Reinforcement Learning via Fenchel-Rockafellar Duality

Authors: Ofir Nachum, Bo Dai

Presenters: Dami Choi, Chris Zhang

Nachum, Ofir, and Bo Dai. "Reinforcement learning via fenchel-rockafellar duality." arXiv preprint arXiv:2001.01866 (2020).

This paper shows …

- How a number RL problems can be expressed as a convex optimization problem
- An overview of how convex duality can be used to transform a problem to be more amenable to optimization
- How recent offline RL algorithms can be derived from this framework

This paper does not show…

- A new algorithm
- New theoretical or experimental results

Outline

1. Background on convex duality
2. Background on reinforcement learning
3. How to apply duality to offline policy evaluation
4. Offline policy optimization teaser
5. Colab notebook

Fenchel conjugates

For some function $f : \Omega \to \mathbb{R}$

The Fenchel conjugate is given as $f_*(y) := \max_{x \in \Omega} \langle x, y \rangle - f(x)$

Under some conditions, we have the **duality** $f_{**} = f$

What's the intuition here?

Fenchel Conjugate : Different way to describe the same function

Describe a function directly:

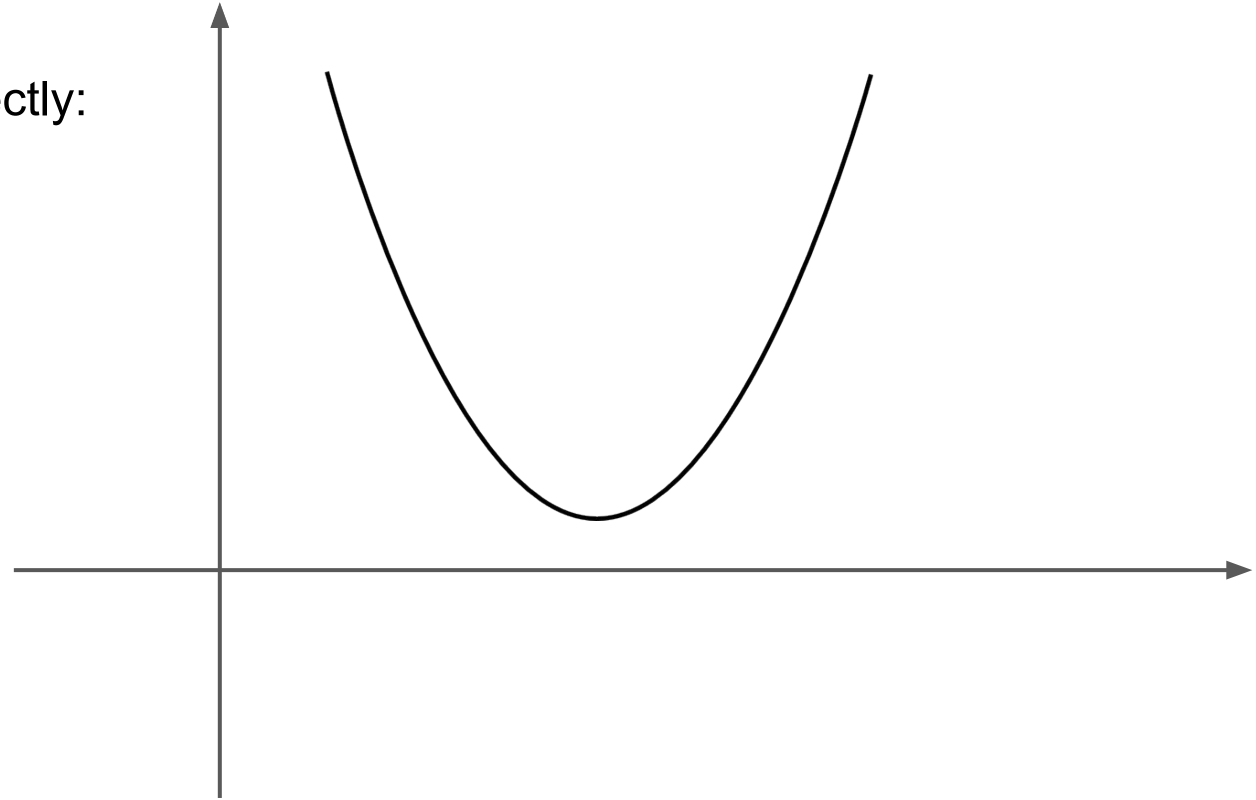


Figure inspired by: D. Bertsekas, “6.253 Convex Analysis and Optimization, Complete Lecture Notes”

Fenchel Conjugate : Different way to describe the same function

Describe a function directly:

- You give me *x*
- I give you *f(x)*

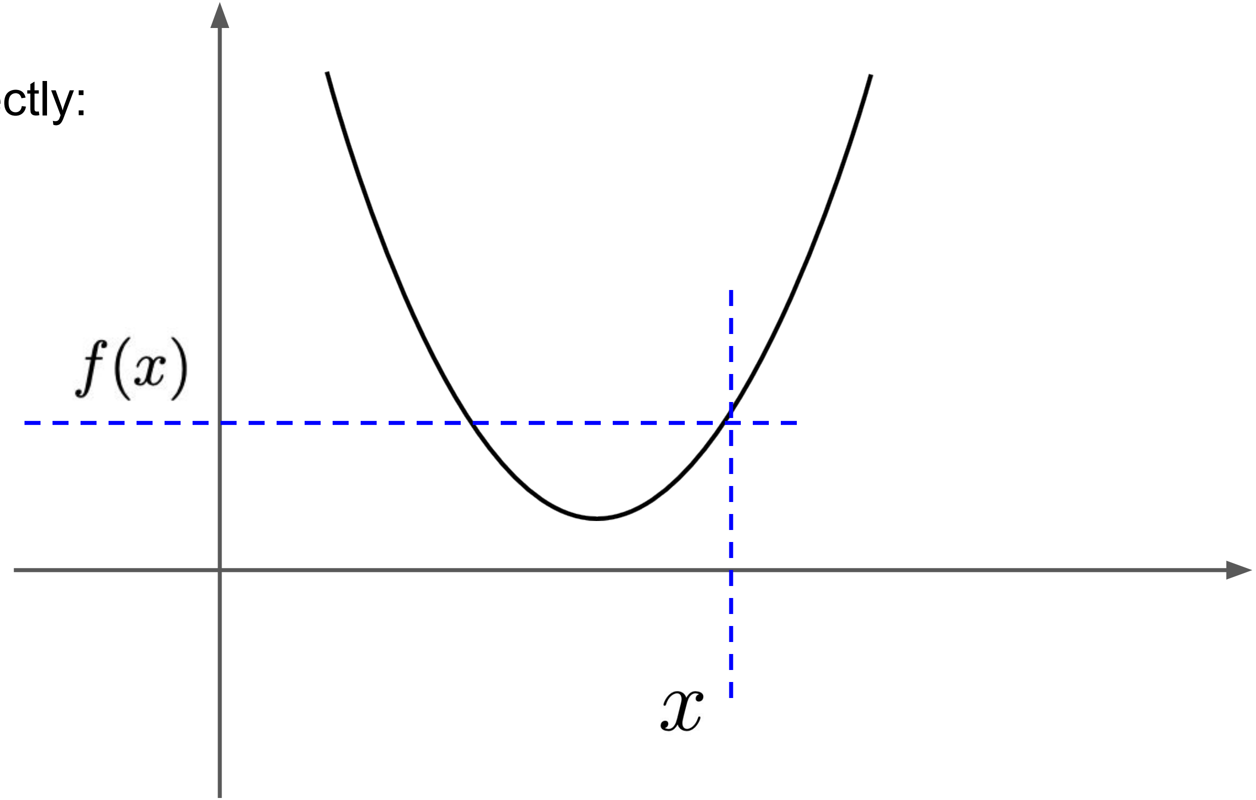


Figure inspired by: D. Bertsekas, “6.253 Convex Analysis and Optimization, Complete Lecture Notes”

