Concentration Inequalities: Hanson-Wright and Self-Normalized Martingales

Bruce Lee (University of Pennsylvania)

Concentration Inequalities: Bound deviation of random variable from some value

Recall statistical model: $Y_t = \theta^\star X_t + V_t, \quad t = 1,\ldots,T$

Recall decomposition:

$$\hat{\theta} - \theta^\star = \frac{1}{\sqrt{T}} \left[\left(\sum_{t=1}^{T} V_t X_t^\top \right) \left(\sum_{t=1}^{T} X_t X_t^\top \right)^{-1/2} \right] \left(\frac{1}{T} \sum_{t=1}^{T} X_t X_t^\top \right)^{-1/2}$$

Self-normalized Martingale

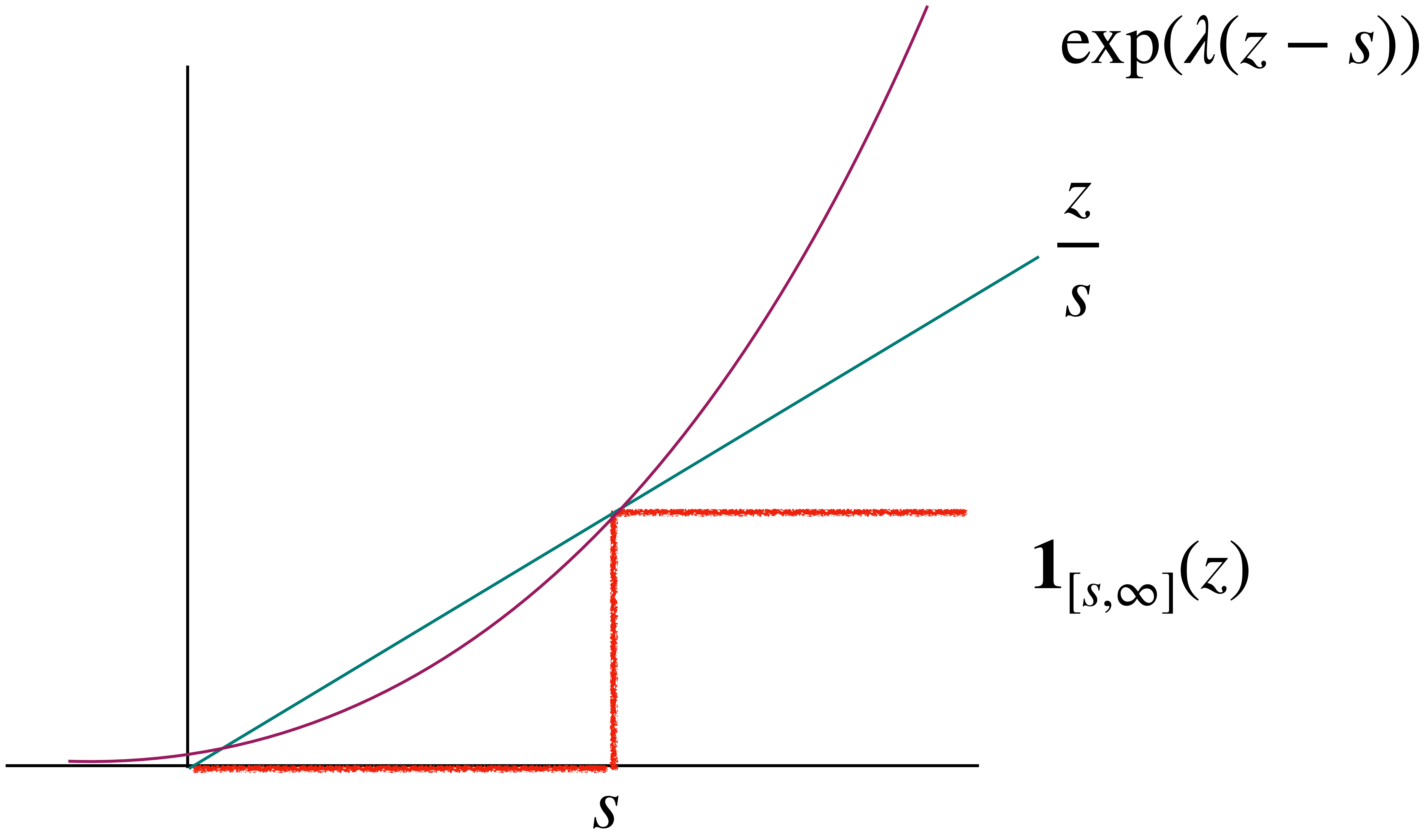
Expected small if V_t zero mean and independent

