Introduction to Reinforcement Learning Theory

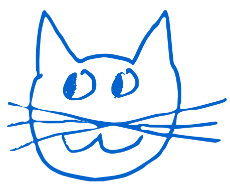
Vanessa Kosoy 2023

Plan of This Talk

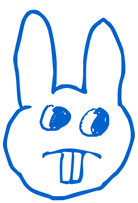
- Binary classification
- Multi-armed bandits
- Decision processes
- Reinforcement learning

Part I

Binary Classification



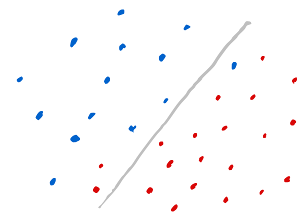
or



Risk

$D \subseteq \Delta X \qquad f : X \to \{0, 1\}$

$\pi : (X \times \{0,1\})^* \times X \to \{0,1\}$

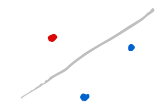


$\mathcal{H} \subseteq \{ X \to \{0,1\}\}$

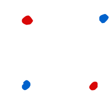
$$R^\pi(n) := \sup_{\substack{D \subseteq \Delta X \\ f \in \mathcal{H}}} \Pr_{\substack{S \sim (D \ltimes f)^n \\ x \sim D}} \left[\pi(S, x) \neq f(x) \right]$$

Vapnik-Chervonenkis Theory

$S \subseteq X$ shatters $\mathcal{H}$ when
$\forall f : S \to \{0,1\} \; \exists h \in \mathcal{H} : f = h|_S$



$$\dim_{VC} \mathcal{H} := \sup_{S \text{ shatters } \mathcal{H}} |S|$$



$$R^*(n) = \tilde{O}\left(\frac{\dim_{VC} \mathcal{H}}{n}\right)$$

Non-Uniform Learning

$$R_f^\pi(n) = \sup_{D \in \Delta X} \Pr_{\substack{S \sim (D \ltimes f)^n \\ x \sim D}} [\pi(S, x) \neq f(x)]$$

Theorem:

$$\exists \pi \, \forall f \in \mathcal{H} : R_f^\pi(n) \xrightarrow[n \to \infty]{} 0 \quad \text{iff}$$

$$\mathcal{H} = \bigcup_{k=0}^{\infty} \mathcal{H}_k \quad \text{s.t.} \quad \dim_{VC} \mathcal{H}_k < \infty$$

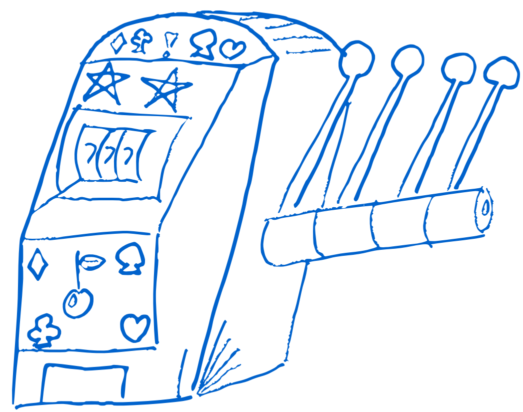
$$\mathcal{H} = \{ 1_{p(x,y) \geq 0} \mid p \in \mathbb{R}[x,y] \}$$

Notions of Dimension in Learning

Binary Classification	VC
General Classification	Natarajan
Online Learning	Littlestone
Singular Learning	RLCT
Reinforcement Learning	Decision-Estimation Coefficient

Part II

Multi-Armed Bandits

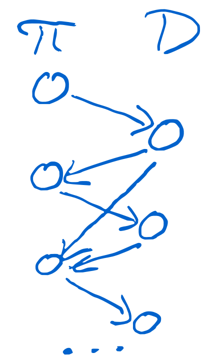


Stochastic Multi-Armed Bandits

$$D : A \to \Delta \mathbb{R}$$

$$\pi : (A \times \mathbb{R})^* \to A$$

$$D\pi \in \Delta (A \times \mathbb{R})^\omega$$



$$Rg^\pi(n) := \max_{a \in A} E[D(a)] - \frac{1}{n} E_{ar \sim D\pi}\left[\sum_{k=1}^{n} r_k\right]$$

$$Rg^*(n) = O\left(\sqrt{\frac{|A|}{n}}\right) \quad \text{distribution independent}$$

$$Rg^*(n) = O\left(c_D \frac{\log n}{n}\right) \quad \text{distribution dependent}$$

