# Reinforcemnet Learning via Fenchel-Rockafellar Duality

Xuhui Liu

LAMDA, NJUAI

January 22, 2021

# Outline

1. Convex Duality
   - Fenchel Conjugate
   - $f$-Divergence
   - Fenchel-Rockafellar Duality
2. Policy Evaluation
3. Policy Optimization
   - The Policy Gradient Theorem
   - Dual Optimization
4. RL with the Linear Programming Form of V

# Table of Contents

# Fenchel Conjugate

The Fenchel conjugate $f^*$ of a function $f : \Omega \to R$ is defined as

$$f^*(y) := \max_{x \in \Omega} < x, y > -f(x)$$

The function is also referred to as the *convex conjugate* or *Legendre-Fenchel transformation* of $f$.

# Fenchel Conjugate

**Definition 1** *We say a function $f$ is proper when $\{x \in \Omega : f(x) < \infty\}$ is non-empty and $f(x) > -\infty$ for all $x \in \Omega$.*

**Definition 2** *We say a function $f$ is lower semi-continuous when $\{x \in \Omega : f(x) > \alpha\}$ is an open set for all $\alpha \in \mathbb{R}$.*

For a proper, convex, lower semi-continuous $f$, its conjugate function $f^*$ is also proper, convex, and lower semi-continuous. Moreover, one has the duality $f^{**} = f$. i.e.,

$$f(x) = \max_{y \in \Omega^*} < x, y > - f^*(y)$$

where $\Omega^*$ denotes the domain of $f^*$.

# Fenchel Conjugate

| 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 $\infty$ 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 $\mathcal{Z}$. |

Table 1: A few common functions and their corresponding Fenchel conjugates.

# $f$-Divergence and its Fenchel Conjugate

For a convex function $f$ and a distribution $p$, the $f$-divergence is defined as,

$$D_f(x||p) = \mathbb{E}_{z\sim p}[f\big(\frac{x(z)}{p(z)}\big)].$$

The conjugate of $D_f(x||p)$ at $y$ is, under mild conditions[1],

$$\begin{aligned} g(y) &= \max_x \sum_z x(z)y(z) - \mathbb{E}_{z\sim p}[f(x(z)/p(z))] \\ &= \mathbb{E}_{z\sim p}[\max_x x(z)y(z)/p(z) - f(x(z)/p(z))] \\ &= \mathbb{E}_{z\sim p}[f^*(y(z))] \end{aligned}$$

[1]Conditions of the interchangeability principle must be satisfied, and p must have sufficient support

# Fenchel-Rockafellar Duality

Consider a primal problem given by

$$\min_{x \in \Omega} J_P(x) := f(x) + g(Ax) \tag{1}$$

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

The dual of this probelm is given by

$$\max_{y \in \Omega^*} J_D := -f^*(-A^* y) - g^*(y) \tag{2}$$

- $A^*$ to denote the adjoint linear operator of $A$; i.e., $A^*$ is the linear operator for which $< y, Ax >=< A^* y, x >$ , for all $x, y$.
- In the common case of $A$ simply being a real-valued matrix, $A^*$ is the transpose of $A$.

# Fenchel-Rockafellar Duality

Under mild conditions, the dual problem may be derived from the primal via

$$\begin{aligned}
\min_{x\in\Omega} f(x) + g(Ax) &= \min_{x\in\Omega} \max_{y\in\Omega^*} f(x) + \langle y, Ax\rangle - g_*(y) \\
&= \max_{y\in\Omega^*} \left\{ \min_{x\in\Omega} f(x) + \langle y, Ax\rangle \right\} - g_*(y) \\
&= \max_{y\in\Omega^*} \left\{ -\max_{x\in\Omega} \langle -A_* y, x\rangle - f(x) \right\} - g_*(y) \\
&= \max_{y\in\Omega^*} -f_*(-A_* y) - g_*(y).
\end{aligned}$$

Thus, we have the duality,

$$\min_{x\in\Omega} J_P(x) = \max_{y\in\Omega^*} J_D(y)$$

# Fenchel-Rockafellar Duality

- The solution to the dual $y^* := \arg\max_y J_D(y)$ can be used to find a solution to the primal.
- If $(f^*)'$ is well-definded, then $x^* = (f^*)'(-A^*y^*)$ is a solution to the primal.
- More generally, one can recover $x^* \in \partial f^*(-A^*y^*) \cap A^{-1}\partial g^*(y^*)$ as the set of all primal solutions.

