X_t = the sum of the dice at time t

$Z_t = \sum_{s=1}^{t} X_s$

$\mathbb{P}(Z_t > 16 \mid \mathcal{F}_0)$ "observed nothing yet"

$\mathbb{P}(Z_t > 16 \mid \mathcal{F}_1)$ "observed first roll only"

$\mathbb{P}(Z_t > 16 \mid \mathcal{F}_t) = \mathbb{1}\{Z_t > 16\}$

$\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3 \subset \dots$

R.V. $X : \Omega \to \mathbb{R}$

$\mathbb{P}(X \in A) = \mathbb{P}(\omega : X^{-1}(A) \ni \omega)$

Let $X_1, X_2, \dots$ be a sequence of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ where $\mathbb{F} = \{\mathcal{F}_t\}_{t=1}^n$ is a filtration of $\mathcal{F}$. We say the sequence $\{X_t\}_{t=1}^n$ is $\mathbb{F}$-adapted if X_t is $\mathcal{F}_t$ measurable for all $1 \le t \le n$.

Definition 2. *An $\mathbb{F}$-adapted sequence of random variables is an $\mathbb{F}$-adapted* martingale *if $\mathbb{E}[X_{t+1}|\mathcal{F}_t] = X_t$ for all t and $\mathbb{E}[|X_t|] < \infty$. Furthermore, if*

- *X_t is a super-martingale if $\mathbb{E}[X_{t+1}|\mathcal{F}_t] \le X_t$*
- *X_t is a sub-martingale if $\mathbb{E}[X_{t+1}|\mathcal{F}_t] \ge X_t$*

Ex. $Y_t = \begin{cases} 1 & \text{w.p. } 1/2 \\ -1 & \text{w.p. } 1/2 \end{cases}$, $X_t = \sum_{s=1}^{t} Y_s$

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

$= \sum_{s=1}^{t} Y_s = X_t$

Definition 3. *Let $\mathbb{F} = \{\mathcal{F}_t\}_{t\in\mathbb{N}}$ be a filtration. A random variable $\tau \in \mathbb{N}$ is a* stopping time *with respect to $\mathbb{F}$ with values in $\mathbb{N} \cup \{\infty\}$ if $\mathbf{1}\{\tau \le t\}$ is $\mathcal{F}_t$ measurable for all $t \in \mathbb{N}$.*

Example 1. *Let $Z_1, Z_2, \dots$ be an $\mathbb{F}$ adapted sequence and define $S_t = \sum_{i=1}^{t} Z_i$. A valid stopping time may be $\tau = \min\{t \in \mathbb{N} : S_t \ge \epsilon\}$ because τ is $\mathcal{F}_t$ measurable: given S_t we can determine whether it is greater than or equal to ϵ or not. An example of a time that is* not *a stopping time is $\tau' = \max\{t \in \mathbb{N} : S_t \ge \epsilon\}$ because given only $\mathcal{F}_t$, the information up to time t, we do not know whether $S_{t'}$ will exceed ϵ again at some future time $t' > t$. Thus, $\mathbf{1}\{\tau' = t\}$ is not meaurable with respect to $\mathcal{F}_t$ and thus, is not a stopping time.*

Lemma 7 (Doob's optional stopping)**.** *Let $\mathbb{F} = \{\mathcal{F}_t\}_{t\in\mathbb{N}}$ be a filtration and $\{X_t\}_t$ be an $\mathbb{F}$-adapted martingale and τ be an $\mathbb{F}$-stopping time. If either of the following two events holds*

- *$\exists N \in \mathbb{N}$ such that $\mathbb{P}(\tau \le N) = 1$, or*
- *$\mathbb{E}[\tau] < \infty$ and $\mathbb{E}[|X_{t+1} - X_t| \,|\mathcal{F}_t] < c$ for all $t < \tau$ for some $c > 0$,*

then X_τ is well-defined and $\mathbb{E}[X_\tau] = \mathbb{E}[X_0]$. Furthermore, if

