Suppose $\mathcal{X}$ is uncountable. Can we still bound

Choose $x_1, \dots, x_T$ measure $y_t = \langle x_t, \theta_* \rangle + \eta_t$

(G-optimal) $\qquad\qquad$ $\uparrow$ 1-sub-Gaussian

$$\hat\theta = \Big(\sum_t x_t x_t^T\Big)^{-1} \sum_t x_t y_t$$

$$= \theta_* + \Big(\sum_t x_t x_t^T\Big)^{-1} \sum_t x_t \eta_t \qquad\qquad A = \sum_t x_t x_t^T$$

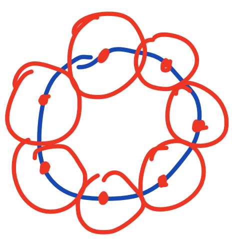
$$\max_{x\in\mathcal{X}} |\langle \hat\theta - \theta_*, x\rangle| = \max_x |(\hat\theta-\theta_*)^T A^{1/2} A^{-1/2} x|$$

$$\le \max_x \|A^{1/2}(\hat\theta-\theta_*)\|_2 \cdot \|A^{-1/2}x\|_2$$

$$= \|\hat\theta-\theta_*\|_A \cdot \underbrace{\max_x \|x\|_{A^{-1}}}_{\le\sqrt{d/T}} \lesssim \sqrt{d+2\log(1/\delta)}\cdot\sqrt{d/T} \approx \sqrt{\frac{d^2}{T}}$$

$$\|\hat\theta-\theta_*\|_A = \|A^{1/2}(\hat\theta-\theta_*)\|_2 = \sup_{u:\|u\|_2\le 1} u^T A^{1/2}(\hat\theta-\theta_*)$$

Lemma $\exists$ cover $C \subset \mathbb{R}^d$: $|C| \le (3/\varepsilon)^d$ and $x \in C$, $\|x\|_2 = 1$ and $\forall x$ w/ $\|x\|_2 = 1$ we have $\|y-x\|_2 \le \varepsilon$ for some $y \in C$.



In other words $\displaystyle\bigcup_{x\in C} B(x,\varepsilon) \supset S^{d-1}$

$\uparrow$ sphere in d-dim

$$\|\hat\theta-\theta_*\|_A = \sup_{u:\|u\|_2\le 1} u^T A^{1/2}(\hat\theta-\theta_*)$$

$$= \sup_u \min_{y\in C} (u-y)^T A^{1/2}(\hat\theta-\theta_*) + y^T A^{1/2}(\hat\theta-\theta_*)$$

$$\le \sup_u \underbrace{\min_{y\in C} \|u-y\|_2}_{\le\varepsilon} \|\hat\theta-\theta_*\|_A + \underbrace{y^T A^{1/2}(\hat\theta-\theta_*)}_{\le\sqrt{2\log(\frac{|C|}{\delta})}}$$

$$\leq \varepsilon \|\hat{\theta}-\theta_*\|_A + \sqrt{2\log\left(\frac{|C|}{\delta}\right)}$$

$$\Longrightarrow \|\hat{\theta}-\theta_*\|_A \leq \frac{1}{1-\varepsilon}\cdot\sqrt{2\log\left(\frac{|C|}{\delta}\right)} \leq \frac{1}{1-\varepsilon}\cdot\sqrt{2d\log(3/\varepsilon)+2\log(1/\delta)}.$$

$$\mathbb{E}\left[\left(y^T A^{1/2}(\hat{\theta}-\theta_*)\right)^2\right] = \mathbb{E}\left[y^T A^{1/2} A^{-1}\left(\sum_t x_t \eta_t\right)\left(\sum_s x_s \eta_s\right)^T A^{-1} A^{1/2} y\right]$$

$$= y^T A^{-1/2} \underbrace{\mathbb{E}\left[\sum_t x_t x_t^T \eta_t^2\right]}_{=A} A^{-1/2} y$$

$$= y^T y$$

$$= 1$$

$\mathbb{E}[\eta_t^2]=1$

$\left\| \mathbb{E}\left[\sum_t x_t x_t^T \eta_t^2\right]^{1/2} A^{-1/2} y \right\|_2$