# The Lagrangian

The Fenchel-Rockafellar duality is general enough that it can be used to derive the Lagrangian duality. Consider the constrained optimization problem

$$\min_x \; f(x) \quad \text{s.t.} \quad Ax \geq b. \tag{14}$$

If we consider this problem expressed as $\min_x f(x) + g(x)$ for $g(x) = \delta_{\mathbb{R}_-}(-Ax + b)$, its Fenchel-Rockafellar dual is given by

$$\max_y \; \langle y, b \rangle - f_*(A_* y) \quad \text{s.t.} \quad y \geq 0. \tag{15}$$

By considering $f_*$ in terms of its Fenchel conjugate (equation (1)), we may write the problem as

$$\min_x \max_{y \geq 0} \; \langle y, b \rangle - \langle x, A_* y \rangle + f(x). \tag{16}$$

Using the fact that $\langle y, Ax \rangle = \langle x, A_* y \rangle$ for any $A$ we may express this as

$$\min_x \max_{y \geq 0} \; \underbrace{\langle y, b - Ax \rangle + f(x)}_{L(x,y)}. \tag{17}$$

The expression $L(x, y)$ is known as the *Lagrangian* of the original problem in (14). One may further derive the well-known Lagrange duality:[4]

$$\max_{y \geq 0} \min_x \; L(x, y) = \min_x \max_{y \geq 0} \; L(x, y). \tag{18}$$

# Table of Contents

# Notation

- Markov Decision Process $M =< \mathcal{S}, \mathcal{A}, P, R, \gamma, \mu_0 >$.
- Policy $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$.
- Value function $V^\pi(s) = \mathbb{E}[\sum_{h=0}^{\infty} \gamma^h r_h]$ and Q-function $Q^\pi(s, a)$.
- $\rho(\pi)$ is the expectation of $V^\pi(s)$ under initial state distribution.
- $P^\pi$ is the policy transition operator,

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

- $d^\pi(s, a)$ is the state-action distribution of policy $\pi$.

# The Linear Programming Form of Q

Q-LP:

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

The optimal $Q^*$ of this LP satisfies $Q^*(s,a) = Q^{\pi}(s,a)$ for all $s,a$ reachable by $\pi$.
The dual of this LP provides us with the visitation perspective on policy evaluation:

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

# Policy Evaluation via the Lagrangian

- Using the Lagrangian of the Q-LP:

$$\rho(\pi) = \min_{Q} \max_{d \geq 0} \quad (1-\gamma) \cdot \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}}[Q(s_0, a_0)] + \sum_{s,a} d(s,a) \cdot (R(s,a) + \gamma \cdot \mathcal{P}^{\pi} Q(s,a) - Q(s,a)).$$

- In an offline setting, where we only have access to a distribution $d^{\mathcal{D}}$, we may make a chage-of-variables via importance sampling, i.e., $\zeta(s,a) = \frac{d(s,a)}{d^{\mathcal{D}}(s,a)}$.

$$\begin{aligned} &\min_{Q} \max_{\zeta \geq 0} \; L(Q, \zeta) \\ &:= \; (1-\gamma) \cdot \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}}[Q(s_0, a_0)] + \mathbb{E}_{(s,a) \sim d^{\mathcal{D}}}[\zeta(s,a) \cdot (R(s,a) + \gamma \cdot \mathcal{P}^{\pi} Q(s,a) - Q(s,a))] \\ &= (1-\gamma) \cdot \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}}[Q(s_0, a_0)] + \mathbb{E}_{\substack{(s,a,s') \sim d^{\mathcal{D}} \\ a' \sim \pi(s')}}[\zeta(s,a) \cdot (R(s,a) + \gamma Q(s',a') - Q(s,a))]. \quad (36) \end{aligned}$$

- Doubly robust property

$$L(Q^*,\zeta) = L(Q,\zeta^*) = L(Q^*,\zeta^*) = \rho(\pi).$$

  Thus, this estimator is robust to errors in at most one of $Q$ and $\zeta$.
- Learning $Q^\pi$ values using rewards turns out to be difficult in practice.
- The bilinear nature of the Lagrangian can lead to instability or poor convergence in optimization[2].

