Learning without mixing: analysis of linear system identification

Max Simchowitz, Horia Mania, Stephen Tu, Michael I. Jordan, Benjamin Recht (2018)

Geelon So, agso@eng.ucsd.edu
DSC291 Sequential decision making — June 10, 2021

Linear regression problem

Problem: suppose we have a data process generating $(X, Y) \in \mathbb{R}^d \times \mathbb{R}^n$,

$$Y = \mathbf{A}X + \eta,$$

where $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $\eta \in \mathbb{R}^n$ is mean-zero noise.

Linear regression problem

Problem: suppose we have a data process generating $(X, Y) \in \mathbb{R}^d \times \mathbb{R}^n$,

$$Y = \mathbf{A}X + \eta,$$

where $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $\eta \in \mathbb{R}^n$ is mean-zero noise.

- Can we use samples $(X_1, Y_1), \ldots, (X_T, Y_T)$ to *identify* $\mathbf{A}$? That is, find $\hat{\mathbf{A}}$,

$$\|\mathbf{A} - \hat{\mathbf{A}}\|_{\text{op}} \approx 0.$$

Linear regression problem

Problem: suppose we have a data process generating $(X, Y) \in \mathbb{R}^d \times \mathbb{R}^n$,

$$Y = \mathbf{A}X + \eta,$$

where $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $\eta \in \mathbb{R}^n$ is mean-zero noise.

- Can we use samples $(X_1, Y_1), \ldots, (X_T, Y_T)$ to *identify* $\mathbf{A}$? That is, find $\hat{\mathbf{A}}$,

$$\|\mathbf{A} - \hat{\mathbf{A}}\|_{\text{op}} \approx 0.$$

  - *Yes*, with i.i.d. samples with $\underset{X \sim p}{\mathbb{E}}\left[XX^\top\right] \succ 0$, (Hsu et al., 2012).

Linear regression problem

Problem: suppose we have a data process generating $(X, Y) \in \mathbb{R}^d \times \mathbb{R}^n$,

$$Y = \mathbf{A}X + \eta,$$

where $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $\eta \in \mathbb{R}^n$ is mean-zero noise.

- Can we use samples $(X_1, Y_1), \ldots, (X_T, Y_T)$ to *identify* $\mathbf{A}$? That is, find $\hat{\mathbf{A}}$,

$$\|\mathbf{A} - \hat{\mathbf{A}}\|_{\text{op}} \approx 0.$$

  - *Yes*, with i.i.d. samples with $\mathbb{E}_{X \sim p}\left[XX^\top\right] \succ 0$, (Hsu et al., 2012).
  - Analysis is difficult if $(X_1, Y_1), \ldots, (X_T, Y_T)$ are not i.i.d.

System identification for linear dynamical systems

An important class of examples where observations $(X_1, Y_1), \ldots, (X_T, Y_T)$ are not i.i.d. are **linear dynamical systems**, which is a system with dynamics:

$$X_{t+1} = \mathbf{A}X_t + \mathbf{B}u_t + \eta_t,$$

- $X_t \in \mathbb{R}^d$ is the state of the system
- u_t is the input to the system
- $\eta_t \in \mathbb{R}^d$ is unobserved noise.

System identification for linear dynamical systems

An important class of examples where observations $(X_1, Y_1), \ldots, (X_T, Y_T)$ are not i.i.d. are **linear dynamical systems**, which is a system with dynamics:[1]

$$X_{t+1} = \mathbf{A}X_t + \eta_t,$$

- $X_t \in \mathbb{R}^d$ is the state of the system
- $\eta_t \in \mathbb{R}^d$ is unobserved noise.

[1] For simplicity, we'll consider systems with no input u_t.

Previous analyses for linear dynamical system identification

Previous analyses could only handle case when the **spectral radius** $\rho(\mathbf{A}) < 1$.

Previous analyses for linear dynamical system identification

Previous analyses could only handle case when the **spectral radius** $\rho(\mathbf{A}) < 1$.

- If $\rho(\mathbf{A}) < 1$, then:
  - deterministic dynamics $\mathbf{A}X_t$ contract state X_{t+1} back toward origin

Previous analyses for linear dynamical system identification

Previous analyses could only handle case when the **spectral radius** $\rho(\mathbf{A}) < 1$.

- If $\rho(\mathbf{A}) < 1$, then:
  - deterministic dynamics $\mathbf{A}X_t$ contract state X_{t+1} back toward origin
  - noise process η_t expands the state X_{t+1} outward

Previous analyses for linear dynamical system identification

Previous analyses could only handle case when the **spectral radius** $\rho(\mathbf{A}) < 1$.

- If $\rho(\mathbf{A}) < 1$, then:
  - deterministic dynamics $\mathbf{A}X_t$ contract state X_{t+1} back toward origin
  - noise process η_t expands the state X_{t+1} outward
- There is a *stationary distribution* where these opposing forces are at equilibrium.

Previous analyses for linear dynamical system identification

Previous analyses could only handle case when the **spectral radius** $\rho(\mathbf{A}) < 1$.

- If $\rho(\mathbf{A}) < 1$, then:
  - deterministic dynamics $\mathbf{A}X_t$ contract state X_{t+1} back toward origin
  - noise process η_t expands the state X_{t+1} outward
- There is a *stationary distribution* where these opposing forces are at equilibrium.
  - As the gap $1 - \rho(\mathbf{A})$ grows, the process $(X_t)_{t=0}^{\infty}$ *mixes* more rapidly.

Previous analyses for linear dynamical system identification

Previous analyses could only handle case when the **spectral radius** $\rho(\mathbf{A}) < 1$.

- If $\rho(\mathbf{A}) < 1$, then:
  - deterministic dynamics $\mathbf{A}X_t$ contract state X_{t+1} back toward origin
  - noise process η_t expands the state X_{t+1} outward
- There is a *stationary distribution* where these opposing forces are at equilibrium.
  - As the gap $1 - \rho(\mathbf{A})$ grows, the process $(X_t)_{t=0}^{\infty}$ *mixes* more rapidly.
  - If $(X_1, Y_1), \ldots, (X_T, Y_T)$ come from stationary distribution with rapid mixing time, they can essentially be treated as independent (Mohri and Rostamizadeh, 2008).

Previous analyses for linear dynamical system identification

Previous analyses could only handle case when the **spectral radius** $\rho(\mathbf{A}) < 1$.

