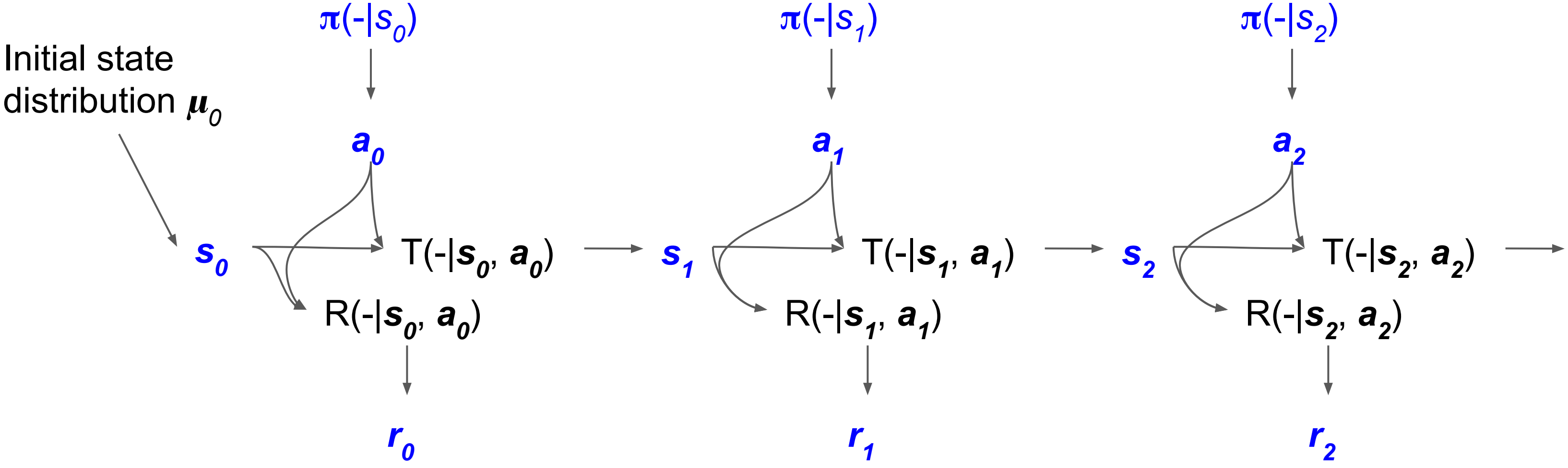
Attacking the Off-Policy Problem with Duality

Ofir Nachum

Off-Policy Reinforcement Learning

- A policy acts on an environment.



- In a general **off-policy** setting, access to the environment is restricted to a fixed dataset of transitions $(s, a, r, s') \sim d^D$.
- But we still want to do RL (policy eval, policy opt, etc.).

The Problem

- **How to do RL in the off-policy setting?**

- **Challenges**:
  - Lack of explicit knowledge of environment dynamics means that correcting for **distribution shift** between on-policy and off-policy state-action distributions is difficult.
  - Limited data can exacerbate **extrapolation and generalization** issues in standard algorithms.

This Talk

- Approach to off-policy RL via convex duality.
- Policy evaluation / optimization can be expressed as linear programs (LPs).
  - Primal LP variables correspond to Q^π.
  - Dual LP variables correspond to d^π.
- **Distribution shift** problem can be attacked by **regularizing dual variables**.
- **Generalization** problem can be attacked by **regularizing primal variables**.

RL As an LP

Many RL problems can be expressed as linear programs (LP)

RL As an LP

This is the Q-LP.

Many RL problems can be expressed as linear programs (LP)

For example, policy evaluation in primal form

$$\rho(\pi) = \min_{Q} \ (1-\gamma) \cdot \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}}[Q(s_0, a_0)]$$
$$\text{s.t. } Q(s,a) \geq R(s,a) + \gamma \cdot \mathcal{P}^{\pi} Q(s,a),$$
$$\forall (s,a) \in S \times A.$$

RL As an LP

This is the Q-LP.

Many RL problems can be expressed as linear programs (LP)

For example, policy evaluation in primal form

$$\rho(\pi) = \min_{Q} \; (1-\gamma) \cdot \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}}[Q(s_0, a_0)]$$

$$\text{s.t. } Q(s,a) \geq R(s,a) + \gamma \cdot \mathcal{P}^{\pi} Q(s,a),$$

$$\forall (s,a) \in S \times A.$$

Q-values

Policy value

Bellman operator

$Q^* = Q^{\pi}$ (Q-values of π)

RL As an LP

This is the Q-LP.

Many RL problems can be expressed as linear programs (LP)

For example, policy evaluation in primal form

$$\rho(\pi) = \min_{Q} \ (1-\gamma) \cdot \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}}[Q(s_0, a_0)]$$

Q-values

Policy value

$$\text{s.t. } Q(s,a) \geq R(s,a) + \gamma \cdot \mathcal{P}^{\pi} Q(s,a),$$

Bellman operator

$$\forall (s,a) \in S \times A.$$

$Q^* = Q^{\pi}$ (Q-values of π)

& dual form

$$\rho(\pi) = \max_{d \geq 0} \sum_{s,a} d(s,a) \cdot R(s,a)$$

$$\text{s.t. } d(s,a) = (1-\gamma)\mu_0(s)\pi(a|s) + \gamma \cdot \mathcal{P}^{\pi}_{*} d(s,a),$$

$$\forall s \in S, a \in A.$$

RL As an LP

This is the Q-LP.

