Toss two coins, come up H or T

- Sample space $\Omega = \{TT, TH, HT, HH\}$
- Outcome $\omega \in \Omega$ $\omega = \{TT\}$
- σ-algebra called $\mathcal{F}$ is a collection of subsets of Ω s.t.
  1. $\Omega \in \mathcal{F}$
  2. If $F \in \mathcal{F}$ then $F^c \in \mathcal{F}$
  3. For a sequence F_n, $\bigcup_n F_n \in \mathcal{F}$

$$\mathcal{F} = \{ \emptyset, \Omega, TT, TH, HT, HH, \{TT, TH\}, \{TT, HT\}, \{ \} . \{TH, HT, HH\}, \ldots \{ \} \}.$$

- Probability measure for a σ-algebra $\mathcal{F}$ on Ω satisfies
  1. $\mathbb{P} : \mathcal{F} \to [0,1]$
  2. $\mathbb{P}(\emptyset) = 0$
  3. $\mathbb{P}(\Omega) = 1$
  4. For a seq $F_n \in \mathcal{F}$ w/ $F_i \cap F_j = \emptyset$, $\mathbb{P}(\bigcup_n F_n) = \sum_n \mathbb{P}(F_n)$
- A probability space is defined as tuple $(\Omega, \mathcal{F}, \mathbb{P})$
- Fix $(\Omega, \mathcal{F}, \mathbb{P})$. A function $g : \Omega \to \mathbb{R}$ is $\mathcal{F}$-measurable if for all Borel sets B we have $\{\omega : g(\omega) \in B\} \in \mathcal{F}$.
- Random variable X on $(\Omega, \mathcal{F}, \mathbb{P})$ is a function $X : \Omega \to \mathbb{R}$ that is $\mathcal{F}$-measurable.
- Let X_t be a set of RV.s indexed by $t \in T$. on space

$(\Omega, \mathcal{F}, \mathbb{P})$, we define $\sigma(\{X_t\}_{t\in T})$ as the natural σ-algebra for $\{X_t\}_{t\in T}$ to be the smallest σ-algebra for which $\{X_t\}_{t\in T}$ are measurable.

X be the R.V. counting # of heads

$$X(\omega) = \begin{cases} 0 & \text{if } \omega \in TT \\ 1 & \text{if } \omega \in \{TH, HT\} \\ 2 & \text{if } \omega \in \{HH\} \end{cases}$$

$\{\omega : X(\omega) = 0\} = \{TT\}$

$\{\omega : X(\omega) = 1\} = \{TH\}, \{H, T\}$

$\{\omega : X(\omega) = 2\} = \{HH\}$

$\{\omega : X(\omega) \notin \{0, 1, 2\}\} = \emptyset$

$\sigma(X) = \emptyset, TT, \{HT, TH\}, \{HH\}, \{TT, HT, TH\}$
$\{TT, HH\}, \{HT, TH, HH\}, \{TT, HT, TH, HH\}$

$$Y(\omega) = \begin{cases} 0 & \text{if } \omega = TT \\ 1 & \text{if } \omega \in \{HT, TH, HH \end{cases}$$

$\sigma(Y) = \emptyset, \{TT\}, \{HT, TH, HH\}, \Omega$

$\sigma(Y) \subset \sigma(X) \subset \mathcal{F}$

Assume X, Y are discrete

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

$$\mathbb{P}(\omega = HT \mid X = 1) = \frac{\mathbb{P}(\omega = HT, X = 1)}{\mathbb{P}(X = 1)} = \frac{1/4}{1/2} = \frac{1}{2}$$

$$\mathbb{E}[X \mid A] := \sum_x x \, \mathbb{P}(X = x \mid A)$$

(ex. $A = \{Y = 1\}$)

In continuous R.V. $\exists$ pathological space

s.t. $\nexists g$ s.t. $g(Y(\omega)) = \mathbb{E}[X \mid Y](\omega)$

Conditional Expectation. Let X be a R.V. on $(\Omega, \mathcal{F}, \mathbb{P})$ and let $\mathcal{H} \subseteq \mathcal{F}$. Then we say $\mathbb{E}[X \mid \mathcal{H}] : \mathcal{H} \to \mathbb{R}$ is the cond. expectation of X given $\mathcal{H}$ if

$$\int_{\omega \in H} \mathbb{E}[X \mid \mathcal{H}](\omega) \, d\mathbb{P}(\omega) = \int_{\omega \in H} X(\omega) \, d\mathbb{P}(\omega). \quad \forall H \in \mathcal{H}$$