$\leq$

Stochastic Processes

Flip two coins

Sample/outcome space $\Omega = \{HH, HT, TH, TT\}$

σ-algebra of a set Ω called $\mathcal{F}$ satisfies

1. $\Omega \in \mathcal{F}$,
2. $F \in \mathcal{F} \Rightarrow F^c \in \mathcal{F}$
3. $F_n \in \mathcal{F}\ \forall n$ then $\bigcup_n F_n \in \mathcal{F}$.

Probability measure $\mathbb{P}$ assign numbers to $\mathcal{F}$ $(\mathbb{P}: \mathcal{F} \to \mathbb{R})$

$\mathbb{P}(F) \geq 0 \quad \forall F \in \mathcal{F}$

$\mathbb{P}(\emptyset) = 0 \qquad \mathbb{P}(\Omega) = 1,$

F_n are disjoint $\forall n$

$\mathbb{P}\left(\bigcup_n F_n\right) = \sum_n \mathbb{P}(F_n)$

Random variable X is a function mapping
$\Omega \to \mathbb{R}$.

Conditional expectation

$$\mathbb{P}(Y|X) = \frac{\mathbb{P}(Y,X)}{\mathbb{P}(X)}$$

(discrete)

If X is finite then

$$\mathbb{E}[Y|X=x] = \sum_y y\,\mathbb{P}(Y=y|X=x]$$

but otherwise

$$\mathbb{E}[Y|X](\omega) = g(\omega) \quad \text{where } g \text{ is any function}$$

satisfying

$$\int_{\omega \in H} g(\omega)\,d\mathbb{P}(\omega) = \int_{\omega \in H} Y(\omega)\,d\mathbb{P}(\omega)$$

for all $H \in \sigma(X\ \ )$

$\exists$ measurable function f: $\quad f(X(\omega)) = \mathbb{E}[Y|X](\omega)$

$X_1, X_2, X_3, \ldots$

$\mathcal{F}_t = \sigma(\{X_s\}_{s \le t})$. $\quad \mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3 \cdots$

Filtration: $\{\mathcal{F}_t\}_{t=1}^{\infty} = \mathbb{F}$

X_t is $\mathbb{F}$-adapted if X_t is $\mathcal{F}_t$ measurable.

$$\mathbb{E}[X_t \mid \mathcal{F}_t] = X_t$$

$$\mathbb{E}[X_{t+1}] = \mathbb{E}[\mathbb{E}[X_{t+1} \mid \mathcal{F}_t]] = \mathbb{E}[X_t] = \mathbb{E}[X_0]$$

An $\mathbb{F}$-adapted sequence of R.V.s is a <u>martingale</u> if $\mathbb{E}[X_{t+1} \mid \mathcal{F}_t] = X_t \; \forall t$, and $\mathbb{E}[|X_t|] < \infty$,

If $\mathbb{E}[X_{t+1} \mid \mathcal{F}_t] \le X_t$ we say its supermartingale

If $\mathbb{E}[X_{t+1} \mid \mathcal{F}_t] \ge X_t$ " sub-martingale.

