Ch6. General Decision Making

- Introduce a unified framework for decision making, including structured bandit problem, contextual bandit problem etc.

- Show that the Decision-Estimation Coefficient (DEC) and associated meta-algorithm (E2D) extend to this general framework and also show that boundness of the DEC is sufficient and necessary for low regret (constitute a fundamental limit)

Setting: Focus on a framework called Decision Making with Structured Observations (DMSO)

- for $t = 1, 2, \cdots, T$ round, the learner select a decision $\pi^t \in \Pi$, the decision space

- Nature selects reward $r^t \in \underbrace{\mathcal{R}}_{\text{reward space}}$ and observation $o^t \in \underbrace{\mathcal{O}}_{\text{observation space}}$ based on π^t

- Both r^t and o^t are observed by the learner

Two key Assumptions :

Assumption 1 (Stochastic Rewards and Observations) :

r^t and o^t are independently generated via

$$(r^t, o^t) \sim M^*(\cdot \mid \pi^t)$$

where $M^*: \Pi \to \Delta(\mathcal{R} \times \mathcal{O})$ is the underlying true model.

Assumption 2 (Realizability): The learner has access to a flexible enough model class $\mathcal{M}$, where $M^* \in \mathcal{M}$

(Here. $\mathcal{M}$ could be of linear models, neural networks, random forest, and other function approximation)

Objectives : For a model $M \in \mathcal{M}$, $\mathbb{E}^{M,\pi}(\cdot)$ is the expectation under $(r, o) \sim M(\pi)$, let

$$f^M(\pi) = \mathbb{E}^{M,\pi}(r)$$

be the mean reward function ; and let

$$\pi_M = \arg\max_{\pi \in \Pi} f^M(\pi)$$

be the optimal decision with maximal expected reward;

and also let

$$\mathcal{F}_{\mathcal{M}} = \{ f^M \mid M \in \mathcal{M} \}$$

be the induced class of mean reward functions.

Finally, for evaluation of the learner's performance, we consider regret to the optimal decision for M^*

$$\text{Reg} = \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t} \left[f^{M^*}(\pi_{M^*}) - f^{M^*}(\pi^t) \right]$$

where $p^t \in \Delta(\Pi)$ is the learner's distribution over decisions at round t. (Shorthand $f^* = f^{M^*}$, $\pi^* = \pi_{M^*}$)

Rmk: Compared to basic bandit problem, $r^t \sim M^*(\cdot \mid \pi^t)$ without the observations, and the mean reward is $\underbrace{\mathbb{E}(r \mid \pi)}_{f^*(\pi)}$ only, while regret is $\sum_{t=1}^{T} f^*(\pi^*) - \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}[f^*(\pi^t)]$

That is, observation provided extra information gain.

Sec 1. Some Examples

(Example 1) Structured Bandits. When $\mathcal{O} = \{\phi\}$, i.e. no observations, DMSO reduces to Structured Bandits problem. We can start with a set of models $\mathcal{M}$ and then define induced class $\mathcal{F}_{\mathcal{M}}$, serving as the class $\mathcal{F}$ of mean reward functions before. Different $\mathcal{F}_{\mathcal{M}}$ will include different structured bandit problems such as linear, nonparametric, etc.

(Example 2) Contextual Bandits. In Context Bandit, reward $r^t \sim M^*(\cdot \mid \pi^t, x^t)$ for some covariate x^t, and $f^*(x, \pi) = \mathbb{E}(r \mid x, \pi)$ for $r \sim M^*(\cdot \mid \pi, x)$.

Think of π^t as functions mapping x^t to an action in $\Pi = [A]$. On round t, the decision-maker selects a mapping $\pi^t : \mathcal{X} \to [A]$, and the context $o^t = x^t$ is observed for each round. This is basically observing x^t, then

Select $\pi^t(x^t) \in [A]$, (which is called behavioral decision rule in decision theory). Let $\mathcal{O} = \mathcal{X}$ be the space of context, $\Pi = [A]$ be the set of actions, and Π: $\mathcal{X} \to [A]$ be the space of decisions. Then $(r, x) \sim M(\cdot|\pi)$ has: $x \sim D^M$ for some context dsn, $r \sim R^M(\cdot|x, \pi(x))$ for some reward dsn R^M. Here, D^M for x is part of M (treating x as O, observation).

(Example 3) Online reinforcement learning. For online reinforcement learning, learner selects a randomized, non-stationary policy $\pi = (\pi_1, \ldots, \underset{\uparrow \pi_h}{\ldots}, \pi_H)$, $\underbrace{\pi_h : S \to \Delta(A)}_{\in \Pi_{rns}}$

Beginning from state $s_1 \sim d_1 \in \Delta(S)$, for $h = 1, \cdots H$

$$a_h \sim \pi_h(s_h)$$

$$r_h \sim R_h^M(s_h, a_h), \quad R_h^M : S \times A \to \Delta(R) \text{ (reward dsn)}$$

and $$s_{h+1} \sim \mathbb{P}_h^M(s_h, a_h), \quad \mathbb{P}_h^M : S \times A \to \Delta(S) \text{ (transition kernel)}$$

then $f^M(\pi) = \mathbb{E}^{M,\pi}\left(\sum_{h=1}^{H} r_h\right)$

stands for MDP $\{S, A, P_h^M, R_h^M\}$

Here, let's take $\Pi = \Pi_{rns}$, $r^t = \sum_{h=1}^{H} r_h^t$, and

$o^t = \tau^t = (s_1^t, a_1^t, r_1^t), \dots, (s_H^t, a_H^t, r_H^t)$, the trajectory of learning. Then

$$(r^t, o^t) = \left(\sum_{h=1}^{H} r_h^t, \tau^t\right) \sim \left(R_h^M, P_h^M \mid \underset{\pi}{A}\right)$$

can be considered as $M(\cdot \mid \pi^t)$ for some dsn M and $\pi^t \in \Pi_{rns}$.

Sec 2. DEC for General Decision Making

Optimally explore and make decisions for $\mathcal{M}$ is cornerstone to understanding the optimal statistical complexity for $\mathcal{M}$.

As seen before, any notion of complexity needs to capture

(i) Simple problems like multi-armed bandit

(ii) problems with structured feedback where observations or structures in the noise can provide extra information

(Decision Estimation Coefficient)

For $\mathcal{M}$, reference model $\hat{M} \in \mathcal{M}$ and $\gamma > 0$ (scalar). DEC for general decision making is

regret of decision

$$\mathrm{dec}_\gamma(\mathcal{M}, \hat{M}) = \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\Big[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D_H^2(M(\pi), \hat{M}(\pi)) \Big]$$

where $D_H^2(\mathbb{P}, \mathbb{Q}) = \int (\sqrt{p} - \sqrt{q})^2 d\nu$ information gain from obs.

for $\mathbb{P}, \mathbb{Q} \ll \nu$. Also, define

$$\mathrm{dec}_\gamma(\mathcal{M}) = \sup_{\hat{M} \in \mathrm{co}(\mathcal{M})} \mathrm{dec}_\gamma(\mathcal{M}, \hat{M})$$

where $\mathrm{co}(\mathcal{M})$ is the convex hull of $\mathcal{M}$. as did for $\mathrm{co}(\mathcal{F})$)

Rmk: Compared to structured bandit problem, the major difference is that instead of "$\max_{f \in \mathcal{F}}$" and considering $f(\pi_f) - f(\pi)$, i.e. restricts on a class of reward functions, here the general DEC is defined over $\mathcal{M}$ (model class for both reward and observations).