$\mathbb{E}[X \mid \sigma(X)] = X$

$\mathbb{E}[X \mid \sigma(Y)] = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 = 1$

Filtrations

Let $X_1, X_2, X_3, \ldots$ be a sequence of R.V. defined on $(\Omega, \mathcal{F}, \mathbb{P})$. $\mathcal{F}_t = \sigma(\{X_s\}_{s \le t})$

We say $\mathbb{F} = \{\mathcal{F}_t\}_t$ is a filtration on $\mathcal{F}$ if $\{\Omega, \emptyset\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$

We say $\{X_t\}_{t=1}^n$ is $\mathbb{F}$ adapted if X_t is $\mathcal{F}_t$ measurable.

We call $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ a filtered probability space.

Martingales

An $\mathbb{F}$-adapted sequence of RV. $\{X_t\}_t$ is an $\mathbb{F}$-adapted martingale if $\mathbb{E}[X_{t+1} \mid \mathcal{F}_t] = X_t$ and $\mathbb{E}[|X_t|] < \infty$ $\forall t$.

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

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

- We say X_t is a supermartingale if $\mathbb{E}[X_{t+1} | \mathcal{F}_t] \leq X_t$
- We say X_t is a submartingale if $\mathbb{E}[X_{t+1} | \mathcal{F}_t] \geq X_t$

$$X_t = \exp\left(\lambda \sum_{s=1}^{t} Z_s - \lambda^2 t/2\right), \quad Z_s \sim \mathcal{N}(0,1)$$

$$\mathbb{E}[X_{t+1} | \mathcal{F}_t] = \mathbb{E}\left[\exp(\lambda Z_{t+1} - \lambda^2/2) \exp\left(\lambda \sum_{s=1}^{t} Z_s - \lambda^2 t/2\right) \Big| \mathcal{F}_t\right]$$

$$= \underbrace{\mathbb{E}[\exp(\lambda Z_{t+1} - \lambda^2/2) | \mathcal{F}_t]}_{\leq 1} \cdot \underbrace{\underbrace{\mathbb{E}\left[\exp\left(\lambda \sum_{s=1}^{t} Z_s - \lambda^2 t/2\right) \Big| \mathcal{F}_t\right]}_{= X_t}}_{X_t}$$

$$\leq X_t$$

A random variable $\tau \in \mathbb{N}$ is a <u>stopping time</u> wrt $\mathbb{F}$ if $\mathbb{1}\{\tau \leq t\}$ is $\mathcal{F}_t$ measurable. $\forall t$

$X_t = \sum_{s=1}^{t} Z_s$, $Z_s \sim \mathcal{N}(0,1)$, $\tau = \min\{t : X_t > \varepsilon\}$

valid stopping time

$\tau' = \max\{t : X_t > \varepsilon\}$

Doob's Optional Stopping Theorem. Let $\mathbb{F}$ be a filtration of $\{X_t\}_t$ martingale sequence and let τ be an $\mathbb{F}$-adapted stopping time. If

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

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

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

" " submartingale then $\mathbb{E}[X_\tau] \ge \mathbb{E}[X_0]$

Ex. Wald's identity. Let $Z_t = \begin{cases} 1 & \text{w.p. } p \\ 0 & \text{w.p. } 1-p \end{cases}$ and

$X_t = \sum_{s=1}^{t} Z_s$. $\tau = \min\{t : X_t \ge \varepsilon\}$. assume $\varepsilon \in \mathbb{N}$

$$M_t = X_t - tp$$

$$\mathbb{E}[M_{t+1} \mid \mathcal{F}_t] = \mathbb{E}[(Z_{t+1} - p) + X_t - tp \mid \mathcal{F}_t]$$
$$= X_t - tp = M_t$$

$$\mathbb{E}[X_\tau - p\tau] = \mathbb{E}[M_\tau] = \mathbb{E}[M_0] = 0$$

$$\Rightarrow \varepsilon = \mathbb{E}[X_\tau] = p \cdot \mathbb{E}[\tau]$$

$$\Rightarrow \quad \mathbb{E}[\tau] = \frac{\varepsilon}{p}$$

<u>Maximal inequality</u> Let X_t be an $\mathbb{F}$-adapted Seq. of R.V. w/ $X_t \geq 0$. Then for any $\varepsilon > 0$

- If X_t is a supermartingale $\mathbb{P}\left(\max_{t \in \mathbb{N}} X_t > \varepsilon\right) \leq \frac{\mathbb{E}[X_0]}{\varepsilon}$.
- If X_t is a submartingale $\mathbb{P}\left(\max_{t=1,\dots,n} X_t > \varepsilon\right) \leq \frac{\mathbb{E}[X_n]}{\varepsilon}$.

