Stochastic Optimal Feedback Control

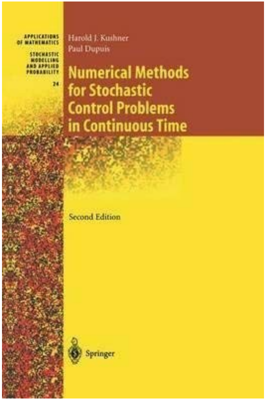


Figure: Kushner, Harold, and Paul G. Dupuis. Numerical methods for stochastic control problems in continuous time. Vol. 24. Springer Science & Business Media, 2013.

Stochastic Optimal Feedback Control

Jongeun Choi[1]
Associate Professor
Mem. ASME
Department of Mechanical Engineering,
Department of Electrical and
Computer Engineering,
Michigan State University,
East Lansing, MI 48864
e-mail: jchoi@egr.msu.edu

Dejan Milutinović
Associate Professor
Mem. ASME
Computer Engineering Department,
University of California,
Santa Cruz,
Santa Cruz, CA 95064
e-mail: dejan@soe.ucsc.edu

Tips on Stochastic Optimal Feedback Control and Bayesian Spatiotemporal Models: Applications to Robotics

This tutorial paper presents the expositions of stochastic optimal feedback control theory and Bayesian spatiotemporal models in the context of robotics applications. The presented material is self-contained so that readers can grasp the most important concepts and acquire knowledge needed to jump-start their research. To facilitate this, we provide a series of educational examples from robotics and mobile sensor networks.

[DOI: 10.1115/1.4028642]

Figure: J. Dyn. Sys., Meas., Control. Mar 2015, 137(3): 030801 (10 pages) Paper No: DS-14-1068 https://doi.org/10.1115/1.4028642 Published Online: October 21, 2014

Stochastic Optimal Feedback Control

Langevin Equation and Îto Integrals

- A stochastic differential equation (SDE) describes the uncertain dynamics
$$\frac{dx}{dt} = a(x,t) + b(x,t)\xi(t) \tag{1}$$
- $a(x,t)$ and $b(x,t)$ are nonlinear functions, i.e., mappings of appropriate dimensions
- $\xi(t)$ is the so-called process noise and is considered to be the zero-mean, unit intensity white noise
- $\mathbb{E}\{\xi(t)\} = 0 \;\text{ and }\; \mathbb{E}\{\xi(t_i)\xi(t_j)\} = I_{m\times m}\delta(t_i - t_j)$
- $I_{m\times m}$ is the unity matrix of dimension $m \times m$, t_i and t_j denote two arbitrary time points, and the function $\delta(t)$ is the Dirac delta function

- Multiply the Langevin equation (1) by dt

$$dx = a(x,t)dt + b(x,t)dw \tag{2}$$

- $dw(t) = \xi(t)dt$ is an increment of the Wiener process $w(t)$ at time point t, i.e., $dw(t) := w(t+dt) - w(t)$
- $\mathbb{E}\{dw(t)^2\} = I_{m\times m}$ dt
- The solution of equation (2) can be expressed as a sum of two integral terms

$$x(t) = x(t_0) + \int_{t_0}^{t} a(x,\tau)d\tau + \int_{t_0}^{t} b(x,\tau)dw$$

- The solution $x(t)$ can be approximated as

$$x(t) \approx x(t_0) + \sum_{k=0}^{N-1} a(x, \tau_k)\Delta t + \sum_{k=0}^{N-1} b(x, \tau_k)\Delta w_k \tag{3}$$

- $\tau_{k+1} - \tau_k = \Delta t$, $\tau_k \in [t_k, t_{k+1}]$, $\Delta w_k = w(t_{k+1}) - w(t_k)$
- $t_k = t_0 + k\Delta t, \Delta t = \frac{t-t_0}{N}$
- If the sampling points are chosen to be $\tau_k = t_k$, (3) can be rewritten in an iterative form as

$$x(t_{k+1}) \approx x(t_k) + a(x, t_k)\Delta t + b(x, t_k)\Delta w_k$$

Stochastic Optimal Feedback Control

Îto calculus chain rule

- Use the second-order Taylor expansion of f(x)
- Substitute $(dw)^2$ with dt
- Ignore every term of the form dt^p with p $>$ 1

$$df(x) = \frac{\partial f(x)}{\partial x} dx + \frac{1}{2} \frac{\partial^2 f(x)}{\partial x^2} dx^2$$

