Optimal Best Markovian Arm Identification with Fixed Confidence

Vrettos Moulos
University of California Berkeley

NEURAL INFORMATION PROCESSING SYSTEMS

Problem Formulation

- Finite state space S, and reward function $f : S \to \mathbb{R}$.
- Family of Markov chains, parametrized by $\theta \in \Theta$.
- **Markov chain** with parameter θ, is driven by **initial distribution** q_θ, and **irreducible stochastic transition matrix** P_θ.
- **Stationary mean reward:** $\mu(\theta) = \sum_x f(x)\pi_\theta(x)$, where π_θ is the **stationary distribution** of P_θ.
- K **Markovian arms** $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$ with unique best stationary arm.
- The **reward process** corresponding to θ_a is $\{Y_n^a\}_{n \in \mathbb{Z}_{\geq 0}} = \{f(X_n^a)\}_{n \in \mathbb{Z}_{\geq 0}}$.
- **best arm:** $\{a^*(\boldsymbol{\theta})\} = \arg\max_a \mu(\theta_a)$.
- **Goal:** find $a^*(\boldsymbol{\theta})$ with probability at least $1 - \delta$, using as few samples as possible.

IID vs Markovian Rewards

- Reward distributions are state dependent, e.g. mean casino.

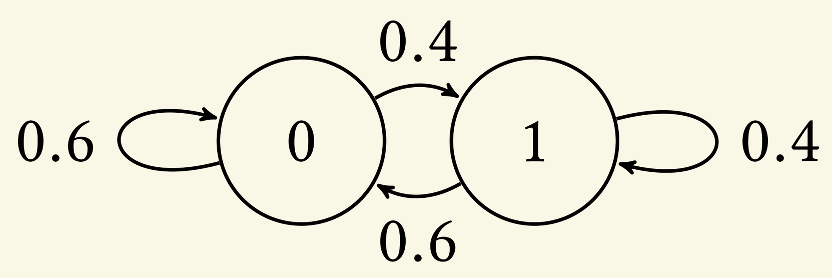


Figure: IID

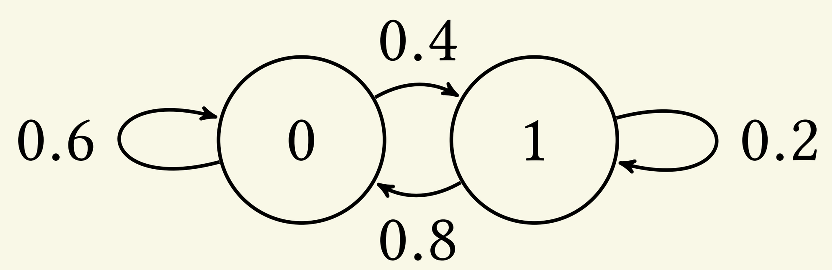


Figure: Markovian

δ-Probably-Correct Strategy

- A triple $\mathcal{A}_\delta = ((A_t)_{t \in \mathbb{Z}_{>0}}, \tau_\delta, \hat{a}_{\tau_\delta})$, with:
  - **sampling rule:** at time t select arm A_t based on the past observations.
  - **stopping rule:** at the stopping time τ_δ stop the data collection, and output estimate for best arm.
  - **decision rule:** our estimate $\hat{a}_{\tau_\delta}$ for the best arm.
  - **uniformly good:** $\mathbb{P}_{\boldsymbol{\lambda}}^{\mathcal{A}_\delta}(\hat{a}_{\tau_\delta} \neq a^*(\boldsymbol{\lambda})) \leq \delta$ for all $\boldsymbol{\lambda}$.