- *X_t is a super-martingale then $\mathbb{E}[X_\tau] \le \mathbb{E}[X_0]$*
- *X_t is a sub-martingale then $\mathbb{E}[X_\tau] \ge \mathbb{E}[X_0]$*

Lemma 9 (Maximal inequality). *Let $\{X_t\}_t$ be an $\mathbb{F}$-adapted sequence of random variables with $X_t \geq 0$ almost surely. Then for any $\epsilon > 0$, if*

- *X_t is a super-martingale then $\mathbb{P}(\max_{t \in \mathbb{N}} X_t \geq \epsilon) \leq \mathbb{E}[X_0]/\epsilon$*
- *X_t is a sub-martingale then $\mathbb{P}(\max_{t \in \{1,\ldots,n\}} X_t \geq \epsilon) \leq \mathbb{E}[X_n]/\epsilon$*

$\tau = \min\{t : X_t \geq \varepsilon\}$

$$\underbrace{\mathbb{P}(X_\tau \geq \varepsilon) \leq \frac{\mathbb{E}[X_\tau]}{\varepsilon}}_{\text{Markov}} \underbrace{\leq \frac{\mathbb{E}[X_0]}{\varepsilon}}_{\text{Doob Optional stopping}}$$

Fix $\lambda \geq 0$

Let $Z_1, Z_2, \ldots$ be iid Bernoulli R.V.s in $[-1,1]$ w/ $\mathbb{E}[Z_i] = 0$.

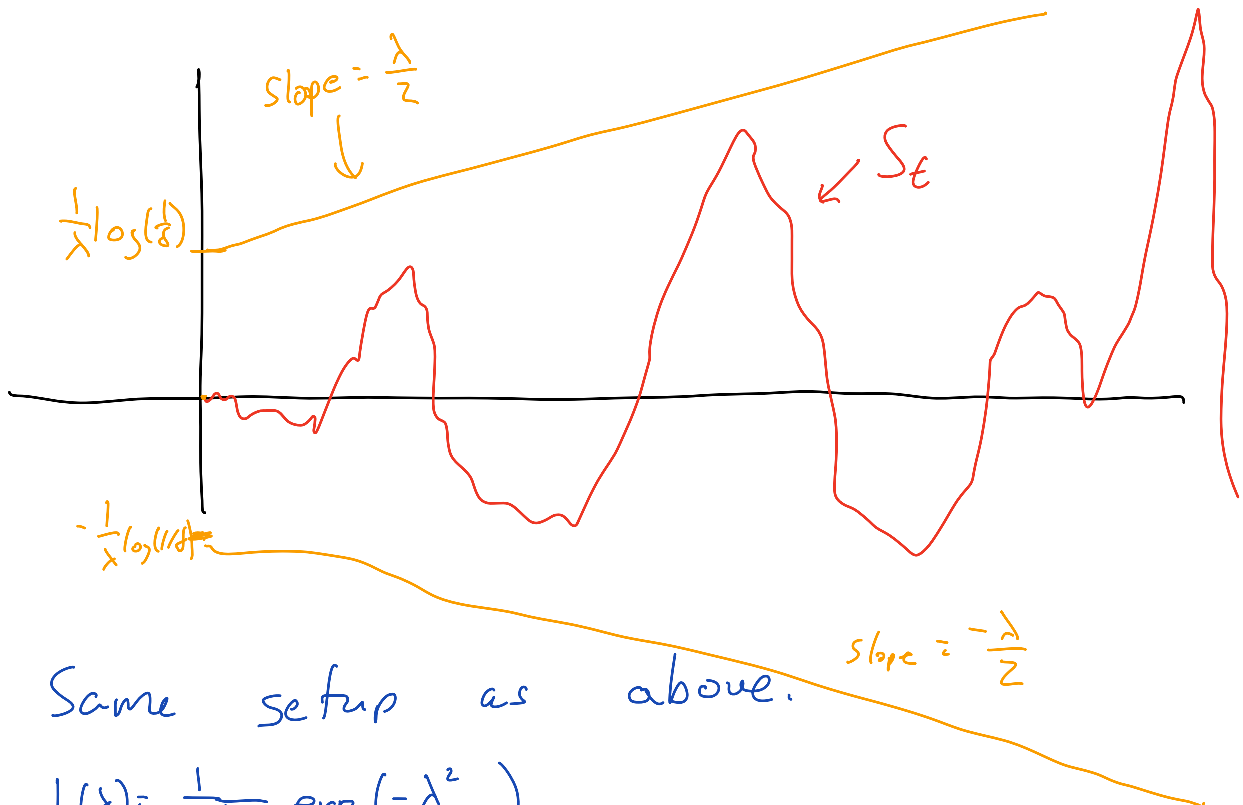
$$M_t(\lambda) := \exp(\lambda S_t - \lambda^2 t/2), \qquad S_t = \sum_{s=1}^{t} Z_s$$