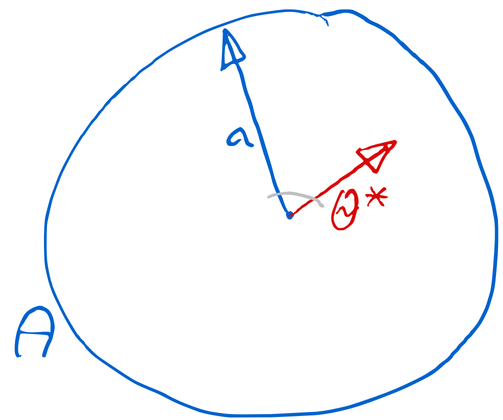
Linear Multi-Armed Bandits

$\mathcal{A} \subseteq \mathbb{R}^d$ $\quad \theta^* \in \mathbb{R}^d$

compact

$$E[r \mid a] = a^t \theta^*$$

$$R_g^*(n) = \tilde{O}\left(\frac{d}{\sqrt{n}}\right)$$

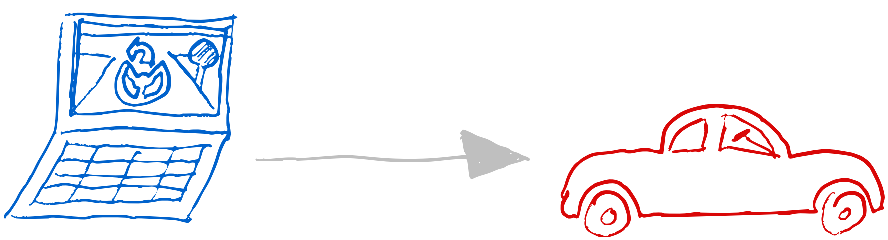


Pure Exploration

$$Rg^{\pi}_{sim}(n) := \max_{a \in A} E[D(a)] - E_{a_r \sim D_{\pi_n}}[r_n]$$

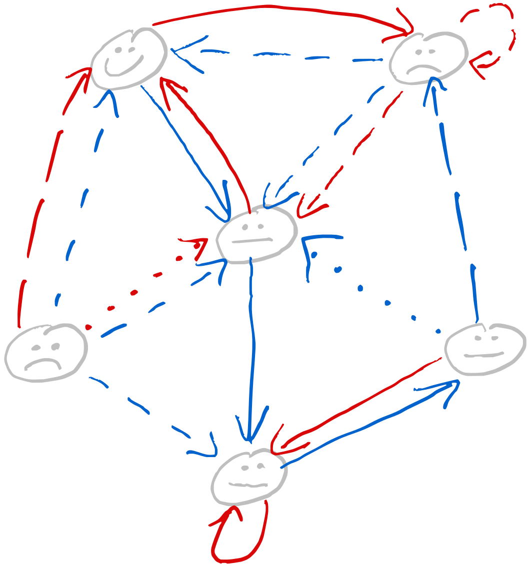
$$Rg^{*}_{sim}(n) = O\left(\sqrt{\frac{|A|}{n}}\right)$$ distribution independent

$$Rg^{*}_{sim}(n) = O\left(C_D e^{-c'_D n}\right)$$ distribution dependent



Part III

Decision Processes



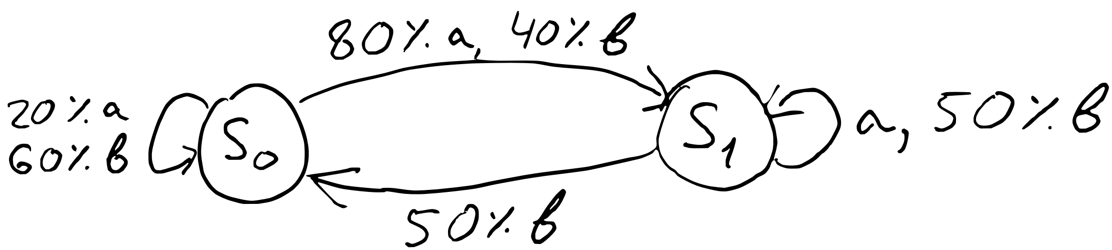
Markov Decision Processes (MDP)