Many RL problems can be expressed as linear programs (LP)

For example, policy evaluation in primal form

$$\rho(\pi) = \min_{Q} \; (1-\gamma) \cdot \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}}[Q(s_0, a_0)]$$

$$\text{s.t. } Q(s,a) \geq R(s,a) + \gamma \cdot \mathcal{P}^{\pi} Q(s,a),$$

$$\forall (s,a) \in S \times A.$$

Policy value

Q-values

Bellman operator

$Q^* = Q^{\pi}$ (Q-values of π)

& dual form

$$\rho(\pi) = \max_{d \geq 0} \sum_{s,a} d(s,a) \cdot R(s,a)$$

$$\text{s.t. } d(s,a) = (1-\gamma)\mu_0(s)\pi(a|s) + \gamma \cdot \mathcal{P}_*^{\pi} d(s,a),$$

$$\forall s \in S, a \in A.$$

Policy value

d is a distribution

$d^* = d^{\pi}$ (on-policy distribution)

Transpose Bellman operator

“Flow” constraints

Beyond LP Duality: Convex Duality

Whether you are in primal or dual, LP has lots of constraints.

Hard to handle all the constraints in stochastic, offline settings. (If we could write down all the constraints, we could just apply standard LP solvers.)

Convex duality enables us to circumvent intractable constraints by applying convex regularizers.

Picking the right regularizer is key!

Attacking Distribution Shift

- **Challenges**:
  - Lack of explicit knowledge of environment dynamics means that correcting for **distribution shift** between on-policy and off-policy state-action distributions is difficult.

- Policy evaluation / optimization can be expressed as linear programs (LPs).
  - Primal LP variables correspond to Q^π.
  - Dual LP variables correspond to d^π.
- **Distribution shift** problem can be attacked by **regularizing dual variables**.

Regularizing the Dual

Regularizing the Dual

Dual LP:

$$\rho(\pi) = \max_{d \geq 0} \sum_{s,a} d(s,a) \cdot R(s,a)$$

$$\text{s.t. } d(s,a) = (1-\gamma)\mu_0(s)\pi(a|s) + \gamma \cdot \mathcal{P}_*^\pi d(s,a),$$

$$\forall s \in S, a \in A.$$

Regularizing the Dual

Dual LP:

$$\rho(\pi) = \max_{d \geq 0} \sum_{s,a} d(s,a) \cdot R(s,a)$$
$$\text{s.t. } d(s,a) = (1-\gamma)\mu_0(s)\pi(a|s) + \gamma \cdot \mathcal{P}_*^\pi d(s,a),$$
$$\forall s \in S, a \in A.$$

LP for regularized policy value:

$$\rho(\pi) - D_f(d^\pi \| d^\mathcal{D}) = \max_{d} \; - D_f(d \| d^\mathcal{D}) + \sum_{s,a} d(s,a) \cdot R(s,a)$$
$$\text{s.t. } d(s,a) = (1-\gamma)\mu_0(s)\pi(a|s) + \gamma \cdot \mathcal{P}_*^\pi d(s,a),$$
$$\forall s \in S, a \in A.$$

Regularizing the Dual

Dual LP:

$$\rho(\pi) = \max_{d \geq 0} \sum_{s,a} d(s,a) \cdot R(s,a)$$
$$\text{s.t. } d(s,a) = (1-\gamma)\mu_0(s)\pi(a|s) + \gamma \cdot \mathcal{P}^{\pi}_{*} d(s,a),$$
$$\forall s \in S, a \in A.$$

LP for regularized policy value:

Reference distribution (fixed)

$$\rho(\pi) - D_f(d^{\pi} \| d^{\mathcal{D}}) = \max_{d} \; -D_f(d \| d^{\mathcal{D}}) + \sum_{s,a} d(s,a) \cdot R(s,a)$$
$$\text{s.t. } d(s,a) = (1-\gamma)\mu_0(s)\pi(a|s) + \gamma \cdot \mathcal{P}^{\pi}_{*} d(s,a),$$
$$\forall s \in S, a \in A.$$

Regularizing the Dual

Dual LP:

$$\rho(\pi) = \max_{d \geq 0} \sum_{s,a} d(s,a) \cdot R(s,a)$$
$$\text{s.t. } d(s,a) = (1-\gamma)\mu_0(s)\pi(a|s) + \gamma \cdot \mathcal{P}^{\pi}_{*} d(s,a),$$
$$\forall s \in S, a \in A.$$

LP for regularized policy value:

Reference distribution (fixed)

$$\rho(\pi) - D_f(d^{\pi} \| d^{\mathcal{D}}) = \max_{d} \; -D_f(d \| d^{\mathcal{D}}) + \sum_{s,a} d(s,a) \cdot R(s,a)$$
$$\text{s.t. } d(s,a) = (1-\gamma)\mu_0(s)\pi(a|s) + \gamma \cdot \mathcal{P}^{\pi}_{*} d(s,a),$$
$$\forall s \in S, a \in A.$$

Note: Regularization doesn't change the fact that d* = d^{π}, because |S|*|A| constraints uniquely determine optimal d* = d^{π} regardless of objective.

