**EXP3($\gamma$): Exponential Weights for Exploration Exploitation**

**Input:** Time horizon $T$, $n$ arms, $\eta > 0$, $\gamma \in [0,1]$, $\lambda \in \triangle_n$.

**Initialize:** Player sets $p_1 = (1/n, \dots, 1/n) \in \triangle_n$. Adversary chooses $\{\ell_t\}_{t=1}^T \subset [-1,1]^n$.

**for:** $t = 1, \cdots, T$

Player draws $I_t \sim q_t := (1-\gamma)p_t + \gamma\lambda$ and suffers (and observes) loss $\ell(I_t, \ell_t) = \ell_{t,I_t}$

Player computes $\widehat{\ell}_{t,i} = \frac{\mathbf{1}\{I_t = i\}}{q_{t,i}} \ell_{t,i}$

Update iterates:

$$w_{t+1,i} = w_{t,i} \exp(-\eta \widehat{\ell}_{t,i}) \qquad p_{t+1,i} = w_{t+1,i} / \sum_{j=1}^n w_{t+1,j}.$$

**Theorem 17.** *Fix any sequence $\ell_t \in [-1,1]$ for all $t$ and assume $\eta, \gamma \geq 0$ satisfy $-\eta\widehat{\ell}_{t,i} \leq 1$ for all $i, t$. Then we have*

$\gamma = 2d$

$$\max_{i=1,\dots,n} \mathbb{E}\Big[\sum_{t=1}^T \ell_{t,I_t} - \ell_{t,i}\Big] \leq 2\gamma T + \frac{\log(n)}{\eta} + (1-\gamma)\eta \mathbb{E}\Big[\sum_{t=1}^T \sum_{j=1}^n p_{t,j} \widehat{\ell}_{t,j}^2\Big].$$

$\frac{dT}{1-\gamma}$

$= 2\gamma dT + \frac{\log(n)}{\gamma} + \gamma dT$

$= 3\gamma dT + \frac{\log(n)}{\gamma}$

$\leq \sqrt{12 dT \log(n)}$

## Linear Bandits - Adversarial Setting

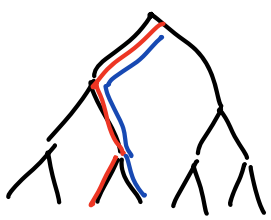
Input $\mathcal{X} = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$

Adversary chooses $\{\theta_t\}_{t=1}^T \subset \mathbb{R}^d$ : $\max_{x \in \mathcal{X}} |\langle x, \theta_t \rangle| \leq 1 \quad \forall t$

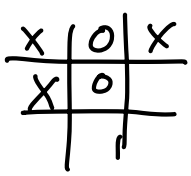
for $t = 1, 2, \dots, T$

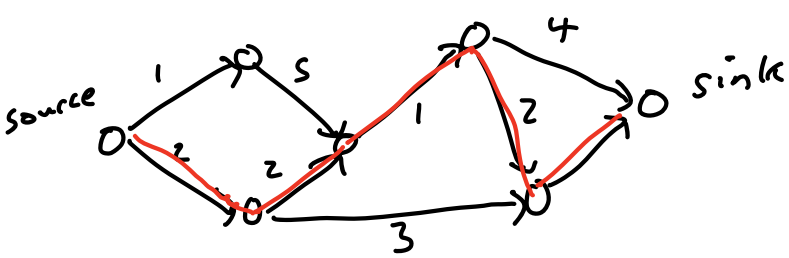
Player chooses $x_t \in \mathcal{X}$

Adversary reveals $\langle \theta_t, x_t \rangle$ (not $\theta_t$!)



$$\text{Regret} = \max_{x \in \mathcal{X}} \sum_{t=1}^T \langle \theta_t, x_t - x \rangle$$