Ex. $Z_t \sim \mathcal{N}(0,1)$, then $X_t = \sum_{s=1}^{t} Z_s$ is a martingale.

$$\mathbb{E}[X_{t+1} \mid \mathcal{F}_t] = \mathbb{E}\left[\sum_{s=1}^{t+1} Z_s \mid \mathcal{F}_t\right] = \mathbb{E}[Z_{t+1} \mid \mathcal{F}_t] + \sum_{s=1}^{t} Z_s$$

Ex. $Z_t = \begin{cases} 2 & \text{w.p. } 1/2 \\ 0 & \text{w.p. } 1/2 \end{cases}$, $X_t = \prod_{s=1}^{t} Z_s$, $X_0 = 1$ $= 0 + X_t$

Ex. $Z_t \sim N(0,1)$, $X_t = \prod_{s=1}^{t} \exp(\lambda Z_s)$, $\lambda > 0$

$$\mathbb{E}[X_{t+1} \mid \mathcal{F}_t] = \prod_{s=1}^{t} \exp(\lambda Z_s)\, \mathbb{E}\underbrace{[\exp(\lambda Z_{t+1}) \mid \mathcal{F}_t]}_{e^{\lambda^2/2}}$$

$X_0 = 1$

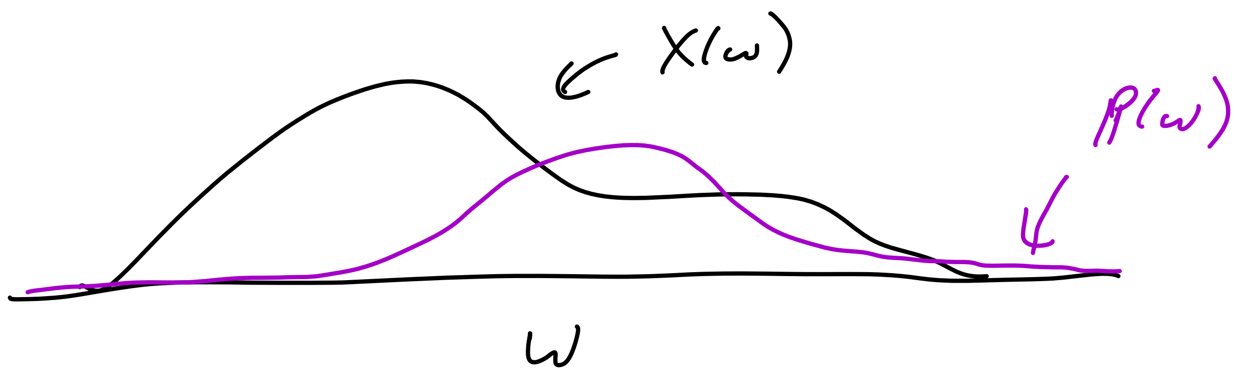
$$= X_t \cdot e^{\lambda^2/2}$$

$$\geq X_t \implies X_t \text{ is sub-martingale}$$

$$X_t = \prod_{s=1}^{t} \exp(\lambda Z_s - \lambda^2/2) \implies X_t \text{ is super-martingale}$$

$X(\omega)$ $\omega \in \Omega$

$$\mathbb{E}[X] = \sum_{\omega} X(\omega)\, \mathbb{P}(\omega)$$



Def A random variable $\tau \subset \mathbb{N}$ is a stopping time if $\mathbb{1}\{\tau \le t\}$ is $\mathcal{F}_t$ measurable $\forall t$.

Ex. $S_t = \sum_{s=1}^{t} Z_s$ $Z_s \sim N(0,1)$ then

$\tau = \min\{t : S_t \ge \varepsilon\}$ is a stopping time

Ex. $\tau = \max\{t : S_t \ge \varepsilon\}$ is <u>not</u> a stopping time.

Lemma Doob's Optional Stopping. Let τ be a stopping time. and X_t be $\mathbb{F}$-adapted martingale. If

- $\exists N : \mathbb{P}(\tau \le N) = 1$, or
- $\mathbb{E}[\tau] < \infty$ and $\mathbb{E}[|X_{t+1} - X_t|] \le c$ $\forall t$ some c

Then X_τ is well-defined and $\mathbb{E}[X_\tau] = \mathbb{E}[X_0]$.

Furthermore if X_t is supermartingale $\mathbb{E}[X_\tau] \le \mathbb{E}[X_0]$.