$M = (S, A, s_0, T)$

$s_0 \in S$

$T: S \times A \longrightarrow \Delta(S \times \mathbb{R})$

$A = \{a, b\}$

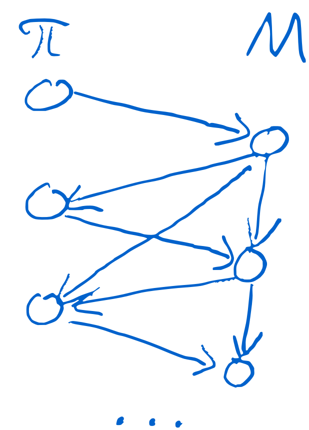


Policies

$$\pi : (A \times S \times \mathbb{R})^* \to A$$

$$M\pi \in \Delta(A \times S \times \mathbb{R})^\omega$$

$$\begin{cases} a_{k+1} = \pi(a_1 s_1 r_1\, a_2 s_2 r_2 \ldots a_k s_k r_k) \\ (s_{k+1}, r_{k+1}) \sim T(s_k, a_{k+1}) \end{cases}$$



Optimal Policies

$$\pi^* := \operatorname{argmax}_\pi E_{M\pi}[\mathcal{U}]$$

$$\mathcal{U} = \sum_{t=0}^{\infty} d_t r_t \qquad \left(\sum_{t=1}^{\infty} d_t = 1\right)$$

$$\pi^* : \mathbb{N} \times S \to A \quad (\text{Markov})$$

$$d_t = \frac{1}{n} \mathbb{1}_{t \leq n}$$ finite-horizon

$$d_t = (1-\gamma)\gamma^t \quad (\gamma \in (0,1))$$ geometric

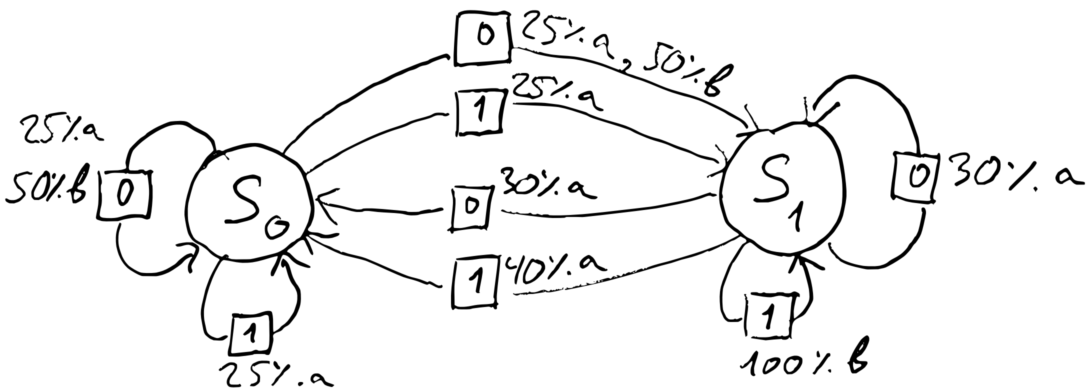
$$\pi^* : S \to A \quad (\text{Stationary})$$

Partially Observable Markov Decision Process (POMDP)

$$\theta_0 \in \Delta S$$

$$T : S \times A \to \Delta(S \times O)$$

$$Re : O \to \mathbb{R}$$



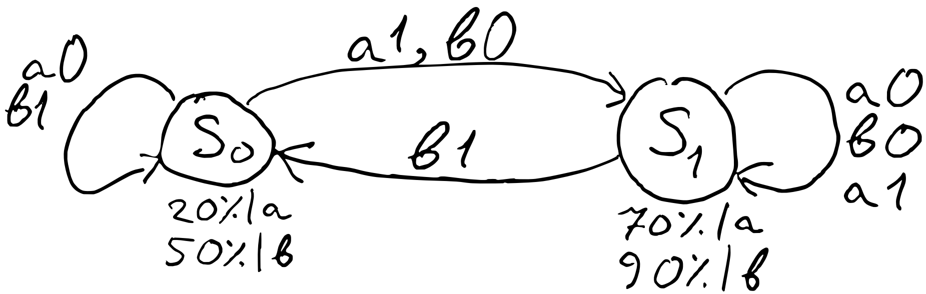
Control problem is PSPACE-hard!

