Active Imitation Learning (+ other IL ideas)

August 2, 2024

Contents

1 Deterministic expert

1.1 Setup

Let Π be a given deterministic policy class. Denote by Π_{RNS} the set of randomized nonstationary Markovian policies, and by Π_{RNM} the set of randomized non-Markovian policies.

For a fixed $\widehat{\pi}$, define $\Pi_\varepsilon(\widehat{\pi}) := \{\pi \in \Pi : \rho(\pi, \widehat{\pi}) \geq \varepsilon\}$ to be the set of policies from Π that disagree with $\widehat{\pi}$ with probability at least ε. Our objective is to find for another policy class $\widetilde{\Pi}$

$$\sup_{p \in \widetilde{\Pi}} \mathcal{L}(p) := \inf_{\pi \in \Pi_\varepsilon(\widehat{\pi})} \mathbb{P}^p[\exists h : \pi(x_h) \neq \widehat{\pi}(x_h)]. \tag{1}$$

The class $\widetilde{\Pi}$ may be $\Pi, \Pi_{\mathrm{RNS}}, \Pi_{\mathrm{RNM}}$, in order of increasing generality. Let $\tau = \{x_1, a_1, \ldots, x_H, a_H, x_{H+1}\}$ denote a trajectory.

1.2 Mirror descent (trajectories)

Let $\mathcal{P} = \left\{\mathbb{P}^\pi : \pi \in \widetilde{\Pi}\right\}$ be the set of admissible laws of trajectories induced by rolling $\pi \in \widetilde{\Pi}$ out in the MDP. We can equivalently consider the problem in Eq. (1) as

$$\sup_{p \in \mathcal{P}} \mathcal{L}(p) = \inf_{\pi \in \Pi_\varepsilon(\widehat{\pi})} \sum_\tau p(\tau) \cdot \mathbb{I}[\exists h : \pi(x_h) \neq \widehat{\pi}(x_h)].$$

If $\widetilde{\Pi}$ is a convex set in policy space, then $\mathcal{P}$ is also a convex set in trajectory space.

For a fixed p, define $\pi_p := \min_{\pi \in \Pi} \mathbb{P}^p[\exists h : \pi(x_h) \neq \widehat{\pi}(x_h)]$. The derivative is

$$\frac{\partial \mathcal{L}(p)}{\partial p(\tau)} = \mathbb{I}[\exists h : \pi_p(x_h) \neq \widehat{\pi}(x_h)].$$

From this it can be observed that $\sup_{p \in \mathcal{P}} \|\nabla \mathcal{L}(p)\|_\infty \leq 1$.

Procedure. The mirror descent procedure is as follows. We abbreviate $\mathbb{P}^t \equiv \mathbb{P}^{p^t}$; similarly, the best-response policy with respect to p^t is $\pi^t \equiv \pi_{p^t}$. Initialize p^1, a law over trajectories. Then for $t = 1, \dots, T$:

1. Obtain the best-response policy w.r.t the current p^t,

$$\pi^t = \operatorname*{argmin}_{\pi \in \Pi_\varepsilon(\widehat{\pi})} \mathbb{P}^t[\exists h : \pi(x_h) \neq \widehat{\pi}(x_h)].$$

2. Solve the mirror descent update

$$p^{t+1} = \operatorname*{argmin}_{p \in \mathcal{P}} \mathbb{P}^p\big[\exists h : \widetilde{\pi}^t(x_h) \neq \widehat{\pi}(x_h)\big] - \beta D_{\mathsf{KL}}\big(p \,\|\, p^t\big). \tag{2}$$

Analysis. Let $p^\star$ be such that $\mathcal{L}(p^\star) = \sup_{p \in \mathcal{P}} \mathcal{L}(p)$. The mirror descent guarantee states that

$$\begin{aligned}
\sum_t \mathcal{L}(p^\star) - \sum_t \mathcal{L}(p^t) &\leq \frac{D_{\mathsf{KL}}\big(p^\star \,\|\, p^1\big)}{\beta} + \frac{\beta}{2}\sum_t \|\nabla \mathcal{L}(p^t)\|_\infty^2 \\
&\leq \frac{\log\big(1 + D_{\chi^2}\big(p^\star \,\|\, p^1\big)\big)}{\beta} + \frac{T\beta}{2}
\end{aligned}$$