Let $Z_1, Z_2, \ldots \in \{-1,1\}$ be Bernoulli$(1/2)$. Define $S_t = \sum_{s \le t} Z_s$.

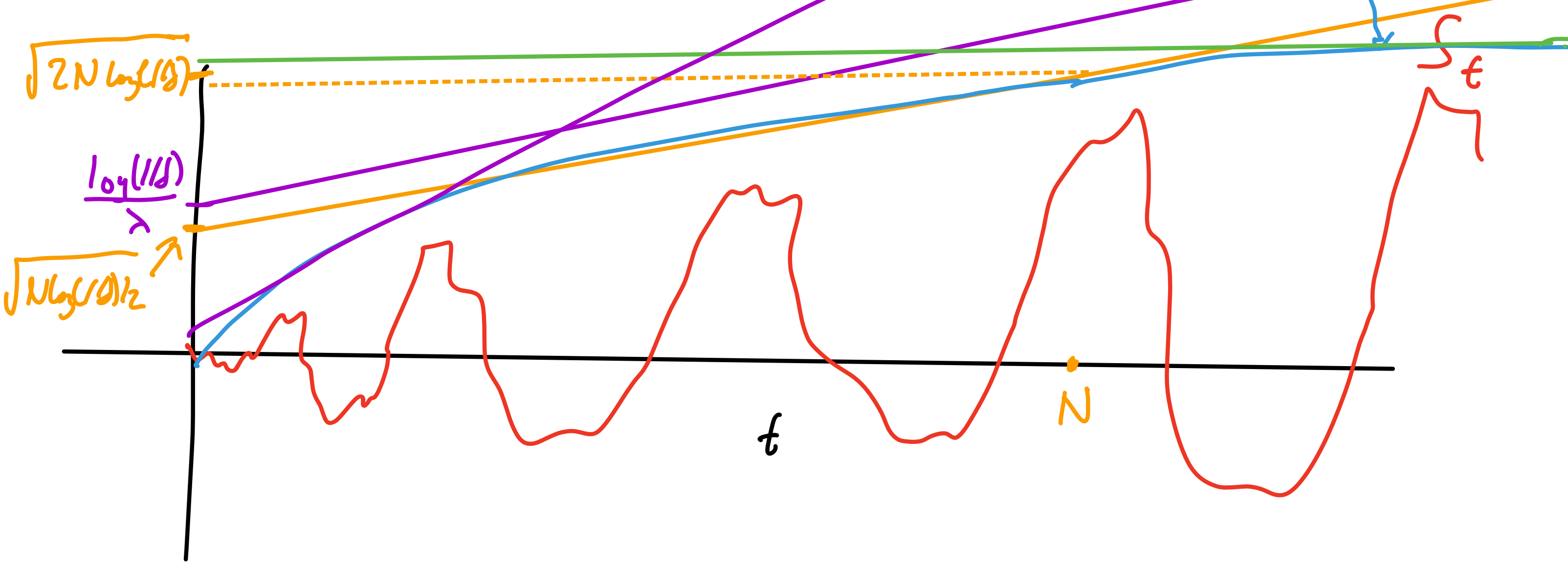
Fix any $\lambda \in \mathbb{R}$ and define

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

$\mathbb{E}[M_{t+1}(\lambda) \mid \mathcal{F}_t] \le M_t(\lambda)$

$$\mathbb{P}(\exists t : M_t(\lambda) \ge 1/\delta) \le \frac{\mathbb{E}[M_0(\lambda)]}{1/\delta} = \delta$$

$$\Rightarrow \quad \mathbb{P}\left(\exists t : S_t \ge \underbrace{t\lambda/2 + \frac{\log(1/\delta)}{\lambda}}\right) \le \delta$$