- If $\rho(\mathbf{A}) < 1$, then:
  - deterministic dynamics $\mathbf{A}X_t$ contract state X_{t+1} back toward origin
  - noise process η_t expands the state X_{t+1} outward
- There is a *stationary distribution* where these opposing forces are at equilibrium.
  - As the gap $1 - \rho(\mathbf{A})$ grows, the process $(X_t)_{t=0}^{\infty}$ *mixes* more rapidly.
  - If $(X_1, Y_1), \ldots, (X_T, Y_T)$ come from stationary distribution with rapid mixing time, they can essentially be treated as independent (Mohri and Rostamizadeh, 2008).

Counterintuitive: if X_t is larger compared to the noise process η_t, it should be easier to recover $\mathbf{A}$ since there is a higher *signal-to-noise* ratio. This happens as $\rho(\mathbf{A})$ increases!

Previous analyses for linear dynamical system identification

Previous analyses could only handle case when the **spectral radius** $\rho(\mathbf{A}) < 1$.

- If $\rho(\mathbf{A}) < 1$, then:
  - deterministic dynamics $\mathbf{A}X_t$ contract state X_{t+1} back toward origin
  - noise process η_t expands the state X_{t+1} outward
- There is a *stationary distribution* where these opposing forces are at equilibrium.
  - As the gap $1 - \rho(\mathbf{A})$ grows, the process $(X_t)_{t=0}^{\infty}$ *mixes* more rapidly.
  - If $(X_1, Y_1), \ldots, (X_T, Y_T)$ come from stationary distribution with rapid mixing time, they can essentially be treated as independent (Mohri and Rostamizadeh, 2008).

Counterintuitive: if X_t is larger compared to the noise process η_t, it should be easier to recover $\mathbf{A}$ since there is a higher *signal-to-noise* ratio. This happens as $\rho(\mathbf{A})$ increases!

- Motivates Simchowitz et al. (2018), identification without appeals to mixing.

Interlude: geometry of linear regression

Geometric view of a matrix

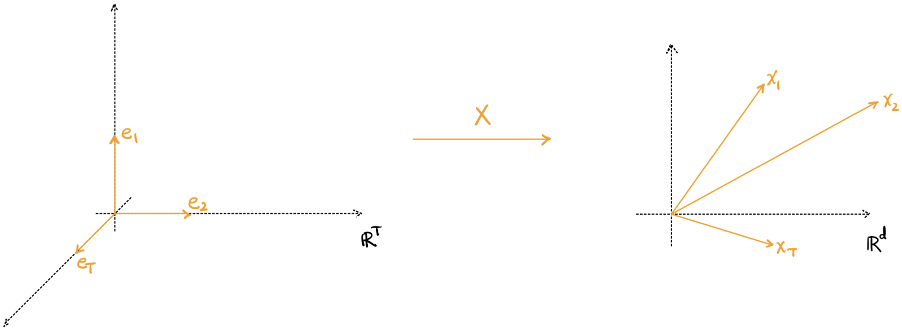

Figure 1: $\mathbf{X} \in \mathbb{R}^{d \times T}$ maps the standard basis of $\mathbb{R}^T$ to the columns of $\mathbf{X}$.

Singular value decomposition

Theorem (SVD)

Let $\mathbf{X} \in \mathbb{R}^{d\times T}$*. There exists orthogonal matrices* $\mathbf{U} \in \mathbb{R}^{d\times d}$ *and* $\mathbf{V} \in \mathbb{R}^{T\times T}$ *and diagonal matrix* $\mathbf{\Sigma} \in \mathbb{R}^{d\times T}$ *such that:*

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top.$$

Singular value decomposition

Theorem (SVD)

Let $\mathbf{X} \in \mathbb{R}^{d \times T}$*. There exists orthogonal matrices* $\mathbf{U} \in \mathbb{R}^{d \times d}$ *and* $\mathbf{V} \in \mathbb{R}^{T \times T}$ *and diagonal matrix* $\mathbf{\Sigma} \in \mathbb{R}^{d \times T}$ *such that:*

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top.$$

- Recall $\mathbf{U}$ is orthogonal if $\mathbf{U}\mathbf{U}^\top = \mathbf{U}^\top\mathbf{U} = \mathbf{I}_{d \times d}$.
  - Geometrically, $\mathbf{U}$ maps the standard orthonormal basis to another orthonormal basis.

Singular value decomposition

Theorem (SVD)

Let $\mathbf{X} \in \mathbb{R}^{d \times T}$. There exists orthogonal matrices $\mathbf{U} \in \mathbb{R}^{d \times d}$ and $\mathbf{V} \in \mathbb{R}^{T \times T}$ and diagonal matrix $\mathbf{\Sigma} \in \mathbb{R}^{d \times T}$ such that:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top.$$

- Recall $\mathbf{U}$ is orthogonal if $\mathbf{U}\mathbf{U}^\top = \mathbf{U}^\top\mathbf{U} = \mathbf{I}_{d \times d}$.
  - Geometrically, $\mathbf{U}$ maps the standard orthonormal basis to another orthonormal basis.
- $\mathbf{\Sigma}$ is diagonal if $\mathbf{\Sigma}_{ij} = 0$ when $i \neq j$. Denote $\mathbf{\Sigma}_{ii} = \sigma_i$.
  - WLOG, choose SVD so that $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$.

Geometric view of singular value decomposition

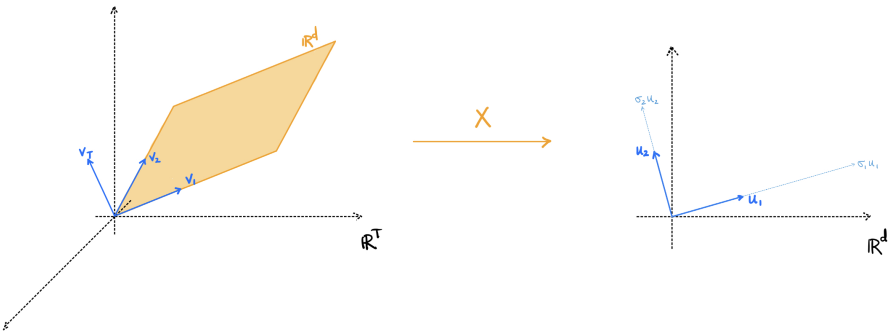

Figure 2: There exists orthonormal bases of $\mathbb{R}^T$ and $\mathbb{R}^d$ such that $\mathbf{X}v_i = \sigma_i u_i$.

