

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 , $\bigcup_n F_n \in \mathcal{F}$

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

$\{TH, HT, HH\}, \dots \}$

• Probability measure For a σ -algebra $\mathcal{F}$ on Ω satisfies

1. $P: \mathcal{F} \rightarrow [0, 1]$

2. $P(\emptyset) = 0$

3. $P(\Omega) = 1$

4. For a seq $F_n \in \mathcal{F}$ w/ $F_i \cap F_j = \emptyset$, $P(\bigcup_n F_n) = \sum_n P(F_n)$

• A probability space is defined as tuple $(\Omega, \mathcal{F}, P)$

• Fix $(\Omega, \mathcal{F}, P)$. A function $g: \Omega \rightarrow \mathbb{R}$ is $\mathcal{F}$ -measurable if for all Borel sets B we have $\{\omega: g(\omega) \in B\} \in \mathcal{F}$.

• Random variable X on $(\Omega, \mathcal{F}, P)$ is a function $X: \Omega \rightarrow \mathbb{R}$ that is $\mathcal{F}$ -measurable.

• Let X_t be a st of R.V.s indexed by $t \in T$ on space

$(\Omega, \mathcal{F}, P)$, 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

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

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

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

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

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

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

$$\{TT, HH\}, \{HT, TH, HH\}, \{TT, HT, TH, HH\}$$

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

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

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

Assume X, Y are discrete

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

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

$$E[X|A] := \sum_x x P(X=x|A)$$

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

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

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

Conditional Expectation. Let X be a R.V. on $(\Omega, \mathcal{F}, P)$ and let $\mathcal{H} \subseteq \mathcal{F}$. Then we say

$E[X|\mathcal{H}]: \mathcal{H} \rightarrow \mathbb{R}$ is the cond. expectation of X given $\mathcal{H}$ if

$$\int_{\omega \in H} E[X|\mathcal{H}](\omega) dP(\omega) = \int_{\omega \in H} X(\omega) dP(\omega), \quad \forall H \in \mathcal{H}$$

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

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

Filtrations

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

We say $\mathbb{F} = \{\mathcal{F}_t\}_t$ is a filtration on $\mathcal{F}$

if $\{\Omega, \emptyset\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots$

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

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

Martingales

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

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

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

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

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

$$\begin{aligned}
\mathbb{E}[X_{t+1} | \mathcal{F}_t] &= \mathbb{E}\left[\exp(\lambda z_{t+1} - \lambda^2 / 2) \exp\left(\lambda \sum_{s=1}^t z_s - \lambda^2 t / 2\right) | \mathcal{F}_t\right] \\
&= \underbrace{\mathbb{E}[\exp(\lambda z_{t+1} - \lambda^2 / 2) | \mathcal{F}_t]}_{\leq 1} \cdot \underbrace{\mathbb{E}[\exp(\lambda \sum_{s=1}^t z_s - \lambda^2 t / 2) | \mathcal{F}_t]}_{= X_t} \\
&\leq X_t
\end{aligned}$$

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

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

valid stopping time

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

Doob's Optional Stopping Theorem. Let $\mathcal{F}$ be a filtration of $\{X_t\}_t$ martingale sequence and let $\mathcal{T}$ be an $\mathcal{F}$ -adapted stopping time. If

- $\exists N \in \mathbb{N} : P(\mathcal{T} < N) = 1$, or

- $E[\mathcal{T}] < \infty$ and $E[|X_{t+1} - X_t|] < \infty \quad \forall t$

then $X_{\mathcal{T}}$ is well-defined and $E[X_{\mathcal{T}}] = E[X_0]$.