Regular Decision Processes (RDP)

$s_0 \in S$

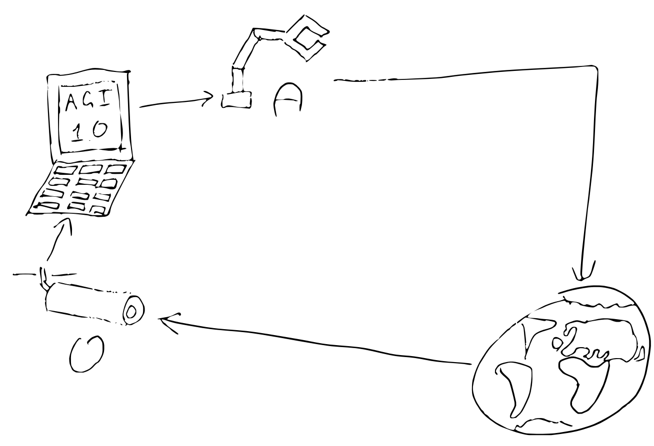
$$\left.\begin{array}{l} O: S \times A \to \Delta O \\ T: S \times A \times O \to S \\ Re: O \to \mathbb{R} \end{array}\right\} T^b: S \times A \to \Delta(S \times \mathbb{R})$$

$A = \{a, b\} \quad O = \{0, 1\}$



Part IV

Reinforcement Learning



Episodic Reinforcement Learning

$$M^H{}_\pi \in \Delta(A \times S \times \mathbb{R})^\omega$$

$$\begin{cases} a_{k+1} = \pi(a_1 s_1 r_1 a_2 s_2 r_2 \dots a_k s_k r_k) \\ (s_{k+1}, r_{k+1}) \sim T_{k \bmod H}(s_k, a_{k+1}) \text{ when } k+1 \not\equiv 0 \pmod H \\ s_{iH} = s_0 \end{cases}$$

$$r^* := \frac{1}{H} \operatorname{argmax}_\pi \mathrm{E}_{M^*\pi}\left[\sum_{k=1}^{H} r_k\right]$$

$$Rg^\pi(n) := r^* - \frac{1}{nH} \mathrm{E}_{(M^*)^H \pi}\left[\sum_{k=1}^{nH} r_k\right]$$

$$Rg^*(n) = \tilde{O}\left(\sqrt{\frac{|S| \cdot |A|}{n}}\right)$$

Lifelong Reinforcement Learning

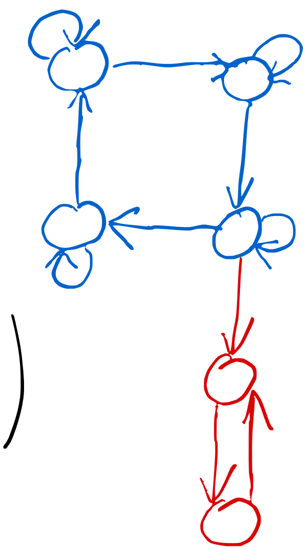
$$D(M) := \max_{x,y \in S} \min_{\pi} \mathbb{E}_{M\pi} \left[\min\{k \mid s_k = y\} \mid s_0 = x \right]$$

$$r^*(n) := \frac{1}{n} \operatorname{argmax}_{\pi} \mathbb{E}_{M^*\pi} \left[\sum_{k=1}^{n} r_k \right]$$

$$Rg^{\pi}(n) := r^*(n) - \frac{1}{n} \mathbb{E}_{M^*\pi} \left[\sum_{k=1}^{n} r_k \right]$$

or $$\lim_{m \to \infty} r^*(m) - \frac{1}{n} \mathbb{E}_{M^*\pi} \left[\sum_{k=1}^{n} r_k \right]$$

$$\Omega\left(\sqrt{\frac{D|S||A|}{n}} \right) \leq Rg^*(n) \leq \tilde{O}\left(D \sqrt{\frac{|S||A|}{n}} \right)$$