Sample Covariance matrix

Expected to concentrate to true covariance with enough samples

Z: random variable s.t $\mathbf{E}[Z]$ exists

Consider $\mathbf{P}[Z \geq s] = \mathbf{E}[1_{[s,\infty]}(Z)]$



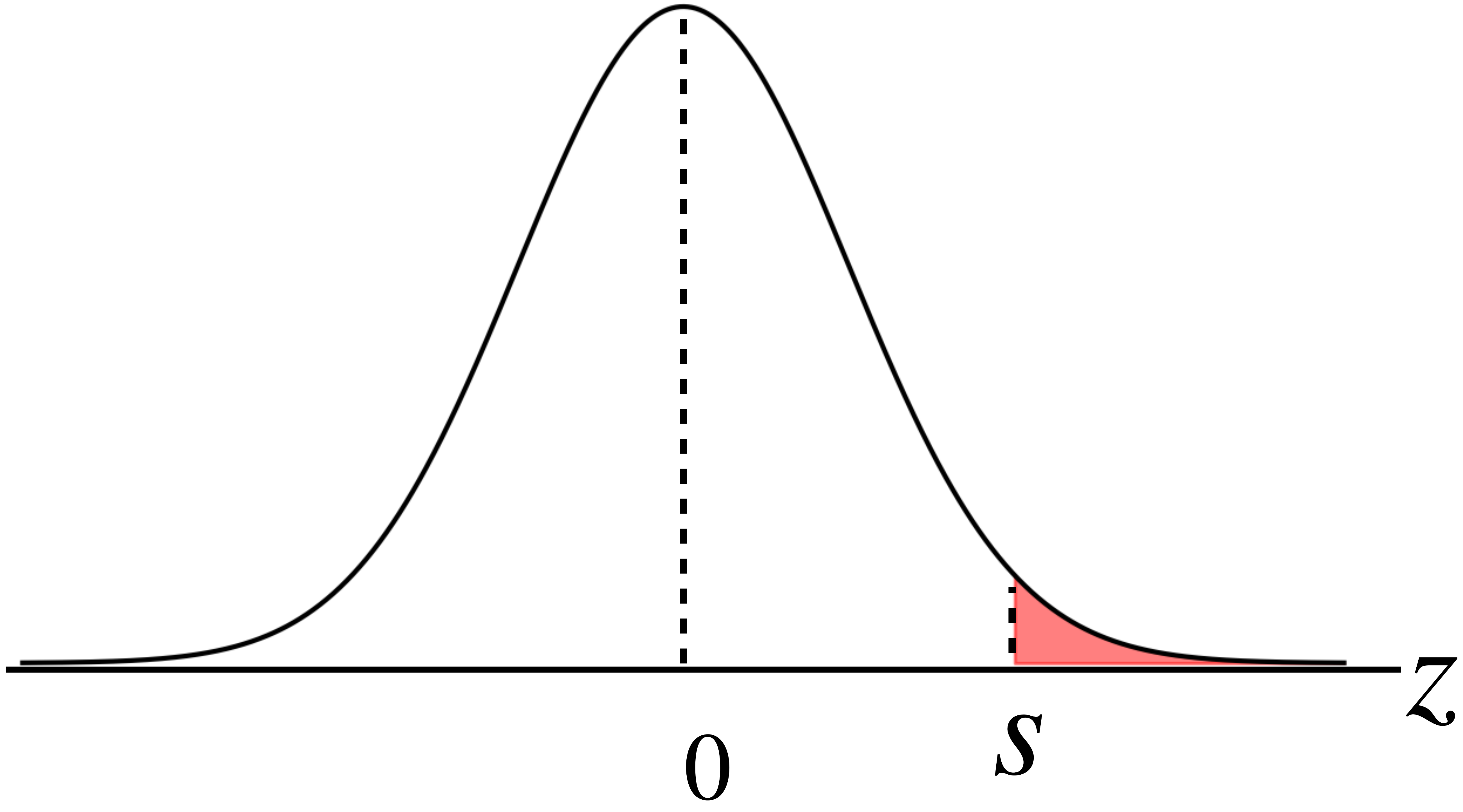
Markov's inequality $\quad \mathbf{P}[Z \geq s] \leq s^{-1}\mathbf{E}\,[Z] \quad$ (Z nonnegative)

Chernoff Bound $\quad \mathbf{P}[Z \geq s] \leq \min_{\lambda \geq 0} \exp(-\lambda s)\mathbf{E}\left[\exp(\lambda Z)\right] \quad$ ($\mathbf{E}[\exp(\lambda Z)]$ exists)