Choosing $\beta = \sqrt{\frac{2\log(1 + D_{\chi^2}(p^\star \,\|\, p^1))}{T}}$,

$$\sum_t \mathcal{L}(p^\star) - \sum_t \mathcal{L}(p^t) \leq \sqrt{2\log\big(1 + D_{\chi^2}(p^\star \,\|\, p^1)\big)T}.$$

Put another way, there exists $t \in [T]$ such that

$$\mathcal{L}(p^\star) - \mathcal{L}(p^t) \leq \sqrt{\frac{2\log\big(1 + D_{\chi^2}(p^\star \,\|\, p^1)\big)}{T}}.$$

This pays for log-coverability of $p^\star$ over p^1. However, we cannot actually solve for the update in Eq. (2) without constructing $\mathcal{P}$.

1.3 DPOing the policy update

Consider again the trajectory-level formulation. To make the DPO substitution, we first need to recover the maximizer to each MD update. Suppose p^t is fixed and admissible, in the sense that $p^t = \mathbb{P}^{\widetilde{\pi}^t}$ for some (possibly non-Markovian) policy $\widetilde{\pi}^t$. Let

$$p_\star^{t+1} = \sup_{p \in \Delta(\tau)} \big\{\mathbb{P}^p\big[\exists h : \pi^t(x_h) \neq \widehat{\pi}(x_h)\big] - \beta D_{\mathsf{KL}}\big(p \,\|\, p^t\big)\big\}. \tag{3}$$

For each τ, this takes the closed form

$$\begin{aligned}
\mathbb{I}\big[\exists h : \pi^t(x_h) \neq \widehat{\pi}(x_h)\big] &= \beta \log\left(\frac{p_\star^{t+1}(\tau)}{p^t(\tau)}\right) + Z \\
&= \beta \log\left(\frac{\widetilde{\pi}_\star^{t+1}(a_{1:H} \mid x_{1:H})}{\widetilde{\pi}^t(a_{1:H} \mid x_{1:H})}\right) + Z,
\end{aligned}$$

where $\widetilde{\pi}_\star^{t+1}$ is the policy that induces $p_\star^{t+1}$, i.e., $p_\star^{t+1} = \mathbb{P}^{\widetilde{\pi}_\star^{t+1}}$. [**TODO:** Needs to be more exact. Does $\widetilde{\pi}_\star^{t+1}$ always exist, e.g., by factoring out the transition probabilities?]

With this substitution, we can consider the following alternative procedure. Given a policy class Π_{DPO}, data in the form of pairs of trajectories drawn as $(\tau, \tau') \sim \pi_{\mathsf{ref}}$, and initial policy $\widetilde{\pi}^1$,

1. Obtain the best-response policy w.r.t the current $\widetilde{\pi}^t$,

$$\pi^t = \underset{\pi \in \Pi_\varepsilon(\widehat{\pi})}{\operatorname{argmin}} \mathbb{P}^{\widetilde{\pi}^t}[\exists h : \pi(x_h) \neq \widehat{\pi}(x_h)].$$

2. Let $g^t(\tau) := \mathbb{I}[\exists h : \pi^t(x_h) \neq \widehat{\pi}(x_h)]$. Solve the DPO update

$$\widetilde{\pi}^{t+1} = \underset{\widetilde{\pi} \in \Pi_{\mathsf{DPO}}}{\operatorname{argmin}} \mathbb{E}_{\tau,\tau' \sim \pi_{\mathsf{ref}}}\left[\left(g^t(\tau) - g^t(\tau') - \beta \log\left(\frac{\widetilde{\pi}(a_{1:H} \mid x_{1:H})}{\widetilde{\pi}^t(a_{1:H} \mid x_{1:H})}\right) + \beta \log\left(\frac{\widetilde{\pi}(a'_{1:H} \mid x'_{1:H})}{\widetilde{\pi}^t(a'_{1:H} \mid x'_{1:H})}\right)\right)^2\right] \quad (4)$$