Moore-Penrose pseudoinverse

The (right) **pseudoinverse** $\mathbf{X}^+$ of $\mathbf{X}$ maps $\mathbb{R}^d$ back to the span$(v_1, \ldots, v_d)$ so that:

$$\mathbf{X}\mathbf{X}^+ = \mathbf{I}_{d\times d}.$$

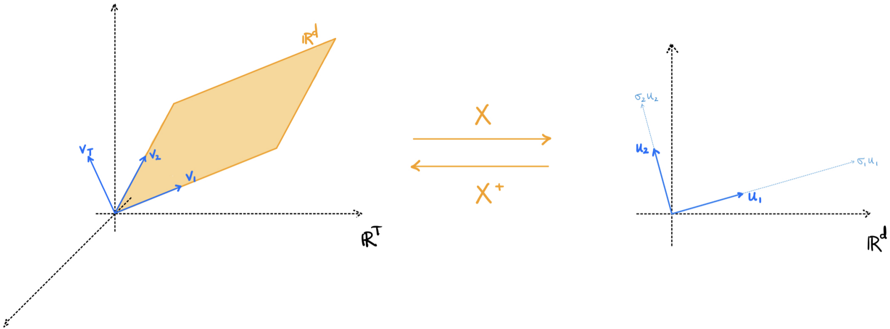

Figure 3: $\mathbf{X}^+$ maps $\sigma_i u_i$ to v_i.

Moore-Penrose pseudoinverse form

Given the singular value decomposition of $\mathbf{X}$, we can compute $\mathbf{X}^+$,

$$\mathbf{X}^+ = \mathbf{V}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\top,$$

where $\boldsymbol{\Sigma}^{-1} \in \mathbb{R}^{T\times d}$ is the diagonal matrix where $\boldsymbol{\Sigma}_{ii}^{-1} = \begin{cases} \sigma_i^{-1} & \sigma_i \neq 0 \\ 0 & \sigma_i = 0. \end{cases}$

[2]Note that this shows that $\mathbf{X}^+$ satisfies $\mathbf{X}\mathbf{X}^+ = \mathbf{I}_{d\times d}$ when $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ is full rank (i.e. $\mathbf{X}$ is surjective).

Moore-Penrose pseudoinverse form

Given the singular value decomposition of $\mathbf{X}$, we can compute $\mathbf{X}^+$,

$$\mathbf{X}^+ = \mathbf{V}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\top,$$

where $\boldsymbol{\Sigma}^{-1} \in \mathbb{R}^{T \times d}$ is the diagonal matrix where $\boldsymbol{\Sigma}_{ii}^{-1} = \begin{cases} \sigma_i^{-1} & \sigma_i \neq 0 \\ 0 & \sigma_i = 0. \end{cases}$

- We do not have to explicitly compute the SVD:[2]

$$\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1} = \left(\mathbf{V}\boldsymbol{\Sigma}^\top\mathbf{U}^\top\right)\left(\mathbf{U}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\mathbf{U}^\top\right)^{-1} = \mathbf{V}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\top = \mathbf{X}^+.$$

[2] Note that this shows that $\mathbf{X}^+$ satisfies $\mathbf{X}\mathbf{X}^+ = \mathbf{I}_{d\times d}$ when $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ is full rank (i.e. $\mathbf{X}$ is surjective).

Linear regression (ordinary least squares)

Let $\mathbf{X} \in \mathbb{R}^{d \times T}$ be a matrix of covariates. Let $\mathbf{Y} \in \mathbb{R}^{n \times T}$ be a matrix of responses. The goal is to find a matrix $\hat{\mathbf{A}} \in \mathbb{R}^{n \times d}$ minimizing:[3]

$$\min_{\mathbf{A} \in \mathbb{R}^{n \times d}} \sum_{i \in [T]} \|Y_i - \mathbf{A}X_i\|^2 = \min_{\mathbf{A} \in \mathbb{R}^{n \times d}} \|\mathbf{Y} - \mathbf{A}\mathbf{X}\|_F^2.$$

[3] This description of linear regression is actually in a transposed form of the standard linear regression literature. There, the design matrix is $\mathbf{X} \in \mathbb{R}^{T \times d}$ and response matrix $\mathbf{Y} \in \mathbb{R}^{T \times n}$. The goal there is often:

$$\min_{\beta \in \mathbb{R}^{d \times n}} \|\mathbf{Y} - \mathbf{X}\beta\|_F^2.$$

Geometric view of linear regressions

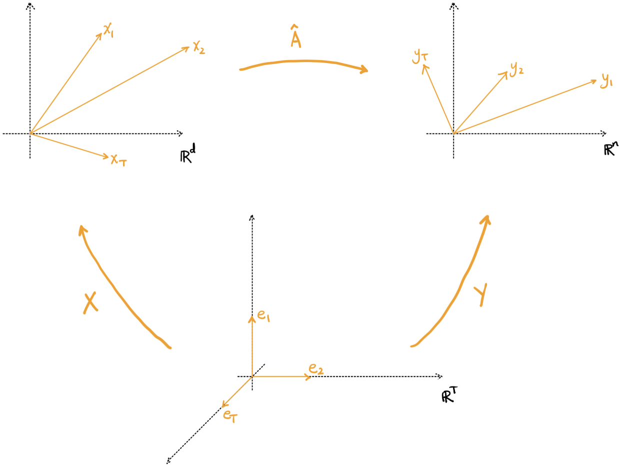

Figure 4: Let $\mathbf{X} \in \mathbb{R}^{d \times T}$ and $\mathbf{Y} \in \mathbb{R}^{n \times T}$ be collections of T vectors in $\mathbb{R}^d$ and $\mathbb{R}^n$.