$x \in \{0,1\}^{|E|}$

For $x_i \in \mathcal{X}$ define $\ell_{t,i} = \langle x_i, \theta_t \rangle$

$$\hat{\ell}_{t,i} = x_i^\top A(q_t)^{-1} x_{I_t} \ell_{t,I_t}, \qquad A(\lambda) = \sum_{i=1}^n \lambda_i x_i x_i^\top$$

$$\mathbb{E}\left[A(q_t)^{-1} x_{I_t} \ell_{t,I_t}\right] = \sum_{j=1}^n q_{t,j} A(q_t)^{-1} x_j \ell_{t,j}$$

$$= \sum_{j=1}^n q_{t,j} A(q_t)^{-1} x_j x_j^\top \theta_t$$

$$= I \cdot \theta_t = \theta_t$$

$$\Rightarrow \quad \mathbb{E}[\hat{\ell}_{t,i}] = \ell_{t,i} \quad \forall i$$

---

Aside $x_t \sim \lambda$, $y_t = \langle x_t, \theta_* \rangle + \eta_t$ $t = 1, \dots, T$

$$\hat{\theta}_{LS} = \left(\frac{1}{T}\sum_{t=1}^T x_t x_t^\top\right)^{-1} \frac{1}{T}\sum_{t=1}^T x_t y_t$$

$$\mathbb{E}[\hat{\theta}_{LS}] = \theta_*, \quad \mathrm{Cov}(\hat{\theta}_{LS}) = \left(\sum_{t=1}^T x_t x_t^\top\right)^{-1}$$

$$\hat{\theta}_{IPS} = \left(\sum_{x \in \mathcal{X}} \lambda_x x x^\top\right)^{-1} \frac{1}{T} \sum_{t=1}^T x_t y_t$$

$$\mathbb{E}[\hat{\theta}_{IPS}] = \theta_*, \quad \mathrm{Cov}(\hat{\theta}_{IPS}) = \left(T \sum_{x \in \mathcal{X}} \lambda_x x x^\top\right)^{-1}$$

$\mathcal{X} = \{\begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}0\\1\end{bmatrix}\}$ $\quad \lambda_1, \lambda_2 = 1-\lambda_1$

$T_1 = \sum_{t=1}^{T} \mathbb{1}\{I_t = 1\}$

$$\hat{\theta}_{LS} = \begin{bmatrix} \frac{1}{T_1} \sum_{t=1}^{T} \mathbb{1}\{I_t=1\}\, y_t \\ \frac{1}{T-T_1} \sum_{t=1}^{T} \mathbb{1}\{I_t=2\}\, y_t \end{bmatrix}$$

$$\hat{\theta}_{IPS} = \begin{bmatrix} \frac{1}{\lambda_1 T} \sum_{t=1}^{T} \mathbb{1}\{I_t=1\}\, y_t \\ \frac{1}{(1-\lambda_1)T} \sum_{t=1}^{T} \mathbb{1}\{I_t=1\}\, y_t \end{bmatrix}$$

---

$$\hat{\ell}_{t,i} = x_i^T A(q_t)^{-1} x_{I_t}\, \ell_{t,I_t}$$

$$-\xi \hat{\ell}_{t,i} \le \xi \left| x_i^T A(q_t)^{-1} x_{I_t} \right| \qquad (|\ell_{t,i}| \le 1)$$

$$= \xi \left| \langle A(q_t)^{-1/2} x_i,\ A(q_t)^{-1/2} x_{I_t} \rangle \right|$$

$$\le \xi \, \|x_i\|_{A(q_t)^{-1}} \cdot \|x_{I_t}\|_{A(q_t)^{-1}}$$

$$\le \max_{j=1,\dots,n} \xi \cdot \|x_j\|^2_{A(q_t)^{-1}}$$

$$\le \max_{j=1,\dots,n} \frac{\xi}{\gamma} \|x_j\|^2_{A(\lambda)^{-1}} \qquad \text{if } \lambda \text{ G-optimal.}$$

$$= \frac{\xi}{\gamma} \cdot d$$

Suggests picking $\gamma = 2d$ s.t. $-\hat{\ell}_{t,i} \le 1 \quad \forall i, t$.

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{n} p_{t,i}\,\hat{\ell}_{t,i}^{2}\right] = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{n} p_{t,i}\, x_i^T A(y_t)^{-1} x_{I_t} x_{I_t}^T A(y_t)^{-1} x_i\, \ell_{t,I_t}^{2}\right]$$

$$\le \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{n} p_{t,i}\, x_i^T A(y_t)^{-1} x_{I_t} x_{I_t}^T A(y_t)^{-1} x_i\right]$$

$$= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{n} p_{t,i}\, x_i^T A(y_t)^{-1} x_i\right]$$

$$\le \frac{1}{1-\gamma}\,\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{n} p_{t,i}\, x_i^T A(p_t)^{-1} x_i\right]$$