Gaussian Tail Bounds

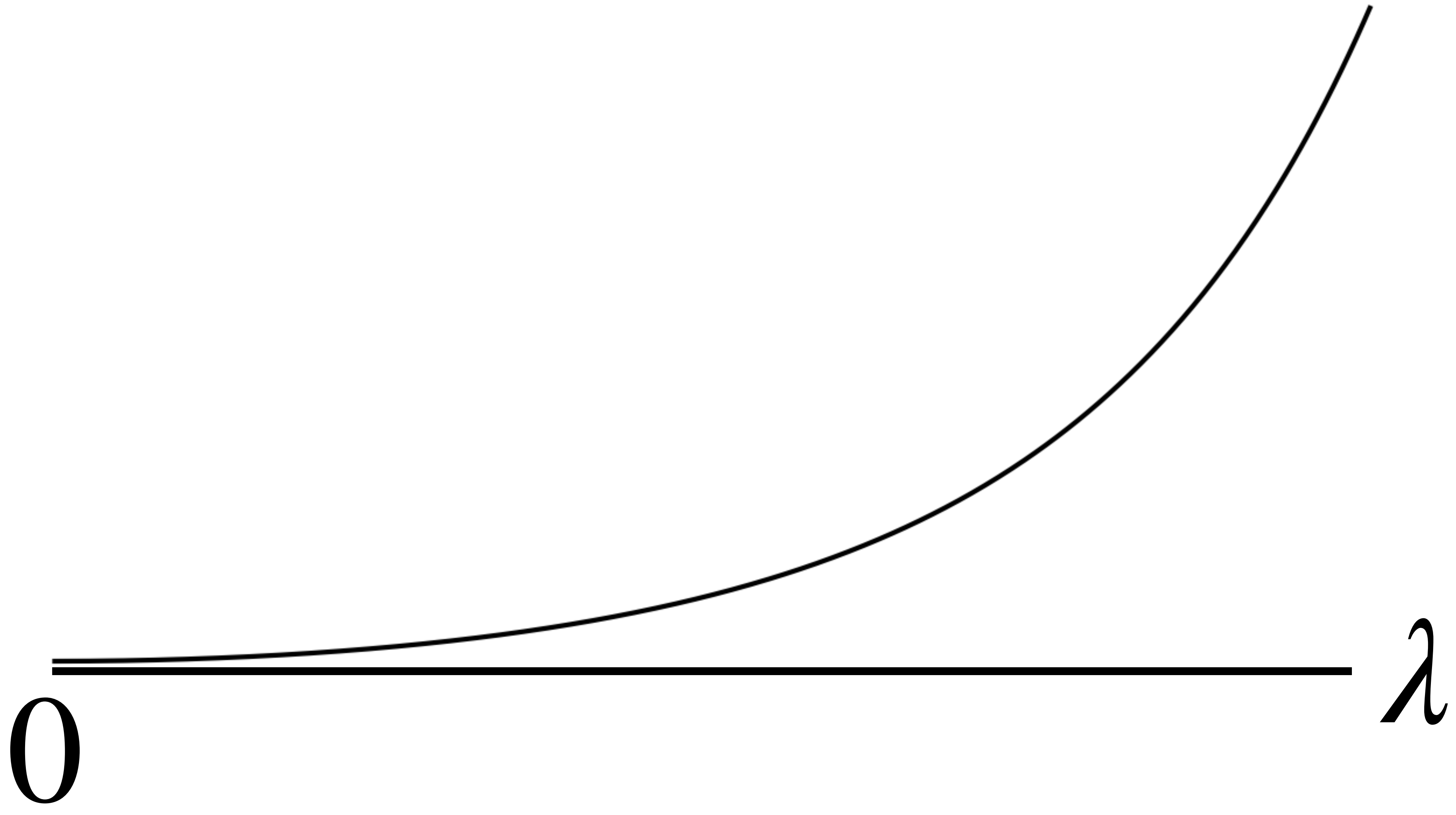
$$Z \sim N(0,\sigma^2)$$

Density



$$\mathbf{E}\left[\exp(\lambda Z)\right] = \exp\left(\frac{1}{2}\sigma^2\lambda^2\right)$$

Moment Generating Function



$$\mathbf{P}[Z \geq s] \leq \min_{\lambda \geq 0} \exp(-\lambda s)\mathbf{E}\left[\exp(\lambda Z)\right] = \min_{\lambda \geq 0} \exp\left(-\lambda s + \frac{1}{2}\sigma^2\lambda^2\right) = \exp\left(\frac{-s^2}{2\sigma^2}\right)$$