Pick $\lambda = \sqrt{\frac{2\log(1/\delta)}{N}}$ $\quad \mathbb{P}\left(\exists t : S_t \ge (t/\sqrt{N} + \sqrt{N})\sqrt{\log(1/\delta)/2}\right) \le \delta$

$$\Rightarrow \quad \mathbb{P}\left(\max_{t=1,\ldots,N} S_t \ge \sqrt{2N\log(1/\delta)}\right) \le \delta$$

Using Hoeffding $\quad \mathbb{P}\left(S_N \ge \sqrt{2N\log(1/\delta)}\right) \le \delta$

$$\mathbb{P}\left(\exists t \in \mathbb{N} : S_t \ge \sqrt{2t\log(2t^2/\delta)}\right) \le \sum_{t=1}^{\infty} \mathbb{P}(\cdot)$$

$$\le \sum_t \frac{\delta}{2t^2} \le \delta$$

Method of Mixtures.

Fix any prob. density $h(\lambda)$ w/ support $\mathbb{R}$

Define $\quad \overline{M}_t = \int_\lambda M_t(\lambda) h(\lambda) d\lambda$

$$\mathbb{E}[\overline{M}_{t+1} \mid \mathcal{F}_t] = \mathbb{E}\left[\int_\lambda M_{t+1}(\lambda) h(\lambda) d\lambda \mid \mathcal{F}_t \right]$$

$$= \int_\lambda \mathbb{E}[M_{t+1}(\lambda) \mid \mathcal{F}_t] \, h(\lambda) \, d\lambda$$

$$= \int_\lambda M_t(\lambda) \, h(\lambda) \, d\lambda$$

$$= \overline{M}_t$$

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

$$\overline{M}_t = \int \frac{1}{\sqrt{2\pi\nu^2}} e^{-\lambda^2/2\nu^2} e^{\lambda S_t - t\lambda^2/2} \, d\lambda$$

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

$$\mathbb{P}(\exists t : \overline{M}_t > 1/\delta) \le \delta$$

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

Convenient choice is $\nu = 1$

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

Interlude: Optimization

Ex. $L(w_t, I_t) = (w_t^T x_{I_t} - y_{I_t})^2$

SGD on a dataset.

L convex.

for $t = 1, 2, \ldots$

$\max_{i,w} \|\nabla L(w,i)\|_2 \leq G$

Draw $I_t \sim \text{uniform}([n])$

$\max_{w \in K} \|w\| \leq R$

$$w_{t+1} = \Pi_K\left(w_t - \gamma \nabla L(w_t, I_t)\right)$$

$$\ell(w) = \mathbb{E}_I[L(w, I)] = \frac{1}{n}\sum_{i=1}^{n} L(w,i)$$

$$\|w_{t+1} - w_*\|_2^2 \leq \|w_t - w_*\|_2^2 - \gamma \nabla L(w_t, I_t)^T(w_t - w_*) + \gamma^2 G^2$$

$$\Rightarrow \quad \nabla L(w_t, I_t)^T(w_t - w_*) \leq \frac{\|w_t - w_*\|_2^2 - \|w_{t+1} - w_*\|_2^2}{\gamma} + \gamma G^2$$

$$\ell(w_*) - \ell(w_t) \geq \nabla \ell(w_t)^T(w_* - w_t)$$

$\varepsilon_t := \nabla \ell(w_t) - \nabla L(w_t, I_t)$

$$\ell(w_t) - \ell(w_*) \leq \nabla \ell(w_t)^T(w_t - w_*)$$

$$= \nabla L(w_t, I_t)^T(w_t - w_*) + \varepsilon_t^T(w_t - w_*)$$

$$\sum_t \ell(w_t)-\ell(w_*) \leq \frac{\|w_1-w_*\|_2^2}{\gamma} + T\gamma G^2 + \sum_t \varepsilon_t(w_t-w_*)$$

$$\mathbb{E}\left[\sum_{s=1}^{t+1} \varepsilon_s^T(w_s-w_*) \,\middle|\, \mathcal{F}_t\right] = \mathbb{E}\left[\varepsilon_{t+1}^T(w_{t+1}-w_*) \mid \mathcal{F}_t\right] + \sum_{s=1}^{t} \varepsilon_s(w_s-w_*)$$

w_{t+1} is $\mathcal{F}_t$-measurable

$$= \sum_{s=1}^{t} \varepsilon_s^T(w_s-w_*)$$

Martingale!