Formal solution to OLS

We can solve for $\hat{\mathbf{A}}$ by taking the derivative of the objective and setting it to zero:

$$
\begin{aligned}
0 &= \frac{d}{d\mathbf{A}} \operatorname{tr}\left(\mathbf{Y}^\top\mathbf{Y} - \mathbf{Y}^\top\mathbf{A}\mathbf{X} - \mathbf{X}^\top\mathbf{A}^\top\mathbf{Y} + \mathbf{X}^\top\mathbf{A}^\top\mathbf{A}\mathbf{X}\right) \\
&= -2\mathbf{X}\mathbf{Y}^\top + 2\mathbf{X}\mathbf{X}^\top\mathbf{A}^\top.
\end{aligned}
$$

This implies:

$$
\hat{\mathbf{A}} = \mathbf{Y}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1} = \mathbf{Y}\mathbf{X}^+.
$$

Geometric view of linear regressions

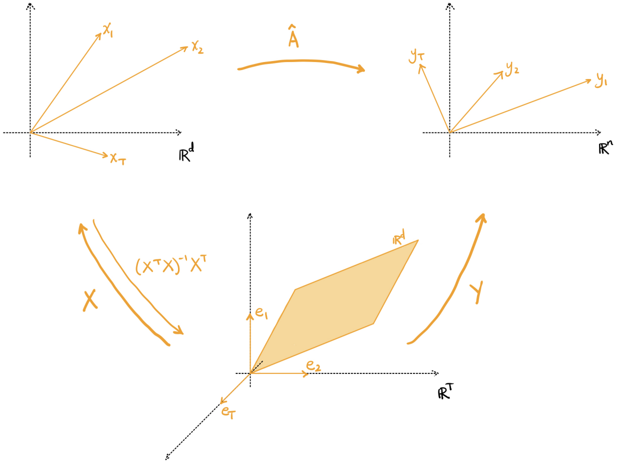

Figure 5: The solution to OLS is $\hat{\mathbf{A}} = \mathbf{Y}\mathbf{X}^+$.

System identification

Assumption 1: model is correct

Suppose that there is some $\mathbf{A}_\star$ such that data is generated:

$$Y = \mathbf{A}_\star X + \eta,$$

where η is noise.

Assumption 1: model is correct

Suppose that there is some $\mathbf{A}_\star$ such that data is generated:

$$Y = \mathbf{A}_\star X + \eta,$$

where η is noise.

- If $\eta \equiv 0$, then with $T \geq d$ full-rank samples $X_1, \ldots, X_T$, can recover $\mathbf{A}_\star$,

$$\|\mathbf{A}_\star - \hat{\mathbf{A}}\|_{\text{op}} = 0.$$

Assumption 1: model is correct

Suppose that there is some $\mathbf{A}_\star$ such that data is generated:

$$Y = \mathbf{A}_\star X + \eta,$$

where η is noise.

- If $\eta \equiv 0$, then with $T \geq d$ full-rank samples $X_1, \ldots, X_T$, can recover $\mathbf{A}_\star$,

$$\|\mathbf{A}_\star - \hat{\mathbf{A}}\|_{\text{op}} = 0.$$

- If $\eta \not\equiv 0$, let $\mathbf{E} \in \mathbb{R}^{n \times T}$ be the error matrix. Then the OLS estimator $\hat{\mathbf{A}}$ satisfies:

$$\|\mathbf{A}_\star - \hat{\mathbf{A}}\|_{\text{op}} = \|(\mathbf{A}_\star \mathbf{X})\mathbf{X}^+ - (\mathbf{A}_\star \mathbf{X} + \mathbf{E})\mathbf{X}^+\|_{\text{op}} = \|\mathbf{E}\mathbf{X}^+\|_{\text{op}}.$$

    - Estimate can be bad if noise $\|\mathbf{E}\|_{\text{op}}$ is large.
    - Estimate can be bad if singular values of $\mathbf{X}^+$ are large ($\mathbf{X}$ has small singular values).

Geometry when estimate is bad: large errors

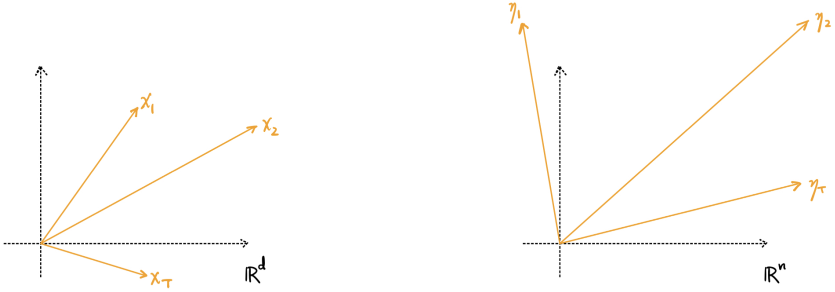

Figure 6: Fitting large noise vectors will throw the estimate off.

Assumption 2: sub-Gaussian noise

Conditioned on the past $\mathcal{F}_t = \sigma(X_1, \eta_1, \ldots, X_t, \eta_t)$ the noise at time $t+1$,

$$\eta_{t+1} \,|\, \mathcal{F}_t \text{ is } \nu^2\text{-sub-Gaussian.}$$

- The noise is highly concentrated about zero:

$$\Pr\left(w^\top \eta_{t+1} \geq \varepsilon \,\middle|\, \mathcal{F}_t\right) \leq \exp\left(-\frac{\varepsilon^2}{2\nu^2}\right),$$

for all unit vectors $w \in S^{n-1}$.

Geometry when estimate is bad: ill-conditioning

Figure 7: High sensitivity to small errors when $\mathbf{X}$ is ill-conditioned, i.e. $\dfrac{\sigma_1(\mathbf{X})}{\sigma_d(\mathbf{X})}$ is large.

Assumption 3: large signal and well-conditioned covariance

We assume that the sample covariance satisfies:

$$\sigma^2 \mathbf{I}_{d\times d} \prec \frac{1}{T}\sum_{i\in[T]} X_i X_i^\top \prec \kappa \cdot \sigma^2 \mathbf{I}_{d\times d}.$$

Assumption 3: large signal and well-conditioned covariance

We assume that the sample covariance satisfies:

$$\sigma^2 \mathbf{I}_{d\times d} \prec \frac{1}{T}\sum_{i\in[T]} X_i X_i^\top \prec \kappa \cdot \sigma^2 \mathbf{I}_{d\times d}.$$

- If noise is ν^2-sub-Gaussian, the first condition ensures a signal-to-noise ratio of $\frac{\sigma}{\nu}$.