- Substitute dx from (2) and $(dw)^2$ with dt,

$$df(x) = \left(\frac{\partial f(x)}{\partial x} a + \frac{1}{2} \frac{\partial^2 f(x)}{\partial x^2} b^2 \right) dt + \frac{\partial f(x)}{\partial x} b dw$$

- For multidimensional SDE in (2), the Îto chain rule is

$$df(x) = \left(\frac{\partial f^T}{\partial x} a(x(t), t) + \frac{1}{2} \mathrm{tr} \left\{ \frac{\partial^2 f}{\partial x^2} b b^T \right\} \right) dt + \frac{\partial f}{\partial x} b(x(t), t) dw$$

Stochastic Optimal Feedback Control

- Fixed velocity two-wheel robot control problems

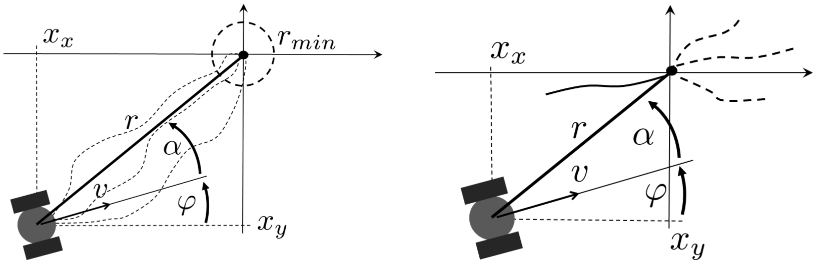


Figure: (a) Minimum expected time control; (b) Distance keeping control; (x_x, x_y) - the robot coordinates relative to the target which is at the origin, r - distance between the robot and the target, φ - robot heading angle, α - bearing angle and v - velocity

Stochastic Optimal Feedback Control

Minimum expected time control

- Robot model

$$
\begin{aligned}
dx_x &= v \cos \varphi dt \\
dx_y &= v \sin \varphi dt \\
d\varphi &= udt + \sigma_r dw
\end{aligned}
$$

- Cost function (expected time to the target)

$$
J(u) = \mathbb{E}\left\{ \int_0^{\tau} 1 \ dt \right\}
$$

- By using the relative coordinates $r = \sqrt{x_x^2 + x_y^2}$ and α, the robot model becomes

$$
\begin{aligned}
dr &= -v \cos \alpha \ dt \\
d\alpha &= \left(\frac{v}{r} \sin \alpha - u \right) \ dt + \sigma_r dw
\end{aligned}
$$

Stochastic Optimal Feedback Control

Minimum expected time control

- Derivation

  By using (1) and (4),

$$
\begin{aligned}
dr &= \begin{bmatrix} \frac{x_x}{\sqrt{x_x^2+x_y^2}} & \frac{x_y}{\sqrt{x_x^2+x_y^2}} \end{bmatrix} \begin{bmatrix} v\cos\varphi \\ v\sin\varphi \end{bmatrix} dt \\
&= \left(\frac{x_x}{\sqrt{x_x^2+x_y^2}} v\cos\varphi + \frac{x_y}{\sqrt{x_x^2+x_y^2}} v\sin\varphi \right) dt \\
&= \left(-\cos(\alpha+\varphi) v\cos\varphi\right) - \sin(\alpha+\varphi) v\sin\varphi)\, dt \\
&= (-(\cos\alpha\cos\varphi - \sin\alpha\sin\varphi) v\cos\varphi - (\sin\alpha\cos\varphi + \cos\alpha\sin\varphi) v\sin\varphi) \\
&= -v\cos\alpha dt
\end{aligned}
$$

Stochastic Optimal Feedback Control

Minimum expected time control

- Derivation
  Similarly, we define $f(x) := \tan^{-1}\left(\frac{x_y}{x_x}\right) - \varphi$.

$$
\begin{aligned}
d\alpha &= \begin{bmatrix} -\frac{x_y}{r^2} & \frac{x_x}{r^2} & -1 \end{bmatrix} \begin{bmatrix} v\cos\varphi \\ v\sin\varphi \\ u \end{bmatrix} dt + \sigma_r dw \\
&= \left(\frac{v}{r^2}(x_x \sin\varphi - x_y \cos\varphi) - u \right) dt + \sigma_r dw \\
&= \left(\frac{v}{r^2}\big(- r\cos(\alpha+\varphi)\sin\varphi + r\sin(\alpha+\varphi)\cos\varphi\big) - u \right) dt + \sigma_r dw \\
&= \left(\frac{v}{r^2}(r\sin\alpha) - u \right) dt + \sigma_r dw \\
&= \left(\frac{v}{r}\sin\alpha - u \right) dt + \sigma_r dw
\end{aligned}
$$

