FoSML: Lecture 21

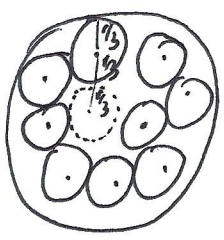
- The previous result about approximate kernels requires understanding covering numbers. In particular, $\mathcal{N}(2R, r)$ appears as a constant driving our bound. In general, $\mathcal{N}(2R, r)$ depends on the geometry and dimension of $\mathcal{X}$.
- In the case $\mathcal{X} \subseteq \mathbb{R}^D$, it can be estimated as follows:

Lemma: Let $\mathcal{X} \subseteq \mathbb{R}^D$ be compact and let $R = \min_{\delta > 0} \{\delta \mid \mathcal{X} \subset B_\delta(0)\}$.

Then $\mathcal{N}(R, r) \leq \left(\frac{3R}{r}\right)^D$.

Proof: Up to constants independent of δ, $\text{Vol}_D(B_\delta(0)) = \delta^D$. Then there is no way to fit more than $R^D/(r/3)^D = (3R/r)^D$ balls of radius $r/3$ in $B_R(0)$ without the balls intersecting; indeed, achieving this bound would require no "wasted space" in the packing.

Now, consider a maximal packing of $\mathcal{X}$ of at most $(3R/r)^D$ balls of radius $r/3$ into the ball $B_R(0)$. Then every element of $B_R(0)$ is distance at most r from the center of one of the radius $\left(\frac{r}{3}\right)$ balls:



The dotted region cannot exist while having a maximal packing.

Then if the balls of radius $\frac{r}{3}$ are inflated to radius r, they necessarily cover $B_R(0)$. Thus, we have a cover of radius r balls of cardinality $\left(\frac{3R}{r}\right)^D \Rightarrow \mathcal{N}(R, r) \leq \left(\frac{3R}{r}\right)^D$. ∎

·Combining these estimates yields:

Theorem: Let K be a continuously differentiable kernel function that satisfies our conditions (shift-invariant, associated non-negative measure μ). Suppose $\sigma_\mu^2 = \mathbb{E}_{w\sim\mu}\left(\|w\|_2^2\right) < \infty$ and $\mathcal{X} \subset \mathbb{R}^D$ is compact and contained in $B_R(0)$. Then for $\Psi : \mathcal{X} \to \mathbb{R}^{2D'}$ defined as before, if $\varepsilon \in (0, 32R\sigma_\mu)$,

$$\mathbb{P}\left(\sup_{x,x'\in\mathcal{X}} \left|\langle \Psi(x), \Psi(x')\rangle - K(x,x')\right| \geq \varepsilon\right)$$

$$\leq \left(\frac{48R\sigma_\mu}{\varepsilon}\right)^2 \exp\left(\frac{-D\varepsilon^2}{4(D'+2)}\right)$$

Proof: The result is immediate from setting $r = \left(\frac{2(6R)^D \exp\left(-\frac{D'\varepsilon^2}{8}\right)}{\left(\frac{4\sigma_\mu}{\varepsilon}\right)^2}\right)^{\frac{2}{D+2}}$ in the lemmas. ∎

=

· Reinforcement learning: planning & learning when an agents wants to maximize some well-defined objective.

· Instead of getting a random sample of training data $\{(x_i, y_i)\}_{i=1}^n$, an agent collects information about its environment through a sequence of actions.

· After engaging in an action, an agent receives two pieces of information:

(a.) Its current state in the environment

(b.) A $\mathbb{R}$-valued reward, which is goal-dependent.

The agent aims to formulate a course of action, or policy, to maximize its

rewards. This policy setting problem needs to manage the exploration-exploitation trade-off, i.e. should the agent exploit the rewards already known, or should the agent risk what is known to explore new possibilities for increased rewards.

- If the environment is known, we have a planning problem. If the environment is unknown, we have a learning problem.

- A classic framework for RL is Markov decision processes (MDP):

Defn: A MDP is defined by:

(a.) A set of states S.

(b.) An initial state $s_0 \in S$.

(c.) A set of actions A.

(d.) A transition probability over destination states $\delta(s,a)$: $\mathbb{P}(s' = \delta(s,a) \mid s,a)$ dist.

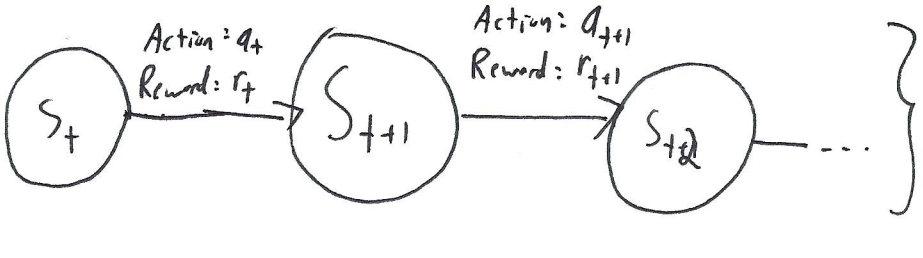
(e.) A reward probability distribution over reward states $r(s,a)$:

$\mathbb{P}(r' = r(s,a) \mid s,a)$.

- Since the destination and reward probabilities depend only on the current state and action (s,a), and not previous states or actions, the process is Markovian.

- We will think of taking discrete time steps, though MDP can be adapted to allow for continuous time parametrization. Let $\{0,\dots,T\}$ denote the decision epochs with finite time horizon T.

- If $|S|, |A| < \infty$, we say the MDP is finite.