[2] Boosting the actor with dual critic
Faster saddle-point optimization for solving large-scale markov decision processes

# Change the Probelm Before Applying Duality

The dual of Q-LP:

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

- The problem is over-constrained: The $|S| \times |A|$ constraints uniquely determine $d^{\pi}$.
- One may replace the objective function without addecting the optimal solution.

# Constant Function

If objective function $h$ is taken to be the constant function $h(d) := 0$.

$$\min_{Q} \max_{\zeta} \; L(Q, \zeta) \\ = (1-\gamma) \cdot \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}}[Q(s_0, a_0)] + \mathbb{E}_{\substack{(s,a,s') \sim d^{\mathcal{D}} \\ a' \sim \pi(s')}}[\zeta(s,a) \cdot (\gamma Q(s', a') - Q(s,a))].$$

- The optimization doesn't involve learning Q-values with repect to environment rewards.
- The Lagrangian is linear in both $Q$ and $\zeta$.

# $f$-Divergence

Let $h(d) := D_f(d||d^{\mathcal{D}})$:

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

# $f$-Divergence

Lagrange Duality:

$$\max_d \min_Q L(Q,d)$$
$$:= -D_f(d \| d^{\mathcal{D}}) + \sum_{s,a} Q(s,a) \cdot ((1-\gamma)\mu_0(s)\pi(a|s) + \gamma \cdot \mathcal{P}_*^\pi d(s,a) - d(s,a))$$
$$= (1-\gamma) \cdot \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}}[Q(s_0,a_0)] - D_f(d \| d^{\mathcal{D}}) + \sum_{s,a} Q(s,a) \cdot (\gamma \cdot \mathcal{P}_*^\pi d(s,a) - d(s,a))$$

Make the change-of-variables:

$$\max_\zeta \min_Q L(Q,\zeta)$$
$$:= (1-\gamma) \cdot \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}}[Q(s_0,a_0)] - \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[f(\zeta(s,a))] + \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[\zeta(s,a) \cdot (\gamma \cdot \mathcal{P}^\pi Q(s,a) - Q(s,a))]$$
$$= (1-\gamma) \cdot \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}}[Q(s_0,a_0)] + \mathbb{E}_{(s,a,s')\sim d^{\mathcal{D}}}[\zeta(s,a) \cdot (\gamma \cdot \mathcal{P}^\pi Q(s,a) - Q(s,a)) - f(\zeta(s,a))]. \quad (45)$$

# $f$-Divergence

We write the problem as

$$\max_d -g(-Ad) - h(d)$$

where $g(-Ad)$ corresponds to the linear constraints with repect to the adjoint Bellman operator:

$$g := \delta_{\{(1-\gamma)\mu_0 \times \pi\}} \text{ and } A := \gamma P_*^\pi - I.$$

The dual problem is therefore given by:

$$\begin{aligned} &\min_Q \; g_*(Q) + h_*(A_* Q) \\ =&\min_Q \; (1-\gamma) \cdot \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}}[Q(s_0, a_0)] + \mathbb{E}_{(s,a) \sim d^{\mathcal{D}}}[f_*(\gamma \cdot \mathcal{P}^\pi Q(s,a) - Q(s,a))]. \end{aligned}$$

# $f$-Divergence

If we set $f = \frac{1}{2}x^2$,

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

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

# Summary

- Q-LP with Lagrangian is MQL.
- The dual of Q-LP with constant function is MWL.
- The dual of Q-LP with $f$-Divergence and using Lagrange Duality is dual form of DualDICE.
- The dual of Q-LP with $f$-Divergence and using Fenchel-Rockafellar Duality is primal form of DualDICE.