$$\varepsilon_s^T(w_s-w_*) \leq \|\varepsilon_s\|_2 \cdot \|w_s-w_*\|_2 \leq 2GR$$

$$\Rightarrow \varepsilon_s^T(w_s-w_*) \mid \mathcal{F}_{s-1} \text{ is } (2GR)^2\text{-sub-Gaussian}$$

$$\mathbb{P}\left(\exists t \leq T: \sum_{s=1}^{t} \varepsilon_s^T(w_s-w_*) \geq (2GR)\sqrt{2T\log(1/\delta)}\right) \leq \delta$$

w.p. $\geq 1-\delta$ $$\sum_{t=1}^{T} \ell(w_t)-\ell(w_*) \leq \underbrace{\frac{4R^2}{\gamma} + T\gamma G^2}_{\lesssim GR\sqrt{T}} + 2GR\sqrt{2T\log(1/\delta)}$$

$$\lesssim c\, GR\sqrt{T\log(1/\delta)}$$

$$\frac{1}{T}\sum_{t=1}^{T} \ell(w_t) - \ell(w_*) \geq \ell(\overline{w}_T) - \ell(w_*)$$

$$\overline{w}_T = \frac{1}{T}\sum_{t=1}^{T} w_t$$

w.p. $\geq 1-\delta$

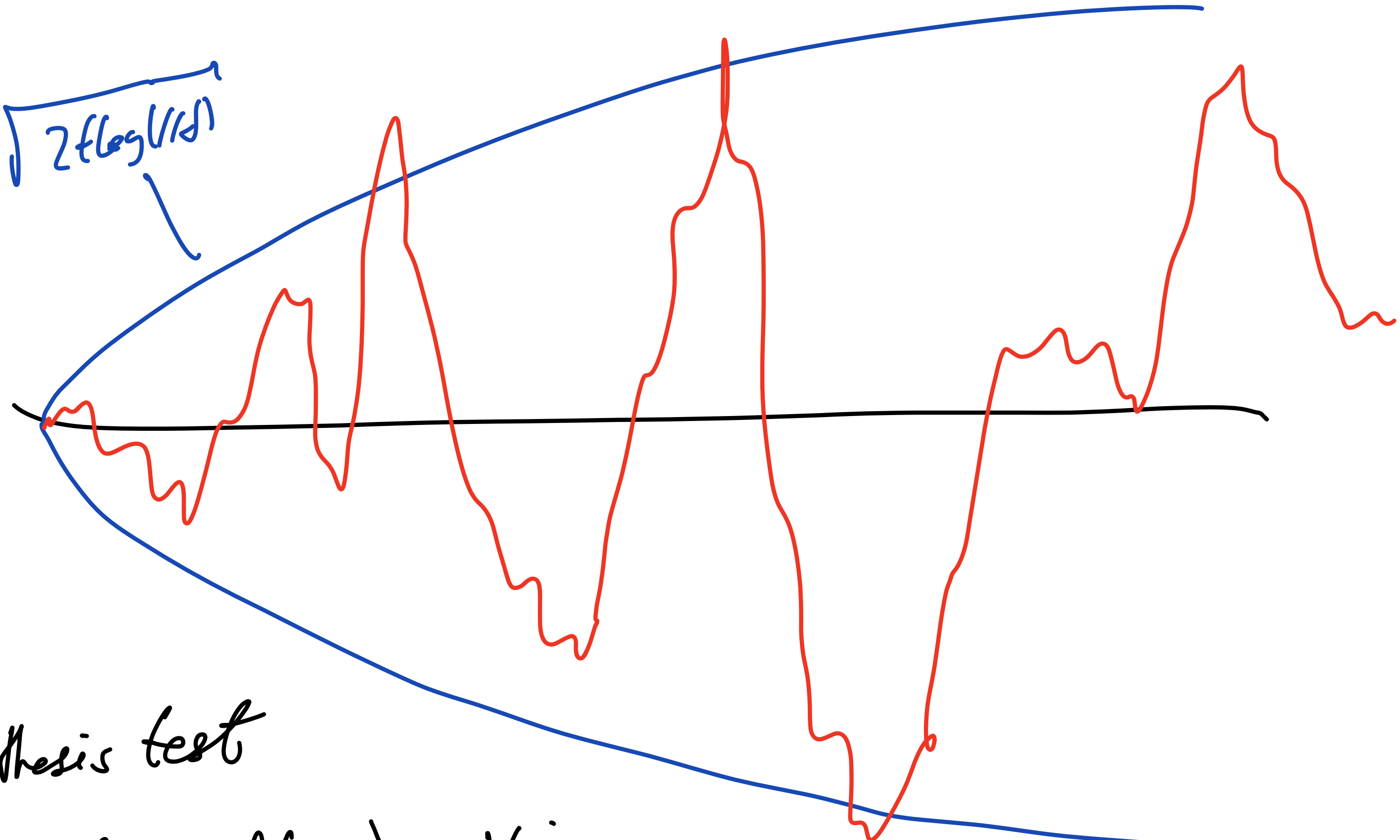
$$\ell(\overline{w}_T) - \ell(w_*) \lesssim_c \sigma R \sqrt{\frac{\log(1/\delta)}{T}}$$