General Lower Bound

For any δ-PC sampling strategy $\mathcal{A}_\delta$ and any parameter configuration $\boldsymbol{\theta}$,

$$T^*(\boldsymbol{\theta}) \leq \liminf_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}[\tau_\delta]}{\log \frac{1}{\delta}},$$

where,

$$T^*(\boldsymbol{\theta})^{-1} = \sup_{\boldsymbol{w} \in \mathcal{M}_1([K])} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\theta})} \sum_{a=1}^K w_a \overline{D}(\theta_a \parallel \lambda_a),$$

and $\overline{D}(\theta \parallel \lambda)$ is the **Kullback-Leibler divergence rate** given by,

$$\overline{D}(\theta \parallel \lambda) = \sum_{x,y} \log \frac{P_\theta(x,y)}{P_\lambda(x,y)} \pi_\theta(x) P_\theta(x,y).$$

- **Change of measure** lemma using renewals:

$$D\left(\mathbb{P}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}|_{\mathcal{F}_{\tau_\delta}} \,\middle\|\, \mathbb{P}_{\boldsymbol{\lambda}}^{\mathcal{A}_\delta}|_{\mathcal{F}_{\tau_\delta}}\right) \leq \sum_{a=1}^K D\left(q_{\theta_a} \parallel q_{\lambda_a}\right) + \sum_{a=1}^K \left(\mathbb{E}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}[N_a(\tau_\delta)] + R_{\theta_a}\right) \overline{D}(\theta_a \parallel \lambda_a),$$

where $R_{\theta_a} = \mathbb{E}_{\theta_a}\left[\inf\{n > 0 : X_n^a = X_0^a\}\right]$, and $N^a(t) = \sum_{s=1}^t I_{\{A_s = a\}} - 1$.

- **Data processing inequality** and δ-PC:

$$D_2(\delta \parallel 1-\delta) \leq D\left(\mathbb{P}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}|_{\mathcal{F}_{\tau_\delta}} \,\middle\|\, \mathbb{P}_{\boldsymbol{\lambda}}^{\mathcal{A}_\delta}|_{\mathcal{F}_{\tau_\delta}}\right).$$

- Combine **multiple changes of measure** at once (Garivier and Kaufmann 2016).

Exponential Family of MCs

- **Exponential tilt:** use P as a **generator**, and for each $\theta \in \mathbb{R}$ set

$$\tilde{P}_\theta(x,y) = P(x,y)e^{\theta f(y)}.$$

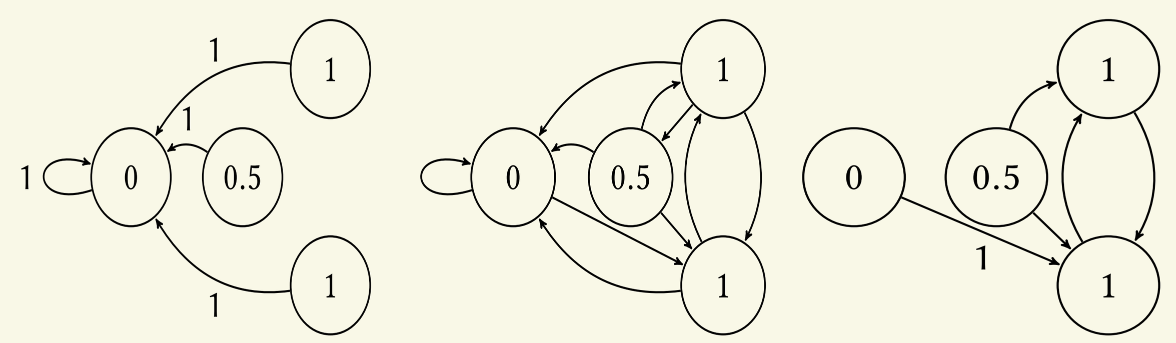
- **Perron-Frobenius theory:** The spectral radius $\rho(\theta)$ of $\tilde{P}_\theta$ is a simple eigenvalue, associated with a unique left u_θ and right v_θ eigenvectors, such that they are both positive, $\|u_\theta\|_1 = 1$, and $\langle u_\theta, v_\theta \rangle = 1$.
- **log-Perron-Frobenius eigenvalue:** $A(\theta) = \log \rho(\theta)$, serves as a **scaled cumulant generating function**.
- **Exponential family:**

$$P_\theta(x,y) = \frac{v_\theta(y)}{v_\theta(x)} e^{\theta f(y) - A(\theta)} P(x,y).$$