Furthermore, if X_t is supermartingale then $E[X_{\mathcal{T}}] \leq E[X_0]$
 " " submartingale then $E[X_{\mathcal{T}}] \geq E[X_0]$

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

$$X_t = \sum_{s=1}^t z_s. \quad \mathcal{T} = \min\{t: X_t \geq \varepsilon\}. \quad \text{assume } \varepsilon \in \mathbb{N}$$

$$M_t = X_t - tp$$

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

$$\mathbb{E}[X_3 - p3] = \mathbb{E}[M_3] = \mathbb{E}[M_0] = 0$$

$$\Rightarrow \varepsilon = \mathbb{E}[X_3] = p \cdot \mathbb{E}[3]$$

$$\Rightarrow \mathbb{E}[3] = \frac{\varepsilon}{p}$$

Maximal inequality 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 $P(\max_{t \in \mathbb{N}} X_t > \varepsilon) \leq \frac{\mathbb{E}[X_0]}{\varepsilon}$.

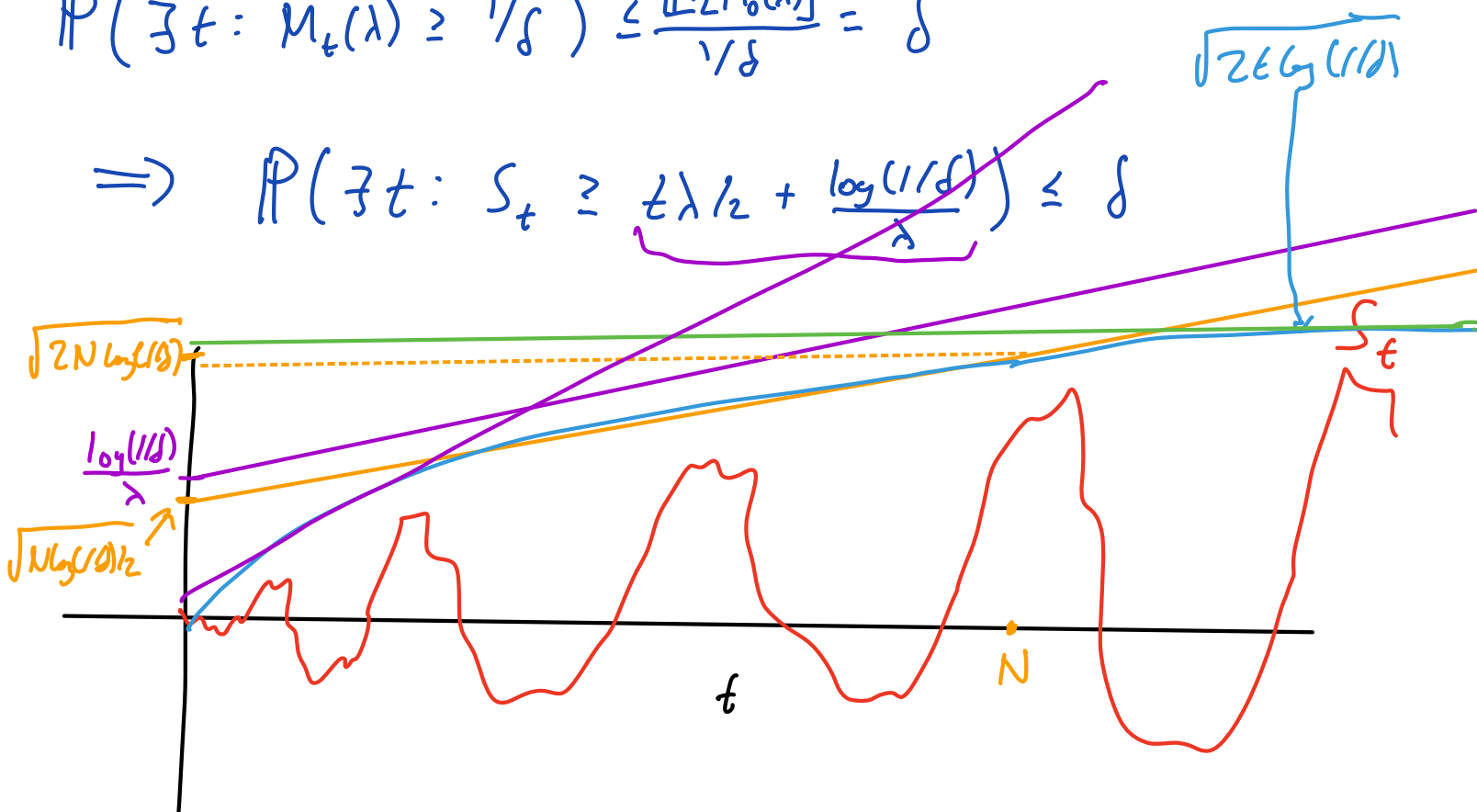
• If X_t is a submartingale $P(\max_{t=1, \dots, n} X_t > \varepsilon) \leq \frac{\mathbb{E}[X_n]}{\varepsilon}$.