" " sub-martingale $\mathbb{E}[X_\tau] \ge \mathbb{E}[X_0]$.

Lemma Maximal inequality. Let X_t be $\mathbb{F}$-adapted sequence w/ $X_t \ge 0$ $\forall t$ a.s. Then for any $\varepsilon > 0$

- $\mathbb{P}\left(\max_{t \in \mathbb{N}} X_t \ge \varepsilon\right) \le \frac{\mathbb{E}[X_0]}{\varepsilon}$ if X_t is supermartingale
- $\mathbb{P}\left(\max_{t \in [1,\dots,n]} X_t \ge \varepsilon\right) \le \frac{\mathbb{E}[X_n]}{\varepsilon}$ if X_t is sub-martingale

Let Z_t be IID mean-zero 1-sub-Gaussian R.V. so that

$$\mathbb{E}[\exp(\lambda Z_t)] \le e^{\lambda^2/2}. \qquad S_t = \sum_{s=1}^{t} Z_s$$

$$M_t(\lambda) = \exp(\lambda S_t - t\lambda^2/2).$$

$M_t(\lambda)$ is supermartingale.

$$\mathbb{E}[M_{t+1}(\lambda) \mid \mathcal{F}_t] = \mathbb{E}[\exp(\lambda S_{t+1} - (t+1)\lambda^2/2) \mid \mathcal{F}_t]$$

$$= \exp(\lambda S_t - t\lambda^2/2)\, \underbrace{\mathbb{E}[e^{\lambda Z_{t+1} - \lambda^2/2} \mid \mathcal{F}_t]}_{\le 1}$$