Assumption 3: large signal and well-conditioned covariance

We assume that the sample covariance satisfies:

$$\sigma^2 \mathbf{I}_{d\times d} \prec \frac{1}{T}\sum_{i\in[T]} X_i X_i^\top \prec \kappa \cdot \sigma^2 \mathbf{I}_{d\times d}.$$

- If noise is ν^2-sub-Gaussian, the first condition ensures a signal-to-noise ratio of $\frac{\sigma}{\nu}$.
- The two conditions together ensure the data is well-conditioned,

$$\mathrm{cond}\left(\mathbf{X}\mathbf{X}^\top\right) \le \kappa.$$

Convergence rate for OLS

Theorem

Let $(X_t, Y_t)_{t=1}^T$ *be a sequence where* $Y_t = \mathbf{A}_\star X_t + \eta_t$ *such that:*

(a) *the noise* $\eta_{t+1} \mid \mathcal{F}_t$ *is* ν^2*-sub-Gaussian,*

(b) *with probability at least* $1 - \delta$*, the sample covariance satisfies:*

$$\sigma^2 \mathbf{I}_{d\times d} \prec \frac{1}{T} \sum_{i\in[T]} X_i X_i^\top \prec \kappa \cdot \sigma^2 \mathbf{I}_{d\times d}.$$

Convergence rate for OLS

Theorem

Let $(X_t, Y_t)_{t=1}^T$ *be a sequence where* $Y_t = \mathbf{A}_\star X_t + \eta_t$ *such that:*

(a) *the noise* $\eta_{t+1} \mid \mathcal{F}_t$ *is* ν^2*-sub-Gaussian,*

(b) *with probability at least* $1 - \delta$*, the sample covariance satisfies:*

$$\sigma^2 \mathbf{I}_{d\times d} \prec \frac{1}{T} \sum_{i\in[T]} X_i X_i^\top \prec \kappa \cdot \sigma^2 \mathbf{I}_{d\times d}.$$

Then, with probability $\geq 1 - 2\delta$*, the OLS estimator* $\hat{\mathbf{A}}$ *satisfies:*

$$\|\mathbf{A}_\star - \hat{\mathbf{A}}\|_{\mathrm{op}} = O\left(\frac{\nu}{\sigma}\sqrt{\frac{1}{T}\left(n + d\log\kappa + \log\frac{1}{\delta}\right)}\right).$$

Proof sketch

We decompose the error as:

$$\|\mathbf{A}_\star - \hat{\mathbf{A}}\|_{\mathrm{op}} = \|\mathbf{E}\mathbf{X}^+\|_{\mathrm{op}} \leq \|\mathbf{E}\|_{\mathrm{op}} \cdot \sigma_d(\mathbf{X})^{-1}.$$

Proof sketch

We decompose the error as:

$$\|\mathbf{A}_\star - \hat{\mathbf{A}}\|_{\mathrm{op}} = \|\mathbf{E}\mathbf{X}^+\|_{\mathrm{op}} \leq \|\mathbf{E}\|_{\mathrm{op}} \cdot \sigma_d(\mathbf{X})^{-1}.$$

- Control $\|\mathbf{E}\|_{\mathrm{op}}$ by concentration of sub-Gaussian martingales to show w.p. $\geq 1-\delta$,

$$\|\mathbf{E}\|_{\mathrm{op}} = O\left(\nu\sqrt{n + d\log\kappa + \log\frac{1}{\delta}}\right).$$

Proof sketch

We decompose the error as:

$$\|\mathbf{A}_\star - \hat{\mathbf{A}}\|_{\mathrm{op}} = \|\mathbf{E}\mathbf{X}^+\|_{\mathrm{op}} \le \|\mathbf{E}\|_{\mathrm{op}} \cdot \sigma_d(\mathbf{X})^{-1}.$$

- Control $\|\mathbf{E}\|_{\mathrm{op}}$ by concentration of sub-Gaussian martingales to show w.p. $\ge 1-\delta$,

$$\|\mathbf{E}\|_{\mathrm{op}} = O\left(\nu\sqrt{n + d\log\kappa + \log\frac{1}{\delta}}\right).$$

- The assumption $\sigma^2 \mathrm{I}_{d\times d} \prec \frac{1}{T}\mathbf{X}\mathbf{X}^\top$ implies that w.p. $\ge 1-\delta$,

$$\sigma_d(\mathbf{X})^{-1} \le \frac{1}{\sigma\sqrt{T}}.$$

Proof sketch

We decompose the error as:

$$\|\mathbf{A}_\star - \hat{\mathbf{A}}\|_{\mathrm{op}} = \|\mathbf{E}\mathbf{X}^+\|_{\mathrm{op}} \leq \|\mathbf{E}\|_{\mathrm{op}} \cdot \sigma_d(\mathbf{X})^{-1}.$$

- Control $\|\mathbf{E}\|_{\mathrm{op}}$ by concentration of sub-Gaussian martingales to show w.p. $\geq 1-\delta$,

$$\|\mathbf{E}\|_{\mathrm{op}} = O\left(\nu\sqrt{n + d\log\kappa + \log\frac{1}{\delta}}\right).$$

- The assumption $\sigma^2 \mathrm{I}_{d\times d} \prec \frac{1}{T}\mathbf{X}\mathbf{X}^\top$ implies that w.p. $\geq 1-\delta$,

$$\sigma_d(\mathbf{X})^{-1} \leq \frac{1}{\sigma\sqrt{T}}.$$

Multiplying the two bounds yields the result. □

Technique: concentration bounds

Euclidean norm of sub-Gaussian random variables

Let $B^r \subset \mathbb{R}^r$ be the unit ball. The operator norm $\|\mathbf{E}\|_{\text{op}}$ for $\mathbf{E} : \mathbb{R}^T \to \mathbb{R}^n$ is defined as:

$$\|\mathbf{E}\|_{\text{op}} = \sup_{\substack{w \in B^T \\ v \in B^n}} v^\top \mathbf{E} w.$$

Euclidean norm of sub-Gaussian random variables

Let $B^r \subset \mathbb{R}^r$ be the unit ball. The operator norm $\|\mathbf{E}\|_{\mathrm{op}}$ for $\mathbf{E} : \mathbb{R}^T \to \mathbb{R}^n$ is defined as:

$$\|\mathbf{E}\|_{\mathrm{op}} = \sup_{\substack{w \in B^T \\ v \in B^n}} v^\top \mathbf{E} w.$$

- In this case, this is equal to $\|\mathbf{E}\|_{\mathrm{op}} = \max_{i \in [T]} \|\eta_i\|$.