Convex Duality with Regularized Dual

Replace LP objective with f-divergence from offline state-action distribution.

$$\rho(\pi) - D_f(d^\pi \| d^{\mathcal{D}}) = \max_d \; - D_f(d \| d^{\mathcal{D}}) + \sum_{s,a} d(s,a) \cdot R(s,a)$$
$$\text{s.t. } d(s,a) = (1-\gamma)\mu_0(s)\pi(a|s) + \gamma \cdot \mathcal{P}_*^\pi d(s,a),$$
$$\forall s \in S, a \in A.$$

Optimal d^* is still d^π.

Convex Duality with Regularized Dual

Replace LP objective with f-divergence from offline state-action distribution.

$$\rho(\pi) - D_f(d^\pi \| d^\mathcal{D}) = \max_d \; -D_f(d \| d^\mathcal{D}) + \sum_{s,a} d(s,a) \cdot R(s,a)$$
$$\text{s.t. } d(s,a) = (1-\gamma)\mu_0(s)\pi(a|s) + \gamma \cdot \mathcal{P}_*^\pi d(s,a),$$
$$\forall s \in S, a \in A.$$

Optimal d* is still d^π.

Take convex dual:

$$\min_Q \; (1-\gamma) \cdot \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}}[Q(s_0, a_0)] + \mathbb{E}_{(s,a)\sim d^\mathcal{D}}[f_*(R(s,a) + \gamma \cdot \mathcal{P}^\pi Q(s,a) - Q(s,a))]$$

Convex Duality with Regularized Dual

Replace LP objective with f-divergence from offline state-action distribution.

$$\rho(\pi) - D_f(d^\pi \| d^\mathcal{D}) = \max_d \; - D_f(d \| d^\mathcal{D}) + \sum_{s,a} d(s,a) \cdot R(s,a)$$
$$\text{s.t. } d(s,a) = (1-\gamma)\mu_0(s)\pi(a|s) + \gamma \cdot \mathcal{P}_*^\pi d(s,a),$$
$$\forall s \in S, a \in A.$$

Optimal d* is still d^π.

Take convex dual:

$$\min_Q \; (1-\gamma) \cdot \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}}[Q(s_0,a_0)] + \mathbb{E}_{(s,a)\sim d^\mathcal{D}}[f_*(R(s,a) + \gamma \cdot \mathcal{P}^\pi Q(s,a) - Q(s,a))]$$

constraints are now penalties

Convex Duality with Regularized Dual

Replace LP objective with f-divergence from offline state-action distribution.

$$\rho(\pi) - D_f(d^\pi \| d^\mathcal{D}) = \max_d \; - D_f(d \| d^\mathcal{D}) + \sum_{s,a} d(s,a) \cdot R(s,a)$$

$$\text{s.t. } d(s,a) = (1-\gamma)\mu_0(s)\pi(a|s) + \gamma \cdot \mathcal{P}_*^\pi d(s,a),$$
$$\forall s \in S, a \in A.$$

Optimal d* is still d^π.

Take convex dual:

Original $h(d) := D_f(d \| d^\mathcal{D}) - \langle d, R \rangle$

Dual $h_*(\cdot) = \mathbb{E}_{d^\mathcal{D}}[f_*(\cdot)]$

$$\min_Q \; (1-\gamma) \cdot \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}}[Q(s_0,a_0)] + \mathbb{E}_{(s,a)\sim d^\mathcal{D}}[f_*(R(s,a) + \gamma \cdot \mathcal{P}^\pi Q(s,a) - Q(s,a))]$$

constraints are now penalties

Convex Duality with Regularized Dual

Replace LP objective with f-divergence from offline state-action distribution.

$$\rho(\pi) - D_f(d^\pi \| d^\mathcal{D}) = \max_d \; - D_f(d \| d^\mathcal{D}) + \sum_{s,a} d(s,a) \cdot R(s,a)$$

$$\text{s.t. } d(s,a) = (1-\gamma)\mu_0(s)\pi(a|s) + \gamma \cdot \mathcal{P}_*^\pi d(s,a),$$

$$\forall s \in S, a \in A.$$

Optimal d* is still d^π.