$$\begin{aligned}
\mathbb{E}[M_{t+1}(\lambda) \mid \mathcal{F}_t] &= \mathbb{E}[\exp(\lambda S_{t+1} - \lambda^2(t+1)/2) \mid \mathcal{F}_t] \\
&= \mathbb{E}[\exp(\lambda S_t - \lambda^2 t/2)\exp(\lambda Z_{t+1} - \lambda^2/2) \mid \mathcal{F}_t] \\
&= \exp(\lambda S_t - \lambda^2 t/2)\, \mathbb{E}[\exp(\lambda Z_{t+1} - \lambda^2/2) \mid \mathcal{F}_t] \\
&\leq \exp(\lambda S_t - \lambda^2 t/2) = M_t(\lambda)
\end{aligned}$$

$\mathbb{E}[\exp(\lambda Z_{t+1})] \leq \exp(\lambda^2/2)$ by Hoeffding.

$$\begin{aligned}
&\mathbb{P}(\exists t \in \mathbb{N} : S_t \geq t\lambda/2 + \log(1/\delta)/\lambda) \\
&= \mathbb{P}(\exists t \in \mathbb{N} : \lambda S_t - t\lambda^2/2 \geq \log(1/\delta)) \\
&= \mathbb{P}(\exists t \in \mathbb{N} : \exp(\lambda S_t - t\lambda^2/2) \geq \tfrac{1}{\delta}) \\
&= \mathbb{P}(\exists t \in \mathbb{N} : M_t(\lambda) \geq \tfrac{1}{\delta}) \\
&\leq \frac{1}{1/\delta} = \delta
\end{aligned}$$

$M_0(\lambda) = \exp(\lambda \sum_{s=1}^{0} Z_s - 0)$
$= \exp(0) = 1$



Same setup as above.

$$h(\lambda) = \frac{1}{\sqrt{2\pi\nu^2}} \exp\left(-\frac{\lambda^2}{2\nu^2}\right)$$

$$\bar{M}_t = \mathbb{E}_{\lambda\sim h}\left[M_t(\lambda)\right] = \int_{\lambda=-\infty}^{\infty} M_t(\lambda) \frac{1}{\sqrt{2\pi\nu^2}} \exp\left(-\frac{\lambda^2}{2\nu^2}\right) d\lambda$$

$$\mathbb{E}\left[\bar{M}_{t+1} \mid \mathcal{F}_t\right] = \mathbb{E}\left[\int \text{—} \quad \text{—} \; d\lambda \mid \mathcal{F}_t\right]$$

$$= \int_\lambda \mathbb{E}\left[M_{t+1}(\lambda) \mid \mathcal{F}_t\right] \frac{1}{\sqrt{2\pi\nu^2}} \exp\left(-\frac{\lambda^2}{2\nu^2}\right) d\lambda$$

$$\leq \int_\lambda M_t(\lambda) \cdot \frac{1}{\sqrt{2\pi\nu^2}} \exp\left(\frac{-\lambda^2}{2\nu^2}\right) d\lambda$$

$$= \bar{M}_t$$

$\bar{M}_t$ is a super-martingale.