Assumption 1.1 (Policy completeness)**.** *Given β, for any $\pi \in \Pi_{\mathsf{DPO}}$ and trajectory-level reward function $g \in \{g^\pi(\tau) = \mathbb{I}[\exists h : \pi(x_h) \neq \widehat{\pi}(x_h)] : \pi \in \Pi\}$, there exists $\pi' \in \Pi_{\mathsf{DPO}}$ such that*

$$g(\tau) = \beta \log\left(\frac{\pi'(a_{1:H} \mid x_{1:H})}{\pi(a_{1:H} \mid x_{1:H})}\right) + Z, \ \forall \tau.$$

Sketch.

- The update in Eq. (4) approximately solves Eq. (3), and pays for all-policy coverage over π_{ref}
- [**TODO:** Mirror descent guarantee involves $p^\star$ computed from what policy class? What kind of policy class is Π_{DPO}?]

1.4 Mirror descent (history-dependent policies)

Here we consider mirror descent in (possibly history-dependent) policy space. Let $\widetilde{x}_h = \{x_1, a_1, \ldots, x_{h-1}, a_{h-1}, x_h\}$ be the history up until time h. Now $p \in \widetilde{\Pi}$ maps $\widetilde{x}_h \to \Delta(\mathcal{A})$ First we calculate the gradient for a fixed p.

$$\frac{\partial \mathcal{L}(p)}{\partial p(a_h|\widetilde{x}_h)} = \mathbb{P}^p(\widetilde{x}_h) \cdot \mathbb{P}^p[\exists h : \pi_p(x_h) \neq \widehat{\pi}(x_h) \mid \widetilde{x}_h, a_h].$$

Procedure. The mirror descent procedure is as follows. Initialize p^1. Then for $t = 1, \ldots, T$:

1. Obtain the best-response policy w.r.t the current p^t,

$$\pi^t = \underset{\pi \in \Pi_\varepsilon(\widehat{\pi})}{\operatorname{argmin}} \mathbb{P}^t[\exists h : \pi(x_h) \neq \widehat{\pi}(x_h)].$$

2. Compute the surrogate value function

$$Q^t(\widetilde{x}_h, a_h) = \mathbb{P}^t\big[\exists h : \pi^t(x_h) \neq \widehat{\pi}(x_h) \mid \widetilde{x}_h, a_h\big].$$

3. Solve the mirror descent update

$$p^{t+1} = \sup_{p \in \widetilde{\Pi}} \sum_h \mathbb{E}^{p^t}\big[Q^t(\widetilde{x}_h, p(\widetilde{x}_h))\big] - \beta \underbrace{D_{\mathsf{KL}}\big(\mathbb{P}^p \,\|\, \mathbb{P}^t\big)}_{= \mathbb{E}^p\left[\sum_h D_{\mathsf{KL}}(p(\widetilde{x}_h) \,\|\, p^t(\widetilde{x}_h))\right]}$$

Analysis. [**TODO:** Conjugate norm?] The regularization term is strongly convex with respect to $p(a_h|\widetilde{x}_h)$ in the $\|\cdot\|_{1,d^p(\widetilde{x}_h)}$ norm...? But there is a mismatch between distributions over which expectations are taken in the two terms.

2 Misspecification

2.1 Insufficiency of log loss

2.2 Misspecification in Hellinger distance for deterministic experts

Is it possible to learn a policy where the error scales with misspecification in Hellinger distance? This is a natural goal to pursue since we care about outputting a policy that's close to $\pi^\star$ in Hellinger distance. Further, $D^2_{\mathsf{H}}(P,Q) \le D_{\mathsf{KL}}(P \,\|\, Q) \le \log(1 + D_{\chi^2}(P \,\|\, Q))$, so this would be a strict improvement on the misspecification error under log loss.