# Table of Contents

# The policy Gradient Theorem

1. By Danskin's theorem:

$$\frac{\partial}{\partial \pi}\rho(\pi) = \frac{\partial}{\partial \pi}\min_{Q}\max_{d \geq 0} L(Q, d; \pi) = \frac{\partial}{\partial \pi} L\left(Q^*, d^*; \pi\right)$$

2. We may compute the gradient of $L(Q^*, d^*, \pi)$ w.r.t. $\pi$ term-by-term.

# The Policy Gradient Theorem

1. For the first term

$$\begin{aligned}
&\frac{\partial}{\partial \pi}(1-\gamma) \cdot \mathbb{E}_{a_0 \sim \pi(s_0), s_0 \sim \mu_0}\left[Q^*\left(s_0, a_0\right)\right] \\
&= (1-\gamma) \cdot \mathbb{E}_{a_0 \sim \pi(s_0), s_0 \sim \mu_0}\left[Q^*\left(s_0, a_0\right) \nabla \log \pi\left(a_0 \mid s_0\right)\right]
\end{aligned}$$

2. For the second term

$$\begin{aligned}
\frac{\partial}{\partial \pi} \mathbb{E}_{(s,a)\sim d^*}[R(s,a) + \gamma \cdot \mathcal{P}^{\pi} Q^*(s,a) - Q^*(s,a)] &= \mathbb{E}_{(s,a)\sim d^*}\left[\gamma \cdot \frac{\partial}{\partial \pi} \mathbb{E}_{\substack{s' \sim T(s,a) \\ a' \sim \pi(s')}}[Q^*(s', a')]\right] \\
&= \gamma \cdot \mathbb{E}_{\substack{(s,a)\sim d^*, s' \sim T(s,a) \\ a' \sim \pi(s')}}[Q^*(s', a') \nabla \log \pi(a'|s')]. \qquad (53)
\end{aligned}$$

# The Policy Gradient Theorem

1. Bellman equation

$$d^{\pi}(s,a) = (1-\gamma)\mu_0(s)\pi(a \mid s) + \gamma\pi(a \mid s)\sum_{\tilde{s},\tilde{a}} T\left(s' \mid \tilde{s},\tilde{a}\right) d^{\pi}(\tilde{s},\tilde{a})$$

2. $$\frac{\partial}{\partial \pi} L\left(Q^*, d^*; \pi\right) = \mathbb{E}_{(s,a)\sim d^{\pi}}\left[Q^{\pi}(s,a)\nabla \log \pi(a \mid s)\right]$$

# Fenchel-Rockafellar Duality for Regularized Optimization

Consider regularizing the max-reward policy objective with the $f$-divergence:

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

Fenchel-Rockafellar duality yeilds the following dual formulation:

$$\begin{aligned}\rho(\pi) - D_f(d^\pi \| d^{\mathcal{D}}) = \min_Q \ (1-\gamma) \cdot \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}}[Q(s_0, a_0)] + \\ \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[f_*(R(s,a) + \gamma \cdot \mathcal{P}^\pi Q(s,a) - Q(s,a))].\end{aligned}$$

# Regularization with the KL-Divergence

The optimization objective can be formulated as

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

For a specific $Q$, the gradient of this objective with respect to $\pi$ is

$$\begin{aligned}(1-\gamma) \cdot \mathbb{E}_{\substack{a_0 \sim \pi(s_0) \\ s_0 \sim \mu_0}} [Q(s_0, a_0) \nabla \log \pi(a_0|s_0)] \\ + \gamma \cdot \mathbb{E}_{\substack{(s,a,s') \sim d^{\mathcal{D}} \\ a' \sim \pi(s')}} [\mathrm{softmax}_{d^{\mathcal{D}}} (R + \gamma \cdot \mathcal{P}^{\pi} Q - Q)(s,a) \cdot Q(s', a') \nabla \log \pi(a'|s')],\end{aligned}$$

· Bears simiarities to max-likelihood policy learning.

# Imitation Learning

- If one ignores rewards, the optimization corresponds to finding a policy $\pi$ which minimizes the $f$-divergence in terms of the state-action occupancies from $d^{\mathcal{D}}$.
- With the same techniques as we applied for offline policy evaluation and offline policy optimization, one can derive offline imitation learning algorithms.