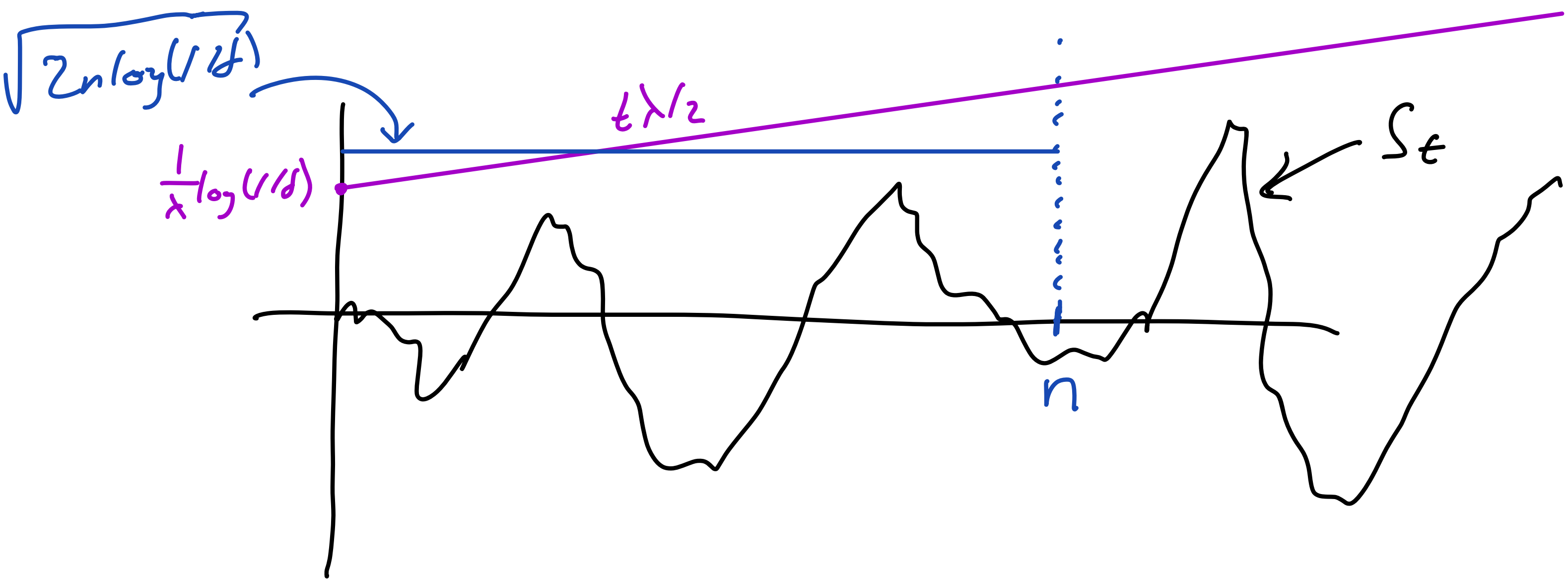
$$\le M_t(\lambda)$$

$$M_0(\lambda) := 1$$

By maximal inequality we have

$$\mathbb{P}\Big(\underbrace{\max_{t \in \mathbb{N}} \exp(\lambda S_t - t\lambda^2/2) \ge \frac{1}{\delta}}\Big) \le \frac{\mathbb{E}[M_0(\lambda)]}{1/\delta} = \delta$$

$$\Longrightarrow \mathbb{P}\Big(\exists t : S_t \ge t\lambda/2 + \frac{1}{\lambda}\log(1/\delta)\Big) \le \delta.$$



Fix $n \in \mathbb{N}$ set $\lambda = \sqrt{\dfrac{2\log(1/\delta)}{n}}$

$$\Rightarrow \mathbb{P}\left(\exists t: S_t \geq (t/\sqrt{n} + \sqrt{n})\sqrt{\log(1/\delta)/2}\right) \leq \delta$$

$$\Rightarrow \mathbb{P}\left(\max_{t=1,\dots,n} : S_t \geq \sqrt{2n\log(1/\delta)}\right) \leq \delta$$

Chernoff Bound says $\mathbb{P}\left(S_n \geq \sqrt{2n\log(1/\delta)}\right) \leq \delta$

Method of Mixtures

Let $h(\lambda)$ be a probability distribution over λ

Define $\overline{M}_t = \mathbb{E}_\lambda[M_t(\lambda)] = \int_\lambda M_t(\lambda)h(\lambda)d\lambda$

$\overline{M}_t$ is a supermartingale if $M_t(\lambda)$ is a superm.

$$\begin{aligned}
\mathbb{E}[\overline{M}_{t+1} \mid \mathcal{F}_t] &= \mathbb{E}\left[\int_\lambda M_{t+1}(\lambda)h(\lambda)d\lambda \,\middle|\, \mathcal{F}_t\right] \\
&\leq \int_\lambda \mathbb{E}[M_{t+1}(\lambda) \mid \mathcal{F}_t]\, h(\lambda)d\lambda \\
&= \int_\lambda M_t(\lambda)h(\lambda)d\lambda \\
&= \overline{M}_t
\end{aligned}$$

S_t is defined as above (sum 1-sub-Gaussian)

and define $h(\lambda) = \frac{1}{\sqrt{2\pi\nu^2}} e^{-\frac{\lambda^2}{2\nu^2}}$

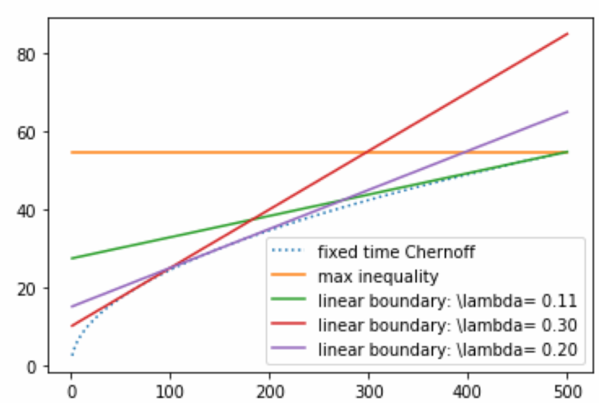
$$\bar{M}_t \doteq \int M_t(\lambda)\, h(\lambda)\, d\lambda$$

$$= \int \exp(\lambda S_t - \lambda^2/2) \frac{1}{\sqrt{2\pi\nu^2}} \exp(-\lambda^2/2\nu^2)\, d\lambda$$