$$\Rightarrow \mathbb{P}(\exists t : \bar{M}_t > 1/\delta) \leq \delta.$$

$$\bar{M}_t = \int_\lambda M_t(\lambda) \frac{1}{\sqrt{2\pi\nu^2}} \exp\left(-\frac{\lambda^2}{2\nu^2}\right) d\lambda$$

$$= \int_\lambda \exp(\lambda S_t - t\lambda^2/2) \cdot \frac{1}{\sqrt{2\pi\nu^2}} \exp\left(\frac{-\lambda^2}{2\nu^2}\right) d\lambda$$

$\vdots$ math

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

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

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

Ex. At each time t 2 people show up and you give first person control and recieve $X_{t,1}$ w/ $\mathbb{E}[X_{t,1}] = \mu_1$, and second person treatment and recieve $X_{t,2}$ w/ $\mathbb{E}[X_{t,2}] = \mu_2$.

Question: Let $\Delta = \mu_2 - \mu_1$. Assuming $\Delta > 0$ when can I declare so? $(X_{t,i} \in \{0,1\})$

$$Z_t = X_{t,2} - X_{t,1} \qquad \mathbb{E}[Z_t] = \Delta$$

$$\mathbb{E}\left[\sum_{t=1}^{n} Z_t\right] = n\Delta \qquad Z_t \in \{-1, 0, 1\}$$

$$S_n = \sum_{t=1}^{n}(Z_t - \Delta) = -n\Delta + \sum_{t=1}^{n} Z_t$$

Test: Declare control arm 2 as "good"

when

Test

$$\sum_{t=1}^{n} Z_t > \sqrt{2(n+\nu^{-2})\left(\log(1/\delta) + \frac{1}{2}\log\left(\frac{n+\nu^{-2}}{\nu^{-2}}\right)\right)}$$

If $\Delta = 0$ then this never occurs w.p. $\geq 1-\delta$.

When does it happen?

$$\sum_{t=1}^{n} Z_t = n\Delta + S_n$$

$$\geq n\Delta - \sqrt{2(n+\nu^{-2})\left(\log(1/\delta) + \frac{1}{2}\log\left(\frac{n+\nu^{-2}}{\nu^{-2}}\right)\right)}$$