Gaussian Tail Bounds

$$\mathbf{P}[Z \geq s] \leq \exp\left(\frac{-s^2}{2\sigma^2}\right)$$

Sub-Gaussian Random Variables

$$\mathbf{P}[Z \geq s] \leq \exp\left(\frac{-s^2}{2\sigma^2}\right)$$

This tail bound holds for any RV Z w/

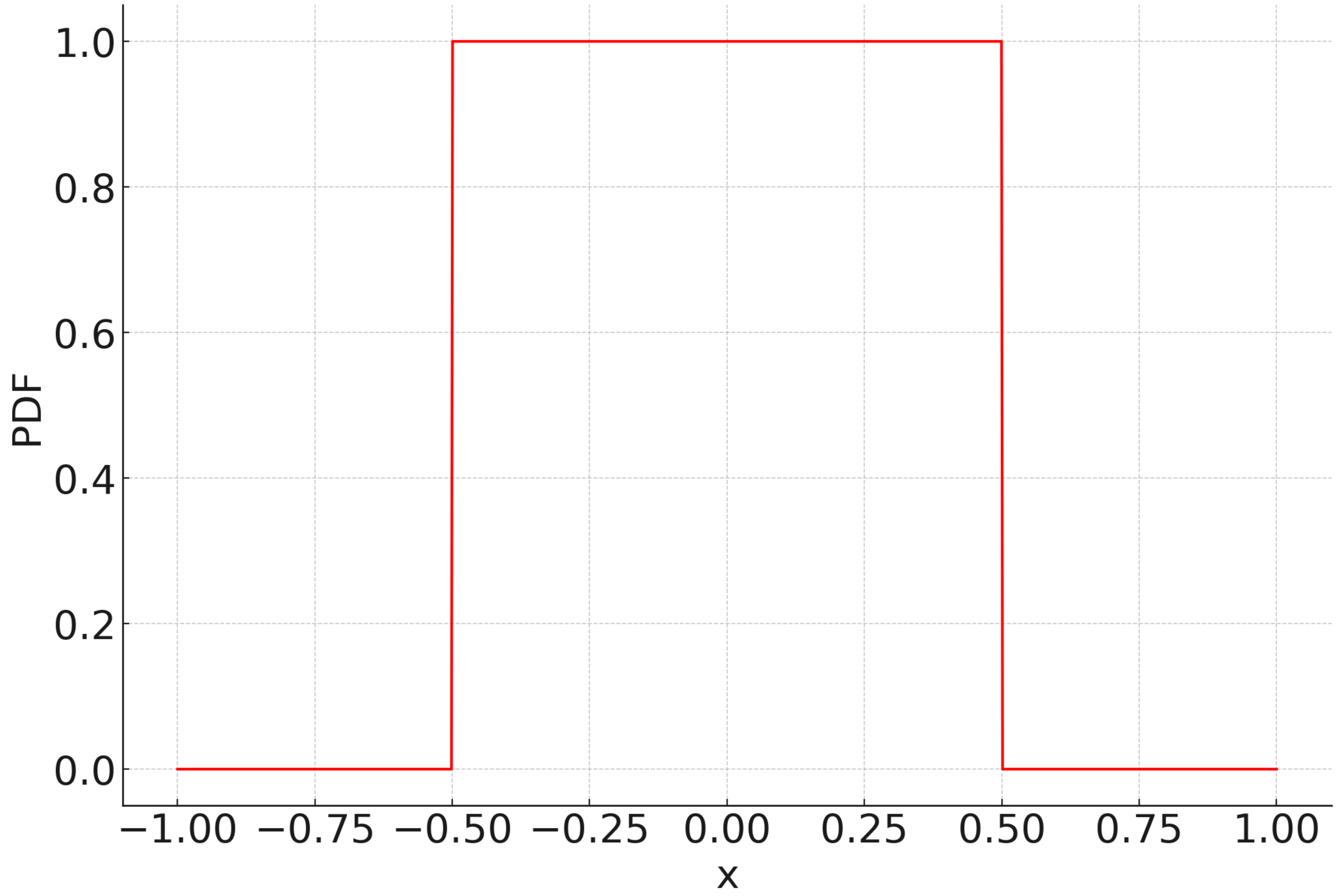
$$\mathbf{E}\left[\exp(\lambda Z)\right] \leq \exp\left(\frac{1}{2}\sigma^2\lambda^2\right)$$