Let $Z_1, Z_2, \ldots$ be an $\mathbb{F}$-adapted sequence and let σ_t be a predictable sequence in the sense that σ_{t+1} is $\mathcal{F}_t$-measurable.

If $\mathbb{E}[Z_{t+1} \mid \mathcal{F}_t] = 0$ and for any λ we have $\mathbb{E}[\exp(\lambda Z_t) \mid \mathcal{F}_{t-1}] \leq e^{\lambda^2 \sigma_t^2/2}$ then $M_t(\lambda) = \exp(\lambda S_t - V_t \lambda^2/2)$ is a supermartingale where $S_t = \sum_{s \leq t} Z_s$, $V_t = \sum_{s \leq t} \sigma_s^2$.

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

and $\mathbb{P}\left(\exists t \in \mathbb{N} : |S_t| \geq \sqrt{2(V_t+1)\log\left(\frac{\sqrt{V_t+1}}{\delta}\right)}\right) \leq \delta.$



Hypothesis test

$H_0 : X_i \sim \mathcal{N}(0,1) \quad \forall i$

$H_1 : X_i \sim \mathcal{N}(\mu,1)$ for $\mu > 0 \quad \forall i$

p-value: the probability of observing an outcome under the null hypothesis

Reject the null if $p < .05$.

Neyman-Pearson testing gives an optimal way to test between two hypotheses using likelihood ratio.

$$\ell_\theta(X_i;\theta) = \frac{1}{\sqrt{2\pi}} e^{-(X_i-\theta)^2/2}$$

LRT: $$\underbrace{\frac{\prod_{i=1}^{t} \ell(X_i;\mu)}{\prod_{i=1}^{t} \ell(X_i;0)}} > \frac{1}{\delta} \Rightarrow \mathbb{P}_0(\text{rejecting null}) \le \delta$$

$\mathbb{P}(\quad) \le \delta$ by Markov's inequality.

Test is equiv. to $S_t > \alpha$

Define p-value $\min\{\delta : S_t \le \sqrt{2t\log(1/\delta)}\}$

Suppose $Z_1, Z_2, \ldots \in \mathbb{R}^d$ is a $\mathbb{F}$-adapted sequence w/ $\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] = 0$, $\mathbb{E}[\exp(\langle \lambda, Z_t\rangle \mid \mathcal{F}_{t-1})] \le \exp(\|\lambda\|^2_{\Sigma_t}/2)$ for all fixed $\lambda \in \mathbb{R}^d$ and Σ_t is a predictabe sequence of PSD matrices. Define $S_t = \sum_{i=1}^t Z_i$, $V_t = \sum_{i=1}^t \Sigma_i$ and $M_t(\lambda) = \exp(\langle \lambda, S_t\rangle - \|\lambda\|^2_{V_t}/2)$.