$$= \frac{1}{1-\gamma}\,\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{n} p_{t,i}\, \mathrm{Trace}\left(x_i^T A(p_t)^{-1} x_i\right)\right]$$

$$= \frac{1}{1-\gamma}\,\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i} p_{t,i}\, \mathrm{Trace}\left(A(p_t)^{-1} x_i x_i^T\right)\right]$$

$$= \frac{dT}{1-\gamma}$$

# Contextual Bandits

for $t = 1, 2, \ldots, T$

Nature reveals context $c_t \overset{iid}{\sim} \nu$

Player chooses $I_t \in [n]$

and gets reward $y_t \in [-1, 1]$, s.t. $\mathbb{E}[y_t | c_t] = r(c_t, I_t)$

Suppose $\text{support}(\nu) = \mathcal{C}$ and $|\mathcal{C}| < \infty$.

Treat each context as indep. bandit.

$$\sum_{t=1}^{T} \max_i \left( r(c_t, i) - r(c_t, I_t) \right)$$

$$= \sum_{c \in \mathcal{C}} \max_i \sum_{t=1}^{T} \mathbb{1}\{c_t = c\} \left( r(c_t, i) - r(c_t, I_t) \right)$$

$$\leq \sum_{c \in \mathcal{C}} \sqrt{n T_c}$$ ← If playing MAB w/ $\sqrt{nT}$ regret

$T_c : \sum_t \mathbb{1}\{c_t = c\}$

$$\leq \sqrt{\left( \sum_{c \in \mathcal{C}} (\sqrt{T_c})^2 \right) \cdot \left( \sum_{c \in \mathcal{C}} (\sqrt{n})^2 \right)}$$

$$= \sqrt{n |\mathcal{C}| T}$$

Alternatively, what if we ignore context?

$$\max_i \sum_{t=1}^{T} r(c_t, i) - r(c_t, I_t) \leq \sqrt{nT}$$

Total Reward:

Bandit per context $\sum_{t=1}^{T} r(c_t, I_t) \geq \sum_{t=1}^{T} \max_i r(c_t, i) - \sqrt{n|C|T}$

$|\Pi| = n^{|C|}$

Ignore context $\sum_{t=1}^{T} r(c_t, I_t) \geq \max_i \sum_{t=1}^{T} r(c_t, i) - \sqrt{nT}$

$|\Pi| = n$

Policy regret $\sum_{t=1}^{T} r(c_t, a_t) \geq \max_{\pi \in \Pi} \sum_{t=1}^{T} \mathbb{E}_{i \sim \pi}[r(c_t, i)] - \sqrt{nT \log|\Pi|}$

Consider a finite set of policies $\Pi$, $\pi \in \Pi$ $\pi : C \to [n]$

Policy regret $\max_{\pi \in \Pi} \sum_{t=1}^{T} r(c_t, \pi) - r(c_t, \hat{\pi}_t)$

---

$\pi(c_t) \in \Delta_d$, $a_t \sim \pi(c_t)$ then $P(j = a_t | c_t) = \pi(j | c_t)$

Suppose contexts are arbitrary (adversarial!)

Adversary chooses $\{c_t\}_{t=1}^{T}$ in advance

for $t = 1, 2, \ldots, T$

Adversary reveals $c_t$

Player chooses $I_t \in [n]$, draws $a_t \in \pi_{I_t}(c_t)$

gets loss $\ell(c_t, a_t)$

Regret $\max_{\pi \in \Pi} \sum_{t=1}^{T} \mathbb{E}_{a_t \sim \pi(c_t)} [\ell(c_t, a_t)] - \mathbb{E}_{a \sim \pi(c_t)} [\ell(c_t, a)]$

**EXP3($\gamma$): Exponential Weights for Exploration Exploitation**

**Input:** Time horizon $T$, $n$ arms, $\eta > 0$, $\gamma \in [0,1]$

**Initialize:** Player sets $p_1 = (1/n, \dots, 1/n) \in \triangle_n$. Adversary chooses $\{\ell_t\}_{t=1}^T \subset [-1,1]^n$.