Fenchel Conjugate : Different way to describe the same function

Describe a function directly:

- You give me *x*
- I give you *f(x)*

Describe a function by its hyperplanes:

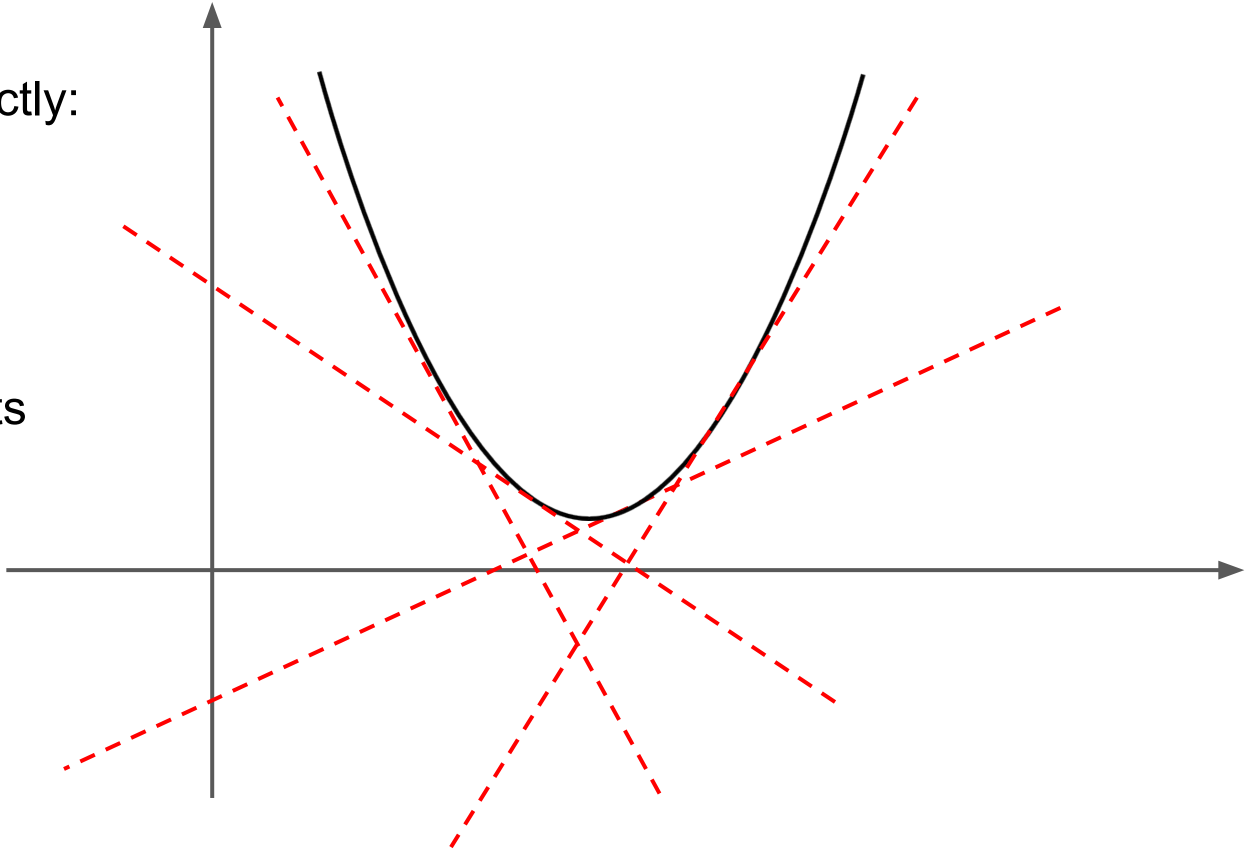


Figure inspired by: D. Bertsekas, “6.253 Convex Analysis and Optimization, Complete Lecture Notes”

Fenchel Conjugate : Different way to describe the same function

Describe a function directly:

- You give me *x*
- I give you *f(x)*

Describe a function by its hyperplanes:

- You give me the slope *y* of some hyperplane

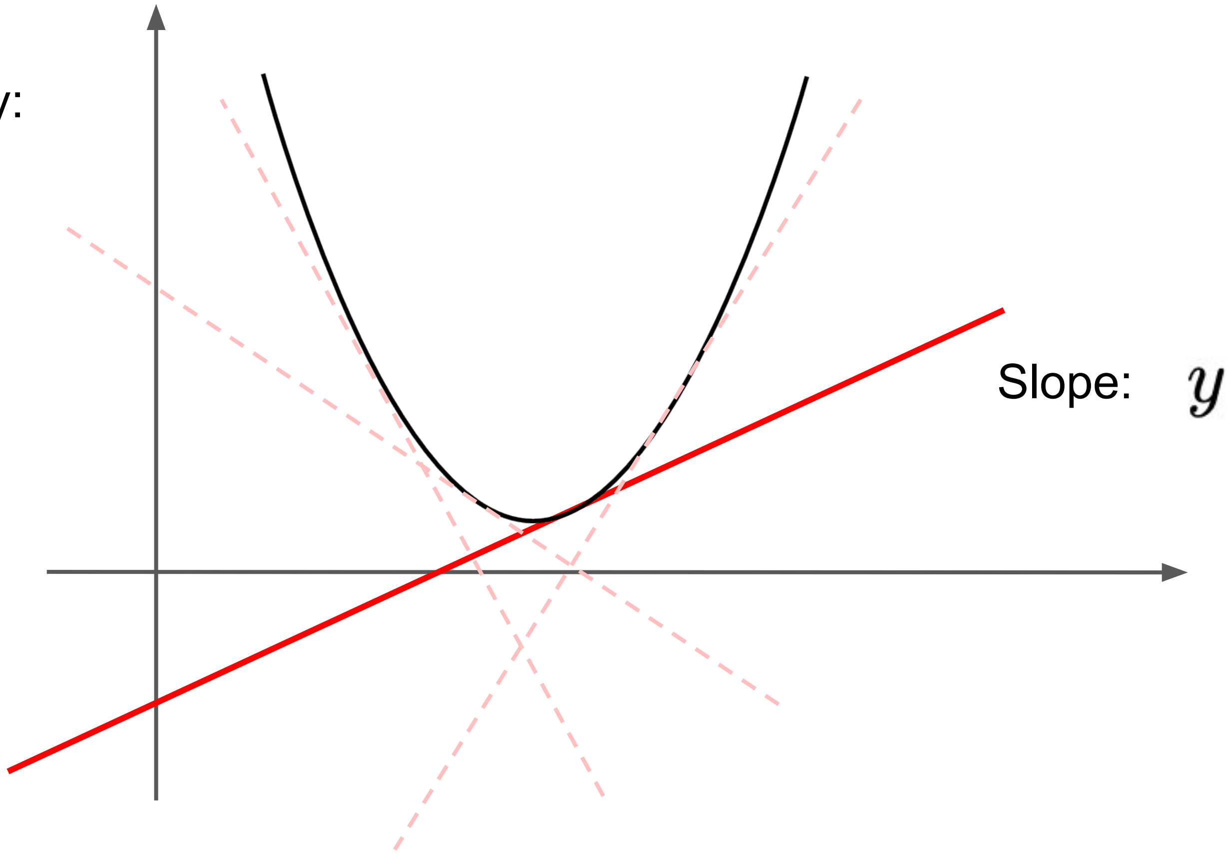


Figure inspired by: D. Bertsekas, “6.253 Convex Analysis and Optimization, Complete Lecture Notes”