# Table of Contents

# V-LP

- V-LP

$$\begin{aligned} \min_{V}\ & (1-\gamma)\cdot \mathbb{E}_{s_0\sim\mu_0}[V(s_0)] \\ \text{s.t. } & V(s) \geq R(s,a) + \gamma\cdot \mathcal{T}V(s,a), \\ & \quad \forall s\in S, a\in A, \end{aligned}$$

- The dual of V-LP

$$\begin{aligned} \max_{d\geq 0}\ & \sum_{s,a} d(s,a)\cdot R(s,a) \\ \text{s.t. } & \sum_{a} d(s,a) = (1-\gamma)\mu_0(s) + \gamma\cdot \mathcal{T}_* d(s), \\ & \forall s\in S, \end{aligned}$$

# V-LP

- The probelm is not over-constraint.
- Cannot ignore constraints $d \geq 0$.
- This lead to a dual objective over two functions: $V : S \to \mathbb{R}$ and $K : S \times A \to \mathbb{R}_+$:

$$\min_{K \geq 0, V} \; (1-\gamma) \cdot \mathbb{E}_{s_0 \sim \mu_0}[V(s_0)] + \mathbb{E}_{(s,a) \sim d^{\mathcal{D}}}[f_*(K(s,a) + R(s,a) + \gamma \cdot \mathcal{T}V(s,a) - V(s))].$$

- This objective only involves a single optimization over $V$ and $K$ s opposed to a max-min optimization over $pi$ and $Q$.
- The solution will give us $V^*$ rather than the policy itself.

# V-LP

- To derive the optimal policy,

$$d^*(s,a) = d^{\mathcal{D}}(s,a) \cdot f'_*(K^*(s,a) + R(s,a) + \gamma \cdot \mathcal{T}V^*(s,a) - V^*(s)).$$

- Using Bayes's rule,

$$\pi^*(a|s) = \frac{d^*(s,a)}{\sum_{\tilde{a}} d^*(s,\tilde{a})} = \frac{d^{\mathcal{D}}(s,a) \cdot f'_*(K^*(s,a) + R(s,a) + \gamma \cdot \mathcal{T}V^*(s,a) - V^*(s))}{\sum_{\tilde{a}} d^{\mathcal{D}}(s,\tilde{a}) \cdot f'_*(K^*(s,a) + R(s,\tilde{a}) + \gamma \cdot \mathcal{T}V^*(s,\tilde{a}) - V^*(s))}.$$

# Max-Likelihood Policy Learning

- Regularization with $D_{\mathrm{KL}}(d||d^{\mathcal{D}})$ for the dual of V-LP yeilds
$$\min_V \ (1-\gamma)\cdot \mathbb{E}_{s_0\sim\mu_0}[V(s_0)] + \log \mathbb{E}_{(s,a)\sim d^{\mathcal{D}}}[\exp\{R(s,a)+\gamma\cdot\mathcal{T}V(s,a)-V(s)\}],$$
- Aviod the numerical instability and ensure the positiveness of $d$.
- The visitations of the optimal policy are now given by the softmax function:
$$d^{\pi^*}(s,a) = d^{\mathcal{D}}(s,a)\cdot \mathrm{softmax}_{d^{\mathcal{D}}}(R+\gamma\cdot\mathcal{T}V^*-V^*)(s,a).$$
- The optimal policy thus has a similar form:
$$\pi^*(a|s) = d^{\mathcal{D}}(a|s)\cdot \mathrm{softmax}_{d^{\mathcal{D}}(\cdot|s)}(R(s,\cdot)+\gamma\cdot\mathcal{T}V^*(s,\cdot)-V^*(s))(a).$$

# Policy Evaluation with the V-LP

- We deconpose $d(s,a) = \mu(s)\pi(a|s)$ for a fixed policy $\pi(a|s)$:

$$\begin{aligned}\max_{\mu} \ & \sum_{s,a} \mu(s)\pi(a|s) \cdot R(s,a) \\ \text{s.t. } & \mu(s) = (1-\gamma)\mu_0(s) + \gamma \cdot \mathcal{T}_*(\mu \times \pi)(s), \\ & \forall s \in S.\end{aligned}$$

- The LP is over-constrained.
- We can replace the objective function as Q-LP.
- This require the knowledge of $d^{\mathcal{D}}(a|s)$.