Euclidean norm of sub-Gaussian random variables

Let $B^r \subset \mathbb{R}^r$ be the unit ball. The operator norm $\|\mathbf{E}\|_{\text{op}}$ for $\mathbf{E} : \mathbb{R}^T \to \mathbb{R}^n$ is defined as:

$$\|\mathbf{E}\|_{\text{op}} = \sup_{\substack{w \in B^T \\ v \in B^n}} v^\top \mathbf{E} w.$$

- In this case, this is equal to $\|\mathbf{E}\|_{\text{op}} = \max_{i \in [T]} \|\eta_i\|$.
- The following tail bound holds (Rinaldo, 2019), w.p. $\geq 1 - \delta$,

$$\|\eta\| = O\left(\nu \sqrt{n + \log \frac{1}{\delta}}\right).$$

Euclidean norm of sub-Gaussian random variables

Let $B^r \subset \mathbb{R}^r$ be the unit ball. The operator norm $\|\mathbf{E}\|_{\mathrm{op}}$ for $\mathbf{E} : \mathbb{R}^T \to \mathbb{R}^n$ is defined as:

$$\|\mathbf{E}\|_{\mathrm{op}} = \sup_{\substack{w \in B^T \\ v \in B^n}} v^\top \mathbf{E} w.$$

- In this case, this is equal to $\|\mathbf{E}\|_{\mathrm{op}} = \max_{i \in [T]} \|\eta_i\|$.
- The following tail bound holds (Rinaldo, 2019), w.p. $\geq 1 - \delta$,

$$\|\eta\| = O\left(\nu \sqrt{n + \log \frac{1}{\delta}}\right).$$

- It follows by a union bound, w.p. $\geq 1 - \delta$, $\|\mathbf{E}\|_{\mathrm{op}} = O\left(\nu \sqrt{n + \log T + \log \frac{1}{\delta}}\right)$.

Proof of tail bound

Assumption: η is ν^2-sub-Gaussian in $\mathbb{R}^n$. The norm is:

$$\|\eta\| = \sup_{v \in B^n} v^\top \eta.$$

Proof of tail bound

Assumption: η is ν^2-sub-Gaussian in $\mathbb{R}^n$. The norm is:

$$\|\eta\| = \sup_{v \in B^n} v^\top \eta.$$

1. Let $\mathcal{N}_{1/2}$ be a $\frac{1}{2}$-net of B^n. The optimal v^* decomposes into:

$$v^* = z + u,$$

where $z \in \mathcal{N}_{1/2}$ and $u \in \frac{1}{2}B^n$.

Proof of tail bound

Assumption: η is ν^2-sub-Gaussian in $\mathbb{R}^n$. The norm is:

$$\|\eta\| = \sup_{v \in B^n} v^\top \eta.$$

1. Let $\mathcal{N}_{1/2}$ be a $\frac{1}{2}$-net of B^n. The optimal v^* decomposes into:

$$v^* = z + u,$$

   where $z \in \mathcal{N}_{1/2}$ and $u \in \frac{1}{2} B^n$.
2. Taking the supremum over $\mathcal{N}_{1/2}$ instead yields upper bound,

$$\|\eta\| = \sup_{v \in B^n} v^\top \eta$$

Proof of tail bound

Assumption: η is ν^2-sub-Gaussian in $\mathbb{R}^n$. The norm is:

$$\|\eta\| = \sup_{v \in B^n} v^\top \eta.$$

1. Let $\mathcal{N}_{1/2}$ be a $\frac{1}{2}$-net of B^n. The optimal v^* decomposes into:

   $$v^* = z + u,$$

   where $z \in \mathcal{N}_{1/2}$ and $u \in \frac{1}{2}B^n$.
2. Taking the supremum over $\mathcal{N}_{1/2}$ instead yields upper bound,

   $$\|\eta\| = \sup_{v \in B^n} v^\top \eta \leq \sup_{z \in \mathcal{N}_{1/2}} z^\top \eta + \sup_{u \in \frac{1}{2}B^n} u^\top \eta$$

Proof of tail bound

Assumption: η is ν^2-sub-Gaussian in $\mathbb{R}^n$. The norm is:

$$\|\eta\| = \sup_{v \in B^n} v^\top \eta.$$

1. Let $\mathcal{N}_{1/2}$ be a $\frac{1}{2}$-net of B^n. The optimal v^* decomposes into:

   $$v^* = z + u,$$

   where $z \in \mathcal{N}_{1/2}$ and $u \in \frac{1}{2}B^n$.
2. Taking the supremum over $\mathcal{N}_{1/2}$ instead yields upper bound,

$$\|\eta\| = \sup_{v \in B^n} v^\top \eta \leq \sup_{z \in \mathcal{N}_{1/2}} z^\top \eta + \sup_{u \in \frac{1}{2}B^n} u^\top \eta = \sup_{z \in \mathcal{N}_{1/2}} z^\top \eta + \frac{1}{2}\|\eta\|.$$

Proof of tail bound

Assumption: η is ν^2-sub-Gaussian in $\mathbb{R}^n$. The norm is:

$$\|\eta\| = \sup_{v \in B^n} v^\top \eta.$$

1. Let $\mathcal{N}_{1/2}$ be a $\frac{1}{2}$-net of B^n. The optimal v^* decomposes into:

   $$v^* = z + u,$$

   where $z \in \mathcal{N}_{1/2}$ and $u \in \frac{1}{2}B^n$.
2. Taking the supremum over $\mathcal{N}_{1/2}$ instead yields upper bound, $\|\eta\| \leq 2 \sup_{z \in \mathcal{N}_{1/2}} z^\top \eta$,

$$\|\eta\| = \sup_{v \in B^n} v^\top \eta \leq \sup_{z \in \mathcal{N}_{1/2}} z^\top \eta + \sup_{u \in \frac{1}{2}B^n} u^\top \eta = \sup_{z \in \mathcal{N}_{1/2}} z^\top \eta + \frac{1}{2}\|\eta\|.$$

Proof of tail bound (cont.)

2. We have an upper bound, $\|\eta\| \leq 2 \sup_{z \in \mathcal{N}_{1/2}} z^\top \eta$.

Proof of tail bound (cont.)