Fenchel Conjugate : Different way to describe the same function

Describe a function directly:

- You give me *x*
- I give you *f(x)*

Describe a function by its hyperplanes:

- You give me the slope *y* of some hyperplane
- I give you that plane's intercept $-f_*(y)$

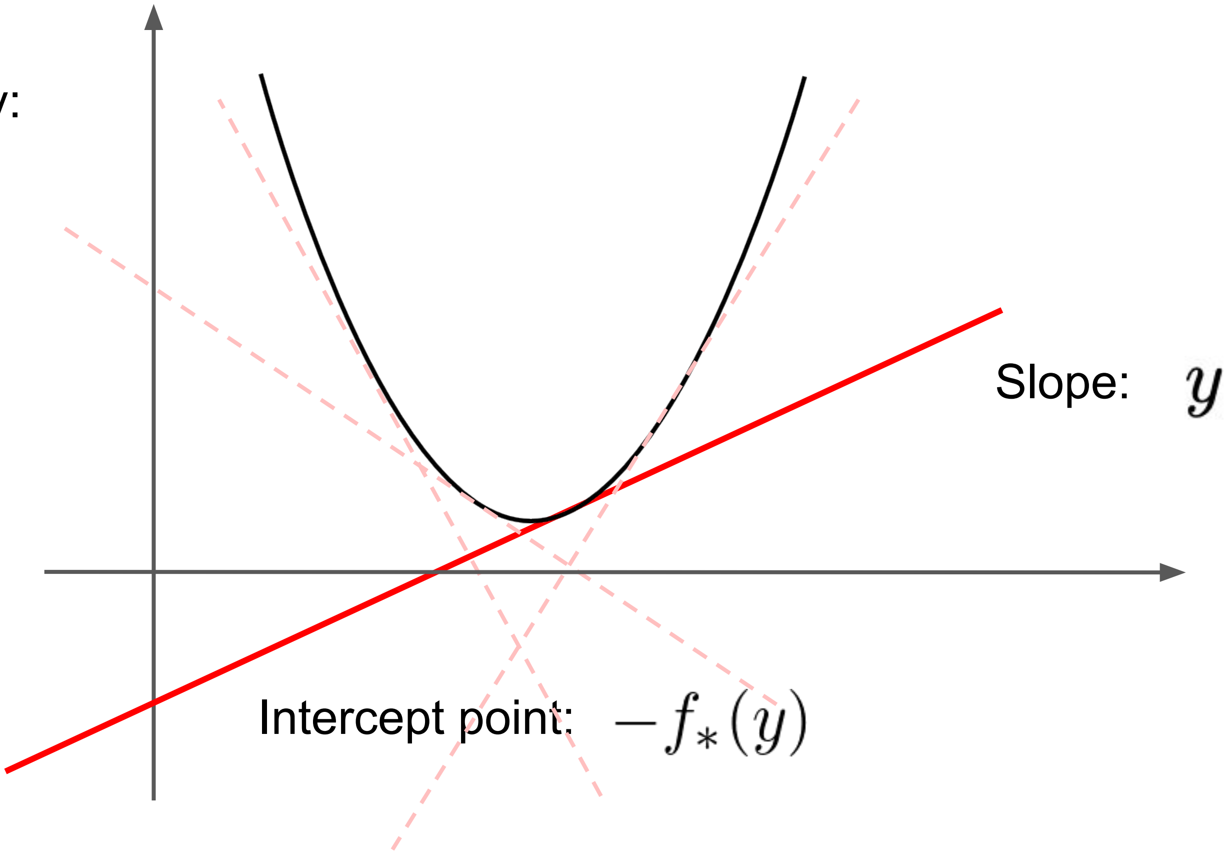


Figure inspired by: D. Bertsekas, "6.253 Convex Analysis and Optimization, Complete Lecture Notes"

Some common functions and their conjugates

Function	Conjugate	Notes
$\frac{1}{2}x^2$	$\frac{1}{2}y^2$	
$\frac{1}{p}\lvert x\rvert^p$	$\frac{1}{q}\lvert y\rvert^q$	For $p, q > 0$ and $\frac{1}{p} + \frac{1}{q} = 1$.
$\delta_{\{a\}}(x)$	$\langle a, y\rangle$	$\delta_C(x)$ is 0 if $x \in C$ and ∞ otherwise.
$\delta_{\mathbb{R}_+}(x)$	$\delta_{\mathbb{R}_-}(y)$	$\mathbb{R}_\pm := \{x \in \mathbb{R} \mid \pm x \geq 0\}$.
$\langle a, x\rangle + b \cdot f(x)$	$b \cdot f_*\left(\frac{y-a}{b}\right)$	
$D_f(x\Vert p)$	$\mathbb{E}_{z\sim p}[f_*(y(z))]$	For $x : \mathcal{Z} \to \mathbb{R}$ and p a distribution over $\mathcal{Z}$.
$D_{\mathrm{KL}}(x\Vert p)$	$\log \mathbb{E}_{z\sim p}[\exp y(z)]$	For $x \in \Delta(\mathcal{Z})$, *i.e.*, a normalized distribution over $\mathcal{Z}$.

What's the use though?

Fenchel-Rockafellar Duality

Consider the primal problem

$$\min_{x\in\Omega} J_{\mathrm{P}}(x) := f(x) + g(Ax)$$

Where $f, g : \Omega \to \mathbb{R}$ are convex and lower semi-continuous, A is a linear map

The corresponding dual problem is

$$\max_{y\in\Omega^*} J_{\mathrm{D}} := -f_*(-A_*y) - g_*(y)$$

Where A_* is the adjoint (transpose) of A, i.e. satisfying $\langle y, Ax\rangle = \langle A_*y, x\rangle$

Fenchel-Rockafellar Duality

$$\min_{x \in \Omega} J_{\mathrm{P}}(x) := f(x) + g(Ax)$$

Primal

$$\max_{y \in \Omega^*} J_{\mathrm{D}} := -f_*(-A_* y) - g_*(y)$$

Dual

Under mild conditions, we have duality

$$\min_{x \in \Omega} J_{\mathrm{P}}(x) = \max_{y \in \Omega^*} J_{\mathrm{D}}(y)$$