Take convex dual:

Original $h(d) := D_f(d\|d^\mathcal{D}) - \langle d, R\rangle$

Dual $h_*(\cdot) = \mathbb{E}_{d^\mathcal{D}}[f_*(\cdot)]$

$$\min_Q \; (1-\gamma) \cdot \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}}[Q(s_0, a_0)] + \mathbb{E}_{(s,a)\sim d^\mathcal{D}}[f_*(R(s,a) + \gamma \cdot \mathcal{P}^\pi Q(s,a) - Q(s,a))]$$

off-policy

constraints are now penalties

Convex Duality for Policy Optimization

Regularized policy optimization via max-min

$$\max_{\pi} \min_{Q} \; (1-\gamma)\cdot \mathbb{E}_{\substack{a_0\sim\pi(s_0)\\ s_0\sim\mu_0}}[Q(s_0,a_0)] + \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[f_*(R(s,a)+\gamma\cdot\mathcal{P}^{\pi}Q(s,a)-Q(s,a))]$$

Convex Duality for Policy Optimization

Regularized policy optimization via max-min

$$\max_{\pi} \min_{Q} \; (1-\gamma) \cdot \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}}[Q(s_0, a_0)] + \mathbb{E}_{(s,a) \sim d^{\mathcal{D}}}\boxed{[f_*(R(s,a) + \gamma \cdot \mathcal{P}^{\pi} Q(s,a) - Q(s,a))]}$$

sort of Q-learning / actor-critic

Convex Duality for Policy Optimization

Regularized policy optimization via max-min

$$\max_{\pi} \min_{Q} \; (1-\gamma) \cdot \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}} [Q(s_0, a_0)] + \mathbb{E}_{(s,a) \sim d^{\mathcal{D}}} [f_*(R(s,a) + \gamma \cdot \mathcal{P}^{\pi} Q(s,a) - Q(s,a))]$$

off-policy   **sort of Q-learning / actor-critic**

Convex Duality for Policy Optimization

Regularized policy optimization via max-min

$$\max_{\pi} \min_{Q} \ (1-\gamma) \cdot \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}} [Q(s_0, a_0)] + \mathbb{E}_{(s,a) \sim d^{\mathcal{D}}} [f_*(R(s,a) + \gamma \cdot \mathcal{P}^{\pi} Q(s,a) - Q(s,a))]$$

What's the gradient of the inner objective w.r.t. π?

Convex Duality for Policy Optimization

Regularized policy optimization via max-min

$$\max_{\pi} \min_{Q} \; (1-\gamma)\cdot \mathbb{E}_{\substack{a_0\sim\pi(s_0)\\ s_0\sim\mu_0}}[Q(s_0,a_0)] + \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[f_*(R(s,a)+\gamma\cdot\mathcal{P}^{\pi}Q(s,a)-Q(s,a))]$$

What's the gradient of the inner objective w.r.t. π?

Chain rule will give us this term:

$$d^{\mathcal{D}}(s,a)\cdot f_*'(R(s,a)+\gamma\mathcal{P}^{\pi}Q^*(s,a)-Q^*(s,a))$$

Convex Duality for Policy Optimization

Regularized policy optimization via max-min

$$\max_{\pi} \min_{Q} \ (1-\gamma) \cdot \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}}[Q(s_0, a_0)] + \mathbb{E}_{(s,a) \sim d^{\mathcal{D}}}[f_*(R(s,a) + \gamma \cdot \mathcal{P}^\pi Q(s,a) - Q(s,a))]$$

What's the gradient of the inner objective w.r.t. π?

Chain rule will give us this term:

$$d^{\mathcal{D}}(s,a) \cdot f_*'(R(s,a) + \gamma \mathcal{P}^\pi Q^*(s,a) - Q^*(s,a))$$

Convex duality tells us this is d*:

$$d^{\mathcal{D}}(s,a) \cdot f_*'(R(s,a) + \gamma \mathcal{P}^\pi Q^*(s,a) - Q^*(s,a)) = d^*(s,a)$$

Convex Duality for Policy Optimization

Regularized policy optimization via max-min

$$\max_{\pi} \min_{Q} \ (1-\gamma)\cdot\mathbb{E}_{\substack{a_0\sim\pi(s_0)\\ s_0\sim\mu_0}}[Q(s_0,a_0)] + \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[f_*(R(s,a)+\gamma\cdot\mathcal{P}^{\pi}Q(s,a)-Q(s,a))]$$

What's the gradient of the inner objective w.r.t. π?

Chain rule will give us this term:

$$d^{\mathcal{D}}(s,a)\cdot f_*'(R(s,a)+\gamma\mathcal{P}^{\pi}Q^*(s,a)-Q^*(s,a))$$

Convex duality tells us this is d*:

$$d^{\mathcal{D}}(s,a)\cdot f_*'(R(s,a)+\gamma\mathcal{P}^{\pi}Q^*(s,a)-Q^*(s,a)) = d^*(s,a) = d^{\pi}(s,a)$$

Convex Duality for Policy Optimization

Regularized policy optimization via max-min

$$\max_{\pi} \min_{Q} \; (1-\gamma)\cdot \mathbb{E}_{\substack{a_0\sim\pi(s_0)\\ s_0\sim\mu_0}}[Q(s_0,a_0)] + \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[f_*(R(s,a)+\gamma\cdot\mathcal{P}^{\pi}Q(s,a)-Q(s,a))]$$

What’s the gradient of the inner objective w.r.t. π?