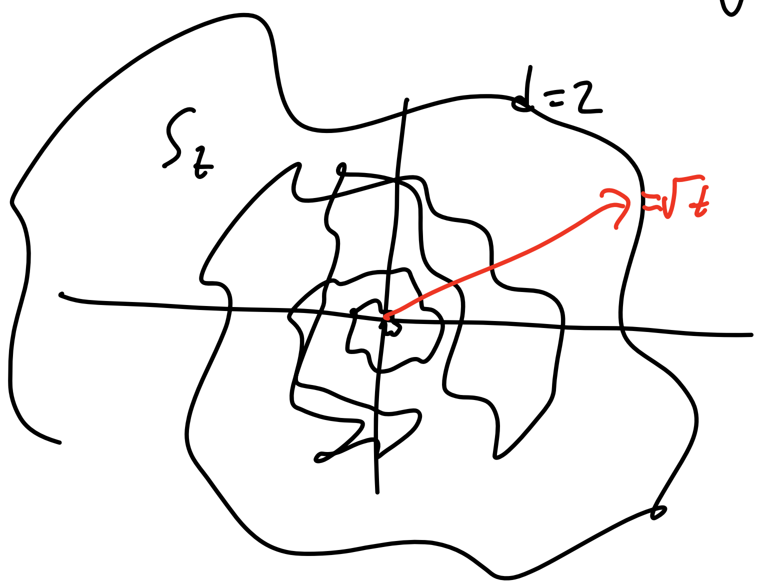
$M_t(\lambda)$ is a super martingale

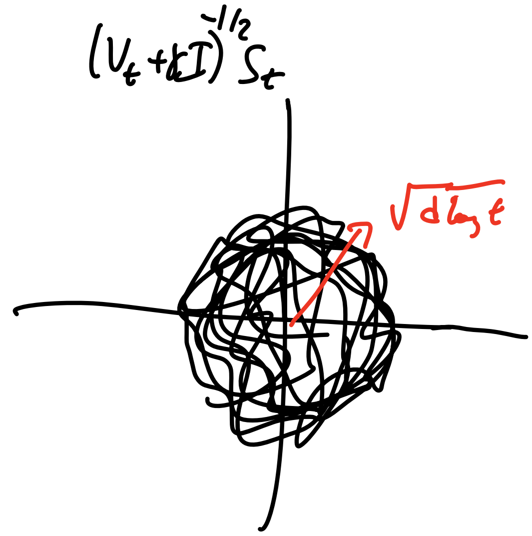
$$h(\lambda) = \frac{1}{(2\pi/\gamma)^{d/2}} \exp\left(-\|\lambda\|_2^2 \gamma / 2\right)$$

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

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

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





Let $x_1, x_2, \ldots \in \mathbb{R}^d$ be and $\mathcal{F}_{t-1}$-measurable sequence (i.e. x_t is predictable). Assume $\exists\, \theta_* \in \mathbb{R}^d$ s.t.

$y_t = \langle x_t, \theta_* \rangle + \eta_t$ is $\mathbb{F}$-adapted sequence w/

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

In the above define $Z_i = x_i \eta_i$, $S_t = \sum_{i=1}^{t} x_i \eta_i$, $V_t = \sum_{i=1}^{t} x_i x_i^T$

$$\mathbb{E}[\exp(\langle \lambda, x_i \eta_i \rangle) \mid \mathcal{F}_{i-1}] = \mathbb{E}[\exp(\langle \lambda, x_i \rangle \eta_i) \mid \mathcal{F}_{i-1}]$$

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

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

$\Rightarrow$ $\exp(\langle \lambda, S_t \rangle - \|\lambda\|_{V_t}^2 / 2)$ is a supermartingale.

$$\hat{\theta}_t = \arg\min_\theta \sum_i (y_i - \langle \theta, x_i \rangle)^2 + \gamma \|\theta\|_2^2$$

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

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

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

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

Note

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

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

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

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

$$= \gamma \sqrt{\theta^T (V_t + \gamma I)^{-1} \theta}$$

$$\leq \gamma \sqrt{\theta^T (\gamma I)^{-1} \theta}$$

$$\leq \sqrt{\gamma} \|\theta\|_2$$

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

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

If $\|x_t\|_2 \leq L$ then $\log|V_t+\gamma I| = d\log(tLd+\gamma)$

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

$$x_t = \underset{x\in\mathcal{X}}{\operatorname{argmax}}\ UCB_t(x)\ \cancel{\alpha}\ \|x\|_{(V_t+\gamma I)^{-1}}$$

$$UCB_t(x) = \max_{\theta\in\mathcal{E}_t} \langle x,\theta\rangle, \quad \mathcal{E}_t = \{\theta: \|\hat{\theta}_t-\theta\|_{V_t+\gamma I} \leq \beta_t\}$$

Thm/ Regret of UCB is $\leq d\sqrt{T}$ (ignoring log factors)

Suppose $x_1,\ldots,x_T$ are chosen in advance of seeing data.

Then for a $z\in\mathbb{R}^d$, $\langle z, \hat{\theta}_t-\theta_*\rangle \leq \|z\|_{V_t^{-1}}\cdot\sqrt{2\log(1/\delta)}$

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

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