Furthermore, the solution to the dual can recover the solution to the primal

$$y^* := \arg\max_y J_{\mathrm{D}}(y)$$

$$x^* = f_*'(-A_* y^*)$$

Duality : Different formulation of the same problem

$$\min_x [f(x) + g(x)]$$

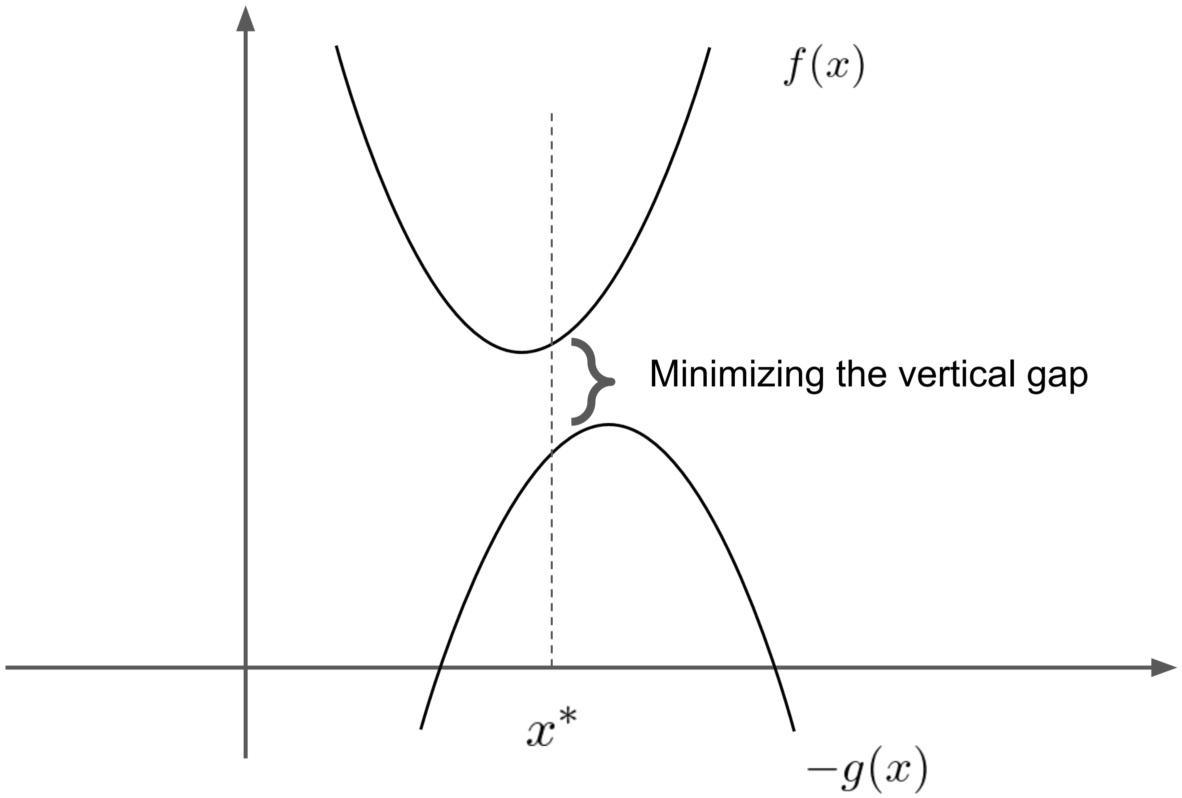


Figure inspired by: D. Bertsekas, “6.253 Convex Analysis and Optimization, Complete Lecture Notes”

Duality : Different formulation of the same problem

$$\min_x [f(x) + g(x)] = \max_y [-f_*(y) - g_*(-y)]$$

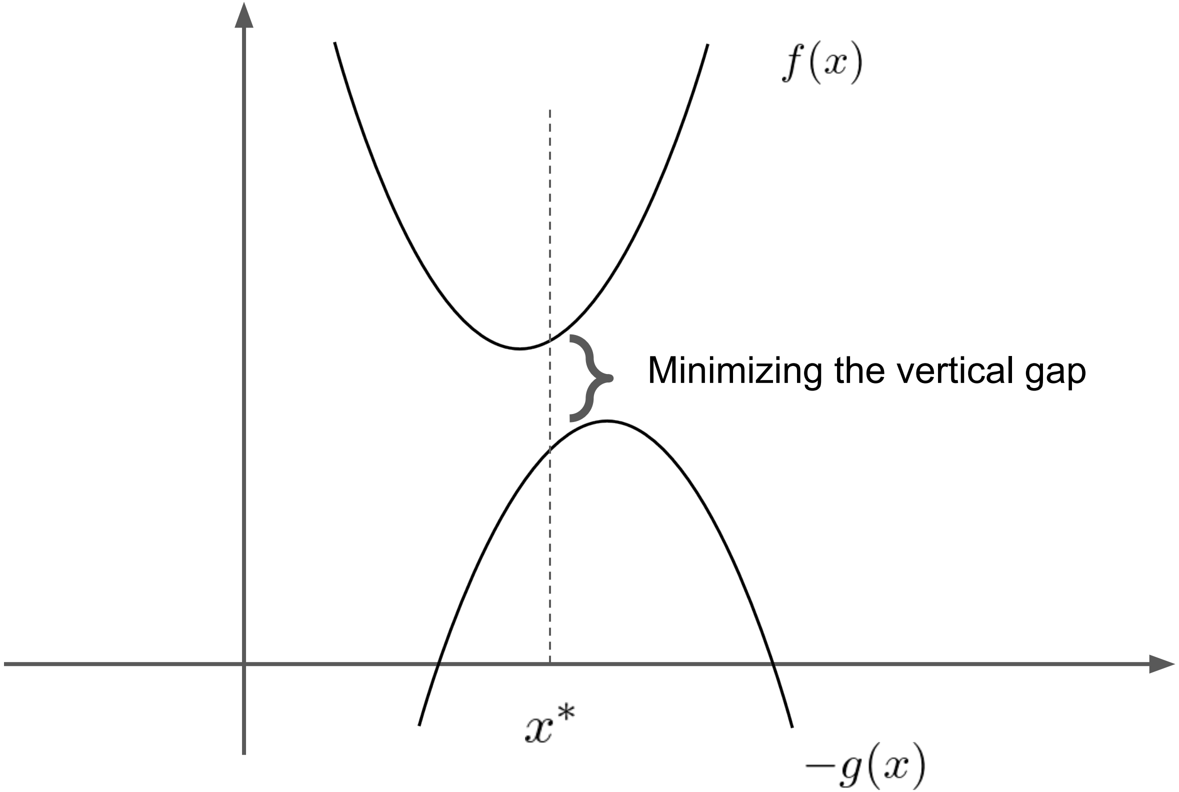


Figure inspired by: D. Bertsekas, “6.253 Convex Analysis and Optimization, Complete Lecture Notes”