Chain rule will give us this term:

$$d^{\mathcal{D}}(s,a)\cdot f_*'(R(s,a)+\gamma\mathcal{P}^{\pi}Q^*(s,a)-Q^*(s,a))$$

Convex duality tells us this is d*:

$$d^{\mathcal{D}}(s,a)\cdot f_*'(R(s,a)+\gamma\mathcal{P}^{\pi}Q^*(s,a)-Q^*(s,a)) = d^*(s,a) = d^{\pi}(s,a)$$

→ **Off-policy correction naturally comes from Q* values.**

Convex Duality for Policy Optimization

Regularized policy optimization via max-min

$$\max_{\pi} \min_{Q} \ (1-\gamma)\cdot \mathbb{E}_{\substack{a_0\sim\pi(s_0)\\ s_0\sim\mu_0}}[Q(s_0,a_0)] + \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[f_*(R(s,a)+\gamma\cdot\mathcal{P}^{\pi}Q(s,a)-Q(s,a))]$$

What's the gradient of the inner objective w.r.t. π?

Chain rule will give us this term:

$$d^{\mathcal{D}}(s,a)\cdot f_*'(R(s,a)+\gamma\mathcal{P}^{\pi}Q^*(s,a)-Q^*(s,a))$$

Convex duality tells us this is d*:

$$d^{\mathcal{D}}(s,a)\cdot f_*'(R(s,a)+\gamma\mathcal{P}^{\pi}Q^*(s,a)-Q^*(s,a)) = d^*(s,a) = d^{\pi}(s,a)$$

→ **Off-policy correction naturally comes from Q* values.**

→ **On-policy gradient from off-policy data.**

Attacking Generalization

- **Challenges**:
  - Limited data can exacerbate **extrapolation and generalization** issues in standard algorithms.

- Policy evaluation / optimization can be expressed as linear programs (LPs).
  - Primal LP variables correspond to Q^π.
  - Dual LP variables correspond to d^π.
- **Generalization** problem can be attacked by **regularizing primal variables**.

Generalization in the Primal LP

Generalization in the Primal LP

Q-LP:

$$\rho(\pi) = \min_{Q}\ (1-\gamma)\cdot \mathbb{E}_{\substack{a_0\sim\pi(s_0)\\ s_0\sim\mu_0}}[Q(s_0,a_0)]$$
$$\text{s.t. } Q(s,a) \geq R(s,a) + \gamma\cdot\mathcal{P}^{\pi}Q(s,a),$$
$$\forall(s,a)\in S\times A.$$

Generalization in the Primal LP

Q-LP:

$$\rho(\pi) = \min_{Q} \; (1-\gamma) \cdot \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}}[Q(s_0, a_0)]$$

$$\text{s.t. } Q(s,a) \geq R(s,a) + \gamma \cdot \mathcal{P}^{\pi} Q(s,a),$$

$$\forall (s,a) \in S \times A.$$

What does generalization mean here?

Generalization in the Primal LP

Q-LP:

$$\rho(\pi) = \min_{Q} \; (1-\gamma) \cdot \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}} [Q(s_0, a_0)]$$

$$\text{s.t. } Q(s,a) \geq R(s,a) + \gamma \cdot \mathcal{P}^{\pi} Q(s,a),$$

$$\forall (s,a) \in S \times A.$$

What does generalization mean here?

Constraint is missing for (s,a) that policy π visits leads to $\rho(\pi) \to -\infty$

Generalization in the Primal LP

Q-LP:

$$\rho(\pi) = \min_{Q} \ (1-\gamma) \cdot \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}}[Q(s_0, a_0)]$$
$$\text{s.t. } Q(s,a) \geq R(s,a) + \gamma \cdot \mathcal{P}^{\pi} Q(s,a),$$
$$\forall (s,a) \in S \times A.$$

What does generalization mean here?

Constraint is missing for (s,a) that policy π visits leads to $\rho(\pi) \to -\infty$

Natural to "regularize" primal by constraining it to some function class $\mathcal{F}$.

Generalization in the Primal LP

Q-LP:
$$\rho(\pi) = \min_{Q} \ (1-\gamma) \cdot \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}}[Q(s_0, a_0)]$$
$$\text{s.t. } Q(s,a) \geq R(s,a) + \gamma \cdot \mathcal{P}^\pi Q(s,a),$$
$$\forall (s,a) \in S \times A.$$

What does generalization mean here?

Constraint is missing for (s,a) that policy π visits leads to $\rho(\pi) \rightarrow -\infty$

Natural to “regularize” primal by constraining it to some function class $\mathcal{F}$.