Stochastic Optimal Feedback Control

Minimum expected time control

- $x = \begin{bmatrix} x_x & x_y & \varphi \end{bmatrix}^T, b = \begin{bmatrix} 0 & 0 & \sigma_r \end{bmatrix}^T$
-

$$\frac{1}{2} tr \left\{ \frac{\partial^2 f}{\partial x^2} b b^T \right\} = \frac{1}{2} tr \left\{ \begin{bmatrix} 0 & -\frac{1}{r^2} & 0 \\ \frac{1}{r^2} & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \sigma_r^2 \end{bmatrix} \right\} = 0$$

Stochastic Optimal Feedback Control

Minimum expected time control

- Cost-to-go function $V(r, \alpha)$
  - Expected cost from the point in space (r, α)
  - The solution of the Hamilton-Jacobi-Bellman equation (HJB)

$$0 = \min_u \left\{ b_1 \frac{\partial V}{\partial r} + b_2(u) \frac{\partial V}{\partial \alpha} + \frac{\sigma_r^2}{2} \frac{\partial^2 V}{\partial \alpha^2} + 1 \right\} \tag{4}$$

  - $b_1 = -v \cos \alpha$, $b_2(u) = \frac{v}{r} \sin \alpha - u$
- Derivation of HJB from the Bellman equation

$$\begin{aligned} V(x,t) &= \min_u \{ \ell(x,u)\Delta t + E[V(x', t+\Delta t)] \} \\ &= \min_u \{ \ell(x,u)\Delta t + E[V(x+\delta x, t+\Delta t)] \} \end{aligned}$$

  - Here, the cost rate $\ell(x,u) = 1$
- Use the Taylor-series expansion of V

$$V(x+\delta x, t+\Delta t) = V(x,t) + \frac{\partial V}{\partial t}\Delta t + \delta x^T \frac{\partial V}{\partial x} + \frac{1}{2}\delta x^T \frac{\partial^2 V}{\partial x^2}\delta x$$

Stochastic Optimal Feedback Control

Hamilton-Jacobi-Bellman equation

- Using the fact that $E[d^T M d] = \text{tr}(\text{cov}[d]M)$, the expectation is

$$E[V(x+\delta x, t+\Delta t)] = V(x,t) + \frac{\partial V}{\partial t}\Delta t + \frac{\partial V}{\partial x} a(x(t),t)^T \Delta t + \frac{1}{2}\text{tr}\left(bb^T \frac{\partial^2 V}{\partial x^2}\right)\Delta t$$

- Substituting $E[V(x+\delta x, t+\Delta t)]$ in the Bellman equation,

$$V(x,t) = \min_u \left\{ \begin{aligned} &\Delta t + V(x,t) + \frac{\partial V}{\partial t}\Delta t \\ &+ \frac{\partial V}{\partial x} a(x(t),t)^T \Delta t + \frac{1}{2}\text{tr}\left(bb^T \frac{\partial^2 V}{\partial x^2}\right)\Delta t \end{aligned} \right\}$$

- Simplifying, dividing by Δt yields the HJB equation

$$-\frac{\partial V}{\partial t} = \min_u \left\{ 1 + \frac{\partial V}{\partial x} a(x(t),t)^T + \frac{1}{2}\text{tr}\left(bb^T \frac{\partial^2 V}{\partial x^2}\right) \right\}$$

- In order to discretize HJB, we substitute (4) as follows.
  - $b_1 \frac{\partial V}{\partial r} = \frac{V(r+\Delta r,\alpha) - V(r,\alpha)}{\Delta r} b_1^+ - \frac{V(r,\alpha) - V(r-\Delta r,\alpha)}{\Delta r} b_1^-$, which is the derivative's upwind approximation
  - $b_1^+ = \max[0, b_1]$, $b_1^- = \max[0, -b_1]$
  - Similarly, $b_2 \frac{\partial V}{\partial \alpha} = \frac{V(r,\alpha+\Delta\alpha) - V(r,\alpha)}{\Delta\alpha} b_2^+ - \frac{V(r,\alpha) - V(r,\alpha-\Delta\alpha)}{\Delta\alpha} b_2^-$
  - $\frac{\partial^2 V}{\partial \alpha^2} = \frac{V(r,\alpha+\Delta\alpha) + V(r,\alpha-\Delta\alpha) - 2V(r,\alpha)}{(\Delta\alpha)^2}$