Duality : Different formulation of the same problem

$$\min_x [f(x) + g(x)] = \max_y [-f_*(y) - g_*(-y)]$$

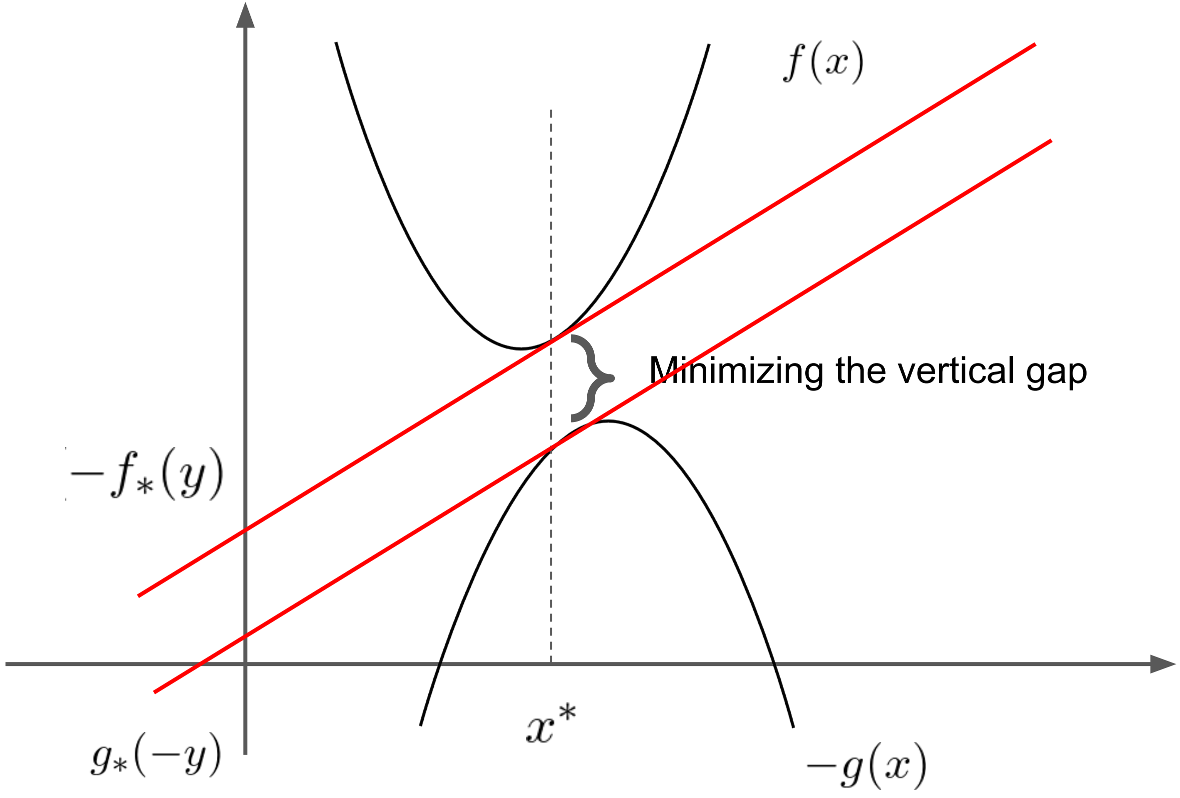


Figure inspired by: D. Bertsekas, “6.253 Convex Analysis and Optimization, Complete Lecture Notes”

Duality : Different formulation of the same problem

$$\min_x [f(x) + g(x)] = \max_y [-f_*(y) - g_*(-y)]$$

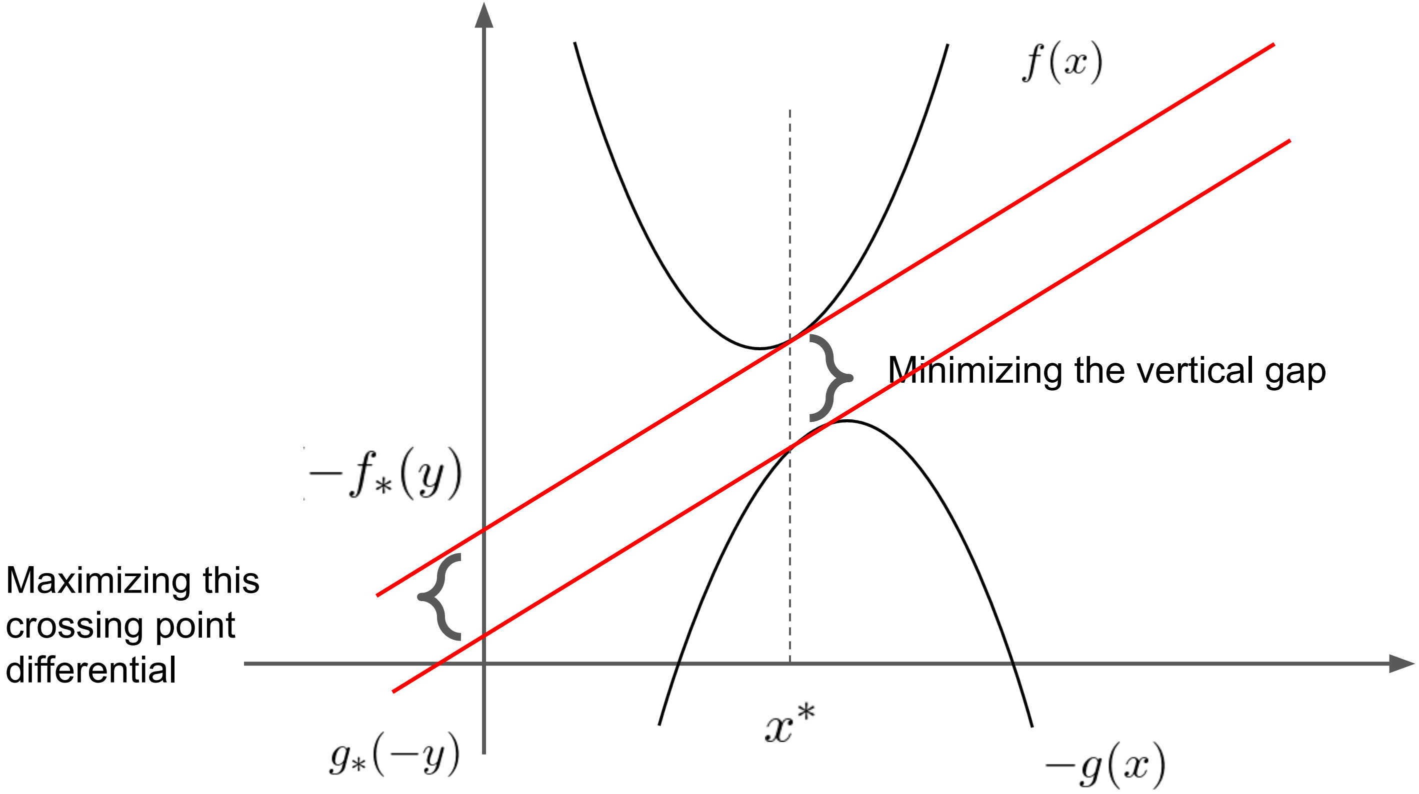


Figure inspired by: D. Bertsekas, “6.253 Convex Analysis and Optimization, Complete Lecture Notes”

Summary

The Fenchel conjugate is another way of describing a function

- Given some slope value, return the crossing point of the corresponding bounding hyperplane

Fenchel-Rockafellar duality allows us to describe an optimization problem in different (possibly more computational friendly) manner.

- This is done by changing the form of the problem to be expressed using the conjugate of a function

Reinforcement Learning

Given an MDP $\mathcal{M} = \langle S, A, R, T, \mu_0, \gamma \rangle$

we are interested in the value of policies w.r.t. to the MDP

$$\rho(\pi) = (1 - \gamma) \cdot \mathbb{E}_{\substack{s_0 \sim \mu_0, \; a_t \sim \pi(s_t) \\ s_{t+1} \sim T(s_t, a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \right]$$

Reinforcement Learning

Given an MDP $\mathcal{M} = \langle S, A, R, T, \mu_0, \gamma \rangle$

we are interested in the value of policies w.r.t. to the MDP

$$\rho(\pi) = (1 - \gamma) \cdot \mathbb{E}_{\substack{s_0 \sim \mu_0,\ a_t \sim \pi(s_t) \\ s_{t+1} \sim T(s_t, a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \right]$$

- Estimating $\rho(\pi)$ of a given policy π <= policy evaluation

Reinforcement Learning

Given an MDP $\mathcal{M} = \langle S, A, R, T, \mu_0, \gamma \rangle$

we are interested in the value of policies w.r.t. to the MDP

$$\rho(\pi) = (1-\gamma) \cdot \mathbb{E}_{\substack{s_0 \sim \mu_0,\ a_t \sim \pi(s_t) \\ s_{t+1} \sim T(s_t, a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \right]$$

- Estimating $\rho(\pi)$ of a given policy π <= policy evaluation
- Maximizing $\rho(\pi)$ w.r.t. π ($\pi^* := \arg\max_\pi \rho(\pi)$) <= policy optimization

Offline RL

This paper focuses on the offline RL setting, where the goal is to estimate $\rho(\pi)$

Using a static dataset of logged experience

$$\mathcal{D} = \{(s^{(i)}, a^{(i)}, r^{(i)}, s^{(i)\prime})\}_{i=1}^{N}$$

$$(s^{(i)}, a^{(i)}) \sim d^{\mathcal{D}} \quad \text{and} \quad s^{(i)\prime} \sim T(s^{(i)}, a^{(i)})$$

unknown distribution

Outline of the paper

1. Formulate RL problems as constrained optimization problems
2. Apply various techniques to make the problem easier to solve

