.1) with \( d = 1 \). Fix any \( t > 0 \). We have that:

\[
\mathbb{P} \left\{ \sum_{i=1}^{N} x_i^2 \leq N - \sqrt{N} t \right\} \leq \exp(-t/4).
\]

### 3.1 Proof of Theorem 3.1

Let \( \{u_t\}, \{v_t\} \) be mutually i.i.d. \( N(0, 1) \) random variables. Let \( \{a_t\}, \{b_t\} \) be real-value processes defined as \( a_{t+1} = \rho u_t + a_t, b_{t+1} = \rho v_t + b_t \), with the base case \( a_1 = u_0 \) and \( b_1 = v_0 \). It is clear that the process \( \{\begin{bmatrix} a_t \\ b_t \end{bmatrix} \} \) has the same distribution as the process \( \{x_t\} \). Define the random variables \( T, D \) as:

\[
T := \sum_{k=1}^{N} a_k^2 + \sum_{k=1}^{N} b_k^2,
\]

\[
D := \left( \sum_{k=1}^{N} a_k^2 \right) \left( \sum_{k=1}^{N} b_k^2 \right) - \left( \sum_{k=1}^{N} a_k b_k \right)^2.
\]

We first calculate \( \mathbb{E}[T] \) and \( \mathbb{E}[D] \). Focusing on \( D \), by the fact that the \( a_t \)'s are independent of the \( b_t \)'s and have the same distribution,

\[
\mathbb{E}[D] = \left( \sum_{k=1}^{N} \mathbb{E}[a_k^2] \right)^2 - \sum_{i,j=1}^{N} \mathbb{E}[a_i a_j] = \sum_{i,j=1}^{N} \left( \mathbb{E}[a_i^2] \mathbb{E}[a_j^2] - \mathbb{E}[a_i a_j]^2 \right)
= 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left( \mathbb{E}[a_i^2] \mathbb{E}[a_j^2] - \mathbb{E}[a_i a_j]^2 \right) = 2 \sum_{i=1}^{N-1} \sum_{k=1}^{N-i} \left( \mathbb{E}[a_i^2] \mathbb{E}[a_{i+k}^2] - \mathbb{E}[a_i a_{i+k}]^2 \right).
\]

Now we have \( a_{i+k} = \rho^k a_i + \sum_{\ell=0}^{k-1} \rho^{k-1-\ell} u_{i+\ell} \) for \( k \geq 0 \). Therefore, we have \( \mathbb{E}[a_i^2] = \sum_{\ell=0}^{i-1} \rho^{2\ell} \). Furthermore, \( \mathbb{E}[a_i a_{i+k}] = \rho^k \mathbb{E}[a_i^2] = \rho^k \mathbb{E}[a_i^2] \) by \( \rho^k \mathbb{E}[a_i^2] \). Therefore:

\[
\mathbb{E}[D] = 2 \sum_{i=1}^{N-1} \sum_{k=1}^{N-i} \left( \rho^{2\ell} \right) \left( \rho^{2\ell} \right) - \left( \rho^k \mathbb{E}[a_i^2] \right)^2
= N^2 \rho^4 - 2N^2 \rho^2 + N^2 - 2N \rho^2 + 2N \rho^2 + 4N^2 + 4N \rho^4 - 2N \rho^2 - N + 2N \rho^2 + 4N^2
= (\rho^2 - 1)^4 \Theta(N \rho^{2(N-2)}) \quad \text{when } \rho \gg 1.
\]

On the other hand, we have

\[
\mathbb{E}[T] = 2 \sum_{i=1}^{N} \mathbb{E}[a_i^2] = 2 \sum_{i=1}^{N} \sum_{\ell=0}^{i-1} \rho^{2\ell} = -N \rho^2 + \rho^2 \left( \rho^{2N} - 1 \right) + N \rho^4 = (\rho^2 - 1)^2 \Theta(\rho^{2(N-1)}) \quad \text{when } \rho \gg 1.
\]
Because $X_N$ is a 2-by-2 matrix, we have that:
\[
\lambda_{\min}(X_N) = \frac{1}{2}(T - \sqrt{T^2 - 4D}) \leq \frac{D}{\sqrt{T^2 - 4D}}.
\]
Above, the last inequality follows from the concavity of $x \mapsto \sqrt{x}$.