Stochastic Optimal Feedback Control

Hamilton-Jacobi-Bellman equation

- If we move all the terms that include $V(r,\alpha)$ to the left side of equation(4), define $|b_1| = b_1^+ + b_1^-$, $|b_2| = b_2^+ + b_2^-$ and $\Delta t = \left(\frac{|b_1|}{\Delta r} + \frac{|b_2|}{\Delta \alpha} + \frac{\sigma_r^2}{(\Delta\alpha)^2}\right)^{-1}$, we obtain

$$\begin{aligned} V(r,\alpha) = &\min_u \{ p_{\Delta r^+} V(r+\Delta r,\alpha) + p_{\Delta r^-} V(r-\Delta r,\alpha) \\ &+ p_{\Delta\alpha^+} V(r,\alpha+\Delta\alpha) + p_{\Delta\alpha^-} V(r,\alpha-\Delta\alpha) + \Delta t \} \end{aligned} \tag{5}$$

- $p_{\Delta r^\pm} = \Delta t \frac{b_1^\pm}{\Delta r}$ and $p_{\Delta\alpha^\pm} = \Delta t \left(\frac{b_2^\pm}{\Delta\alpha} + \frac{\sigma_r^2}{2\Delta\alpha^2}\right)$ that can be interpreted as the discrete Markov-chain transition probabilities

Stochastic Optimal Feedback Control

Minimum expected time control

- Numerically solve (5) using value iteration

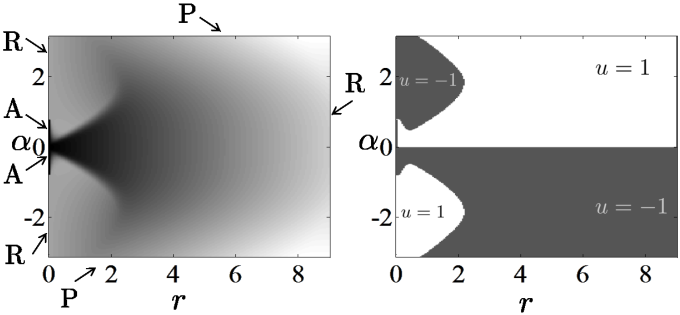


Figure: Solution of the minimum expected time problem (P1): (left panel) gray colored map of the value function $V(r, \alpha)$; black color at the absorbing boundary (A) indicates $V(r, \alpha) = 0$ and the lighter shades depict longer expected times. The type of the boundary conditions is labeled by P-periodic, R-reflective, A-absorbing; (right panel) optimal feedback control; white $u = 1$ and gray $u = -1$.

Stochastic Optimal Feedback Control

Minimum expected time control

- Simulated trajectories of the optimal feedback control

Simulation results of the minimum expected time problem (P1): $x_0 = \begin{bmatrix} 0 & 0 & \frac{\pi}{18} \end{bmatrix}^T$.

Stochastic Optimal Feedback Control

Minimum expected time control

- Simulated trajectories of the optimal feedback control

Simulation results of the minimum expected time problem (P1):
$x_0 = \begin{bmatrix} 0 & 0 & \frac{\pi}{4} \end{bmatrix}^T$;

Stochastic Optimal Feedback Control

Minimum expected time control

- Simulated trajectories of the optimal feedback control

Simulation results of the minimum expected time problem (P1): starting from an opposite heading angle.

Dynamic Programming

A method for solving complex problems by breaking them dowm into subproblems.