It happens before

is bigger than

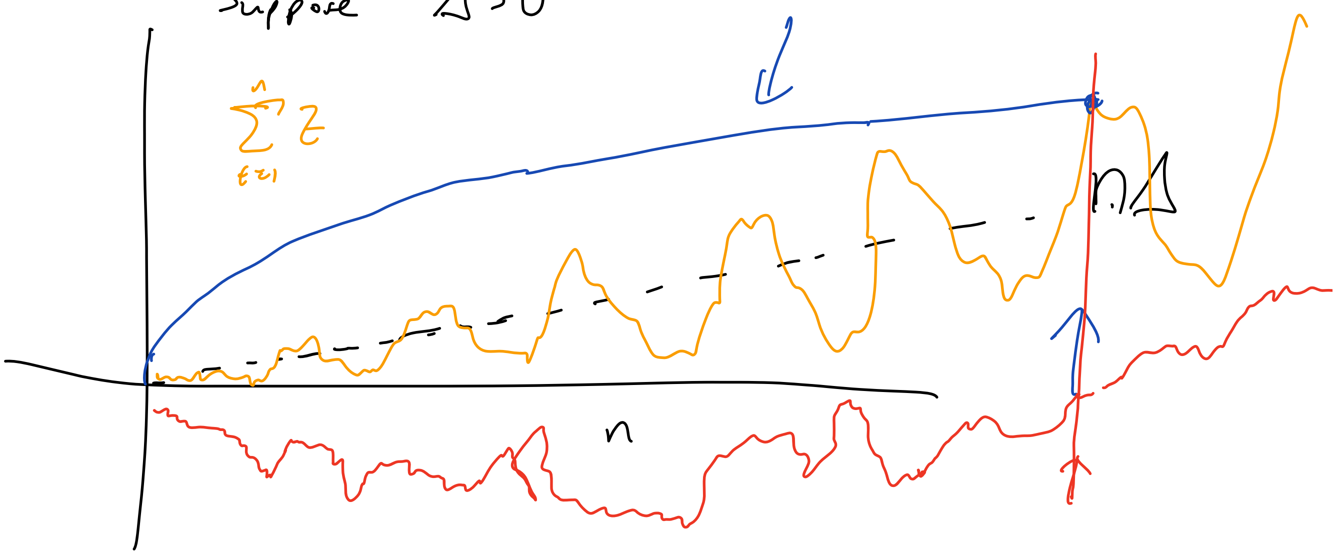
The value of n when this happens

$$\approx \left(\frac{\Delta^{-2}+\nu^{-2}}{\nu^{-2}}\right)\log\left(\left(\frac{\Delta^{-2}+\nu^{-2}}{\nu^{-2}}\right)/\delta\right)$$

(about)

$$\sqrt{2(n+\nu^{-2})\left(\log(1/\delta)+\frac{1}{2}\log\left(\frac{n+\nu^{-2}}{\nu^{-2}}\right)\right)}$$

Suppose $\Delta > 0$

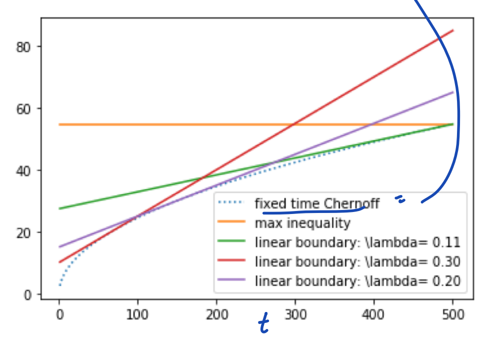


$$\sum_{t=1}^{n} z_t - \sqrt{\qquad} \geq 0$$

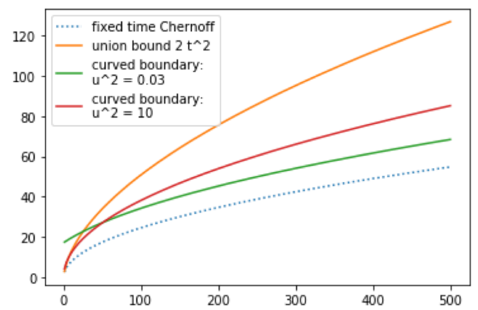
$\mathbb{P}(\exists t : S_t > \sqrt{2t\log(1/\delta)}) = 1$

Fix $t \in \mathbb{N}$.

$\mathbb{P}(S_t > \sqrt{2t\log(1/\delta)}) \leq \delta$



(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 $\sqrt{2\log(2t^2/\delta)}$.

(i.e. z_t is $\mathcal{F}_t$-meas.)

Let $z_1, z_2, \ldots$ be an $\mathcal{F}_t$ adapted sequence.

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

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

$$S_t = \sum_{s=1}^{t} z_s \qquad V_t = \sum_{s=1}^{t} \sigma_t^2$$

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

One can show $M_t(\lambda)$ is a super-martingale

$$\mathbb{P}(\exists t: S_t \geq \lambda V_t/2 + \log(1/\delta)/\lambda)$$
$$= \mathbb{P}(\exists t: M_t(\lambda) > 1/\delta) \leq \delta$$

Also define $\overline{M}_t$ as above to get

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

Vector-valued martingales.