Functional Approximation

Model-based: $\mathcal{H}_{MB} \subseteq \{ S \times A \to \Delta(S \times \mathbb{R}) \}$

Model-free: $\mathcal{H}_V \subseteq \{ S \to \mathbb{R} \}$

$$\mathcal{H}_Q \subseteq \{ S \times A \to \mathbb{R} \}$$

$$V(s, \gamma) = \max_{\pi} \; E_{M\pi}\left[(1-\gamma) \sum_{n=0}^{\infty} \gamma^n r_n \,\middle|\, s_0 = s \right]$$

$$Q(s, a, \gamma) = \max_{\pi : \pi(\lambda) = a} E_{M\pi}\left[(1-\gamma) \sum_{n=0}^{\infty} \gamma^n r_n \,\middle|\, s_0 = s \right]$$

Linear Reinforcement Learning

$$\varphi : S \times A \to \mathbb{R}^d \qquad \theta^* \in \mathbb{R}^d \qquad \eta^* \in \mathbb{R}^{d \times S}$$

$$pr_S T(s,a) = \varphi(s,a)^t \eta^*$$

$$E[pr_{\mathbb{R}} T(s,a)] = \varphi(s,a)^t \theta^*$$

$$\mathrm{Reg}^*(n) \leq \tilde{O}\left(\sqrt{\frac{d^3 H}{n}}\right)$$

Learning the State Representation

$$\Phi \subseteq \{\varphi : (A \times O)^* \to S_\varphi\} \qquad \varphi^* \in \Phi$$

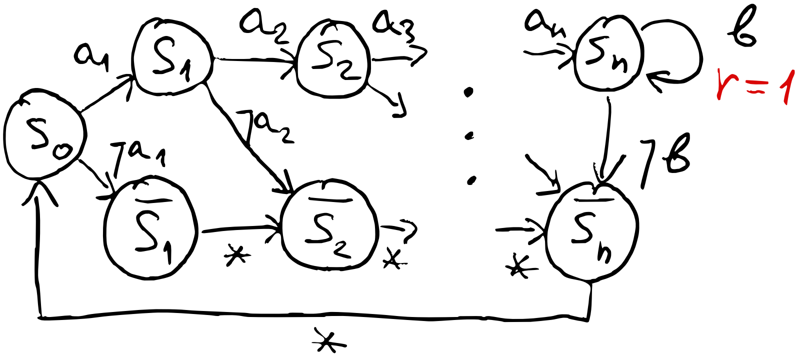
$$Re : (A \times O)^* \to \mathbb{R}$$

$$\Pr[\varphi^*(hao) = s_2, Re(hao) \in [x, y] \mid ha] =$$

$$T^*(s_2 \times [x, y] \mid \varphi^*(h), a)$$

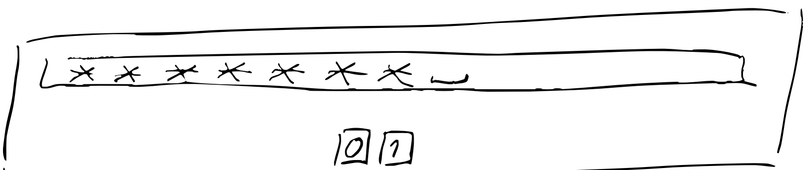
$$Rg^*(n) \leq \tilde{O}\left(D \max_{\varphi \in \Phi} |S_\varphi| \sqrt{\frac{|A||\Phi|}{n}}\right)$$

Regular Reinforcement Learning



$$|S| = 2n+1$$

$$D = n$$



✓ $r = 1$ forever

X $r = 0$ start over

$$XT(M) := \max_{x,y \in S} \mathbb{E}_{M\pi_0}\left[\min\{k \mid s_k = y\} \mid s_0 = x\right]$$