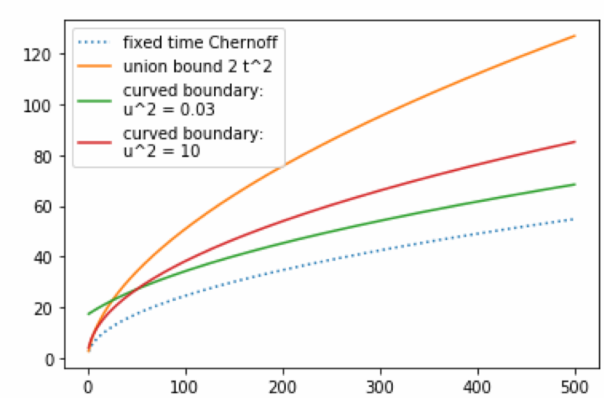
$$= \sqrt{\frac{\nu^{-2}}{t+\nu^{-2}}} \cdot \exp\left(S_t^2 (t+\nu^{-2})^{-1} / 2 \right).$$

$$\mathbb{P}\left(\exists t : |S_t| \geq \sqrt{2(t+\nu^{-2})\left(\log(1/\delta) + \frac{1}{2}\log\left(\frac{t+\nu^{-2}}{\nu^{-2}}\right)\right)}\right)$$

$$= \mathbb{P}\left(\exists t : \bar{M}_t \geq \frac{1}{\delta}\right) \leq \delta.$$



(a) Fix $\delta = 0.05$. The 'fixed time Chernoff' represents $\sqrt{2t\log(1/\delta)}$ which holds at each t but not all $t \leq 500$ simultaneously (which is why it is dotted). The 'max inequality' holds for all $t \leq 500$, and the linear boundaries hold for all $t \in \mathbb{N}$ simultaneously.



(b) Fix $\delta = 0.05$. The 'fixed time Chernoff' represents $\sqrt{2t\log(1/\delta)}$ which holds at each t but not all $t \in \mathbb{N}$ simultaneously (which is why it is dotted). All other curves do hold for all $t \in \mathbb{N}$ simultaneously. "union bound $2t^2$" plots $\sqrt{2\log(2t^2/\delta)}$.

$Z_1, Z_2, \ldots$ is an $\mathbb{F}$-adapted sequence. We say

σ_t is <u>predictable</u> if σ_t is $\mathcal{F}_{t-1}$-measurable.

—

Suppose for any λ we have $\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] = 0$

$$\mathbb{E}[\exp(\lambda Z_t) \mid \mathcal{F}_{t-1}] \leq \exp(\lambda^2 \sigma_t^2 / 2)$$

and define $S_t = \sum_{s=1}^{t} Z_s$. Then $V_t = \sum_{s=1}^{t} \sigma_s^2$

$$M_t(\lambda) = \exp(\lambda S_t - \lambda^2 V_t / 2)$$

is a super martingale.

$$\mathbb{P}\left(\exists t : S_t \geq \lambda V_t / 2 + \frac{1}{\lambda} \log(1/\delta)\right) \leq \delta$$

Define $\bar{M}_t = \mathbb{E}_{\lambda \sim h}[M_t(\lambda)]$ $\quad h(\lambda) = \frac{1}{\sqrt{2\pi}} e^{-\lambda^2/2}$

$$\mathbb{P}\left(\exists t : |S_t| \geq \sqrt{(V_t + 1) \log\left(\frac{V_t + 1}{\delta}\right)}\right) \leq \delta.$$