$z_1, z_2, \ldots \in \mathbb{R}^d$ is $\mathcal{F}_t$ adapted w/ $\mathbb{E}[z_t \mid \mathcal{F}_{t-1}] = 0$

and let $\Sigma_t \in \mathbb{R}^{d \times d}$ be predictable sequence

s.t. for any $\lambda \in \mathbb{R}^d$

$$\mathbb{E}[\exp(\langle \lambda, z_t \rangle) \mid \mathcal{F}_{t-1}] \leq \exp(\|\lambda\|_{\Sigma_t}^2/2)$$

$S_t = \sum_{i=1}^{t} Z_i \qquad V_t = \sum_{i=1}^{n} \Sigma_i$

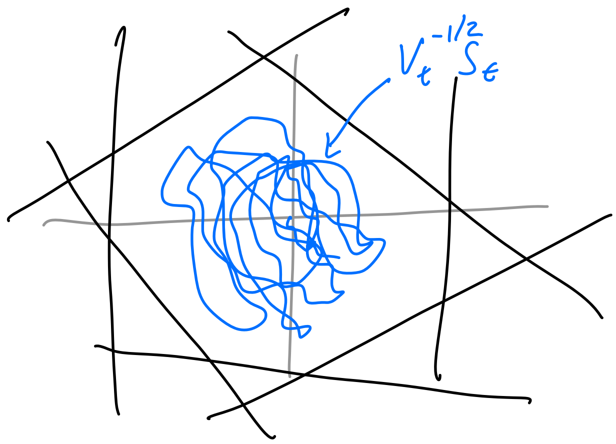
$$M_t(\lambda) = \exp\left(\langle \lambda, S_t \rangle - \|\lambda\|_{V_t}^2 / 2\right)$$

is a super martingale.

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

$$\arg\max_{\lambda} \langle \lambda, S_t \rangle - \|\lambda\|_{V_t}^2 / 2 = V_t^{-1} S_t$$

$d=2$



Define $h(\lambda) = \frac{1}{(2\pi/\gamma)^{d/2}} e^{-\gamma \|\lambda\|^2/2},$

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

$$\vdots$$

$$= \frac{|V_t + \gamma I|^{-1/2}}{\gamma^{-d/2}} \exp\left(\tfrac{1}{2} \|S_t\|^2_{(V_t + \gamma I)^{\gamma}}\right)$$

$$\mathbb{P}\left(\exists t \in \mathbb{N} : \underbrace{\|S_t\|_{(V_t+\gamma I)^{-1}}}_{= \|(V_t+\gamma I)^{-1/2} S_t\|_2} \geq \sqrt{2\log(1/\delta) + \log\left(\frac{|V_t+\gamma I|}{\gamma^d}\right)}\right) \leq \delta.$$

Application: Online linear regression

Let $x_1, x_2, \ldots \in \mathbb{R}^d$ be predictable sequence (x_t is $\mathcal{F}_{t-1}$ measurable) and let

$y_t \in \mathbb{R}$ be $\mathcal{F}_t$ adapted (y_t is $\mathcal{F}_t$ meas.)

And assume $y_t = \langle x_t, \theta_* \rangle + \eta_t$

where η_t is $\mathcal{F}_t$-meas and

$$\mathbb{E}[\exp(s\eta_t) \mid \mathcal{F}_{t-1}] \leq \exp(s^2/2).$$

In prev example, let

$$Z_t = x_t \eta_t \qquad S_t = \sum_{i=1}^{t} x_i \eta_i$$

$$\mathbb{E}[\exp(\langle \lambda, z_t \rangle) \mid \mathcal{F}_{t-1}]$$

$$= \mathbb{E}[\exp(\langle \lambda, x_t \eta_t \rangle) \mid \mathcal{F}_{t-1}]$$

$$= \mathbb{E}[\exp(\langle \lambda, x_t \rangle \eta_t) \mid \mathcal{F}_{t-1}]$$