For deterministic expert policies, this is possible with the L_{max} loss, and it doesn't require known transitions. Here, the L_{max} loss is equivalent to Hellinger distance up to a constant, and doesn't suffer from the same blow-up issues when a candidate policy is off-support relative to $\pi^\star$.

Recall that for a deterministic policy $\pi^\star$,

$$L_{\mathsf{max}}(\pi) = \mathbb{E}^{\pi^\star}\,\mathbb{E}_{a'_{1:H}\sim\pi(x_{1:H})}[\mathbb{I}[\exists h : a'_h \neq a_h]],$$

and $\widehat{L}_{\mathsf{max}}$ is the empirical version. It can be observed that

$$\frac{1}{2}L_{\mathsf{max}}(\pi) \le D^2_{\mathsf{H}}\left(\mathbb{P}^{\pi}, \mathbb{P}^{\pi^\star}\right) \le 2L_{\mathsf{max}}(\pi).$$

2.3 Stochastic experts, starting point: Scheffe with TV distance

Let $\mathcal{P} = \{\mathbb{P}^\pi : \pi \in \Pi\}$. For any $P, Q \in \mathcal{P}$, define the witness function

$$g_{P,Q} = \operatorname*{argmax}_{|g|\le\frac{1}{2}} \mathbb{E}_P[g] - \mathbb{E}_Q[g]$$

and the set of discriminator functions as

$$\mathcal{G} = \{g_{P,Q} : P, Q \in \mathcal{P}, P \neq Q\}.$$

Output the policy

$$\widehat{\pi} = \operatorname*{argmin}_{\pi\in\Pi} \max_{g\in\mathcal{G}} \widehat{\mathbb{E}}[g] - \mathbb{E}_{\mathbb{P}^\pi}[g]. \tag{5}$$

Proposition 2.1. *The output of Eq. (5) satisfies*

$$D_{\mathsf{TV}}\left(\mathbb{P}^{\pi^\star}, \mathbb{P}^{\widehat{\pi}}\right) \le 3\min_{\pi\in\Pi} D_{\mathsf{TV}}\left(\mathbb{P}^{\pi^\star}, \mathbb{P}^{\pi}\right) + 2\varepsilon_{\text{stat}},$$

where $\varepsilon_{\text{stat}} := \max_{g\in\mathcal{G}}\left|\widehat{\mathbb{E}}[g] - \mathbb{E}_{\pi^\star}[g]\right|$.

[**TODO:** comparison to hellinger bound; incomparable]

Proof of Proposition 2.1. Fix any $\bar{\pi} \in \Pi$. Using the triangle inequality,

$$D_{\mathsf{TV}}\left(\mathbb{P}^{\pi^\star}, \mathbb{P}^{\widehat{\pi}}\right) \le D_{\mathsf{TV}}\left(\mathbb{P}^{\pi^\star}, \mathbb{P}^{\bar{\pi}}\right) + D_{\mathsf{TV}}\left(\mathbb{P}^{\bar{\pi}}, \mathbb{P}^{\widehat{\pi}}\right).$$

Let $\widetilde{g} = g_{\mathbb{P}^{\bar{\pi}},\mathbb{P}^{\widehat{\pi}}}$. By construction, $\widetilde{g} \in \mathcal{G}$ so