Ex. Gradient Descent Analysis.

$$w_{t+1} = w_t - \eta \nabla \ell_t(w_t) \qquad \mathbb{E}[\ell_t(w)] = \ell(w).$$

$$\mathbb{E}[f(w_t) - f(w_*)] \leq \frac{RG}{\sqrt{t}} \quad \text{where} \quad R = \|w_1 - w_*\|_2, \quad G = \|\nabla \ell(w)\|_2$$

$Z_t = \ell(w_t) - \mathbb{E}[\ell(w_t) \mid \mathcal{F}_{t-1}]$ is a martingale.

$$Z_t = \nabla \ell_t(w_t)(w_t - w_*) - \nabla \ell(w_t)(w_t - w_*)$$

$$\|\lambda\|_{\Sigma_t}^2 = \lambda^T \Sigma_t \lambda$$

Suppose that $Z_1, Z_2, \ldots \in \mathbb{R}^d$ is $\mathbb{F}$-adapted sequence satisfying $\mathbb{E}[\exp(\langle \lambda, Z_t \rangle) | \mathcal{F}_{t-1}] \leq \exp(\|\lambda\|_{\Sigma_t}^2 / 2)$ for all $\lambda \in \mathbb{R}^d$, where $\Sigma_t \in \mathbb{R}^{d \times d}$ is a predictable sequence.

$$S_t = \sum_{s=1}^{t} Z_s$$

$$V_t = \sum_{s=1}^{t} \Sigma_s$$

$$M_t(\lambda) = \exp(\langle \lambda, S_t \rangle - \|\lambda\|_{V_t}^2 / 2)$$

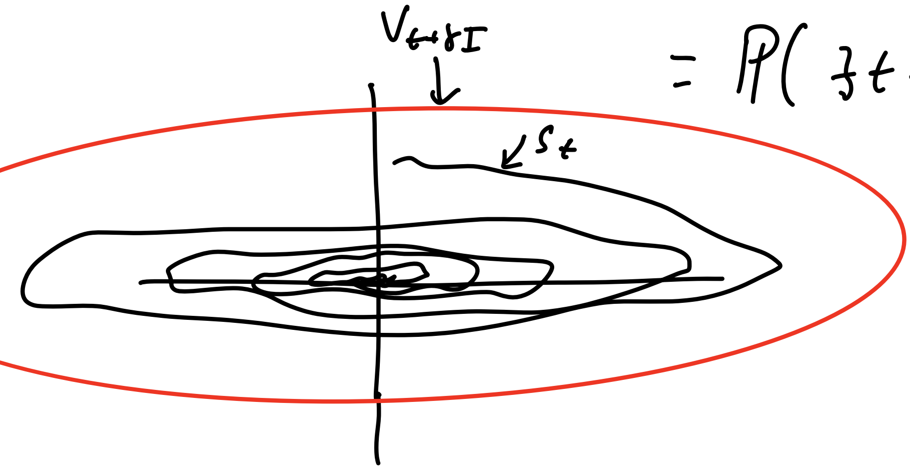
$$\mathbb{P}(\exists t : \langle \lambda, S_t \rangle - \|\lambda\|_{V_t}^2 / 2 \geq \log(1/\delta)) \leq \delta$$

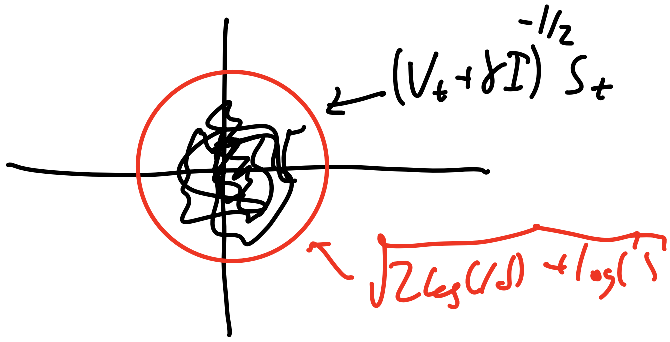
$$\bar{M}_t = \mathbb{E}_{\lambda \sim h}[M_t(\lambda)] \qquad h(\lambda) = \frac{1}{(2\pi/\gamma)^{d/2}} \exp(-\|\lambda\|^2 \gamma / 2)$$