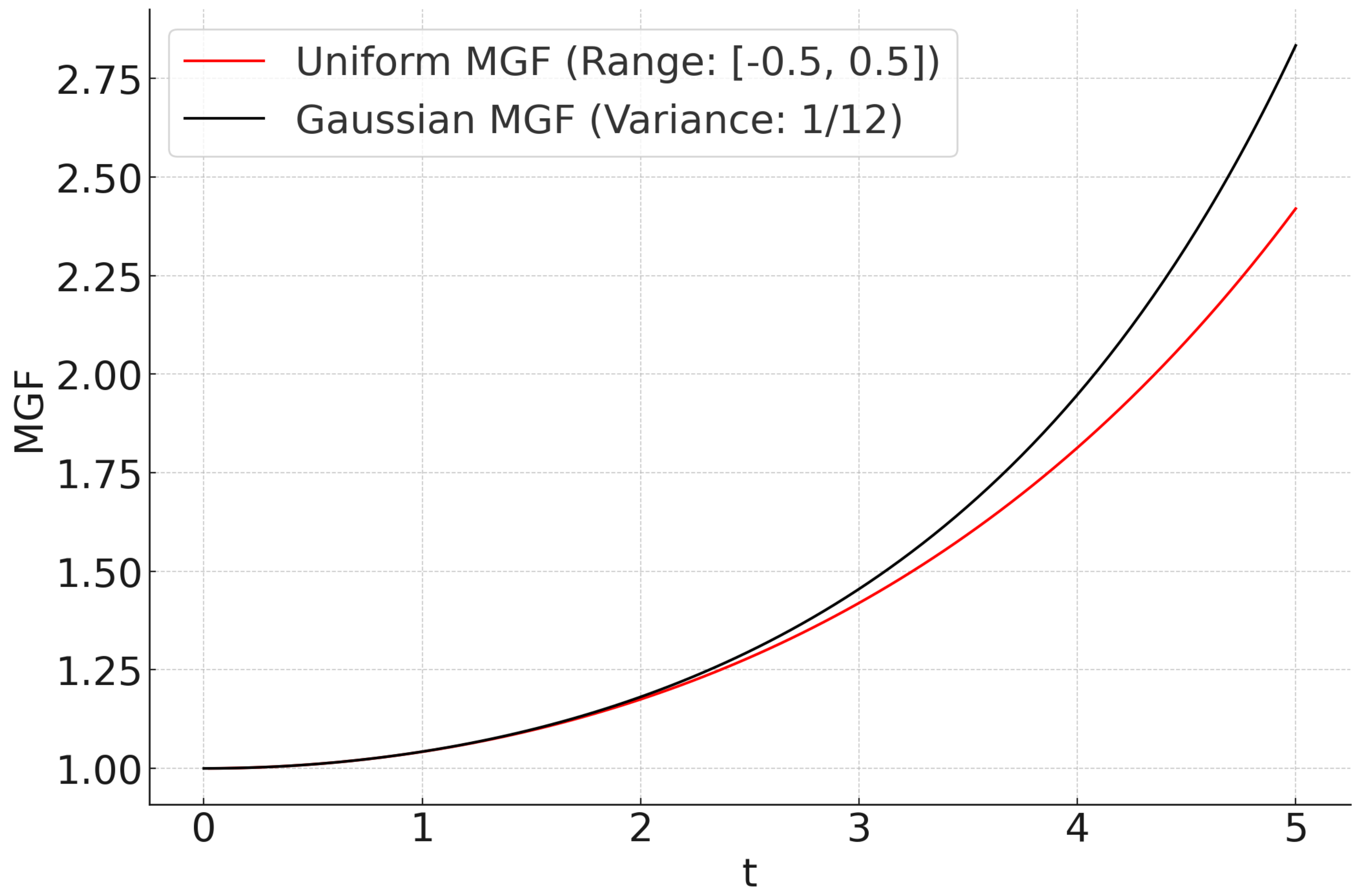
Definition: Such RVs are called σ^2-*sub-Gaussian*

Example:

$Z \sim U[-0.5,0.5]$

$$\mathbf{E}\left[\exp(\lambda Z)\right] \leq \exp\left(\frac{1}{2}\left(\frac{1}{12}\right)\lambda^2\right)$$





Sub-Gaussian Random Variables

Definition: More generally, a d dimensional random vector Z with

$$\mathbf{E}\left[\exp(\lambda v^\top Z)\right] \leq \exp\left(\frac{1}{2}\lambda^2\sigma^2\right) \quad \forall v : \|v\|_2 = 1$$

is σ^2*-sub-Gaussian*

Example: if Z is a zero mean RV assuming values in $\mathbb{R}^d$ with $a \leq Z_i \leq b$ for $i = 1,\ldots,d$, then Z is $\frac{(b-a)^2}{4}$-sub-Gaussian

Sub-Gaussian Random Variables

Definition: More generally, a d dimensional random vector Z with

$$\mathbf{E}\left[\exp(\lambda v^\top Z)\right] \leq \exp\left(\frac{1}{2}\lambda^2\sigma^2\right) \quad \forall v : \|v\|_2 = 1$$

is σ^2-*sub-Gaussian*

SubG concentration: If Z is a σ^2-sub-Gaussian vector assuming values in $\mathbb{R}^d$, then for any unit vector $v \in \mathbb{R}^d$

$$\mathbf{P}[v^\top Z \geq s] \leq \exp\left(\frac{-s^2}{2\sigma^2}\right)$$

Generality offered by sub-Gaussianity makes it a useful assumption for the noise V_t and updates of X_t from our statistical model

$$Y_t = \theta^\star X_t + V_t, \quad t = 1,\ldots,T$$

Recall our empirical covariance matrix: $\frac{1}{T}\sum_{t=1}^{T} X_t X_t^\top$

$$\mathbf{P}\left[\left\|\frac{1}{T}\sum_{t=1}^{T} X_t X_t^\top - \mathbf{E}\left[\frac{1}{T}\sum_{t=1}^{T} X_t X_t^\top\right]\right\|_{\mathrm{op}} \geq s\right]$$

Sub-Gaussian tail bounds are not enough to bound quadratic forms

Hanson-Wright Inequality

- $M \in \mathbb{R}^{d\times d}$
- W is RV assuming values in $\mathbb{R}^d$ with independent, σ^2-sub-Gaussian elements

$$\mathbf{P}\left(|W^\top MW - \mathbf{E}W^\top MW| > s\right) \le 2\exp\left(-\min\left(\frac{s^2}{144\sigma^2\|M\|_F^2}, \frac{s}{16\sqrt{2}\sigma^2\|M\|_{\mathrm{op}}}\right)\right)$$