Also, rather than measuring the information gain via

$f(\pi) - \hat{f}(\pi)$, here we consider information gain from matching the dsn over rewards and observations of M and $\hat{M}$ (for learner's decision π), $\mathbb{E}_{\pi\sim p}[D_H^2(M(\pi), \hat{M}(\pi))]$

(i) incorporates observations o^t or τ^t in reinforcement learning

(ii) even for bandit problems. it measures the distance between dsns rather than the means.

Sec 2.1 E2D Algorithm for General Decision Making

Estimation to Decisions (E2D) for general decision making is readily extended from structured bandit problems.

Parameter $\gamma > 0$ (given)

For $t = 1, \cdots T$ do

Obtain $\hat{M}^t$ from online estimation oracle with

$(\pi^1, r^1, o^1), \cdots (\pi^{t-1}, r^{t-1}, o^{t-1})$

by minimizing $\mathrm{dev}_\gamma(M, \hat{M}^t)$. Compute

$$p^t = \arg\min_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D_H^2(M(\pi), \hat{M}(\pi))]$$

(*)

Sample decision $\pi^t \sim p^t$ and update estimation with (π^t, r^t, o^t).

Rmk: i) Rather attempting to estimate the reward function f^*, one estimate the underlying model M^* (details later)

ii) $$\text{OEO} \to \hat{M}^t \xrightarrow{(\ddagger)} p^t \to \pi^t \to \text{OEO}$$

$$\mathcal{H}^{t-1} = (\pi^1, r^1, o^1), \cdots (\pi^{t-1}, r^{t-1}, o^{t-1})$$

Proposition 1 (26 in the note): Running E2D, the regret is bdd by DEC and estimation error, which is defined here as

$$\text{Est}_H = \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}\left[D_H^2(M^*(\pi^t), \hat{M}^t(\pi^t))\right]$$

That is, for $\gamma > 0$, E2D admits

$$\text{Reg} \leq \sup_{\hat{M} \in \hat{\mathcal{M}}} \text{dec}_\gamma(\mathcal{M}, \hat{M}) \cdot T + \gamma \cdot \text{Est}_H$$

almost surely, where $\hat{\mathcal{M}}$ is any set s.t. $\hat{M}^t \in \hat{\mathcal{M}}$ for all $t \in [T]$

proof: $$\text{Reg} = \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}\left[f^*(\pi_{M^*}) - f^{M^*}(\pi^t)\right]$$

$$= \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t}\left[f^*(\pi_{M^*}) - f^{M^*}(\pi^t) \right]$$

$$- \gamma\, \mathbb{E}_{\pi^t \sim p^t}\left[D_H^2(M^*(\pi^t), \hat{M}^t(\pi^t)) \right]$$

$$+ \gamma\, \mathrm{Est}_H$$

For each t, as $M^* \in \mathcal{M}$ (Assumption 2)

$$\mathbb{E}_{\pi^t \sim p^t}\left[f^*(\pi_{M^*}) - f^{M^*}(\pi^t) \right] - \gamma\, \mathbb{E}_{\pi^t \sim p^t}\left[D_H^2(M^*(\pi^t), \hat{M}^t(\pi^t)) \right]$$

$$\leq \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi^t \sim p^t}\left[f^M(\pi_M) - f^M(\pi^t) \right] - \gamma\, \mathbb{E}_{\pi^t \sim p^t}\left[D_H^2(M(\pi^t), \hat{M}^t(\pi^t)) \right]$$

by def of p^t in E2D

$$= \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}\left[f^M(\pi_M) - f^M(\pi) - \gamma \cdot D_H^2(M(\pi), \hat{M}^t(\pi)) \right]$$

$= \mathrm{dec}_\gamma(\mathcal{M}, \hat{M}^t)$. Summing over t,

$$\mathrm{Reg} \leq \sup_{\hat{M} \in \hat{\mathcal{M}}} \mathrm{dec}_\gamma(\mathcal{M}, \hat{M}) \cdot T + \gamma\, \mathrm{Est}_H \quad \#$$

Rmk: One can optimize over γ above, to yield

$$\mathrm{Reg} \leq \inf_{\gamma} \left\{ \sup_{\hat{M} \in \hat{\mathcal{M}}} \mathrm{dec}_\gamma(\mathcal{M}, \hat{M}) \cdot T + \gamma \cdot \mathrm{Est}_H \right\}$$

$$\leq 2 \inf_{\gamma>0} \left\{ \sup_{\hat{M}\in\hat{\mathcal{M}}} \mathrm{dec}_\gamma(\mathcal{M}, \hat{M})\cdot T,\ \gamma\cdot \mathrm{Est}_H \right\}$$

For any finite class $\mathcal{M}$, the averaged exponential weights (*) algorithm with the log loss achieves $\mathrm{Est}_H \lesssim \log(|\mathcal{M}|/\delta)$ w.p. at least $1-\delta$. We can take $\hat{\mathcal{M}} = \mathrm{co}(\mathcal{M})$, WLOG.

Then, for any finite class, w.p. $1-\delta$,

$$\mathrm{Reg} \lesssim \mathrm{dec}_\gamma(\mathcal{M})\cdot T + \gamma \log(|\mathcal{M}|/\delta).$$

Review: Exponential Weights is a main online learning algorithm, applicable to finite class. At each time, the algorithm compute a dsn $q^t \in \Delta(\mathcal{F})$, where

$\Delta(\mathcal{F}) \propto \exp\{-\eta \sum_{i=1}^{t-1} \ell(f(x^i), y^i)\}$ for $\eta>0$. Specifically,

for $t=1,\dots,T$ do
  compute q^t above.
  Let $\hat{f}^t = \mathbb{E}_{f\sim q^t}(f)$
  Observe (x^t, y^t), incur $\ell(\hat{f}^t(x^t), y^t)$

loss function

Sec 2.2 Examples

(Example 4, Multi-armed bandit with Gaussian-rewards) Let $\Pi = [A]$, $R = \mathbb{R}$, $O = \{\phi\}$ (no observations). Define

$$\mathcal{M}_{MAB-G} = \{ M : M(\pi) = N(f(\pi), 1) \\ f : \Pi \to [0,1] \}$$

Consider $\hat{M} \in \mathcal{M}$. From results of multi-armed bandit problem, $dec_\gamma(\mathcal{F}, \hat{f}) \propto \frac{A}{\gamma}$, it is thus sufficient to argue that the squared Hellinger-distance for Gaussian reduces to square difference between the means.