Conjugate Duality

- Let $M = \max_x f(x)$, $S_M = \{x : f(x) = M\}$, $m = \min_x f(x)$, $S_m = \{x : f(x) = m\}$, and **assume** that:
  1. The submatrix of P with rows and columns in S_M is irreducible.
  2. $\forall x \in S - S_M$, $\exists y \in S_M$ such that $P(x,y) > 0$.
  3. The submatrix of P with rows and columns in S_m is irreducible.
  4. $\forall x \in S - S_m$, $\exists y \in S_m$ such that $P(x,y) > 0$.
- **Conjugate Duality:**
  1. $A'(\theta) = \mu(\theta)$, and is strictly increasing in θ.
  2. $\lim_{\theta \to -\infty} A'(\theta) = m$, and $\lim_{\theta \to \infty} A'(\theta) = M$.

Example: $\theta = -\infty, \ \theta = 0, \ \theta = \infty$



$$\mathbb{R} \underset{(A')^{-1}(\mu)}{\overset{A'(\theta)}{\rightleftarrows}} (0,1)$$

Concentration for MCs

Let $\{X_n\}_{n \in \mathbb{Z}_{>0}}$ be a Markov chain driven by q and P, corresponding to the parameter $\theta = 0$. Then for $\mu \geq \mu(0)$,

$$\mathbb{P}_{(q,P)}\left(\frac{1}{n}\sum_{k=1}^n f(X_k) \geq \mu\right) \leq C e^{-n\overline{D}((A')^{-1}(\mu) \parallel 0)},$$

where $C = C(P, f)$, and if P is positive, then $C = \max_{x,y,z}\{P(y,z)/P(x,z)\}$.

- **Asymptotic analysis** of P_θ.
- **Uniform convergence:**

$$\sup_\theta \left|\frac{1}{n}\log \mathbb{E}_{(q,p)} e^{\theta \sum_{k=1}^n f(X_k)} - A(\theta)\right| \leq \frac{\log C}{n}.$$

(α, δ)-Track-and-Stop Strategy

- **Unknown optimal weights:** $\boldsymbol{w}^*(\boldsymbol{\theta}) = \boldsymbol{w}^*(\boldsymbol{\mu})$, where $\boldsymbol{\mu} = (\mu(\theta_1), \dots, \mu(\theta_K))$.
- **Empirical means:** $\hat{\mu}_a(t) = \frac{1}{N_a(t)} \sum_{n=1}^{N_a(t)} Y_n^a$
- **Sampling rule:** (Garivier and Kaufmann 2016)

$$A_{t+1} \in \begin{cases} \arg\min_a N_a(t), & \text{if } \exists a : N_a(t) < \sqrt{t} - K/2 \text{ (explore)}, \\ \arg\max_a \left\{w_a^*(\hat{\boldsymbol{\mu}}(t)) - \frac{N_a(t)}{t}\right\}, & \text{otherwise (track)}, \end{cases}$$

this way $\hat{\mu}_a(t) \overset{a.s.}{\to} \mu(\theta_a)$, $w_a^*(\hat{\boldsymbol{\mu}}(t)) \overset{a.s.}{\to} w_a^*(\boldsymbol{\mu})$, and so $\frac{N_a(t)}{t} \overset{a.s.}{\to} w_a^*(\boldsymbol{\mu})$.

- **Stopping rule:**
  $\tau_{\alpha,\delta} = \inf\left\{t : \exists a \, \forall b \neq a, \ Z_{a,b}(t) > 2\log\frac{Dt^\alpha}{\delta}\right\}$
  where $Z_{a,b}(t)$ is a statistic which for large values represents large confidence that $\mu(\theta_a) > \mu(\theta_b)$.
- **Decision rule:** $\hat{a}_{\tau_{\alpha,\delta}} = \arg\max_a \hat{\mu}_a(N_a(\tau_{\alpha,\delta}))$.
- **Guarantee:** For $\alpha > 1$ and $\delta \in (0,1)$, the (α,δ)-Track-and-Stop strategy is δ-PC, and

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\theta}}^{\mathcal{A}_\delta}[\tau_{\alpha,\delta}]}{\log \frac{1}{\delta}} \leq 4\alpha T^*(\boldsymbol{\theta}).$$