Take ***F*** to be unit ball in RKHS.

$$\mathcal{F} := \{Q \in \text{RKHS}, \ \text{s.t. } ||Q||_{\mathcal{H}} \leq 1\}$$

Regularizing the Primal

Q-LP:

$$\min_{Q} \ (1-\gamma) \cdot \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}}[Q(s_0, a_0)] + \delta_{||\cdot||_{\mathcal{H}} \leq 1}(Q)$$

$$\text{s.t. } Q(s,a) \geq R(s,a) + \gamma \cdot \mathcal{P}^{\pi} Q(s,a),$$

$$\forall (s,a) \in S \times A.$$

Regularizing the Primal

Q-LP:

$$\min_{Q} \ (1-\gamma) \cdot \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}}[Q(s_0, a_0)] + \delta_{||\cdot||_{\mathcal{H}} \leq 1}(Q)$$

$$\text{s.t. } Q(s,a) \geq R(s,a) + \gamma \cdot \mathcal{P}^{\pi} Q(s,a),$$

$$\forall (s,a) \in S \times A.$$

Apply convex duality:

$$\max_{d \geq 0} \ \langle d, R \rangle - ||d - (1-\gamma) \cdot \mu_0 \pi - \gamma \cdot \mathcal{P}_*^{\pi} d||_{\mathcal{H}}$$

Regularizing the Primal

Q-LP:

$$\min_{Q} \; (1-\gamma)\cdot \mathbb{E}_{\substack{a_0\sim\pi(s_0)\\ s_0\sim\mu_0}}[Q(s_0,a_0)] + \delta_{||\cdot||_{\mathcal{H}}\leq 1}(Q)$$
$$\text{s.t. } Q(s,a) \geq R(s,a) + \gamma\cdot\mathcal{P}^{\pi}Q(s,a),$$
$$\forall(s,a)\in S\times A.$$

Apply convex duality:

$$\max_{d\geq 0} \; \langle d, R\rangle - \boxed{||d-(1-\gamma)\cdot\mu_0\pi - \gamma\cdot\mathcal{P}_*^{\pi}d||_{\mathcal{H}}}$$

constraints are now penalties

Regularizing the Primal

Q-LP:

$$\min_{Q} \; (1-\gamma) \cdot \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}} [Q(s_0, a_0)] + \delta_{||\cdot||_{\mathcal{H}} \leq 1}(Q)$$

$$\text{s.t. } Q(s,a) \geq R(s,a) + \gamma \cdot \mathcal{P}^{\pi} Q(s,a),$$

$$\forall (s,a) \in S \times A.$$

Norm constraint → Norm penalty

Apply convex duality:

$$\max_{d \geq 0} \; \langle d, R \rangle - ||d - (1-\gamma) \cdot \mu_0 \pi - \gamma \cdot \mathcal{P}_*^{\pi} d||_{\mathcal{H}}$$

constraints are now penalties

Why Did We Choose RKHS?

$$\max_{d \geq 0} \ \langle d, R \rangle - ||d - (1-\gamma) \cdot \mu_0 \pi - \gamma \cdot \mathcal{P}_*^\pi d||_{\mathcal{H}}$$

Why Did We Choose RKHS?

$$\max_{d \geq 0} \; \langle d, R \rangle - || d - (1-\gamma) \cdot \mu_0 \pi - \gamma \cdot \mathcal{P}_*^\pi d ||_{\mathcal{H}}$$

Kernel trick:

$$\max_{d \geq 0} \; \langle d, R \rangle - \left(\mathbb{E}_{\substack{(s,a)\sim d \\ (\tilde{s},\tilde{a})\sim d}} [k(s,a,\tilde{s},\tilde{a})] - 2\mathbb{E}_{\substack{(s,a)\sim d \\ (\tilde{s},\tilde{a})\sim \mathcal{B}_*^\pi d}} [k(s,a,\tilde{s},\tilde{a})] + \mathbb{E}_{\substack{(s,a)\sim \mathcal{B}_*^\pi d \\ (\tilde{s},\tilde{a})\sim \mathcal{B}_*^\pi d}} [k(s,a,\tilde{s},\tilde{a})] \right)^{1/2}$$

Why Did We Choose RKHS?

$$\max_{d \geq 0} \langle d, R \rangle - ||d - (1-\gamma) \cdot \mu_0 \pi - \gamma \cdot \mathcal{P}_*^\pi d||_{\mathcal{H}}$$

Kernel trick:

$$\max_{d \geq 0} \langle d, R \rangle - \left(\mathbb{E}_{\substack{(s,a)\sim d \\ (\tilde{s},\tilde{a})\sim d}}[k(s,a,\tilde{s},\tilde{a})] - 2\mathbb{E}_{\substack{(s,a)\sim d \\ (\tilde{s},\tilde{a})\sim \mathcal{B}_*^\pi d}}[k(s,a,\tilde{s},\tilde{a})] + \mathbb{E}_{\substack{(s,a)\sim \mathcal{B}_*^\pi d \\ (\tilde{s},\tilde{a})\sim \mathcal{B}_*^\pi d}}[k(s,a,\tilde{s},\tilde{a})] \right)^{1/2}$$