In fact, $D_H^2(M(\pi), \hat{M}(\pi))$

$$\le D_{KL}(M(\pi) \| \hat{M}(\pi))$$

(*)

$$= \frac{1}{2}(f^M(\pi) - f^{\hat{M}}(\pi))^2$$

On the other hand, $D_H^2(M(\pi), \hat{M}(\pi))$

$$= 1 - \exp\{-\frac{1}{8}(f^M(\pi) - f^{\hat{M}}(\pi))^2\}$$

So that $D_H^2(M(\pi), \hat{M}(\pi))$

$$\geq c \cdot \left(f^M(\pi) - f^{\hat{M}}(\pi)\right)^2$$

as $1 - e^{-x} \geq (1 - e^{-1})x$ for $x \in [0,1]$. Hence these two inequalities give

$$\mathrm{dec}_\gamma(\mathcal{M}_{MAB\text{-}G}) \propto \frac{A}{\gamma}$$

By Proposition 1, then for $\mathcal{M}_{MAB\text{-}G}$.

$$\mathrm{Reg} \lesssim \frac{AT}{\gamma} + \gamma \, \mathrm{Est}_H$$

Review: (*) As densities of any two dens $\mathbb{P}$ and $\mathbb{Q}$, p. q. satisfies $|p-q| = |(\sqrt{p}+\sqrt{q})(\sqrt{p}-\sqrt{q})| \geq (\sqrt{p}-\sqrt{q})^2$ and $TV(\mathbb{P}, \mathbb{Q}) = \frac{1}{2}\int |p-q| \, d\nu$, it follows that

$$H^2(\mathbb{P}, \mathbb{Q}) \leq TV(\mathbb{P}, \mathbb{Q})$$

so $D_H^2(\mathbb{P}, \mathbb{Q}) \leq D_{KL}(\mathbb{P} \| \mathbb{Q})$ by Pinsker's inequality.

Let $\gamma = \sqrt{AT/\mathrm{Est}_H}$, then

$$\mathrm{Reg} \lesssim \sqrt{AT \cdot \mathrm{Est}_H}$$

(Example 5. Bandit problems with structured noise)

Let $\Pi = [A]$, $\mathcal{R} = \mathbb{R}$, $\mathcal{O} = \{\phi\}$. Define

$$\mathcal{M}_{MAB-SN} = \{ M_1, \dots M_A \} \cup \{\hat{M}\}$$

where $M_i(\pi) := \mathcal{N}(\frac{1}{2}, 1)$ for $\pi \neq i$ and $M_i(\pi) = \mathrm{Ber}(3/4)$ for $\pi = i$. Define $\hat{M}(\pi) = \mathcal{N}(\frac{1}{2}, 1)$ for all $\pi \in \Pi = [A]$.

The valuable information contained in the reward dsn. is reflected in the Hellinger divergence, which attains the maximum when comparing a continuous dsn to a discrete one,

$$D_H^2(M_i(\pi), \hat{M}(\pi)) = 2\mathbb{I}(\pi = i)$$

Notice that the maximum over M in the definition of $\mathrm{dec}_\gamma(\mathcal{M}_{MAB-SN}, \hat{M})$ is NOT attained at $M = \hat{M}$, as both the divergence and "regret" term will be zero, regardless of P. Take $p = \text{unif}[A]$. for any $M \in \{M_1, \cdots M_A\}$

$$\mathbb{E}_{\pi \sim p}[f^M(\pi_M) - f^M(\pi)] = \left(1 - \frac{1}{A}\right)\left(\frac{3}{4} - \frac{1}{2}\right)$$

as $f^M(\pi) = \mathbb{E}(\mathbb{E}(r \mid \pi)) = \left(1 - \frac{1}{A}\right) \cdot \frac{1}{2} + \frac{1}{A} \cdot \frac{3}{4}$

and $\pi_M = \arg\max_{\Pi = [A]} f^M(\pi) = 3/4$

and $\text{dec}_\gamma(\mathcal{M}, \hat{M}) \lesssim \left(1 - \frac{1}{A}\right)\left(\frac{3}{4} - \frac{1}{2}\right) - \gamma \frac{2}{A}$

$$\lesssim \mathbb{I}\left(\gamma \le \frac{A}{4}\right)$$

as $\sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}[D_H^2(M_i(\pi), \hat{M}(\pi))] = \frac{1}{A}$.

Hence $\text{dec}_\gamma(\mathcal{M}_{MAB-SN}, \hat{M}) \lesssim \mathbb{I}\left(\gamma \le \frac{A}{4}\right)$

With Proposition 1,

$$\text{Reg} \lesssim \mathbb{I}\left(\gamma \le \frac{A}{4}\right) \cdot T + \gamma \cdot \text{Est}_H$$

Let $\gamma = A$, we have

$$\text{Reg} \lesssim A \cdot \text{Est}_H.$$

Sec 2.3 Some structural Properties of DEC

Without proofs, we list a few structural properties of DEC, which are useful in practice for computing the DEC of specific model class.

Property 1 (Proposition 41 in Note. Square loss is sufficient for structural bandit problem)

Consider any structural bandit problem with decision space Π, function class $\mathcal{F} \subseteq (\Pi \to [0,1])$, and $\mathcal{O} = \{\phi\}$ (no observs). Denote

$$\mathcal{M}_{\mathcal{F}} = \{M \mid f^M \in \mathcal{F},\ M(\pi) \text{ is } 1\text{-sub-Gaussian } \forall \pi\}$$

Let

$$\mathrm{dec}_\gamma^{sq}(\mathcal{F}, \hat{f}) = \inf_{p \in \Delta(\Pi)} \sup_{\mathcal{F}} \mathbb{E}_{\pi \sim p}\left[f(\pi_f) - f(\pi) - \gamma\left(f(\pi) - \hat{f}(\pi)\right)^2\right]$$

Then.

$$\mathrm{dec}_{c_1\gamma}^{sq}(\mathcal{F}) \le \mathrm{dec}_\gamma(\mathcal{M}_{\mathcal{F}}) \le \mathrm{dec}_{c_2\gamma}^{sq}(\mathcal{F})$$

for $c_1, c_2 \geq 0$.

Property 2 (Proposition 42 in Note, Filtering irrelevant information)

Adding observations that are unrelated to the model does not change the DEC!

Consider $\mathcal{M}$ with observation space $\mathcal{O}_1$, and a class of conditional dsn $\mathcal{D}$ over another observation space $\mathcal{O}_2$, where $\forall D \in \mathcal{D}$ has $D(\pi) \in \Delta(\mathcal{O}_2)$. For $M \in \mathcal{M}$, and $D \in \mathcal{D}$, let $(M \otimes D)(\pi)$ be the model that given $\pi \in \Pi$, $(r_1, o_1) \sim M(\pi)$, $o_2 \sim D(\pi)$, then $(r, (o_1, o_2))$ is obtained. Set $\mathcal{M} \otimes \mathcal{D} = \{M \otimes D : M \in \mathcal{M}, D \in \mathcal{D}\}$

Then for $\forall \hat{M} \in \mathcal{M}$, $\hat{D} \in \mathcal{D}$

$$\operatorname{dec}_\gamma(\mathcal{M} \otimes \mathcal{D}, \hat{M} \otimes \hat{D}) = \operatorname{dec}_\gamma(\mathcal{M}, \hat{M})$$

Property 3 (Proposition 43 in Note, Data Processing)

Passing observations through a channel never reduce

DEC. Consider $\mathcal{M}$ with $\mathcal{O}$. Let

$$\rho : \mathcal{O} \to \mathcal{O}'$$

be a given mapping. Define $\rho \circ M$ as the model that given decision π, sample $(r, o) \sim M(\pi)$ or $(r, o) \sim M(\cdot \mid \pi)$, then observes/ predicts $(r, \rho(o))$. Let

$$\rho \circ \mathcal{M} = \{\rho \circ M \mid M \in \mathcal{M}\}$$

For all $\hat{M} \in \mathcal{M}$, we have

$$\operatorname{dec}_\gamma(\mathcal{M}, \hat{M}) \leq \operatorname{dec}_\gamma(\rho \circ \mathcal{M}, \rho \circ \hat{M})$$

This is an immediate consequence of the data processing inequality for Hellinger - distance that

$$D_H^2((\rho \circ M)(\pi), (\rho \circ \hat{M})(\pi)) \leq D_H^2(M(\pi), \hat{M}(\pi)).$$

Sec 2.4 Online Estimation with D_H^2

Eestimation of model M is more challege comparing to regression problem, such as estimating the reward function. As we have seen before, estimating M^* w.r.t the Hellinger distance can be solved using online conditional density estimation with log loss.