**for:** $t = 1, \cdots, T$ Adversary reveals $c_t$

Player defines $\lambda_t \in \triangle_n$ and plays $I_t \sim q_t := (1-\gamma)p_t + \gamma\lambda_t$

Player suffers (and observes) loss $\ell_{t,I_t}$ (but does *not* observe $\ell_{t,i}$ for $i \neq I_t$)

Player computes loss estimator $\widehat{\ell}_{t,i}$ for all $i \in [n]$

Update iterates:

$$w_{t+1,i} = w_{t,i}\exp(-\eta\widehat{\ell}_{t,i}) \qquad p_{t+1,i} = w_{t+1,i}/\sum_{j=1}^n w_{t+1,j}.$$

**Theorem 17.** *Fix any sequence $\ell_t \in [-1,1]$ for all $t$. For any $\widehat{\ell}_{t,i}$ and $\eta, \gamma \geq 0$ that satisfy $\mathbb{E}[\widehat{\ell}_{t,i}|\mathcal{F}_{t-1}] = \ell_{t,i}$ and $-\eta\widehat{\ell}_{t,i} \leq 1$ for all $i, t$ we have*

$\gamma = 2d$

d-actions
n-policies

$$\max_{i=1,\dots,n} \mathbb{E}\Big[\sum_{t=1}^T \ell_{t,I_t} - \ell_{t,i}\Big] \leq 2\gamma T + \frac{\log(n)}{\eta} + (1-\gamma)\eta\mathbb{E}\Big[\sum_{t=1}^T\sum_{j=1}^n p_{t,j}\widehat{\ell}_{t,j}^2\Big].$$

$\frac{dT}{1-\gamma}$

$\ell_{t,i} := \mathbb{E}_{a\sim\pi_i}[\ell(c_t, a)]$

$\leq 3\gamma dT + \frac{\log(n)}{\gamma}$

I observe $\ell(c_t, a_t)$

$\leq \sqrt{12\, dT \log(n)}.$

$$\hat{\ell}_{t,i} = \frac{\pi_i(a_t|c_t)}{\sum_{k=1}^n q_{t,k}\,\pi_k(a_t|c_t)}\,\ell(c_t, a_t)$$

$a_t \sim \pi_{I_t}(c_t) \qquad I_t \sim q_t$

$$\mathbb{E}[\hat{\ell}_{t,i}] = \mathbb{E}\left[\frac{\pi_i(a_t|c_t)\,\ell(c_t,a_t)}{\sum_h q_{t,h}\,\pi_h(a_t|c_t)}\right]$$

$$= \sum_{j=1}^{d} \mathbb{E}\left[ \mathbb{P}(a_t = j) \frac{\pi_i(j \mid c_t)\, \ell(c_t, j)}{\sum_h q_{t,h}\, \pi_h(j \mid c_t)} \right]$$

$$= \sum_{s=1}^{n} \mathbb{P}(a_t = j \mid I_t = s)\, \mathbb{P}(I_t = s)$$

$$= \sum_{s=1}^{n} \pi_s(j \mid c_t)\, q_{t,s}$$

$$= \sum_{s=1}^{d} \mathbb{E}\left[ \pi_i(j \mid c_t)\, \ell(c_t, j) \right]$$

$$= \mathbb{E}_{a \sim \pi(c_t)}\left[ \ell(c_t, a) \right] = \ell_{t,i}$$

---

$$-\eta \hat{\ell}_{t,i} \le \eta \frac{\pi_i(a_t \mid c_t)}{\sum_{k=1}^{n} q_{t,k}\, \pi_k(a_t \mid c_t)}$$

$$\le \eta \sum_{j=1}^{d} \frac{\pi_i(j \mid c_t)}{\sum_{u=1}^{h} q_{t,u}\, \pi_u(a_t \mid c_t)}$$

$$\leq \frac{2}{\gamma} \sum_{j=1}^{d} \frac{\pi_i(j|(\epsilon)}{\sum_{h=1}^{n} \lambda_{t,h} \pi_h(j|(\epsilon)}$$

$$\leq \frac{2d}{\gamma} \leq 1 \quad \text{if} \quad \gamma = 2d.$$

**Proposition** For $i = 1, \ldots, n$ let $p_i \in \Delta_d$. Then

$$\min_{\lambda \in \Delta_n} \max_{i=1,\ldots,n} \sum_{j=1}^{d} \frac{p_{i,j}}{\sum_{k} \lambda_k p_{k,j}} = d.$$