2. We have an upper bound, $\|\eta\| \leq 2 \sup_{z \in \mathcal{N}_{1/2}} z^\top \eta$.
3. Each $z^\top \eta$ is ν^2-sub-Gaussian, so apply union bound:

$$\Pr\left(\|\eta\| \geq \varepsilon\right) \leq \Pr\left(\sup_{z \in \mathcal{N}_{1/2}} z^\top \eta \geq 2\varepsilon\right) \leq |\mathcal{N}_{1/2}| \cdot \exp\left(-\frac{\varepsilon^2}{8\nu^2}\right).$$

Proof of tail bound (cont.)

2. We have an upper bound, $\|\eta\| \leq 2 \sup_{z \in \mathcal{N}_{1/2}} z^\top \eta$.
3. Each $z^\top \eta$ is ν^2-sub-Gaussian, so apply union bound:

$$\Pr\left(\|\eta\| \geq \varepsilon\right) \leq \Pr\left(\sup_{z \in \mathcal{N}_{1/2}} z^\top \eta \geq 2\varepsilon\right) \leq |\mathcal{N}_{1/2}| \cdot \exp\left(-\frac{\varepsilon^2}{8\nu^2}\right).$$

4. By a covering number bound $|\mathcal{N}_{1/2}| \leq 5^n \leq e^{2n}$, this implies:

$$\Pr\left(\|\eta\| \geq \nu\sqrt{16n + 8\log\frac{1}{\delta}}\right) \leq \delta.$$

□

Technical issue

We already remarked that a union bound shows that w.p. $\geq 1 - \delta$,

$$\|\mathbf{E}\|_{\text{op}} = O\left(\nu\sqrt{n + \log T + \log\frac{1}{\delta}}\right).$$

- But we actually want a bound that is independent of T,

$$\|\mathbf{E}\|_{\text{op}} = O\left(\nu\sqrt{n + d\log\kappa + \log\frac{1}{\delta}}\right).$$

First attempt at bound independent of T

To analyze $\|\mathbf{E}\mathbf{X}^{+}\|_{\mathrm{op}}$, we can restrict the domain of $\mathbf{E}$ to at most d dimensions,

$$\mathbf{E} : \mathrm{Im}(\mathbf{X}^{+}) \to \mathbb{R}^{n}.$$

First attempt at bound independent of T

To analyze $\|\mathbf{E}\mathbf{X}^{+}\|_{\text{op}}$, we can restrict the domain of $\mathbf{E}$ to at most d dimensions,

$$\mathbf{E} : \text{Im}(\mathbf{X}^{+}) \to \mathbb{R}^{n}.$$

- If we fix a d-dimensional subspace $V \subset \mathbb{R}^{T}$ and restricted the domain $\mathbf{E} : V \to \mathbb{R}^{n}$, the operator norm is now:

$$\|\mathbf{E}\|_{\text{op}} = \sup_{\substack{w \in B^{T} \cap V \\ v \in B^{n}}} v^{\top} \mathbf{E} w.$$

First attempt at bound independent of T

To analyze $\|\mathbf{E}\mathbf{X}^+\|_{\text{op}}$, we can restrict the domain of $\mathbf{E}$ to at most d dimensions,

$$\mathbf{E} : \text{Im}(\mathbf{X}^+) \to \mathbb{R}^n.$$

- If we fix a d-dimensional subspace $V \subset \mathbb{R}^T$ and restricted the domain $\mathbf{E} : V \to \mathbb{R}^n$, the operator norm is now:

$$\|\mathbf{E}\|_{\text{op}} = \sup_{\substack{w \in B^T \cap V \\ v \in B^n}} v^\top \mathbf{E} w.$$

- We can apply the same covering argument to both B^n and $B^T \cap V$ to obtain:

$$\|\mathbf{E}\|_{\text{op}} = O\left(\nu \sqrt{n + d + \log \frac{1}{\delta}}\right).$$

First attempt at bound independent of T

To analyze $\|\mathbf{E}\mathbf{X}^+\|_{\mathrm{op}}$, we can restrict the domain of $\mathbf{E}$ to at most d dimensions,

$$\mathbf{E} : \mathrm{Im}(\mathbf{X}^+) \to \mathbb{R}^n.$$

- If we fix a d-dimensional subspace $V \subset \mathbb{R}^T$ and restricted the domain $\mathbf{E} : V \to \mathbb{R}^n$, the operator norm is now:

$$\|\mathbf{E}\|_{\mathrm{op}} = \sup_{\substack{w \in B^T \cap V \\ v \in B^n}} v^\top \mathbf{E} w.$$

- We can apply the same covering argument to both B^n and $B^T \cap V$ to obtain:

$$\|\mathbf{E}\|_{\mathrm{op}} = O\left(\nu \sqrt{n + d + \log \frac{1}{\delta}}\right).$$

- However, $\mathrm{Im}(\mathbf{X}^+)$ is data-dependent while V is not; argument does not apply.

Concentration bound independent of T

When the domain is restricted $\mathbf{E} : \mathrm{Im}(\mathbf{X}^+) \to \mathbb{R}^n$, then:

$$\|\mathbf{E}\|_{\mathrm{op}} = \sup_{\substack{w \in \mathbb{R}^d \\ v \in B^n}} \frac{v^\top \mathbf{E} \mathbf{X}^+ w}{\|\mathbf{X}^+ w\|}.$$

Concentration bound independent of T

When the domain is restricted $\mathbf{E} : \mathrm{Im}(\mathbf{X}^+) \to \mathbb{R}^n$, then:

$$\|\mathbf{E}\|_{\mathrm{op}} = \sup_{\substack{w \in \mathbb{R}^d \\ v \in B^n}} \frac{v^\top \mathbf{E} \mathbf{X}^+ w}{\|\mathbf{X}^+ w\|}.$$

- If we know that $\sigma^2 \mathbf{I}_{d\times d} \prec \frac{1}{T}\mathbf{X}\mathbf{X}^\top \prec \kappa \cdot \sigma^2 \mathbf{I}_{d\times d}$, it suffices to consider:

$$\mathcal{N}_{\sigma/2} = \text{ a } \frac{\sigma}{2}\text{-covering of the } \sqrt{\kappa}\cdot\sigma\text{-ball in } \mathbb{R}^d.$$