Given (π^t, r^t, o^t), log loss for M is

$$\ell_{\log}^t(M) = \log\left(\frac{1}{m^M(r^t, o^t \mid \pi^t)}\right)$$

where $m^M(\cdot \mid \pi)$ is the conditional density for (r, o) under model M. Define

$$\mathrm{Reg}_{KL} = \sum_{t=1}^{T} \ell_{\log}^t(\hat{M}^t) - \inf_{M \in \mathcal{M}} \sum_{t=1}^{T} \ell_{\log}^t(M)$$

Then, a bound on the log-loss regret yields an immediate bound on the Hellinger estimation error as follows.

Lemma 1 (21 in lecture): For any online estimation algorithm such as averaged weights, whenever Assumption 2 holds,

$$\mathbb{E}[\mathrm{Reg}_{KL}] \geq \mathbb{E}\left[\sum_{t=1}^{T} D_{KL}\left(M^*(\pi^t) \,\|\, \hat{M}^t(\pi^t)\right)\right]$$

So that

$$\mathbb{E}[\mathrm{Est}_H] \leq \mathbb{E}[\mathrm{Reg}_{KL}] \qquad (\ddagger)$$

Also, $\forall \delta \in (0,1)$, w.p. at least $1-\delta$

$$\mathrm{Est}_H \leq \mathrm{Reg}_{KL} + \underbrace{2\log(1/\delta)}_{} \qquad (\ddagger\ddagger)$$

Does not scale with T! due to mgf of log-loss.

Proof: By assumption,

$$\sum_{t=1}^{T} \ell^t_{\log}(\hat{M}^t) - \sum_{t=1}^{T} \ell^t_{\log}(M^*) \leq \mathrm{Reg}_{KL}$$

So by Assumption 2 that $M^* \in \mathcal{M}$,

$$\sum_{t=1}^{T} \mathbb{E}\left[D_{KL}\left(M^*(\pi^t) \,|\, \hat{M}^t(\pi^t)\right)\right] \leq \mathbb{E}[\mathrm{Reg}_{KL}]$$

By Definition of Est_H, and fact that $D_H^2(\mathbb{P}, \mathbb{Q}) \leq D_{KL}(\mathbb{P} \| \mathbb{Q})$, $(\ddagger)$ follows.

To prove (##), we employ the tail bound for martingales (Review: $\forall$ real-valued r.v. $(X_t)_{t\le T}$ adapted to Filtration $(\mathcal{F}_t)_{t\le T}$, w.p. $1-\delta$, $\forall T'\le T$

$$\sum_{t=1}^{T'} X_t \le \sum_{t=1}^{T'} \log\left[\mathbb{E}_{t-1}\left(e^{X_t}\right)\right] + \log(1/\delta))$$

Define $Z_t = \frac{1}{2}\left(\ell_{\log}^t(\hat{M}^t) - \ell_{\log}^t(M^*)\right)$.

Applying tail bound for martingales to $(-Z_t)_{t\le T}$ w.p. at least $1-\delta$,

$$\sum_{t=1}^{T} -\log\left(\mathbb{E}_{t-1}\left(e^{-Z_t}\right)\right) \le \sum_{t=1}^{T} Z_t + \log(1/\delta)$$

$$= \frac{1}{2}\sum_{t=1}^{T}\left(\ell_{\log}^t(\hat{M}^t) - \ell_{\log}^t(M^*)\right) + \log(1/\delta)$$

Fix t, define

$$z^t = (r^t, o^t),$$

let $\nu(\cdot \mid \pi)$ be any conditional dominating measure for $m^{\hat{M}^t}$ and m^{M^*}. Notice

$$\mathbb{E}_{t-1}\left(e^{-Z_t} \mid \pi^t\right) = \mathbb{E}_{t-1}\left[\sqrt{\frac{m^{\hat{M}^t}(z_t \mid \pi^t)}{m^{M^*}(z_t \mid \pi^t)}} \;\middle|\; \pi^t\right]$$

$$= \int m^{M^*}(y|\pi^t) \sqrt{\frac{m^{\hat{M}^t}(y|\pi^t)}{m^{M^*}(y|\pi^t)}}\, \nu(dy|\pi^t)$$

$$= \int \sqrt{m^{M^*}(y|\pi^t)\, m^{\hat{M}^t}(y|\pi^t)}\, \nu(dy|\pi^t)$$

$$= 1 - \frac{1}{2} D_H^2\left(M^*(\pi^t), \hat{M}^t(\pi^t)\right)$$

Hence, $\mathbb{E}_{t-1}(e^{-Z_t}) = 1 - \frac{1}{2}\mathbb{E}_{t-1}\left[D_H^2\left(M^*(\pi^t), \hat{M}^t(\pi^t)\right)\right]$

and by $-\log(1-x) \geq x$ for $x \in [0,1]$,

$$\frac{1}{2}\sum_{t=1}^{T} \mathbb{E}_{t-1}\left[D_H^2\left(M^*(\pi^t), \hat{M}^t(\pi^t)\right)\right]$$

$$\leq \frac{1}{2}\sum_{t=1}^{T}\left(\ell_{\log}^t(\hat{M}^t) - \ell_{\log}^t(M^*)\right) + \log(1/\delta)$$

#

Rmk: Lemma I is useful as regret minimization w.r.t. log-loss is well-studied, such as averaged weights algorithm which admits

$$\mathrm{Reg}_{KL} \leq \log|\mathcal{M}| \text{ for finite class } \mathcal{M}.$$

Also. for linear model where $m^M(r,o|\pi) = \langle \phi(r,o,\pi), \theta \rangle$ for some feature map $\phi \in \mathbb{R}^d$.

$$\mathrm{Reg}_{KL} = O(d \log(T))$$

Sec 3. Optimality for General Decision Making — DEC: lower bound on Regret

Classical question: for a given class of models $\mathcal{M}$. What is the best regret that can be achieved by ANY algorithm?