$$h(\lambda) \equiv \mathcal{N}(0, \tfrac{1}{\gamma} I_d)$$

$$\mathbb{P}\left(\exists t : \|S_t\|_{(V_t + \gamma I)^{-1}} \geq \sqrt{2\log(1/\delta) + \log\left(\frac{|V_t + \gamma I|}{\gamma^d}\right)}\right)$$

$$= \mathbb{P}(\exists t : \bar{M}_t \geq \tfrac{1}{\delta}) \leq \delta$$





Online Linear Regression.

for $t = 1, 2, \ldots$

Observe $x_t \in \mathbb{R}^d$

Player predicts $\hat{y}_t = \langle x_t, \hat{\theta}_t \rangle$

Observes $y_t = \langle x_t, \theta_* \rangle + \eta_t$ $\leftarrow$ η_t mean-zero sub-Gaussian.

$$\hat{\theta}_{t+1} = \underset{\theta}{\operatorname{argmin}} \sum_{s=1}^{t} (y_s - \langle \theta, x_s \rangle)^2 + \gamma \|\theta\|_2^2$$

$$= \Big(\underbrace{\sum_{s=1}^{t} x_s x_s^T}_{V_t} + \gamma I \Big)^{-1} \sum_{s=1}^{t} x_s y_s$$

$$S_t = \sum_{s=1}^{t} x_s \eta_s$$

$$= (V_t + \gamma I)^{-1} V_t \theta_* + (V_t + \gamma I)^{-1} S_t$$

x_t is $\mathcal{F}_{t-1}$ measurable

from above

$$Z_s \equiv x_s \eta_s \qquad \mathbb{E}[\exp(\langle \lambda, x_s \eta_s \rangle) \mid \mathcal{F}_{s-1}] = \mathbb{E}[\exp(\langle \lambda, x_s \rangle \eta_s) \mid \mathcal{F}_{s-1}]$$

$$\Sigma_s \equiv x_s x_s^T \qquad \leq \exp(\langle \lambda, x_s \rangle^2 / 2)$$

$$= \exp(\lambda^T x_s x_s^T \lambda / 2)$$

$$= \exp(\|\lambda\|^2_{x_s x_s^T} / 2)$$

$$\|\hat{\theta}_{t+1} - \theta_*\|_{(V_t + \gamma I)} = \|S_t - \gamma \theta_*\|_{(V_t + \gamma I)^{-1}}$$

$$\leq \|S_t\|_{(V_t + \gamma)^{-1}} + \gamma \|\theta_*\|_{(V_t + \gamma I)^{-1}}$$

$$\leq \|S_t\|_{(V_t+\gamma)^{-1}} + \sqrt{\gamma}\,\|\theta_*\|_2$$

$$\mathbb{P}\left(\exists t : \|\hat{\theta}_{t+1} - \theta_*\|_{(V_t+\gamma I)} \geq \sqrt{\gamma}\,\|\theta_*\|_2 + \sqrt{2\log(1/\delta) + \log\left(\frac{|V_t+\gamma I|}{\gamma^d}\right)}\right) \leq \delta$$

UCB:

$$x_t = \underset{x \in \mathcal{X}}{\arg\max}\ UCB_t(x)$$

$$UCB_t(x) = \underset{\theta \in C_t}{\arg\max}\ \langle \theta, x \rangle$$

$$C_t = \{\theta : \|\theta - \hat{\theta}_t\|_{(V_t+\gamma I)} \leq \beta_t\}$$