Proved using sub-Gaussian concentration along with a *decoupling* technique:

- f convex
- M diagonal free
- W is RV with independent, zero-mean elements

$$\mathbf{E}f(W^\top MW) \le \mathbf{E}f(4W^\top MW')$$

where W' is an independent copy of W

Hanson-Wright Inequality

$M \in \mathbb{R}^{d \times d}$, W is RV assuming values in $\mathbb{R}^d$ with independent, σ^2-sub-Gaussian elements

$$\mathbf{P}\left(|W^\top M W - \mathbf{E} W^\top M W| > s\right) \leq 2\exp\left(-\min\left(\frac{s^2}{144\sigma^2\|M\|_F^2}, \frac{s}{16\sqrt{2}\sigma^2\|M\|_{\text{op}}}\right)\right)$$

$$\mathbf{P}\left[\left\| \frac{1}{T}\sum_{t=1}^{T} X_t X_t^\top - \mathbf{E}\left[\frac{1}{T}\sum_{t=1}^{T} X_t X_t^\top\right]\right\|_{\text{op}} \geq s\right]$$

Reduce operator norm to difference of scalar quantities

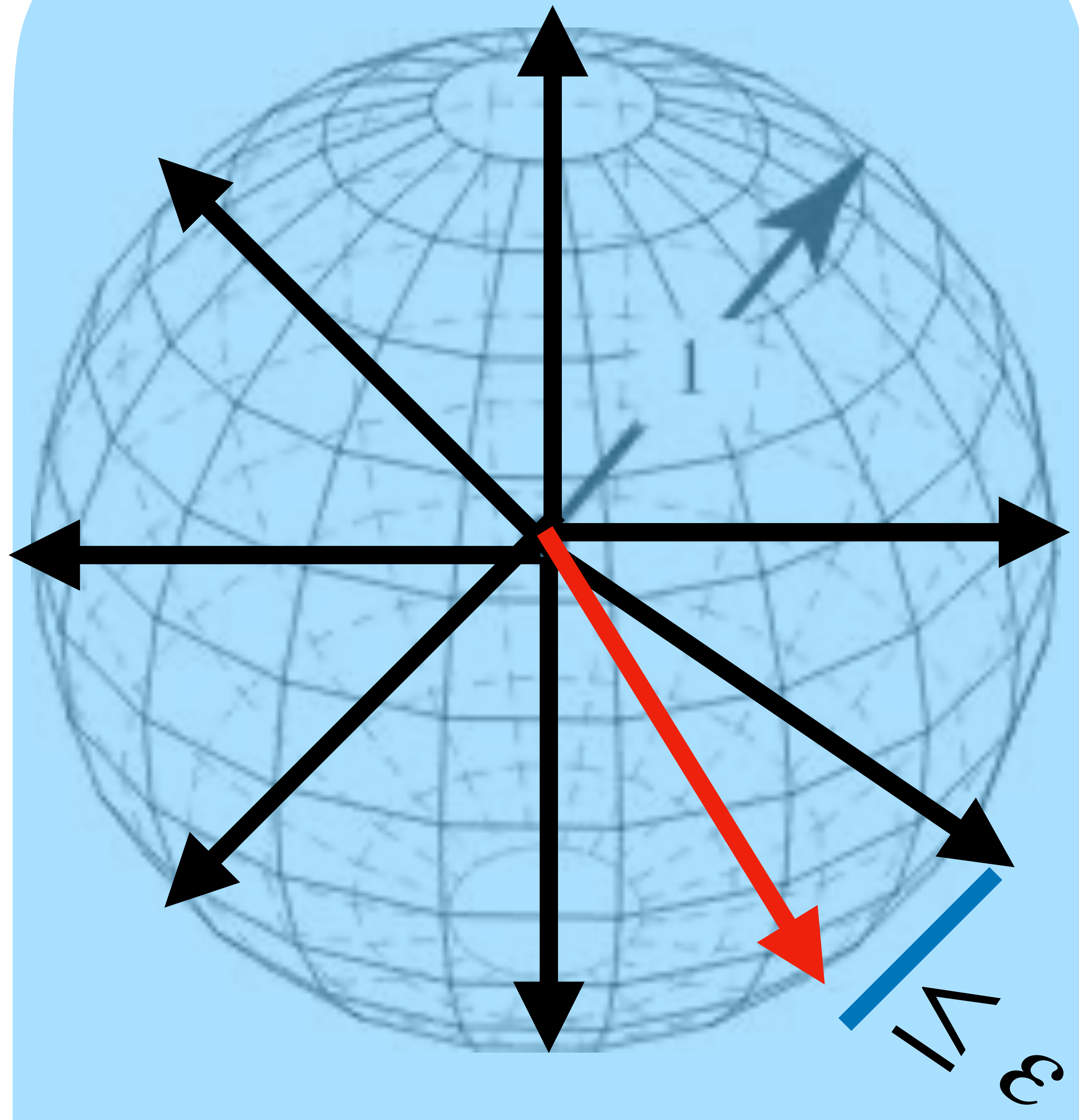
$H \in \mathbb{R}^{d\times d}$ is symmetric, $\|H\|_{\text{op}} = \sup_{v:\|v\|_2=1} |v^\top H v|$