Answer: Minimax optimality — for a model class $\mathcal{M}$. define minimax regret as

$$\mathfrak{M}(\mathcal{M},T) = \inf_{p^1,\dots,p^T} \sup_{M^* \in \mathcal{M}} \mathbb{E}^{M^*,p}[\mathrm{Reg}(T)]$$

where: i) $p^t = p^t(\cdot | \mathcal{H}^{t-1})$ is the algorithm for step t as a function of history $\mathcal{H}^{t-1}$

ii) $\mathrm{Reg}(T)$ makes its dependence on T explicitly.

An algorithm is <u>minimax optimal</u> if it achieves $\mathfrak{M}(\mathcal{M}, T)$ up to a constant free from $\mathcal{M}$ and T.

Sec 3.1. Constrained DEC

How to lower bound the minimax regret for any model class $\mathcal{M}$ in terms of DEC for $\mathcal{M}$?

Working on "<u>constrained DEC</u>" instead of $\mathrm{dec}_\gamma(\mathcal{M})$ in Proposition 1, which is called the <u>offset DEC</u>. Here for $\varepsilon > 0$, Constrained DEC is defined by

$$\mathrm{dec}_\varepsilon^c(\mathcal{M}, \hat{M}) = \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \left\{ \mathbb{E}_{\pi \sim p}\left[f^M(\pi_M) - f^M(\pi)\right] \,\middle|\, \mathbb{E}_{\pi \sim p}\left[D_H^2(M(\pi), \hat{M}(\pi))\right] \le \varepsilon^2 \right\},$$

where $\mathrm{dec}_\varepsilon^c(\mathcal{M}) = \sup_{\hat{M} \in \mathrm{co}(\mathcal{M})} \mathrm{dec}_\varepsilon^c(\mathcal{M} \cup \{\hat{M}\}, \hat{M})$

Rmk: Similar to $\mathrm{dec}_\gamma(\mathcal{M})$, instead of subtracting the information gain due to observations, $\mathrm{dec}_\varepsilon^c(\mathcal{M})$ puts a hard constraint on the information gain. Both of them bias the max learner/player towards model where the gain is small.

Offset/traditional DEC can be viewed as a Lagrangian relaxation of DEC with constraints. as

$$\mathrm{dec}^c_\varepsilon(\mathcal{M}, \hat{M}) = \inf_{p\in\Delta(\Pi)} \sup_{M} \Big\{ \mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] \ \Big|\ \mathbb{E}_{\pi\sim p}[D_H^2(M(\pi), \hat{M}(\pi))] \le \varepsilon^2 \Big\}$$

$$= \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \inf_{\gamma\ge 0} \Big\{ \mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)] - \gamma\big(\mathbb{E}_{\pi\sim p}[D_H^2(M(\pi), \hat{M}(\pi))] - \varepsilon^2\big) \Big\} \vee 0$$

$$\le \inf_{\gamma\ge 0} \inf_{p\in\Delta(\Pi)} \sup_{M} \Big\{ \underline{\mathbb{E}_{\pi\sim p}[f^M(\pi_M) - f^M(\pi)]} \underline{- \gamma\cdot\big(\mathbb{E}_{\pi\sim p}[D_H^2(M(\pi), \hat{M}(\pi))] - \varepsilon^2\big)} \Big\} \vee 0$$

$$= \inf_{\gamma\ge 0} \big\{ \underline{\mathrm{dec}_\gamma(\mathcal{M}, \hat{M})} + \gamma\varepsilon^2 \big\} \vee 0$$

It is easy to see that $\mathrm{dec}_\gamma(\mathcal{M}) \lesssim \mathrm{dec}^c_{\gamma^{-1/2}}(\mathcal{M})$

Rmk: Some classes the constrained DEC is meaningfully smaller than the offset DEC. However, if we restrict to a "localized" sub-class of models that are not "too far" from $\hat{M}$, we may have

Proposition 2 (26 in Note): Given a model $\hat{M}$ and parameter α, define

$$\mathcal{M}_\alpha(\hat{M}) = \{M \in \mathcal{M} : f^{\hat{M}}(\pi_{\hat{M}}) \geq f^M(\pi_M) - \alpha\}$$

For all $\varepsilon > 0$ and $\gamma \geq c_1/\varepsilon$,

$$\mathrm{dec}_\varepsilon^c(\mathcal{M}) \leq c_3 \sup_{\gamma \geq c_1/\varepsilon} \sup_{\hat{M} \in \mathrm{co}(\mathcal{M})} \mathrm{dec}_\gamma\left(\mathcal{M}_{\alpha(\varepsilon,\gamma)}(\hat{M}), \hat{M}\right)$$

with $\alpha(\varepsilon, \gamma) = c_2 \cdot \gamma \varepsilon^2$.

(The lengthy proof is referred to Foster, Golowich & Han, 2023)

Key message of Proposition 2 is that for well-behaved model class such as multi-armed bandits, linear bandits,

$\mathrm{dec}_\gamma(\mathcal{M}_{\alpha(\varepsilon;\gamma)}(\hat{M}), \hat{M}) \approx \mathrm{dec}_\gamma(\mathcal{M}, \hat{M})$

whenever $\mathrm{dec}_\gamma(\mathcal{M}, \hat{M}) \approx \gamma\varepsilon^2$. i.e, localisation does not change the complexity. So. lower bound in terms of constrained DEC immediately implies that in terms of the offset DEC. (Though refined E2D may lead to tighter upper bound for some cases).

Sec 3.2 Lower Bound

Proposition 3 (20 in Note, DEC lower bound) : Let $\varepsilon_T = \frac{c}{\sqrt{T}}$ for $c>0$ sufficiently small. For all T s.t.

$$\mathrm{dec}^c_{\varepsilon_T}(\mathcal{M}) \geq 10\,\varepsilon_T$$

for any algorithm. $\exists$ $M \in \mathcal{M}$ that

$$\mathbb{E}[\mathrm{Reg}(T)] \gtrsim \mathrm{dec}^c_{\varepsilon_T}(\mathcal{M}) \cdot T.$$

Rmk: ① For any algorithm & model class $\mathcal{M}$. the optimal.

regret must scale with the constrained DEC in the worst-case. For example, by Example 4. (multi-armed bandit with A actions).

$$\mathrm{dec}_\gamma(\mathcal{M}, \hat{M}) \lesssim \frac{A}{\gamma}$$

By results in Sec 3.1

$$\mathrm{dec}^c_\varepsilon(\mathcal{M}, \hat{M}) = \inf_\gamma \{\mathrm{dec}_\gamma(\mathcal{M}, \hat{M}) + \gamma\varepsilon^2\} \vee 0$$

So $\mathrm{dec}^c_\varepsilon(\mathcal{M}) \propto \varepsilon\sqrt{A}$. Then.

$$\mathbb{E}[\mathrm{Reg}(T)] \gtrsim \sqrt{AT} \quad \text{for } \varepsilon = \frac{c}{\sqrt{T}}.$$

(2) Combining Propositions 2 & 3, we have

Corollary 1 (1 in Note, lower bound based on localized offset DEC): Fix $T \in \mathbb{N}$. for any algorithm. there exists model $M \in \mathcal{M}$ for which.

$$\mathbb{E}[\mathrm{Reg}(T)] \gtrsim \sup_{\gamma \geq \sqrt{T}} \sup_{\hat{M} \in \mathrm{co}(\mathcal{M})} \mathrm{dec}_\gamma(\mathcal{M}_{\alpha(T,\gamma)}(\hat{M}), \hat{M})$$

with $\alpha(T, \gamma) = c \cdot \gamma / T$.