**proof** $f(\lambda) := \sum_{j=1}^{d} \log\left( \sum_{k=1}^{d} \lambda_k p_{k,j} \right)$, $\lambda^* = \operatorname{argmax}_{\lambda} f(\lambda)$

for any $\lambda$ $\quad f(\lambda) \leq f(\lambda_*) + \nabla f(\lambda^*)^\top (\lambda - \lambda^*)$

$$\Rightarrow \quad 0 \geq \nabla f(\lambda^*)^T (\lambda - \lambda^*)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{d} \frac{p_{i,j}}{\sum_{k=1}^{n} p_{k,j} \lambda_k^*} (\lambda_i - \lambda_i^*)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{d} \frac{p_{i,j} \lambda_i}{\sum_{k=1}^{n} p_{k,j} \lambda_k^*} - d$$

Take $\lambda = e_i$

$$\Rightarrow \quad \sum_{j=1}^{d} \frac{P_{i,j}}{\sum_{k=1}^{n} P_{k,j} \lambda_k^*} \leq d.$$

$$\min_{\lambda \in \Delta_n} \max_{i=1,\dots,n} \sum_{j=1}^{d} \frac{P_{i,j}}{\sum_{k} \lambda_k P_{k,j}} \leq \max_{i=1,\dots,n} \sum_{j=1}^{d} \frac{P_{i,j}}{\sum_{k} \lambda_k^* P_{k,j}} \leq d$$

$$\min_{\lambda \in \Delta_n} \max_{i=1,\dots,n} \sum_{j=1}^{d} \frac{P_{i,j}}{\sum_{k} \lambda_k P_{k,j}} \geq \min_{\lambda} \sum_{i=1}^{n} \lambda_i \sum_{j=1}^{d} \frac{P_{i,j}}{\sum_k \lambda_k P_{k,j}} = d$$

∎

$$\mathbb{E}\left[ \sum_{t=1}^{T} \sum_{i=1}^{n} p_{t,i} \hat{\ell}_{t,i}^2 \right] \leq \frac{dT}{1-\gamma}$$

Input policy set $\Pi$, $\pi \in \Pi$, $\pi : \mathcal{C} \to \mathcal{X}$

for $t = 1, 2, \ldots, T$

Nature reveals $c_t \overset{iid}{\sim} \nu$

Player chooses $x_t \in \mathcal{X}$ observes $r(c_t, x_t) + \eta_t$

Collection of policies $\Pi$, $\pi : \mathcal{C} \to \mathcal{X}$

Off-policy Estimation

Suppose we play a logging policy $\mu(\cdot \mid c_t) \in \Delta_{\mathcal{X}}$

$x_t \sim \mu(\cdot \mid c_t)$ $\forall t$ observe $y_t = r(c_t, x_t) + \eta_t$

we get dataset $\{(c_t, x_t, y_t, p_t)\}_{t=1}^{T}$

$p_t = \mu(x_t \mid c_t)$

$$V(\pi) = \mathbb{E}_{c_t \sim \nu}\left[ r(c_t, \pi(c_t)) \right].$$

$$\hat{V}(\pi) = \frac{1}{T} \sum_{t=1}^{T} \frac{\mathbb{1}\{x_t = \pi(c_t)\}}{p_t} y_t$$

$$\mathbb{E}\left[ \hat{V} \mid \pi \right] = V(\pi)$$

$$\mathbb{E}\left[(\hat{V}(\pi) - V(\pi))^2\right] = \mathbb{E}\left[\mathbb{E}\left[(\hat{V}(\pi) - V(\pi))^2 \mid \{c_t\}_{t=1}^T\right]\right]$$

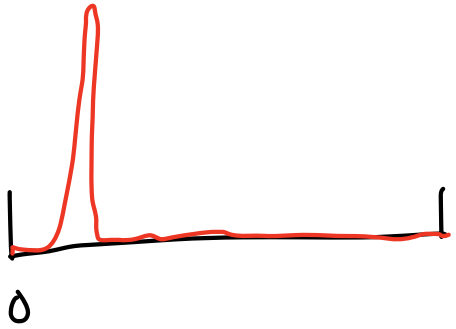
$$\leq \mathbb{E}\left[\frac{1}{T^2}\sum_{t=1}^{T}\frac{1}{p_t}\right]$$

$$\leq \frac{1}{T}\mathbb{E}_{c\sim\nu}\left[\frac{1}{\mu(\pi(c)|c)}\right].$$

**Bernstein's Inequality** Let $X_1, \ldots X_m$ be iid R.V. w/ $\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[(X_i - \mathbb{E}[X_i])^2\right] \leq \sigma^2$ and $|X_i| \leq B$ then

$$\left|\frac{1}{m}\sum_{i=1}^{m} X_i - \mathbb{E}[X_i]\right| \leq \sqrt{\frac{2\sigma^2 \log(2/\delta)}{m}} + \frac{2B\log(2/\delta)}{3m}.$$