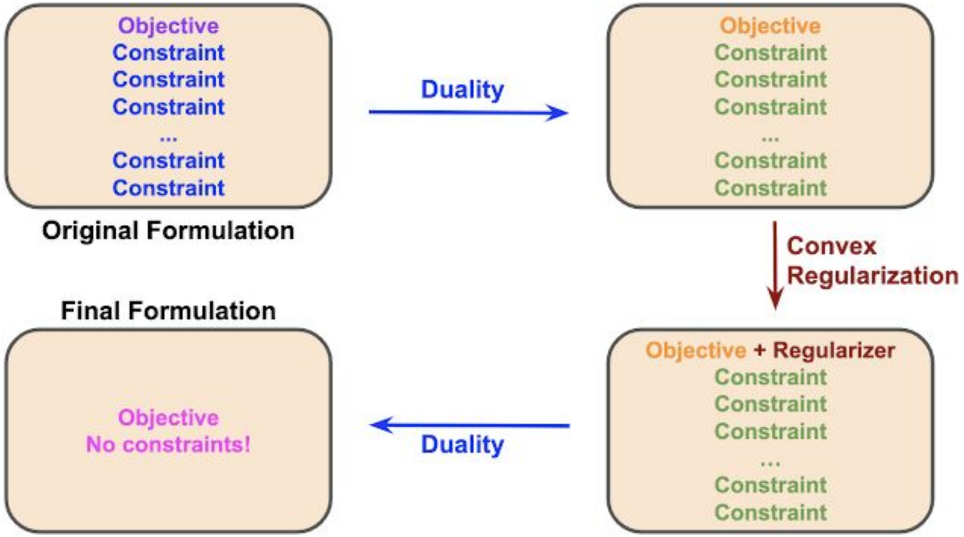


Figure from https://ai.googleblog.com/2020/07/duality-new-approach-to-reinforcement.html

Outline of the paper

1. Formulate RL problems as constrained optimization problems
2. Apply various techniques to make the problem easier to solve

} Policy evaluation

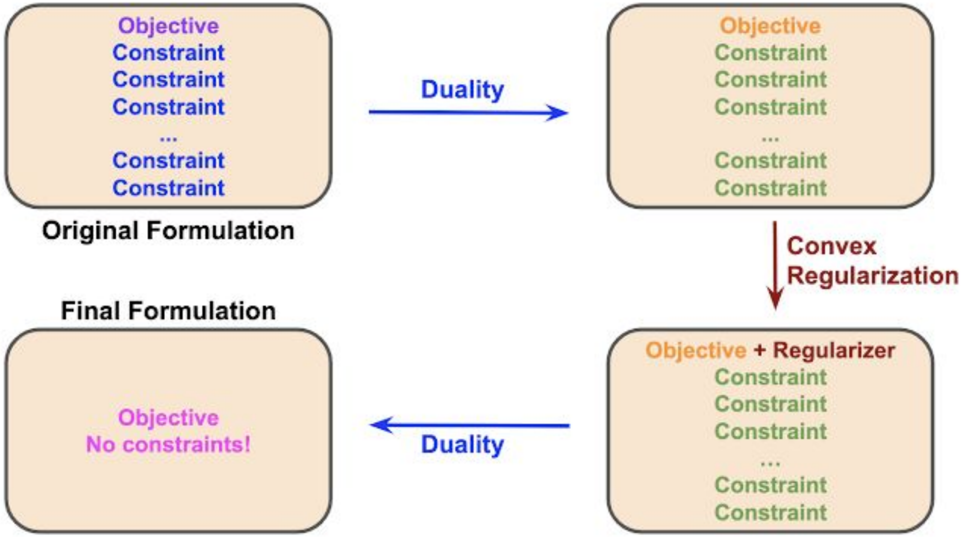


Figure from https://ai.googleblog.com/2020/07/duality-new-approach-to-reinforcement.html

First step

Introduce linear programming formulation of policy evaluation

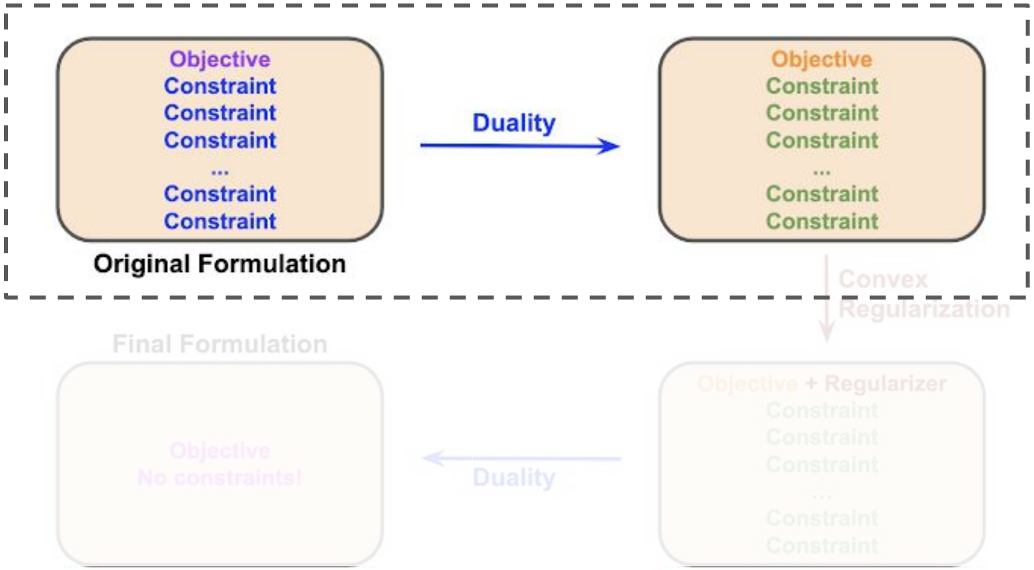


Figure from https://ai.googleblog.com/2020/07/duality-new-approach-to-reinforcement.html

The value $\rho(\pi)$ can be expressed in two different ways

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

$$\rho(\pi) = \quad \mathbb{E}_{(s,a) \sim d^\pi}[R(s,a)]$$

The value $\rho(\pi)$ can be expressed in two different ways

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

$$\rho(\pi) = \quad \mathbb{E}_{(s,a) \sim d^\pi}[R(s,a)]$$

The value $\rho(\pi)$ can be expressed in two different ways

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

$$Q^\pi(s,a) = \mathbb{E}_{\substack{a_t \sim \pi(s_t) \\ s_{t+1} \sim T(s_t, a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \,\middle|\, s_0 = s, a_0 = a \right]$$

Future discounted sum of rewards of following π starting at s, a

$$\rho(\pi) = \mathbb{E}_{(s,a) \sim d^\pi}[R(s,a)]$$

The value $\rho(\pi)$ can be expressed in two different ways

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

$$Q^\pi(s,a) = R(s,a) + \gamma \cdot \mathcal{P}^\pi Q^\pi(s,a)$$

$$\rho(\pi) = \mathbb{E}_{(s,a) \sim d^\pi}[R(s,a)]$$

The value $\rho(\pi)$ can be expressed in two different ways

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

$$Q^\pi(s,a) = R(s,a) + \gamma \cdot \mathcal{P}^\pi Q^\pi(s,a)$$

$$\mathcal{P}^\pi Q(s,a) := \mathbb{E}_{s' \sim T(s,a), a' \sim \pi(s')}[Q(s',a')]$$

$$\rho(\pi) = \quad \mathbb{E}_{(s,a) \sim d^\pi}[R(s,a)]$$

The value $\rho(\pi)$ can be expressed in two different ways

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

$$\rho(\pi) = \qquad \mathbb{E}_{(s,a) \sim d^{\pi}}[R(s,a)]$$

The value $\rho(\pi)$ can be expressed in two different ways

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

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

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

Solution

$$Q^*(s,a) = Q^{\pi}(s,a)$$

$$\rho(\pi) = \quad \mathbb{E}_{(s,a)\sim d^{\pi}}[R(s,a)]$$

The value $\rho(\pi)$ can be expressed in two different ways

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

Solution

$$Q^*(s,a) = Q^{\pi}(s,a)$$

$$\rho(\pi) = \qquad \mathbb{E}_{(s,a) \sim d^{\pi}}[R(s,a)]$$

The value $\rho(\pi)$ can be expressed in two different ways

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

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

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

Solution

$$Q^*(s,a) = Q^\pi(s,a)$$

$$\rho(\pi) = \mathbb{E}_{(s,a)\sim d^\pi}[R(s,a)]$$

$$d^\pi(s,a) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \Pr(s_t = s, a_t = a | \pi)$$