Requirements for dynamic programming:

- Optimal substructure
  - Principle of optimality applies
  - Optimal solution can be decomposed into subproblems
- Overlapping subproblems
  - Subproblems recur many times
  - Solutions can be cached and reused

Markov decision process satisfy both properties

- Bellman equation gives recursive decomposition
- Value function stores and reuses solutions

Value Iteration

Repeatedly update an estimate of the optimal value function according to Bellman optimality equation

1. Initialize an estimate for the value function arbitrarily

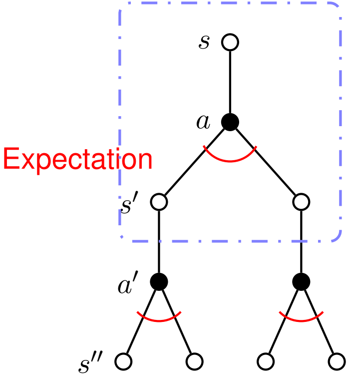
$$\hat{V}(s) \leftarrow 0, \quad \forall s \in \mathcal{S}$$

2. Repeat, update:

$$\hat{V}(s) \leftarrow \max_{a \in \mathcal{A}} \left[\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a \hat{V}(s') \right], \quad \forall s \in \mathcal{S}$$

Dynamic Programming Policy Evaluation

$$\hat{V}(s) \leftarrow \max_{a \in \mathcal{A}} \left[\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a \hat{V}(s') \right], \quad \forall s \in \mathcal{S}$$



DP compute this, bootstraping the rest of the expected return by the value estimate $\hat{V}$.

Contraction Mapping Theorem

Definition 1

Let (X, d) be a complete metric space. Then a map $T : X \to X$ is called a contraction mapping on X if there exists $q \in [0, 1)$ such that

$$d(T(x), T(y)) \leq qd(x, y), \quad \forall x, y \in X$$

Theorem 2 (Contraction Mapping Theorem)

Let (X, d) be a non-empty complete metric space with a contraction mapping $T : X \to X$. Then T admits a unique fixed-point $x^* \in X$ (i.e. $T(x^*) = x^*$). Furthermore, x^* can be found as follows: start with an arbitrary element $x_0 \in X$ and define a sequence $\{x_n\}$ by $x_n = T(x_{n-1})$ for $n \geq 1$. Then $x_n \to x^*$

Convergence of Value Iteration

Theorem 3

Value iteration converges to optimal value: $\hat{V} \to V^*$

Proof.

For any estimate of the value function $\hat{V}$, we define the Bellman backup operator $B : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$

$$B(\hat{V}) = \max_{\pi}(\mathcal{R}^{\pi} + \gamma \mathcal{P}^{\pi} \hat{V})$$

We will show that Bellman operator is a γ-contraction, that for any value function estimates V_1, V_2

$$\begin{aligned}
\|B(V_1) - B(V_2)\|_{\infty} &= \| \max_{\pi}(\mathcal{R}^{\pi} + \gamma \mathcal{P}^{\pi} V_1) - \max_{\pi}(\mathcal{R}^{\pi} + \gamma \mathcal{P}^{\pi} V_2)\|_{\infty} \\
&\leq \gamma \max_{\pi} \|\mathcal{P}^{\pi}(V_1 - V_2)\|_{\infty} \\
&\leq \gamma \max_{\pi} \|\mathcal{P}^{\pi}\|_{\infty} \|V_1 - V_2\|_{\infty}
\end{aligned}$$

$\square$

Since $\|\mathcal{P}^\pi\|_\infty = \max_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'} = 1$ and $\max_\pi \|V_1 - V_2\|_\infty = \|V_1 - V_2\|_\infty$,

$$\|B(V_1) - B(V_2)\|_\infty \leq \gamma \|V_1 - V_2\|_\infty$$

From the contraction mapping theorem, a unique fixed point V^* satisfies $B(V^*) = V^*$

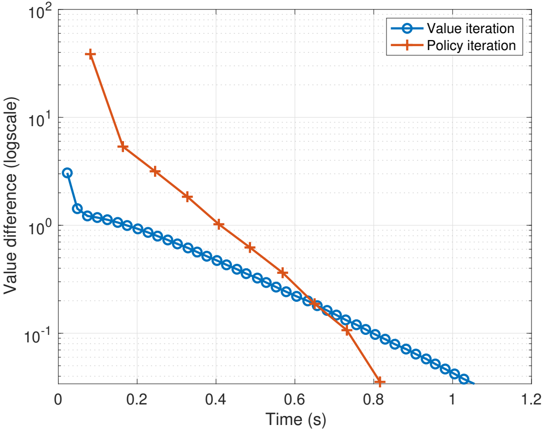
$$\|B(\hat{V}) - V^*\|_\infty \leq \gamma \|\hat{V} - V^*\|_\infty \Rightarrow \hat{V} \to V^*$$

Policy Iteration

Repeatedly update an estimate of the optimal value function according to Bellman optimality equation

1. Initialize random policy $\hat{\pi}$
2. Compute the value of the policy, V^π via solving the linear system
$$V^\pi = (I - \gamma \mathcal{P}^\pi)^{-1} \mathcal{R}^\pi$$
3. Update π to be greedy policy w.r.t V^π
$$\pi(s) \leftarrow \arg\max_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} \mathcal{P}^a_{ss'} V^\pi(s')$$
4. If policy π changed in last iteration, return to step 2.