To bound $\|\cdot\|_{\text{op}}$ for a random matrix H:

- Union bound over scalar concentration events ~~for all $v : \|v\|_2 = 1$?~~
- for all $v \in \varepsilon$-net and account for the error

Covering Argument

ε - net



There exists an ε-net w/ $\leq \left(1 + \frac{2}{\varepsilon}\right)^d$ elements

If $\mathcal{N}$ is a minimum cardinality ε-net for the unit sphere,

$$\mathbf{P}\left(\|H\|_{\text{op}} > \rho\right) \leq \left(1 + \frac{2}{\varepsilon}\right)^d \max_{v\in\mathcal{N}} \mathbf{P}\left(|v^\top H v| > (1-2\varepsilon)\rho\right)$$

Covariance Concentration for Stochastic System Identification

Consider d dimensional system $X_{t+1} = A^\star X_t + W_t$ where W_t has independent σ^2-sub-Gaussian elements

$$\begin{bmatrix} X_1 \\ \vdots \\ X_T \end{bmatrix} = \mathrm{Toep}_T(A^\star) \begin{bmatrix} W_1 \\ \vdots \\ W_{T-1} \end{bmatrix}$$

$$v : \|v\|_2 = 1, \quad \frac{1}{T}\sum_{t=1}^{T} v^\top X_t X_t^\top v = \begin{bmatrix} W_1 \\ \vdots \\ W_{T-1} \end{bmatrix}^\top M_{v,A^\star} \begin{bmatrix} W_1 \\ \vdots \\ W_{T-1} \end{bmatrix}$$

1. Apply covering to reduce tail bounding $\left\| \frac{1}{T}\sum_{t=1}^{T} X_t X_t^\top - \mathbf{E}\left[\frac{1}{T}\sum_{t=1}^{T} X_t X_t^\top\right] \right\|_{\mathrm{op}}$ to tail bounding $\left| \frac{1}{T}\sum_{t=1}^{T} v^\top X_t X_t^\top v - \mathbf{E}\left[\frac{1}{T}\sum_{t=1}^{T} v^\top X_t X_t^\top v\right] \right|$

2. Apply Hanson-Wright to bound

$$\left| \frac{1}{T}\sum_{t=1}^{T} v^\top X_t X_t^\top v - \mathbf{E}\left[\frac{1}{T}\sum_{t=1}^{T} v^\top X_t X_t^\top v\right] \right| = \left| \begin{bmatrix} W_1 \\ \vdots \\ W_{T-1} \end{bmatrix}^\top M_{v,A^\star} \begin{bmatrix} W_1 \\ \vdots \\ W_{T-1} \end{bmatrix} - \mathbf{E} \begin{bmatrix} W_1 \\ \vdots \\ W_{T-1} \end{bmatrix}^\top M_{v,A^\star} \begin{bmatrix} W_1 \\ \vdots \\ W_{T-1} \end{bmatrix} \right|$$

Covariance Concentration for Stochastic System Identification

Consider d dimensional system $X_{t+1} = A^\star X_t + W_t$

where W_t has independent σ^2-sub-Gaussian elements

$$\begin{bmatrix} X_1 \\ \vdots \\ X_T \end{bmatrix} = \mathrm{Toep}_T(A^\star) \begin{bmatrix} W_1 \\ \vdots \\ W_{T-1} \end{bmatrix}$$

Covariance Concentration

$$\mathbf{P}\left[\left\| \frac{1}{T}\sum_{t=1}^{T} X_t X_t^\top - \mathbf{E}\left[\frac{1}{T}\sum_{t=1}^{T} X_t X_t^\top\right] \right\|_{\text{op}} \le s \right] \ge 1 - 2\exp\left(\frac{-c_1 \lambda_{\min}\left(\mathbf{E}\left[\frac{1}{T}\sum_{t=1}^{T} X_t X_t^\top\right]\right) s^2 T}{\sigma^2 \left\| \mathbf{E}\left[\frac{1}{T}\sum_{t=1}^{T} X_t X_t^\top\right] \right\|^2 \left\| \mathrm{Toep}_T(A^\star) \right\|_{\text{op}}^2} + c_2 d \right)$$

where c_1 and c_2 are universal positive constants

For T sufficiently large, empirical covariance $\approx$ true covariance with high probability*

* if $\rho(A^\star) < 1$

Recall decomposition: $$\hat{\theta} - \theta^\star = \frac{1}{\sqrt{T}} \left[\left(\sum_{t=1}^{T} V_t X_t^\top \right) \left(\sum_{t=1}^{T} X_t X_t^\top \right)^{-1/2} \right] \left(\frac{1}{T} \sum_{t=1}^{T} X_t X_t^\top \right)$$