Defn: For a space of actions $\mathcal{A}$, let $\mathcal{P}(\mathcal{A}) = \{\text{probability measures on } \mathcal{A}\}$.

Defn: Given an MDP with state space S and action space $\mathcal{A}$, a policy is a function $\pi : S \to \mathcal{P}(\mathcal{A})$. A policy is deterministic if $\|\text{ess sup}(\pi(s))\| = 1$ $\forall s \in S$.

Remark: This definition of policy is time-independent; one could consider a time-dependent variant with $\pi_t : S \to \mathcal{P}(\mathcal{A})$ changing with t. We say a policy that doesn't depend on time is stationary.

What's a good policy? One that maximizes rewards across time:

ex: If $T < \infty$ and each time is equally valuable, we may want π to maximize

$$\sum_{t=0}^{T} r\big(\underbrace{S_t}_{\text{state}}, \underbrace{\pi(S_t)}_{\text{distribution over actions, given that the agent is at state } S_t \text{ at time } t}\big)$$

ex: If $T = +\infty$, it is necessary to discount in time. This can be done geometrically for example, by considering $\gamma \in (0,1)$ and choosing π to maximize

$$\sum_{t=0}^{\infty} \underbrace{\gamma^t}_{\text{discounting in time.}} \cdot r(S_t, \pi(S_t))$$

Defn: A policy π for an MDS at state s has value $V_\pi(s)$ defined as

(a.) $T < \infty$: $\mathbb{E}_{a_t \sim \pi(s_t)} \left(\sum_{t=0}^{T} r(s_t, a_t) \mid s_0 = s \right)$

(b.) $T = +\infty$, discounting factor γ: $\mathbb{E}_{a_t \sim \pi(s_t)} \left(\sum_{t=0}^{\infty} \gamma^t \cdot r(s_t, a_t) \mid s_0 = s \right)$

- So, we want to make $V_\pi(s)$ large, for a given s, or perhaps averaged over all $s \in S$.

Defn: A policy π^* for an MDP is optimal if $\forall s \in S$ and $\forall \pi : S \to P(A)$, $V_{\pi^*}(s) \geq V_\pi(s)$, i.e. π^* yield maximal reward for all initial states.

Q: When does such a π^* exist? It appears optimality is a hard condition. We will show such states exist, as long as $|A|, |S| < \infty$.

Defn: The state-action function Q for a policy π is

$$Q_\pi(s,a) = \mathbb{E}(r(s,a)) + \mathbb{E}_{a_t \sim \pi(s_t)} \left(\sum_{t=1}^{\infty} \gamma^t \cdot r(s_t, a_t) \mid s_0 = s, a_0 = a \right)$$

$$= \mathbb{E}\left(r(s,a) + \gamma V_\pi(s_1) \mid s_0 = s, a_0 = a \right).$$

Note that $\mathbb{E}_{a \sim \pi(s)} \left(Q_\pi(s,a) \right) = V_\pi(s)$.

Theorem (policy improvement): Let π, π' be policies for a common MDP.

Then: $\left\{ \forall s \in S, \; \mathop{\mathbb{E}}_{a \sim \pi'(s)} \left(Q_\pi(s,a) \right) \geq \mathop{\mathbb{E}}_{a \sim \pi(s)} \left(Q_\pi(s,a) \right) \right\}$ (a.)

$$\Downarrow$$

$$\left\{ \forall s \in S, \; V_{\pi'}(s) \geq V_\pi(s) \right\}. \qquad \text{(b.)}$$

Moreover, strict inequality for at least one $s \in S$ on LHS $\Rightarrow$ strict inequality for at least one $s \in S$ on RHS.

Proof: Suppose π, π' satisfy (a). Then, $\forall s \in S$,

$$\begin{aligned}
V_\pi(s) &= \mathop{\mathbb{E}}_{a \sim \pi(s)} \left(Q_\pi(s,a) \right) \\
&\leq \mathop{\mathbb{E}}_{a \sim \pi'(s)} \left(Q_\pi(s,a) \right) \\
&= \mathop{\mathbb{E}}_{a \sim \pi'(s)} \left(r(s,a) + \gamma V_\pi(s_1) \mid s_0 = s \right) \\
&= \mathop{\mathbb{E}}_{a \sim \pi'(s)} \left(r(s,a) + \gamma \mathop{\mathbb{E}}_{a_1 \sim \pi(s_1)} \left(Q_\pi(s_1, a_1) \right) \mid s_0 = s \right) \\
&\leq \mathop{\mathbb{E}}_{a \sim \pi'(s)} \left(r(s,a) + \gamma \mathop{\mathbb{E}}_{a_1 \sim \pi'(s)} \left(Q_\pi(s_1, a_1) \right) \mid s_0 = s \right) \\
&= \mathop{\mathbb{E}}_{\substack{a \sim \pi'(s) \\ a_1 \sim \pi'(s_1)}} \left(r(s,a) + \gamma r(s_1, a_1) + \gamma^2 V_\pi(s_2) \mid s_0 = s \right).
\end{aligned}$$

Proceeding inductively, for any $T \geq 1$,

$$V_\pi(s) \leq \mathop{\mathbb{E}}_{a_t \sim \pi'(s_t)} \left(\sum_{t=0}^{T} \gamma^t \mathbb{E}\left(r(s_t, a_t)\right) + \gamma^{T+1} V_\pi(s_{T+1}) \mid s_0 = s \right)$$

Taking $T \to \infty$ gives a convergent sum equal to $V_{\pi'}(s)$. ∎