$$\leq \exp(\langle \lambda, x_t \rangle^2 / 2)$$

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

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

$$\Sigma_t = x_t x_t^T \qquad V_t = \sum_i x_i x_i^T$$

Consider estimator (Ridge regression)

$$\hat{\theta}_t = \underset{\theta}{\arg\min} \sum_{i=1}^{t} (y_i - \langle \theta, x_i \rangle)^2 + \gamma \|\theta\|_2^2$$

$$= \left(\sum_{i=1}^{t} x_i x_i^T + \gamma I \right)^{-1} \sum_i x_i y_i$$

$$= \left(\sum_{i=1}^{t} x_i x_i^T + \gamma I \right)^{-1} \sum_i x_i \eta_i + \left(\sum_{i=1}^{t} x_i x_i^T + \gamma I \right)^{-1} \sum_i x_i x_i^T \theta_*$$

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

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

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

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

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

$$\leq \|S_t\|_{(V_t + \gamma I)^{-1}} + \gamma \underbrace{\|\theta_*\|_{(\gamma I)^{-1}}}_{= \left\| \frac{1}{\sqrt{\gamma}} \theta_* \right\|_2 = \frac{1}{\sqrt{\gamma}} \|\theta_*\|_2}$$

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

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

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

$$\leq \delta .$$

Chernoff: If η_t were indep

and we defined $\tilde{\theta}_t = V_t^{-1} S_t$

Fix any $z \in \mathbb{R}^d$.

$$\mathbb{P}\left(\langle z, \tilde{\theta} - \theta_* \rangle \geq \|z\|_{V_t^{-1}} \cdot \sqrt{2\log(1/\delta)} \right) \leq \delta .$$

Now fix $\mathcal{Z} \subset \mathbb{R}^d$

$$\mathbb{P}\left(\bigcup_{z \in \mathcal{Z}} \langle z, \tilde{\theta} - \theta_* \rangle \geq \|z\|_{V_t^{-1}} \cdot \sqrt{2\log(|\mathcal{Z}|/\delta)} \right) \leq \delta$$

Above we showed a bound on $\|\hat{\theta}_t - \theta_*\|_{(V_t+\gamma I)}$

For any z

$$\langle z, \hat{\theta}_t - \theta_* \rangle \leq \|z\|_{(V_t+\gamma I)^{-1}} \cdot \|\hat{\theta}_t - \theta_*\|_{(V_t+\gamma I)}$$

$$\leq \|z\|_{(V_t+\gamma I)^{-1}} \cdot \sqrt{2\log(1/\delta) + d\log(t/\gamma)}$$

<u>Lemma</u> Wald's identity. Let $z_1, z_2, \ldots$ be IID R.V.s w/ $\mathbb{E}[z_i] = \mu$. If τ is a stopping time w/ $\mathbb{E}[\tau] < \infty$ then

$$\mathbb{E}\left[\sum_{i=1}^{\tau} z_i\right] = \mu\, \mathbb{E}[\tau].$$

Hypothesis testing.

$X_1, X_2, \ldots$ R.V.s IID

$$H_0 : X_t \sim P_0$$

$$H_1 : X_t \sim P_1$$

Define
$$L_t = \prod_{i=1}^{t} \frac{P_1(X_i)}{P_0(X_i)}$$

$$\tau = \min\{t : L_t > 1/\delta\}$$

$$\log\left(\frac{1}{\delta}\right) \approx \mathbb{E}_1[\log(L_\tau)] = \mathbb{E}_1\left[\sum_{t=1}^{\tau} \log\left(\frac{P_1(X_t)}{P_0(X_t)}\right)\right] = \mathbb{E}_1[\tau]\,\mathbb{E}_1\left[\log\left(\frac{P_1(X_i)}{P_0(X_i)}\right)\right]$$
$$= \mathbb{E}_1[\tau]\, KL(P_1 \| P_0)$$

$$\mathbb{E}_1[\tau] \approx \frac{\log(1/\delta)}{KL(P_1 \| P_0)}.$$