Measures how likely π is to encounter s, a when interacting with the MDP

The value $\rho(\pi)$ can be expressed in two different ways

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

Solution

$$Q^*(s,a) = Q^{\pi}(s,a)$$

$$\rho(\pi) = \qquad \mathbb{E}_{(s,a) \sim d^{\pi}}[R(s,a)]$$

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

The value $\rho(\pi)$ can be expressed in two different ways

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

Solution

$$Q^*(s,a) = Q^{\pi}(s,a)$$

$$\rho(\pi) = \quad \mathbb{E}_{(s,a)\sim d^{\pi}}[R(s,a)]$$

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

$$\mathcal{P}_*^{\pi} d(s,a) := \pi(a|s) \sum_{\tilde{s},\tilde{a}} T(s|\tilde{s},\tilde{a}) d(\tilde{s},\tilde{a})$$

The value $\rho(\pi)$ can be expressed in two different ways

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

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

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

Solution

$$Q^*(s,a) = Q^{\pi}(s,a)$$

$$\rho(\pi) = \mathbb{E}_{(s,a)\sim d^{\pi}}[R(s,a)]$$

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

$$\mathcal{P}^{\pi}_{*} d(s,a) := \pi(a|s) \sum_{\tilde{s},\tilde{a}} T(s|\tilde{s},\tilde{a}) d(\tilde{s},\tilde{a})$$

Adjoint / transpose relationship!

$$\langle y, \mathcal{P}^{\pi} x \rangle = \langle \mathcal{P}^{\pi}_{*} y, x \rangle$$

The value $\rho(\pi)$ can be expressed in two different ways

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

Solution

$$Q^*(s,a) = Q^{\pi}(s,a)$$

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

The value $\rho(\pi)$ can be expressed in two different ways

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

Solution

$$Q^*(s,a) = Q^{\pi}(s,a)$$

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

Solution

$$d^*(s,a) = d^{\pi}(s,a)$$

Primal (Q-values perspective)

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

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

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

Solution

$$Q^*(s,a) = Q^{\pi}(s,a)$$

Dual (visitation perspective)

$$\rho(\pi) = \max_{d \geq 0} \ \mathbb{E}_{(s,a) \sim d^{\pi}}[R(s,a)]$$

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

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

Solution

$$d^*(s,a) = d^{\pi}(s,a)$$

Primal (Q-values perspective)

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

Solution

$$Q^*(s,a) = Q^{\pi}(s,a)$$

Dual (visitation perspective)

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

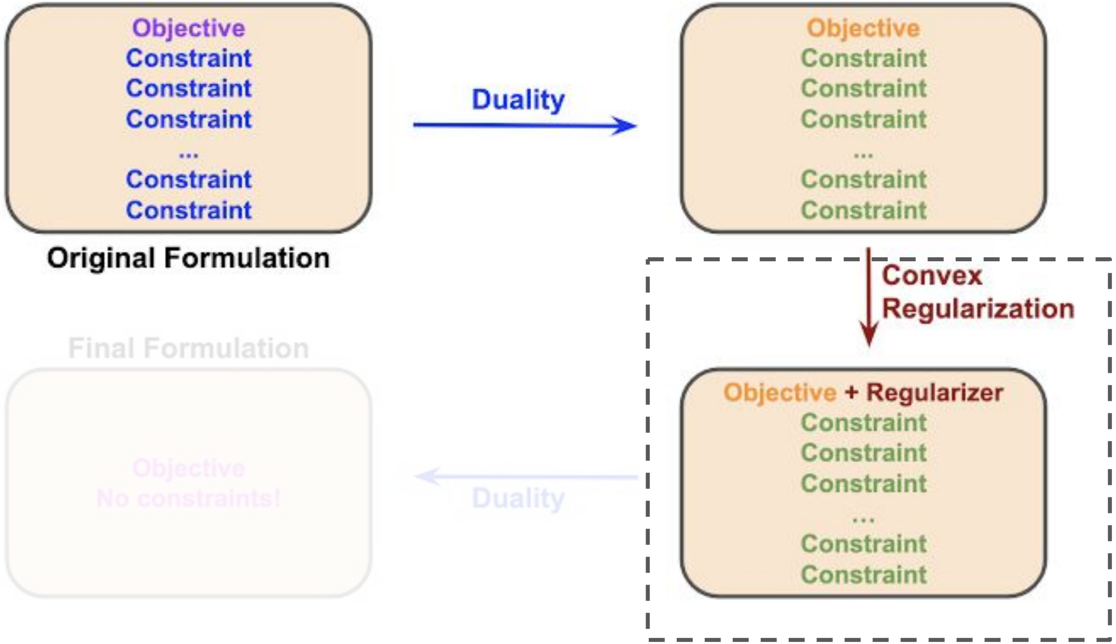
Solution

$$d^*(s,a) = d^{\pi}(s,a)$$

For both problems, number of constraints equal to the product of state and action space!

Second step

Change the dual problem



Changing the problem

Our current dual LP

$$d^* = \arg\max_{d \geq 0} \sum_{s,a} d(s,a) \cdot R(s,a)$$

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

is over-constrained-- equality constraints uniquely determine d regardless of the objective.

Changing the problem

Our current dual LP

$$d^* = \arg\max_{d\geq 0} \cancel{\sum_{s,a} d(s,a)\cdot R(s,a)} \quad \boxed{h(d)}$$

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

is over-constrained-- equality constraints uniquely determine d regardless of the objective.

Idea: Replace original objective with some other function $h(d)$ such that the dual of this problem is easy to optimize.