Now because $D \geq 0$ by Cauchy-Schwarz, by Markov’s inequality we have $\mathbb{P}(D \geq \mathbb{E}[D]/\delta) \leq \delta$ for any $\delta \in (0, 1)$. Hence $D \leq O(N\rho^{2(N-2)})$ with probability at least 0.95. The more difficult part is to control $T^2 - 4D$ from below. To do this, we use a powerful Gaussian anti-concentration result.

**Theorem 3.3** (Special case of Theorem 8, Carbery and Wright [1]). Let $p : \mathbb{R}^n \to \mathbb{R}$ be a degree $d$ polynomial, and $\mu$ be a log-concave measure. We have that:
\[
\mu\{|p| \leq \varepsilon \mathbb{E}_{\mu}|p|\} \leq C d \varepsilon^{1/d},
\]
where $C$ is a universal constant, and $\mathbb{E}_{\mu}|p| = \int |p| \, d\mu$.

By construction, we have $T^2 - 4D$ is a non-negative degree four polynomial of $(w_0, \ldots, w_{N-1}, v_0, \ldots, v_{N-1})$. Hence by Theorem 3.3 with probability at least 0.95, we have
\[
T^2 - 4D \geq c\mathbb{E}[T^2 - 4D] \geq c(\mathbb{E}[T]^2 - 4\mathbb{E}[D]) = \Omega(\rho^{4(N-1)}) \quad \text{when } \rho \gg 1.
\]
for a universal $c$, where the last inequality is Jensen’s inequality. The claim now follows by union bounding over the upper bound for $D$ and the lower bound for $T^2 - 4D$.

### 3.2 Proof of Theorem 3.2

First, an elementary calculation shows that if $\mu$ is fixed, $w \sim \mathcal{N}(0, 1)$, and $\theta < 0$,
\[
\mathbb{E} \exp(\theta(\mu + w)^2) = \frac{1}{\sqrt{1-2\theta}} \exp\left\{\frac{\theta}{1-2\theta} \mu^2\right\} \leq \frac{1}{\sqrt{1-2\theta}}.
\]
Therefore by iterating expectations, for $\theta < 0$ we have:
\[
\mathbb{E} \exp\left\{\theta \sum_{i=1}^N x_i^2\right\} \leq \frac{1}{(1-2\theta)^{N/2}}.
\]
The rest of the proof follows from standard $\chi_k^2$ concentration bounds [7, Lemma 1]. Define the random variable $Z = \sum_{i=1}^N x_i^2 - N$. By a Chernoff bound for any $v > 0$ and $\theta < 0$,
\[
\mathbb{P}(Z \leq -v) = \mathbb{P}(\theta Z \geq -\theta v) \leq \exp(\theta v)\mathbb{E}\exp(\theta Z).
\]
Now define $\psi(\theta) := -\theta - \frac{1}{2} \log(1-2\theta)$. Observe that:
\[
\log \mathbb{E} \exp(\theta Z) = -N\theta + \log \mathbb{E} \exp\left\{\theta \sum_{i=1}^N x_i^2\right\} \leq -N\theta - \frac{N}{2} \log(1-2\theta) = N\psi(\theta).
\]
It is elementary to show that $\psi(\theta) \leq \theta^2$ for $\theta < 0$. Therefore combining with the Chernoff bound:
\[
\mathbb{P}(Z \leq -v) \leq \inf_{\theta < 0} \exp(\theta v + N\psi(\theta)) \leq \inf_{\theta < 0} \exp(\theta v + N\theta^2).
\]
We set $\theta = -v/(4N)$ and therefore $\mathbb{P}(Z \leq -v) \leq \exp(-v^2/(4N))$. Now set $v = \sqrt{Nt}$ for any $t > 0$ which yields the result.
Acknowledgements

We thank Horia Mania for pointing us towards the direction of showing negative results in Section 3. We also thank Mahdi Soltanolkotabi for helpful discussions around the Gaussian min-max theorem.

References