Energy distance:

$$\max_{d \geq 0} \langle d, R \rangle - \left(2\mathbb{E}_{\substack{(s,a)\sim d \\ (\tilde{s},\tilde{a})\sim \mathcal{B}_*^\pi d}}[||(s,a) - (\tilde{s},\tilde{a})||_2] - \mathbb{E}_{\substack{(s,a)\sim d \\ (\tilde{s},\tilde{a})\sim d}}[||(s,a) - (\tilde{s},\tilde{a})||_2] - \mathbb{E}_{\substack{(s,a)\sim \mathcal{B}_*^\pi d \\ (\tilde{s},\tilde{a})\sim \mathcal{B}_*^\pi d}}[||(s,a) - (\tilde{s},\tilde{a})||_2] \right)^{1/2}$$

Why Did We Choose RKHS?

$$\max_{d\geq 0} \langle d, R\rangle - ||d - (1-\gamma)\cdot \mu_0\pi - \gamma \cdot \mathcal{P}_*^\pi d||_{\mathcal{H}}$$

Kernel trick:

$$\max_{d\geq 0} \langle d, R\rangle - \left(\mathbb{E}_{\substack{(s,a)\sim d\\(\tilde{s},\tilde{a})\sim d}}[k(s,a,\tilde{s},\tilde{a})] - 2\mathbb{E}_{\substack{(s,a)\sim d\\(\tilde{s},\tilde{a})\sim \mathcal{B}_*^\pi d}}[k(s,a,\tilde{s},\tilde{a})] + \mathbb{E}_{\substack{(s,a)\sim \mathcal{B}_*^\pi d\\(\tilde{s},\tilde{a})\sim \mathcal{B}_*^\pi d}}[k(s,a,\tilde{s},\tilde{a})]\right)^{1/2}$$

Energy distance:

$$\max_{d\geq 0} \langle d, R\rangle - \left(2\mathbb{E}_{\substack{(s,a)\sim d\\(\tilde{s},\tilde{a})\sim \mathcal{B}_*^\pi d}}[||(s,a)-(\tilde{s},\tilde{a})||_2] - \mathbb{E}_{\substack{(s,a)\sim d\\(\tilde{s},\tilde{a})\sim d}}[||(s,a)-(\tilde{s},\tilde{a})||_2] - \mathbb{E}_{\substack{(s,a)\sim \mathcal{B}_*^\pi d\\(\tilde{s},\tilde{a})\sim \mathcal{B}_*^\pi d}}[||(s,a)-(\tilde{s},\tilde{a})||_2]\right)^{1/2}$$

Implicitly constraints Q-values to be smooth, especially when data is missing.

Why Did We Choose RKHS?

$$\max_{d\geq 0} \langle d, R\rangle - ||d - (1-\gamma)\cdot \mu_0\pi - \gamma \cdot \mathcal{P}_*^\pi d||_{\mathcal{H}}$$

Kernel trick:

$$\max_{d\geq 0} \langle d, R\rangle - \left(\mathbb{E}_{\substack{(s,a)\sim d\\(\tilde{s},\tilde{a})\sim d}}[k(s,a,\tilde{s},\tilde{a})] - 2\mathbb{E}_{\substack{(s,a)\sim d\\(\tilde{s},\tilde{a})\sim \mathcal{B}_*^\pi d}}[k(s,a,\tilde{s},\tilde{a})] + \mathbb{E}_{\substack{(s,a)\sim \mathcal{B}_*^\pi d\\(\tilde{s},\tilde{a})\sim \mathcal{B}_*^\pi d}}[k(s,a,\tilde{s},\tilde{a})]\right)^{1/2}$$

Energy distance:

$$\max_{d\geq 0} \langle d, R\rangle - \left(2\mathbb{E}_{\substack{(s,a)\sim d\\(\tilde{s},\tilde{a})\sim \mathcal{B}_*^\pi d}}[||(s,a)-(\tilde{s},\tilde{a})||_2] - \mathbb{E}_{\substack{(s,a)\sim d\\(\tilde{s},\tilde{a})\sim d}}[||(s,a)-(\tilde{s},\tilde{a})||_2] - \mathbb{E}_{\substack{(s,a)\sim \mathcal{B}_*^\pi d\\(\tilde{s},\tilde{a})\sim \mathcal{B}_*^\pi d}}[||(s,a)-(\tilde{s},\tilde{a})||_2]\right)^{1/2}$$

Good representation is key!

Implicitly constraints Q-values to be smooth, especially when data is missing.