Let $Z_1, Z_2, \dots \in \{-1, 1\}$ be Bernoulli $(1/2)$. Define $S_t = \sum_{s \leq t} Z_s$.
 Fix any $\lambda \in \mathbb{R}$ and define

$$M_t(\lambda) = \exp(\lambda S_t - t\lambda^2/2) \quad \mathbb{E}[M_{t+1}(\lambda) | \mathcal{F}_t] \leq M_t(\lambda)$$

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

$$\Rightarrow P(\exists t: S_t \geq \underbrace{t\lambda/2 + \frac{\log(1/\delta)}{\lambda}}_{\sqrt{2t \log(1/\delta)}}) \leq \delta$$



Pick $\lambda = \sqrt{\frac{2 \log(1/\delta)}{N}}$ $P(\exists t: S_t \geq (t/\sqrt{N} + \sqrt{N})\sqrt{\log(1/\delta)/2}) \leq \delta$

$$\Rightarrow P(\max_{t=1, \dots, N} S_t \geq \sqrt{2N \log(1/\delta)}) \leq \delta$$

Using Hoeffding $P(S_N \geq \sqrt{2N \log(1/\delta)}) \leq \delta$

$$P(\exists t \in N: S_t \geq \sqrt{2t \log(2t^2/\delta)}) \leq \sum_{t=1}^{\infty} P(\cdot) \leq \sum_{t=1}^{\infty} \frac{\delta}{2t^2} \leq \delta$$

Method of Mixtures.

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

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

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

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

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

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

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

$$\Rightarrow P(\exists t: |S_t| \geq \sqrt{2(t+v^2)} (\log(1/\delta) + \log(\sqrt{\frac{t+v^2}{v^2}}))) \leq \delta$$

Convenient choice is $\nu=1$

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

Interlude: Optimization

$$\mathbb{E}_x. L(w_t, I_t) = (w_t^T x_{I_t} - y_{I_t})^2$$

SGD on a dataset.

L convex.

for $t=1, 2, \dots$

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

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

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

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

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

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

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

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

$$\mathcal{E}_t := \nabla \ell(w_t) - \nabla L(w_t, I_t)$$

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

$$= \nabla L(w_t, I_t)^T (w_t - w_*) + \mathcal{E}_t^T (w_t - w_*)$$

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

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

$$w_{t+1} \text{ is } \mathcal{F}_t\text{-measurable} \quad = \sum_{s=1}^t \mathcal{E}_s^T(w_s - w_*)$$

Martingale!

$$\mathcal{E}_s^T(w_s - w_*) \leq \|\mathcal{E}_s\|_2 \cdot \|w_s - w_*\|_2 \leq 2GR$$

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

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

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

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

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

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

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

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

Let $Z_1, Z_2, \dots$ 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 $\mathbb{E}[Z_{t+1} | \mathcal{F}_t] = 0$ and for any
 λ we have $\mathbb{E}[\exp(\lambda Z_{t+1}) | \mathcal{F}_t] \leq e^{\lambda^2 \sigma_t^2 / 2}$

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

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