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



**Corollary** $|\hat{V}(\pi) - V(\pi)| \le \sqrt{\frac{2\log(2|\Pi|/\delta)\cdot \mathbb{E}\left[\frac{1}{\mu(\pi(c_t)|c_t)}\right]}{T}} + \frac{2\bar{V}\log(2|\Pi|/\delta)}{3T}$

w.p. $\ge 1-\delta$ where $\bar{V} = \max_{c,x} \frac{1}{\mu(x|c)}$. $C_T(\pi)$

Uniform exploration

$\mu(x|c) = \frac{1}{|\mathcal{X}|} \Rightarrow \mathbb{E}\left[\frac{1}{\mu(\cdot|\cdot)}\right] = |\mathcal{X}|$

$\bar{V} = |\mathcal{X}|$

$$\Rightarrow |\hat{V}(\pi) - V(\pi)| \le \sqrt{\frac{2\log(2|\Pi|/\delta)|\mathcal{X}|}{T}} + \frac{2|\mathcal{X}|\log(2|\Pi|/\delta)}{T}$$

$$\le \sqrt{\frac{4|\mathcal{X}|\log(2|\Pi|/\delta)}{T}}.$$

$$\hat{\pi}_{MLE} = \operatorname*{argmax}_{\pi\in\Pi} \hat{V}(\pi)$$

$$= \operatorname*{argmax}_{\pi\in\Pi} \frac{1}{T}\sum_{t=1}^{T} \frac{\mathbb{1}\{\pi(c_t) = x_t\}}{p_t} y_t$$

$$= \operatorname*{argmax}_{\pi} \frac{1}{T}\sum_{t=1}^{T} (1 - \mathbb{1}\{\pi(c_t) \neq x_t\}) \frac{y_t}{p_t}$$

$$= \operatorname*{argmin}_{\pi\in\Pi} \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\{\pi(c_t) \neq x_t\} \frac{y_t}{p_t}$$

Consider generic machine learning problem $\{(x_i, y_i)\}_{i=1}^n$

$x_i \in \mathbb{R}^d$, $y_i \in [k]$. Find a classifier that predicts

$y_i$ from $x_i$. Equiv. find a classifier that minimizes

mis-takes $\frac{1}{n} \sum_{i=1}^n \mathbb{1}\{h(x_i) \neq y_i\}$.

$$\approx \frac{1}{n} \sum_{i=1}^n -\log\left( \frac{\exp([f_\theta(x_i)]_{y_i})}{\sum_h \exp([f_\theta(x_i)]_h)} \right)$$

$f_\theta : \mathbb{R}^d \to \mathbb{R}^k$

Predicting $\mathbb{P}(y_i = j | x_i) = \dfrac{\exp([f_\theta(x_i)]_j)}{\sum_k \exp([f_\theta(x_i)]_k)}$

$$V(\hat{\pi}_{MLE}) = \hat{V}(\hat{\pi}_{MLE}) + V(\hat{\pi}_{MLE}) - \hat{V}(\hat{\pi}_{MLE})$$