Self-normalized Martingale

Sample Covariance matrix

Self-Normalized Martingales

Definition: Martingale A process $S_1, S_2, S_3, \ldots$ is called a martingale if

$$\mathbf{E}\left[S_t \,|\, \text{past randomness}\right] = S_{t-1}$$

If V_t is mean zero, and independent of $X_1, \ldots, X_t$ and $V_1, \ldots V_{t-1}$

then $\mathbf{E}\left[\sum_{s=1}^{t} V_s X_s^\top \,|\, V_1, \ldots V_{t-1}, X_1, \ldots X_{t-1}\right] = \sum_{s=1}^{t-1} V_s X_s^\top$ so the process $\sum_{s=1}^{t} V_s X_s^\top$ is a martingale

Definition: Self-Normalized Martingale

The process $\left(\sum_{t=1}^{T} V_t X_t^\top\right)\left(\sum_{t=1}^{T} X_t X_t^\top\right)^{-1/2}$ is called a *self-normalized martingale* due to the normalization by $\left(\sum_{t=1}^{T} X_t X_t^\top\right)^{-1/2}$, which counteracts the growth of the process due to large X_t

Self-Normalized Martingale Bound

- Suppose V_t are independent σ^2-sub-Gaussian random variables and that X_t are independent from V_k for $k \geq t$
- Let d_X be the dimension of X_t and d_Y be the dimension of Y_t and V_t
- Let Σ be a $d_\mathsf{X} \times d_\mathsf{X}$ dimensional positive definite matrix

With probability at least $1 - \delta$

$$\left\| \left(\sum_{t=1}^{T} V_t X_t^\top \right) \left(\Sigma + \sum_{t=1}^{T} X_t X_t^\top \right)^{-1/2} \right\|_{\mathsf{op}}^2 \leq 4\sigma^2 \log \left(\frac{\det(\Sigma + \sum_{t=1}^{T} X_t X_t^\top)}{\det(\Sigma)} \right) + 8 d_\mathsf{Y} \sigma^2 \log(5) + 8\sigma^2 \log \frac{1}{\delta}$$

For T sufficiently large, empirical covariance $\approx$ true covariance with high probability*

With high probability, the self-normalized martingale term satisfies

* if $\rho(A^\star) < 1$

$$\left\| \left(\sum_{t=1}^{T} V_t X_t^\top \right) \left(\sum_{t=1}^{T} \Sigma + X_t X_t^\top \right)^{-1/2} \right\|_{\text{op}}^2 \leq 4\sigma^2 \log \left(\frac{\det(\Sigma + \sum_{t=1}^{T} X_t X_t^\top)}{\det(\Sigma)} \right) + 8 d_{\mathsf{Y}} \sigma^2 \log(5) + 8\sigma^2 \log \frac{1}{\delta}$$

Recall decomposition: $\hat{\theta} - \theta^\star = \frac{1}{\sqrt{T}} \left[\left(\sum_{t=1}^{T} V_t X_t^\top \right) \left(\sum_{t=1}^{T} X_t X_t^\top \right)^{-1/2} \right] \left(\frac{1}{T} \sum_{t=1}^{T} X_t X_t^\top \right)^{-1/2}$

Self-normalized Martingale

Sample Covariance matrix

C

where W has independent

ements

$\sum_{t=1}^{T} X_t X_t^\top$

Limitation:

If $\rho(A^\star) = 1$, as $T \to \infty$ $\|\mathrm{Toep}_T(A^\star)\| \to \infty$, $\|\Gamma_T(A^\star)\| \to \infty$

If $\rho(A^\star) < 1$, $\|\mathrm{Toep}_T(A^\star)\|$ and $\|\Gamma_T(A^\star)\|$ increase with $\rho(A^\star)$

St

Universal constant

$$\text{Suppose } T \geq c(\log(1/\delta) + d_X) \frac{\left\| \mathrm{Toep}_T(A^\star) \right\|_{op}^2 \left\| \mathbf{E}\frac{1}{T}\sum_{t=1}^{T} X_t X_t^\top \right\|^2}{\lambda_{\min}\left(\mathbf{E}\frac{1}{T}\sum_{t=1}^{T} X_t X_t^\top\right)^3}$$

System theoretic constants

Then with probability at least $1-\delta$,

$$\left\| \hat{A} - A^\star \right\|_{op}^2 \leq 32 \frac{\sigma^2 (d_X^2 + d_X \log \frac{1}{\delta})}{T \lambda_{\min}\left(\mathbf{E}\frac{1}{T}\sum_{t=1}^{T} X_t X_t^\top\right)}$$

Recap

- Basic concentration inequalities (Markov and Chernoff Bounds)
- Sub-Gaussian random variables
- Hanson-Wright Inequality for concentration of quadratic random variables
- ε-nets and covering arguments
- Self-normalized martingales
- The sample complexity of stochastic system identification

Next up: build on these tools to surpass some limitations of this analysis

Thank you!