Value Iteration vs. Policy Iteration



- 64×64 gridworld example with randomly given reward and transition probabilities, stopping criteria is $\|V_{k+1} - V_k\|_2 < 0.03$.
- Value iteration converges in 42 steps.
- Policy iteration converges in 10 steps.

Value Iteration vs. Policy Iteration

- Policy iteration is desirable because of its finite-time convergence to the optimal policy (since the value is acquired analytically by solving the linear system).
- However, policy iteration requires solving possibly large linear systems: each iteration takes $\mathcal{O}(|\mathcal{S}|^3)$ time.
- Value iteration requires $\mathcal{O}(|\mathcal{S}| \times |\mathcal{A}|)$ time at each iteration.
- Typically, policy iteration converges faster, in spite of the larger computation time in a single iteration.

Linear Programming Solution Methods

Consider the following optimization problem

$$\begin{aligned}
&\underset{V}{\text{minimize}} \ \sum_{s \in \mathcal{S}} p(s)V(s) \\
&\text{subject to } V(s) \geq \max_{a \in \mathcal{A}} \left[\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a V(s') \right], \quad \forall s \in \mathcal{S}
\end{aligned}$$

The optimal solution of above problem will satisfies Bellman optimality equation for all $s \in \mathcal{S}$, which means the solution will be V^*. But it is hard to deal with those nonlinear inequality constraints.

Linear Programming Solution Methods

We can capture the constraint

$$V(s) \geq \max_{a \in \mathcal{A}} \left[\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a V(s') \right]$$

via the set of $|\mathcal{A}|$ linear constraints

$$V(s) \geq \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a V(s'), \quad \forall a \in \mathcal{A}.$$

Now consider the linear program

$$\begin{aligned} &\underset{V}{\text{minimize}} \sum_{s \in \mathcal{S}} p(s) V(s) \\ &\text{subject to } V(s) \geq \mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a V(s'), \quad \forall s \in \mathcal{S}, a \in \mathcal{A} \end{aligned}$$

Linear Programming Dual Problem

Primal problem

$$\begin{aligned} &\underset{\mathbf{x}}{\text{minimize}}\ \mathbf{c}^\top \mathbf{x} \\ &\text{subject to } A\mathbf{x} \preceq \mathbf{b} \end{aligned}$$

$$\Downarrow$$

Dual problem

$$\begin{aligned} &\underset{\lambda}{\text{maximize}}\ -\mathbf{b}^\top \lambda \\ &\text{subject to } A^\top \lambda + \mathbf{c} = 0 \\ &\lambda \succeq 0 \end{aligned}$$

Linear Programming Dual Problem

Adding dual variables $\lambda(s, a)$ for each constraint, dual problem is

$$\begin{aligned}
\underset{\lambda(s,a)}{\text{maximize}} \quad & \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \mathcal{R}_s^a \lambda(s,a) \\
\text{subject to} \quad & \sum_{a'\in\mathcal{A}} \lambda(s',a') = p(s') + \gamma \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \mathcal{P}_{ss'}^a \lambda(s,a), \quad \forall s' \in \mathcal{S} \\
& \lambda(s,a) \geq 0, \quad \forall s \in \mathcal{S}, a \in \mathcal{A}
\end{aligned}$$

These have the interpretation that

$$\lambda(s,a) = \sum_{t=0}^{\infty} \gamma^t p(s_t = s, a_t = a).$$

Linear Programming Dual Problem

Dual problem is equivalent to policy iteration:

- Objective:

$$\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\mathcal{R}_s^a\lambda(s,a) = \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\mathcal{R}_s^a\sum_{t=0}^{\infty}\gamma^t p(s_t = s, a_t = a)$$

$$= \mathbb{E}[\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t)]$$

- Optimal policy:

$$\pi^*(s) = \arg\max_a \lambda(s,a)$$