$$\geq \hat{V}(\hat{\pi}_{MLE}) - C_T(\hat{\pi}_{MLE})$$

$$\geq \hat{V}(\hat{\pi}_\alpha) - C_T(\hat{\pi}_{MLE})$$

$$\geq V(\pi_{\hat{a}}) - C_T(\hat{\pi}_{\hat{a}}) - C_T(\hat{\pi}_{MLE})$$

$$\geq V(\pi_{\hat{a}}) - \max_{\pi \in \Pi} 2C(\pi)$$

Principle of Pessimism

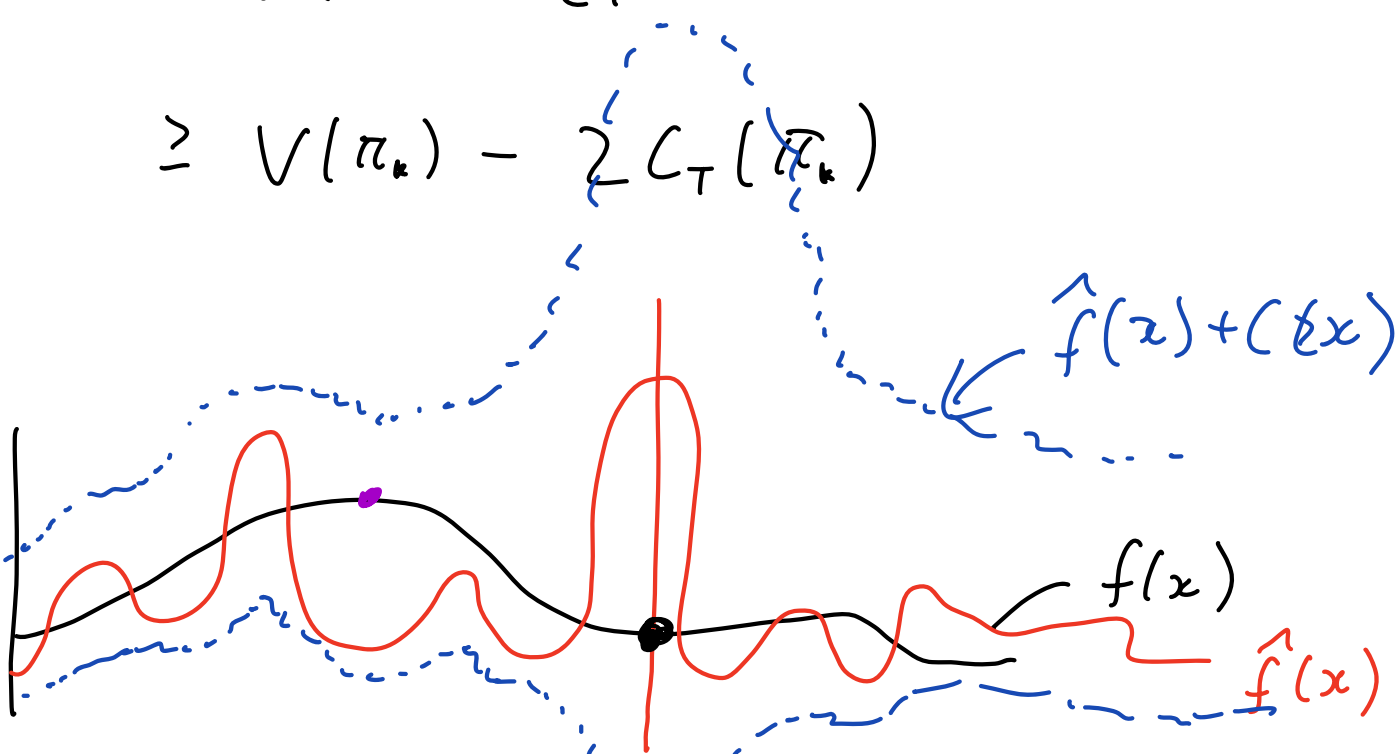
$$\hat{\pi}_{pes} = \operatorname{argmax}_{\pi \in \Pi} \hat{V}(\pi) - C_T(\pi)$$

$$V(\hat{\pi}_{pes}) = \hat{V}(\hat{\pi}_{pes}) + V(\hat{\pi}_{pes}) - \hat{V}(\hat{\pi}_{pes})$$

$$\geq \hat{V}(\hat{\pi}_{pes}) - C_T(\hat{\pi}_{pes})$$

$$\geq \hat{V}(\pi_{\hat{a}}) - C_T(\pi_{\hat{a}})$$

$$\geq V(\pi_{\hat{a}}) - 2C_T(\pi_{\hat{a}})$$



$x$ ↑ $f(x) - C(x)$

$$\hat{x} = \underset{x}{\operatorname{argmax}}\ \hat{f}(x)$$

$$\mathbb{E}[f(\hat{x})] \leq \max_x f(x)$$

$$\mathbb{E}[\hat{f}(\hat{x})] >> \max_x f(x)$$