③ In Foster, Golowich & Han (2023), the authors design an algorithm based on a refined variant of E2D, s.t. the upper bound on regret is based on the constrained DEC

Proposition 4 (29 in Note): For a finite class $\mathcal{M}$.
$\bar{\varepsilon}_T = c \cdot \sqrt{\frac{\log(|\mathcal{M}|/\delta)}{T}}$ with sufficiently small c. Under technical conditions, $\exists$ an algorithm s.t

$$\mathbb{E}[\mathrm{Reg}(T)] \lesssim \mathrm{dec}^{c}_{\bar{\varepsilon}_T}(\mathcal{M}) \cdot T .$$

w.p. at least $1-\delta$.

(Though there exists a $\log|\mathcal{M}|$ gap, for class with finite $\log|\mathcal{M}|$, $\mathrm{dec}^{c}_{\varepsilon}$ is necessary & sufficient to lower bound regret)

Proof of Proposition 3: Basic idea of establishing any lower bound is similar: finding a pair of models M and $\hat{M}$ s.t.

i) any algorithm achieving low regret must be able to distinguish M and $\hat{M}$

ii) M and $\hat{M}$ are difficult to distinguish statistically

i.e, some information-theoretic difference between them is small

$\Rightarrow$ Algorithm must have large regret on either M or $\hat{M}$ (Similar idea in Hájek 1973)

Some simplifications:

1) $\exists$ C s.t $D_{KL}(M(\pi) \| M'(\pi)) \leq C D_H^2(M(\pi), M'(\pi))$ for all $M, M' \in \mathcal{M}$ and $\pi \in \Pi$

2) Rather than proving a lower bound scaling with $\mathrm{dec}_\varepsilon^c(\mathcal{M}) = \sup_{\hat{M} \in \mathrm{co}(\mathcal{M})} \mathrm{dec}_\varepsilon^c(\mathcal{M} \cup \{\hat{M}\}, \hat{M})$, one focuses on a weaker one that scales with $\sup_{\hat{M} \in \mathcal{M}} \mathrm{dec}_\varepsilon^c(\mathcal{M}, \hat{M})$

Fix T and an algorithm, defined by a sequence of mappings $p^1, \dots, p^T$ where $p^t = p^t(\cdot \mid \mathcal{H}^{t-1})$. Let $\mathbb{P}^M$ denote the dsn over $\mathcal{H}^T$ for the algorithm when M is the true model, and denote $\mathbb{E}^M$ the expectation.

Each p^t is a RV as a fction of $\mathcal{H}^{t-1}$, we can consider its expected value under M. For any $M \in \mathcal{M}$, let

$$p_M = \mathbb{E}^M\left[\frac{1}{T}\sum_{t=1}^{T} p^t\right] \in \Delta(\Pi)$$

be the algorithm's average action dsn when M is the true model. Our goal is to show that we can find a model M for which the algorithm's regret is at least as large as the lower bound $\sup\limits_{\gamma \geq \sqrt{T}} \sup\limits_{\hat{M} \in \mathrm{co}(\mathcal{M})} \mathrm{dec}_\gamma(\mathcal{M}_{\alpha(T,\gamma)}(\hat{M}), \hat{M})$.

Fix $\varepsilon > 0$ and arbitrary model $\hat{M} \in \mathcal{M}$, set

$$M = \arg\max_{\mathcal{M}} \left\{ \underbrace{\mathbb{E}_{\pi \sim p_{\hat{M}}}\left[f^M(\pi_M) - f^M(\pi)\right]}_{\mathrm{dec}^c_\varepsilon} \;\middle|\; \mathbb{E}_{\pi \sim p_{\hat{M}}}\left[D_H^2(M(\pi), \hat{M}(\pi))\right] \leq \varepsilon^2 \right\} \quad (\ddagger)$$

Model M should be considered as the "worst-case alternative" to $\hat{M}$, but only for the algorithm fixed now. Next, we'll show that the algorithm needs to have large regret on either M or $\hat{M}$. To this end, define $g^M(\pi) = f^M(\pi_M) - f^M(\pi)$, we will establish:

① for all models M, (as $Reg(T) = \sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t} \cdots$)

$$(1) \quad \frac{1}{T} \mathbb{E}^M [Reg(T)] = \mathbb{E}_{\pi \sim p_M} [g^M(\pi)]$$

So, to prove lower bound on Reg, we need to show that either $\mathbb{E}_{\pi \sim p_M}[g^M(\pi)]$ or $\mathbb{E}_{\pi \sim p_{\hat{M}}}[g^{\hat{M}}(\pi)]$ is large.

② $\mathbb{E}_{\pi \sim p_{\hat{M}}} [g^M(\pi)] \geq \mathrm{dec}_\varepsilon^c(\mathcal{M}, \hat{M}) := \Delta$ (2)

by the definition of $\mathrm{dec}_\varepsilon^c$ and by the construction of M above in (‡), M is the best response to a potentially sub-optimal choice $p_{\hat{M}}$. Then, it remains to fill in the gap that g^M is about M and $p_{\hat{M}}$ is about $\hat{M}$.

③ Using the chain rule of KL-divergence ($P_{\hat{M}}$ and P_M)

$$D_{KL}(\mathbb{P}^{\hat{M}} \| \mathbb{P}^M) = \mathbb{E}^{\hat{M}} \left[\sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t} D_{KL}(\hat{M}(\pi^t) \| M(\pi^t)) \right]$$

simplication ①

$$\le C \cdot \mathbb{E}^{\hat{M}}\left[\sum_{t=1}^{T} \mathbb{E}_{\pi^t \sim p^t} D_H^2(\hat{M}(\pi^t), M(\pi^t))\right]$$

$$= CT \cdot \mathbb{E}_{\pi \sim p_{\hat{M}}}\left[D_H^2(\hat{M}(\pi), M(\pi))\right]$$

Here, the first equality follows from the chain rule.

Apply the chain rule to sequence $\pi^1, z^1, \cdots \pi^T, z^T$, with $z^t = (r^t, o^t)$. Then

$$D_{KL}(\mathbb{P}^{\hat{M}} \| \mathbb{P}^{M})$$

$$= \mathbb{E}^{\hat{M}}\left[\sum_{t=1}^{T} D_{KL}(\mathbb{P}^{\hat{M}}(z^t | \mathcal{H}^{t-1}, \pi^t) \| \mathbb{P}^{M}(z^t | \mathcal{H}^{t-1}, \pi^t)) + D_{KL}(\mathbb{P}^{\hat{M}}(\pi^t | \mathcal{H}^{t-1}) \| \mathbb{P}^{M}(\pi^t | \mathcal{H}^{t-1}))\right]$$

$$= \mathbb{E}^{\hat{M}}\left[\sum_{t=1}^{T} D_{KL}(\hat{M}(\pi^t) \| M(\pi^t))\right]$$

as π^t is free from the model if conditional on $\mathcal{H}^{t-1}$.