Concentration bound independent of T

When the domain is restricted $\mathbf{E} : \mathrm{Im}(\mathbf{X}^+) \to \mathbb{R}^n$, then:

$$\|\mathbf{E}\|_{\mathrm{op}} = \sup_{\substack{w \in \mathbb{R}^d \\ v \in B^n}} \frac{v^\top \mathbf{E} \mathbf{X}^+ w}{\|\mathbf{X}^+ w\|}.$$

- If we know that $\sigma^2 \mathbf{I}_{d\times d} \prec \frac{1}{T}\mathbf{X}\mathbf{X}^\top \prec \kappa \cdot \sigma^2 \mathbf{I}_{d\times d}$, it suffices to consider:

$$\mathcal{N}_{\sigma/2} = \text{ a } \frac{\sigma}{2}\text{-covering of the } \sqrt{\kappa}\cdot\sigma\text{-ball in } \mathbb{R}^d.$$

  - The set $\mathbf{X}^+\mathcal{N}_{\sigma/2}$ is therefore a $\frac{1}{2}$-covering of $\mathrm{Im}(\mathbf{X}^+)$.

Concentration bound independent of T

When the domain is restricted $\mathbf{E} : \mathrm{Im}(\mathbf{X}^+) \to \mathbb{R}^n$, then:

$$\|\mathbf{E}\|_{\mathrm{op}} = \sup_{\substack{w \in \mathbb{R}^d \\ v \in B^n}} \frac{v^\top \mathbf{E} \mathbf{X}^+ w}{\|\mathbf{X}^+ w\|}.$$

- If we know that $\sigma^2 \mathbf{I}_{d\times d} \prec \frac{1}{T}\mathbf{X}\mathbf{X}^\top \prec \kappa \cdot \sigma^2 \mathbf{I}_{d\times d}$, it suffices to consider:

$$\mathcal{N}_{\sigma/2} = \text{ a } \frac{\sigma}{2}\text{-covering of the } \sqrt{\kappa}\cdot\sigma\text{-ball in } \mathbb{R}^d.$$

  - The set $\mathbf{X}^+\mathcal{N}_{\sigma/2}$ is therefore a $\frac{1}{2}$-covering of $\mathrm{Im}(\mathbf{X}^+)$.
- Since $\log|\mathcal{N}_{\sigma/2}| = O\left(d \log \kappa\right)$, we obtain:

$$\|\mathbf{E}\|_{\mathrm{op}} = O\left(\nu\sqrt{n + d\log\kappa + \log\frac{1}{\delta}}\right).$$

Geometric view of covering

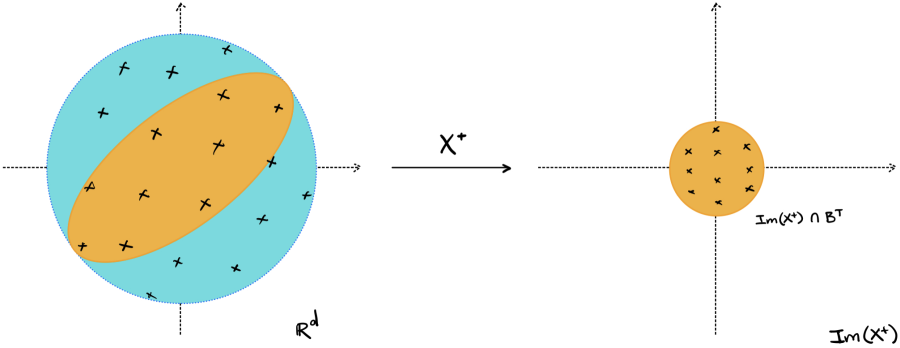

Figure 8: $\mathcal{N}_{\sigma/2}$ is a $\frac{\sigma}{2}$-covering (black x's) of the $\sqrt{\kappa} \cdot \sigma B^d$ (cyan ball). As long as $\frac{1}{T}\mathbf{X}\mathbf{X}^\top$ is sufficiently well-conditioned with eigenvalues greater than σ^2 (i.e. orange ellipse is contained in $\sqrt{\kappa} \cdot \sigma B^d$ and contains σB^d), then $\mathbf{X}^+\mathcal{N}_{\sigma/2}$ is a $\frac{1}{2}$-covering of $\mathrm{Im}(\mathbf{X}^+) \cap B^T$.

Application to linear dynamical systems

Identification of linear dynamical systems

Corollary

Consider a linear dynamical system $X_{t+1} = \mathbf{A}_\star X_t + \eta_t$ *where* $\mathbf{A}_\star$ *is marginally stable (i.e.* $\rho(\mathbf{A}_\star) \leq 1$*),* $X_0 = 0$*, and* $\eta_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \nu^2 \mathbf{I}_{n\times n})$*. For sufficiently large* T*, w.p.* $\geq 1-\delta$*,*

$$\|\mathbf{A}_\star - \hat{\mathbf{A}}\|_{\text{op}} = O\left(\sqrt{\frac{1}{T\lambda_{\min}(\mathbf{\Gamma}_k)}\left(d\log\frac{d}{\delta} + \log\det\left(\mathbf{\Gamma}_T \mathbf{\Gamma}_k^{-1}\right)\right)}\right),$$

where $\mathbf{\Gamma}_t := \sum_{s=0}^{t-1} (\mathbf{A}_\star^s)(\mathbf{A}_\star^s)^\top$ *and* $\lambda_{\min}$ *yields the smallest eigenvalue.*

- Note that $X_t = \sum_{s=1}^{t} \mathbf{A}_\star^{t-s} \eta_{s-1}$, so that $\mathbb{E}\left[X_t X_t^\top\right] = \nu^2 \mathbf{\Gamma}_t$.

References

Daniel Hsu, Sham M Kakade, and Tong Zhang. Random design analysis of ridge regression. In *Conference on learning theory*, pages 9–1. JMLR Workshop and Conference Proceedings, 2012.

Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for non-iid processes. 2008.

Alessandro Rinaldo. Lecture 8 – euclidean norm of sub-gaussian random vectors. In *Lecture notes for Advanced Statistical Theory*, February 2019. URL *http://www.stat.cmu.edu/~arinaldo/Teaching/36709/S19/Scribed_Lectures/Feb21_Shenghao.pdf*.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pages 439–473. PMLR, 2018.