$$\begin{aligned}
D_{\mathsf{TV}}\left(\mathbb{P}^{\bar{\pi}}, \mathbb{P}^{\widehat{\pi}}\right) &= \mathbb{E}_{\bar{\pi}}[\widetilde{g}] - \mathbb{E}_{\widehat{\pi}}[\widetilde{g}] \\
&= \mathbb{E}_{\bar{\pi}}[\widetilde{g}] - \widehat{\mathbb{E}}[\widetilde{g}] + \widehat{\mathbb{E}}[\widetilde{g}] - \mathbb{E}_{\widehat{\pi}}[\widetilde{g}] \\
&\le \mathbb{E}_{\bar{\pi}}[\widetilde{g}] - \widehat{\mathbb{E}}[\widetilde{g}] + \max_{g\in\mathcal{G}}\left\{\widehat{\mathbb{E}}[\widetilde{g}] - \mathbb{E}_{\widehat{\pi}}[\widetilde{g}]\right\}
\end{aligned}$$

$$\leq \mathbb{E}_{\overline{\pi}}[\widetilde{g}] - \widehat{\mathbb{E}}[\widetilde{g}] + \max_{g \in \mathcal{G}} \left\{ \widehat{\mathbb{E}}[\widetilde{g}] - \mathbb{E}_{\overline{\pi}}[\widetilde{g}] \right\},$$

since $\widehat{\pi} = \mathrm{argmin}_{\pi \in \Pi} \max_{g \in \mathcal{G}} \left\{ \widehat{\mathbb{E}}[\widetilde{g}] - \mathbb{E}_{\pi}[\widetilde{g}] \right\}$. Next, define $\varepsilon_{\mathrm{stat}} = \max_{g \in \mathcal{G}} \left| \widehat{\mathbb{E}}[g] - \mathbb{E}_{\pi^\star}[g] \right|$. Letting $\bar{g} = \max_{g \in \mathcal{G}} \widehat{\mathbb{E}}[g] - \mathbb{E}_{\overline{\pi}}[g]$, we have

$$\begin{aligned}
D_{\mathsf{TV}}\left(\mathbb{P}^{\overline{\pi}}, \mathbb{P}^{\widehat{\pi}}\right) &\leq \mathbb{E}_{\overline{\pi}}[\widetilde{g}] - \widehat{\mathbb{E}}[\widetilde{g}] + \widehat{\mathbb{E}}[\bar{g}] - \mathbb{E}_{\overline{\pi}}[\bar{g}] \\
&\leq \mathbb{E}_{\overline{\pi}}[\widetilde{g}] - \mathbb{E}_{\pi^\star}[\widetilde{g}] + \mathbb{E}_{\pi^\star}[\bar{g}] - \mathbb{E}_{\overline{\pi}}[\bar{g}] + 2\varepsilon_{\mathrm{stat}} \\
&\leq 2 \sup_{|g| \leq \frac{1}{2}} \left\{ \mathbb{E}_{\overline{\pi}}[g] - \mathbb{E}_{\pi^\star}[g] \right\} + 2\varepsilon_{\mathrm{stat}} \\
&= 2 D_{\mathsf{TV}}\left(\mathbb{P}^{\overline{\pi}}, \mathbb{P}^{\pi^\star}\right) + 2\varepsilon_{\mathrm{stat}}.
\end{aligned}$$

□

For the Proposition 2.1 to hold for general f-divergences, we need (1) a general version of the triangle inequality to isolate the misspecification term in the first step; (2) a concentration inequality from $\widehat{\mathbb{E}}$ to $\mathbb{E}_{\pi^\star}$ that holds for all $g \in \mathcal{G}$, which means that $\mathcal{G}$ must be bounded; and (3) possibly symmetry of the discriminator set. These properties should be satisfied by the Hellinger distance and triangular discrimination, in addition to TV.

Questions.

- **Imitation learning (known dynamis):** What objective should we use for the triangular discrimination metric to get a fast rate? What happens when you try to use Hellinger directly?
- **Imitation learning (unknown dynamics):** The objective in Eq. (5) requires known dynamics. How many queries to the model are required if the dynamics are not known? Can we lower bound the number of queries?
- **Distribution learning:** For the analysis in Proposition 2.1 to go through, we need the divergence to be symmetric and bounded. Can either of these be relaxed to extend this result to general f-divergences?
- **Online imitation learning:** What questions can we generate in this settting?