[Review: Chain Rule for KL Divergence:

Let $(\mathcal{X}_1, \mathcal{F}_1), \cdots (\mathcal{X}_n, \mathcal{F}_n)$ be a sequence of measurable spaces. Let $\mathcal{X}^i = \prod_{t=1}^{i} \mathcal{X}_t$, $\mathcal{F}^i = \bigotimes_{t=1}^{i} \mathcal{F}_t$

For each i, let $\mathbb{P}^i(\cdot|\cdot)$ and $\mathbb{Q}^i(\cdot|\cdot)$ be probability kernels from $(\mathcal{X}^{i-1}, \mathcal{F}^{i-1})$ to $(\mathcal{X}^i, \mathcal{F}^i)$.

Let $\mathbb{P}$ and $\mathbb{Q}$ be the dsns of $X_1, \dots X_n$ under $X_i \sim \mathbb{P}^i(\cdot | X_{1:i-1})$ and $X_i \sim \mathbb{Q}^i(\cdot | X_{1:i-1})$, respectively.

Then, it holds that

$$D_{KL}(\mathbb{P} \| \mathbb{Q}) = \mathbb{E}_{\mathbb{P}}\left[\sum_{i=1}^{n} D_{KL}(\mathbb{P}^i(\cdot | X_{1:i-1}) \| \mathbb{Q}^i(\cdot | X_{1:i-1}))\right]$$

An easier way to understand chain rule of KL divergence is considering $\mathbb{P}(X,Y)$, $\mathbb{Q}(X,Y)$, then

$$D_{KL}(\mathbb{P}(X,Y) \| \mathbb{Q}(X,Y)) = \mathbb{E}_{\mathbb{P}}[D_{KL}(\mathbb{P}(Y|X=x) \| \mathbb{Q}(Y|X=x))] + D_{KL}(\mathbb{P}(X) \| \mathbb{Q}(X))$$

$$= D_{KL}(\mathbb{P}(Y|X) \| \mathbb{Q}(Y|X)) + D_{KL}(\mathbb{P}(X) | \mathbb{Q}(X))$$

(in $\mathrm{dec}_\varepsilon^c$)

Now, we can choose $\varepsilon = c_1 / \sqrt{cT}$ for $c_1 > 0$ sufficiently small s.t. (3) $TV^2(\mathbb{P}^{\hat{M}}, \mathbb{P}^M) \le D_{KL}(\mathbb{P}^M \| \mathbb{P}^M) \le \frac{1}{100}$.

That is, with non-trivial probability, the algorithm fixed here fails to separate M and $\hat{M}$.

④ Finally, as $r \in [0,1]$,

$$\mathbb{E}_{\pi \sim p_{\hat{M}}}\left[\underset{\|}{r}\; f^{M}(\pi) - f^{\hat{M}}(\pi)\right]$$

$$\leq \mathbb{E}_{\pi \sim p_{\hat{M}}}\left[TV(M(\pi), \hat{M}(\pi))\right]$$

(4) $$\leq \sqrt{\mathbb{E}_{\pi \sim p_{\hat{M}}}\left[D_H^2(M(\pi), \hat{M}(\pi))\right]}$$

$$\leq \varepsilon.$$

Step 1. Define $G_M = \{\pi \in \Pi \mid g^M(\pi) \leq \Delta/10\}$, where $\Delta = \mathrm{dec}_\varepsilon^c(\mathcal{M}, \hat{M})$. Notice that

$$\mathbb{E}_{\pi \sim p_M}[g^M(\pi)] \underset{\text{(Markov)}}{\geq} \frac{\Delta}{10} P_M(\pi \notin G_M)$$

$$\geq \frac{\Delta}{10}\left(P_{\hat{M}}(\pi \notin G_M) - TV(P_M, P_{\hat{M}})\right)$$

(5) $$\geq \frac{\Delta}{10}\left(P_{\hat{M}}(\pi \notin G_M) - 1/10\right)$$

as $TV(P_M, P_{\hat{M}}) \leq TV(\mathbb{P}^M, \mathbb{P}^{\hat{M}}) \leq \frac{1}{10}$ by above choice of ε in (3) and data-processing inequality

(Review: Data processing inequality in Notes: $D_H^2(M(\pi), \hat{M}(\pi))$

$\geq D_H^2((\rho \circ M)(\pi), (\rho \circ \hat{M})(\pi))$, fixed ρ; also held for TV
e.g. Proposition 43)

Next, assume that

(6) $$\mathbb{E}_{\pi \sim p_{\hat{M}}}[g^{\hat{M}}(\pi)] \leq \frac{\Delta}{10},$$

otherwise, we are done by (1). Our goal is to show that under (6), $p_{\hat{M}}(\pi \notin G_M) \geq 1/2$, which will imply that $\mathbb{E}_{\pi \sim p_M}[g^M(\pi)] \gtrsim \Delta$ by (5).

Step 2: Adding (6) and (2)

$$f^M(\pi_M) - f^{\hat{M}}(\pi_{\hat{M}})$$

$$= \mathbb{E}_{\pi \sim p_{\hat{M}}}[f^M(\pi_M) - f^{\hat{M}}(\pi_{\hat{M}})]$$

$$= \mathbb{E}_{\pi \sim p_{\hat{M}}}[f^M(\pi_M) - f^M(\pi) + f^M(\pi) - f^{\hat{M}}(\pi) + f^{\hat{M}}(\pi) - f^{\hat{M}}(\pi_{\hat{M}})]$$

$$= \mathbb{E}_{\pi \sim p_{\hat{M}}}[\underbrace{f^M(\pi_M) - f^M(\pi)}_{g^M(\pi)} - \underbrace{f^{\hat{M}}(\pi_{\hat{M}}) + f^{\hat{M}}(\pi)}_{g^{\hat{M}}(\pi)}] + \mathbb{E}_{\pi \sim p_{\hat{M}}}[f^M(\pi) - f^{\hat{M}}(\pi)]$$

$$\geq \mathbb{E}_{\pi \sim p_{\hat{M}}}[g^M(\pi) - g^{\hat{M}}(\pi)]$$

$$- \mathbb{E}_{\pi \sim p_{\hat{M}}}[|f^M(\pi) - f^{\hat{M}}(\pi)|]$$

$$\geq \frac{9}{10}\Delta - \mathbb{E}_{\pi \sim p_{\hat{M}}}[|f^M(\pi) - f^{\hat{M}}(\pi)|].$$

By (4), $\mathbb{E}_{\pi \sim p_{\hat{M}}}[|f^M(\pi) - f^{\hat{M}}(\pi)|] \leq \varepsilon$, so

$$f^M(\pi_M) - f^{\hat{M}}(\pi_{\hat{M}}) \geq \frac{9}{10}\Delta - \varepsilon$$

As long as $\varepsilon \leq \frac{1}{10}\Delta$, which is by the condition that $\mathrm{dec}^c_{\varepsilon_T}(\mathcal{M}) \geq 10\,\varepsilon_T$,

$$f^M(\pi_M) - f^{\hat{M}}(\pi_{\hat{M}}) \geq \frac{4}{5}\Delta.$$

Step 3: Observe that if $\pi \in G_M$,

$$|f^M(\pi) - f^{\hat{M}}(\pi)|_+ \geq |f^M(\pi_M) - f^{\hat{M}}(\pi) - \frac{\Delta}{10}|_+$$

$$\geq |f^M(\pi_M) - f^{\hat{M}}(\pi_{\hat{M}}) - \frac{\Delta}{10}|_+$$

$$\geq \frac{7}{10}\Delta$$

by step 2. Using (4) again,

$$\varepsilon \geq \mathbb{E}_{\pi \sim p_{\hat{M}}}\left[\left|f^{M}(\pi) - f^{\hat{M}}(\pi)\right|_+\right]$$

$$\geq \frac{7}{10}\Delta \cdot P_{\hat{M}}(\pi \in G_M) \qquad \text{(Markov)}$$