Regularizing the Primal - Making it Off-Policy

$$\max_{d \geq 0} \langle d, R \rangle - ||d - (1-\gamma) \cdot \mu_0 \pi - \gamma \cdot \mathcal{P}_*^\pi d||_{\mathcal{H}}$$

Regularizing the Primal - Making it Off-Policy

$$\max_{d \geq 0} \langle d, R \rangle - ||d - (1-\gamma) \cdot \mu_0 \pi - \gamma \cdot \mathcal{P}_*^\pi d||_{\mathcal{H}}$$

Kernel trick:

$$\max_{d \geq 0} \langle d, R \rangle - \left(\mathbb{E}_{\substack{(s,a)\sim d \\ (\tilde{s},\tilde{a})\sim d}}[k(s,a,\tilde{s},\tilde{a})] - 2\mathbb{E}_{\substack{(s,a)\sim d \\ (\tilde{s},\tilde{a})\sim \mathcal{B}_*^\pi d}}[k(s,a,\tilde{s},\tilde{a})] + \mathbb{E}_{\substack{(s,a)\sim \mathcal{B}_*^\pi d \\ (\tilde{s},\tilde{a})\sim \mathcal{B}_*^\pi d}}[k(s,a,\tilde{s},\tilde{a})] \right)^{1/2}$$

Regularizing the Primal - Making it Off-Policy

$$\max_{d \geq 0} \langle d, R \rangle - ||d - (1-\gamma) \cdot \mu_0 \pi - \gamma \cdot \mathcal{P}_*^\pi d||_{\mathcal{H}}$$

Kernel trick:

$$\max_{d \geq 0} \langle d, R \rangle - \left(\mathbb{E}_{\substack{(s,a)\sim d \\ (\tilde{s},\tilde{a})\sim d}}[k(s,a,\tilde{s},\tilde{a})] - 2\mathbb{E}_{\substack{(s,a)\sim d \\ (\tilde{s},\tilde{a})\sim \mathcal{B}_*^\pi d}}[k(s,a,\tilde{s},\tilde{a})] + \mathbb{E}_{\substack{(s,a)\sim \mathcal{B}_*^\pi d \\ (\tilde{s},\tilde{a})\sim \mathcal{B}_*^\pi d}}[k(s,a,\tilde{s},\tilde{a})] \right)^{1/2}$$

Off-policy:

$$\zeta(s,a) := d(s,a)/d^{\mathcal{D}}(s,a)$$

Regularizing the Primal - Making it Off-Policy

$$\max_{d \geq 0} \langle d, R \rangle - \| d - (1-\gamma) \cdot \mu_0 \pi - \gamma \cdot \mathcal{P}_*^\pi d \|_{\mathcal{H}}$$

Kernel trick:

$$\max_{d \geq 0} \langle d, R \rangle - \left(\mathbb{E}_{\substack{(s,a)\sim d \\ (\tilde{s},\tilde{a})\sim d}} [k(s,a,\tilde{s},\tilde{a})] - 2\mathbb{E}_{\substack{(s,a)\sim d \\ (\tilde{s},\tilde{a})\sim \mathcal{B}_*^\pi d}} [k(s,a,\tilde{s},\tilde{a})] + \mathbb{E}_{\substack{(s,a)\sim \mathcal{B}_*^\pi d \\ (\tilde{s},\tilde{a})\sim \mathcal{B}_*^\pi d}} [k(s,a,\tilde{s},\tilde{a})] \right)^{1/2}$$

Off-policy:

$$\zeta(s,a) := d(s,a)/d^{\mathcal{D}}(s,a)$$

$$\max_{\zeta \geq 0} \mathbb{E}_{d^{\mathcal{D}}}[\zeta(s,a) \cdot R(s,a)] -$$

$$\left(\mathbb{E}_{\substack{(s,a)\sim d^{\mathcal{D}} \\ (\tilde{s},\tilde{a})\sim d^{\mathcal{D}}}} [\zeta(s,a)\zeta(\tilde{s},\tilde{a})k(s,a,\tilde{s},\tilde{a})] - 2\mathbb{E}_{\substack{(s,a)\sim d^{\mathcal{D}} \\ (\tilde{s},\tilde{a},\tilde{s}',\tilde{a}')\sim d^{\mathcal{D}}\times\pi}} [\zeta(s,a)\zeta(\tilde{s},\tilde{a})k(s,a,\tilde{s}',\tilde{a}')] + \mathbb{E}_{\substack{(s,a,s',a')\sim d^{\mathcal{D}}\times\pi \\ (\tilde{s},\tilde{a},\tilde{s}',\tilde{a}')\sim d^{\mathcal{D}}\times\pi}} [\zeta(s,a)\zeta(\tilde{s},\tilde{a})k(s',a',\tilde{s}',\tilde{a}')] \right)^{1/2}$$

(for $\gamma = 1$; case of $\gamma < 1$ is slightly different)

Summary and Looking Ahead

- **Distribution shift** problem can be attacked by **regularizing dual variables**.
  - Application to policy evaluation: "DualDICE" (Nachum, et al. 2019)
  - Application to policy optimization: "AlgaeDICE" (Nachum, et al. 2019), "REPS" (Peters 2010)
  - Application to imitation learning: "ValueDICE" (Kostrikov, et al. 2019)
  - Other applications?
- **Generalization** problem can be attacked by **regularizing primal variables**.
  - Application to policy evaluation: "MWL" (Uehara, et al. 2019); also, Liu/Li/Tang/Zhou (2018)
  - Application to policy optimization: Liu/Swaminathan/Agarwal/Brunskill (2019)
  - Other applications?

- Choice of regularizer is key! What choices are we overlooking?