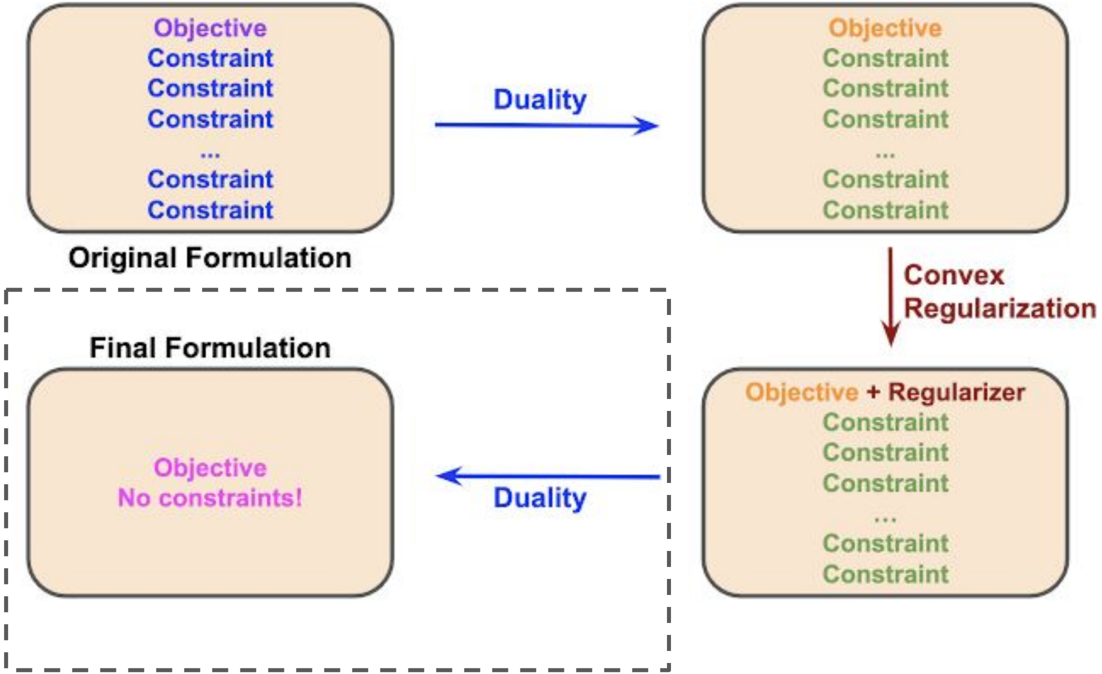
Changing the problem

Choosing $h(d) = -D_f(d||d^{\mathcal{D}})$ reproduces results from DualDICE (Nachum et al. 2019) :

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

Last step

Apply duality once more



Apply duality once more

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

We can write the above problem into a form that we can apply Fenchel-Rockafellar duality to:

$$\max_{d} \ -g(-Ad) - h(d) \quad \Longrightarrow \quad \min_{Q} \ g_*(Q) + h_*(A_* Q)$$

$$\boxed{g := \delta_{\{(1-\gamma)\mu_0 \times \pi\}} \text{ and } A := \gamma \cdot \mathcal{P}_*^{\pi} - I}$$

Final Form

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

- Using the f-divergence w.r.t $d^{\mathcal{D}}$ naturally led to an offline problem with expectations over offline data.

Final Form

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

- Using the f-divergence w.r.t $d^{\mathcal{D}}$ naturally led to an offline problem with expectations over offline data.
- There are no constraints! More amenable to optimization
  - Can use standard gradient-based techniques to find Q^*

Final Form

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

- Using the f-divergence w.r.t $d^{\mathcal{D}}$ naturally led to an offline problem with expectations over offline data.
- There are no constraints! More amenable to optimization
  - Can use standard gradient-based techniques to find Q^*
- We can show that $f_*'(\gamma\cdot\mathcal{P}^\pi Q^*(s,a) - Q^*(s,a)) = \frac{d^\pi(s,a)}{d^{\mathcal{D}}(s,a)}$

  Which allows us to compute the value of π with offline data:

$$\rho(\pi) = \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\left[\frac{d^\pi(s,a)}{d^{\mathcal{D}}(s,a)} R(s,a)\right]$$

DualDICE

If we set $f(x) = \frac{1}{2}x^2$ we can obtain:

$$Q^* = \arg\min_{Q} \ (1-\gamma) \cdot \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}}[Q(s_0, a_0)] + \frac{1}{2}\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[(\gamma \cdot \mathcal{P}^\pi Q(s,a) - Q(s,a))^2]$$

$$\Rightarrow \gamma \cdot \mathcal{P}^\pi Q^*(s,a) - Q^*(s,a) = \frac{d^\pi(s,a)}{d^{\mathcal{D}}(s,a)}, \quad \forall s \in S, a \in A.$$

Optimal Bellman residuals are exactly equal to $\frac{d^\pi(s,a)}{d^{\mathcal{D}}(s,a)}$ when we

DualDICE

If we set $f(x) = \frac{1}{2}x^2$ we can obtain:

$$Q^* = \arg\min_Q \;(1-\gamma)\cdot \mathbb{E}_{\substack{a_0\sim\pi(s_0)\\ s_0\sim\mu_0}}[Q(s_0,a_0)] + \frac{1}{2}\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\boxed{[(\gamma\cdot\mathcal{P}^\pi Q(s,a) - Q(s,a))^2]}$$

$$\Rightarrow \gamma\cdot\mathcal{P}^\pi Q^*(s,a) - Q^*(s,a) = \frac{d^\pi(s,a)}{d^{\mathcal{D}}(s,a)}, \quad \forall s\in S, a\in A.$$

Optimal Bellman residuals are exactly equal to $\frac{d^\pi(s,a)}{d^{\mathcal{D}}(s,a)}$ when we

- Minimize the squared Bellman residuals w.r.t. zero reward and

DualDICE

If we set $f(x) = \frac{1}{2}x^2$ we can obtain:

$$Q^* = \arg\min_{Q} \ (1-\gamma) \cdot \boxed{\mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}}[Q(s_0, a_0)]} + \frac{1}{2}\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[(\gamma \cdot \mathcal{P}^\pi Q(s,a) - Q(s,a))^2]$$

$$\Rightarrow \gamma \cdot \mathcal{P}^\pi Q^*(s,a) - Q^*(s,a) = \frac{d^\pi(s,a)}{d^{\mathcal{D}}(s,a)}, \quad \forall s \in S, a \in A.$$

Optimal Bellman residuals are exactly equal to $\frac{d^\pi(s,a)}{d^{\mathcal{D}}(s,a)}$ when we

- Minimize the squared Bellman residuals w.r.t. zero reward and
- Minimize the initial Q-values

DualDICE

If we set $f(x) = \frac{1}{2}x^2$ we can obtain:

$$Q^* = \arg\min_{Q} \ (1-\gamma) \cdot \boxed{\mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}}[Q(s_0, a_0)]} + \frac{1}{2}\mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}\boxed{[(\gamma \cdot \mathcal{P}^\pi Q(s,a) - Q(s,a))^2]}$$

$$\Rightarrow \gamma \cdot \mathcal{P}^\pi Q^*(s,a) - Q^*(s,a) = \frac{d^\pi(s,a)}{d^{\mathcal{D}}(s,a)}, \quad \forall s \in S, a \in A.$$

Optimal Bellman residuals are exactly equal to $\frac{d^\pi(s,a)}{d^{\mathcal{D}}(s,a)}$ when we

- Minimize the squared Bellman residuals w.r.t. zero reward and
- Minimize the initial Q-values

We verify this in our Colab notebook!

Summary of Policy Evaluation

- Policy evaluation can be expressed as LPs
  - Primal solution is Q^π
  - Dual solution is d^π
- Changing the objective of the dual does not affect the solution
- Using f-divergence as the new dual objective and applying Fenchel-Rockafellar duality results in a more easy problem to optimize
- Solution to problem can be used for offline policy evaluation

Policy Optimization Teaser

- Can apply many of the same techniques used for policy evaluation
- Caveat: In this setting, modifying the objective changes the solution
  - However, solution to a regularized problem can still be valuable
- Depending on exact form of regularization, we can get a method reminiscent of offline actor critic algorithms
- However, the more principled formulation allows us to get true on-policy policy gradients using only offline data

If any of this sounds interesting, you can learn more from the paper!

Conclusion

- When presented with a problem that appears difficult to solve, we can write the problem as a constrained convex optimization problem and solve its Fenchel-Rockafellar dual
- If the dual is still difficult to solve, we can modify the original objective by either replacing it (policy evaluation) or applying a convex regularizer (policy optimization)

Limitations:

- Gap between theory and practice
- Importance weights $\frac{d^\pi(s,a)}{d^\mathcal{D}(s,a)}$ are not reliable when $d^\mathcal{D}$ is too different from d^π