Since $\varepsilon \leq \Delta/10$ by assumption, we have

$$\frac{\Delta}{10} \geq \frac{7}{10}\Delta \cdot P_{\hat{M}}(\pi \in G_M).$$

i.e. $P_{\hat{M}}(\pi \in G_M) \leq 1/7$. Combining this with (5) gives

$$\frac{1}{T}\mathbb{E}^{M}[\mathrm{Reg}(T)] = \mathbb{E}_{\pi \sim p_M}[g^{M}(\pi)]$$

$$\geq \frac{\Delta}{10}\left(1 - \frac{1}{7} - \frac{1}{10}\right)$$

$$\geq \frac{\Delta}{20}$$

Finally, notice that the choice of $\hat{M} \in \mathcal{M}$ is arbitrary, we are free to choose $\hat{M}$ to maximize $\mathrm{dec}_{\varepsilon}^{c}(\mathcal{M}, \hat{M})$

#.

Sec 3.3 Examples

Consider a few concrete model classes to demonstrate.

(Example 4 continued, Multi-armed bandit with Gaussian award) What is the Contrained DEC for this case?

Set $\hat{M}(\pi) = N(\frac{1}{2}, 1)$, let $\{M_1, \cdots M_A\} \subseteq \mathcal{M}$ be a sub-family of models that $M_i(\pi) = N(f^{M_i}(\pi), 1)$

where $f^{M_i}(\pi) = \frac{1}{2} + \Delta \mathbb{I}(\pi = i)$ for parameter Δ.

For all i, $\mathbb{E}_{\pi \sim p}[D_H^2(M_i(\pi), \hat{M}(\pi))] \leq \frac{1}{2}\Delta^2 p(i)$

(Hellinger distance for Gaussian is squared difference in means as seen before). and

$$\mathbb{E}_{\pi \sim p}[f^{M_i}(\pi_{M_i}) - f^{M_i}(\pi)] = (1 - p(i))\Delta$$

Then. $\mathrm{dec}_\varepsilon^c(\mathcal{M}, \hat{M})$

$$= \inf_{p \in \Delta(\Pi)} \sup_{M} \left\{ \mathbb{E}_{\pi \sim p}[f^M(\pi_M) - f^M(\pi)] \;\middle|\; \mathbb{E}_{\pi \sim p}[D_H^2(M(\pi), \hat{M}(\pi))] \leq \varepsilon^2 \right\}$$

$$\geq \inf_{p \in \Delta(\Pi)} \max_{i} \left\{ (1-p(i))\Delta \;\middle|\; p(i)\frac{\Delta^2}{2} \leq \varepsilon^2 \right\}$$

For any p, $\exists\, i$ s.t $p(i) \leq 1/A$. Choose $\Delta = \varepsilon\sqrt{2A}$

then this choice for i satisfies $p(i)\frac{\Delta^2}{2} \leq \varepsilon^2$.

Hence, $\quad \operatorname{dec}^c_\varepsilon(\mathcal{M}, \hat{M}) \geq (1-p(i))\Delta$

$$\geq \varepsilon\sqrt{A/2}$$

as $1-p(i) \geq 1/2$. By Proposition 3,

$$\mathbb{E}[\mathrm{Reg}] \geq \tilde{\Omega}(\sqrt{AT})$$

Rmk: One can generalize this to any $\mathcal{M}$ that "embeds" the multi-armed bandit problem in a certain sense

Proposition 5: Given reference model $\hat{M}$, suppose that a class $\mathcal{M}$ contains a sub-class $\{M_1, \ldots, M_N\}$, and a collection of decisions $\pi_1, \ldots \pi_N$, that for each i

(i) $D_H^2(M_i(\pi), \hat{M}(\pi)) \leq \beta^2 \mathbb{I}(\pi = \pi_i)$

(ii) $f^{M_i}(\pi_{M_i}) - f^{M_i}(\pi) \geq \alpha\, \mathbb{I}(\pi \neq \pi_i)$

Then. $\mathrm{dec}_\varepsilon^c(\mathcal{M}, \hat{M}) \gtrsim \alpha \cdot \mathbb{I}(\varepsilon \gtrsim \beta/\sqrt{N})$

Notice conditions (i) & (ii) are exactly the two basic teeths to derive any lower bound.

(Example 5 Continued, Bandits with Structured noise)

Recall that $\mathcal{M} = \{M_1, \cdots, M_A\}$ with $M_i(\pi) = N(\frac{1}{2}, 1)$ $\mathbb{I}(i \neq \pi) + \mathrm{Ber}(3/4)\,\mathbb{I}(i = \pi)$. If we consider reference model $\hat{M}(\pi) = N(\frac{1}{2}, 1)$. then by Proposition 5 above, $\alpha = 1/4$, and $\beta^2 = 2$.

$(f^{M_i}(\pi_{M_i}) - f^{M_i}(\pi) \geq 1/4$ if $\pi \neq i)$

Thus, $\mathrm{dec}_\varepsilon^c(\mathcal{M}_{MAB\text{-}SN}) \gtrsim \mathbb{I}(\varepsilon \geq \sqrt{2/A})$, yielding

$$\mathbb{E}(\mathrm{Reg}) \gtrsim O(A)$$

by Proposition 3.

(Example 6. Linear Bandit and Lipschitz bandit)

Linear: $\mathcal{F} = \{\pi \to \langle \theta, \phi(\pi) \rangle \mid \theta \in \Theta\}$

$\Theta \subseteq B_2^d(1)$. $\phi : \Pi \to \mathbb{R}^d$ is known.
feature map

$\mathcal{M}$ is set of all reward dsns with $f^M \in \mathcal{F}$.
and 1-sub-Gaussian noise. Then

$$\mathrm{dec}_\varepsilon^c(\mathcal{M}) \gtrsim \varepsilon \sqrt{d}$$

and $\mathbb{E}(\mathrm{Reg}) \gtrsim \sqrt{dT}$

Lipschitz: $\mathcal{F} = \{ f : \Pi \to [0,1] \mid f \text{ is } 1\text{-Lip w.r.t } \rho \}$

Π is a metric space with metric ρ.

$\mathcal{M}$ is set of all reward dsns with $f^M \in \mathcal{F}$.
and 1-sub-Gaussian noise. Assume
covering s.t. $N_\rho(\Pi, \varepsilon) \geq 1/\varepsilon^d$ for $d>0$

Then, $\mathrm{dec}_\varepsilon^c(\mathcal{M}) \gtrsim \varepsilon^{\frac{2}{d+2}}$

and $\mathbb{E}(\mathrm{Reg}) \gtrsim T^{\frac{d+1}{d+2}}$