9

Policy gradient methods

9.1 Set-up

Given some initial distribution ξ_{in} over the state space and a policy π, we can consider the objective function

$$\pi \mapsto V^{\pi}(\xi_{\text{in}}) = \mathbb{E}_{\xi_{\text{in}}}[V^{\pi}(X)].$$

Policy-based estimates are procedures for attempting to maximize this function over some space of (parameterized) policies.

Let us consider some examples to illustrate.

- tabular policies over probability simplex
- linear policies for continuous state-action problems
- linear soft-max policies using feature vectors

9.2 Parameterized policies and gradient estimation

Consider a parameterized family of policies, say of the form $\pi_\theta(x \mid u)$.

9.2.1 Basic importance sampling estimate

Let $f_{\pi_\theta, \xi_{\text{in}}}$ denote the density of trajectories $\tau = \{(x_t, u_t)\}_{t=0}^{\infty}$ that are drawn with

$$\text{Initial state } x_0 \sim \xi_{\text{in}}, \quad \text{actions } u_t \sim \pi_\theta(u_t \mid x_t), \quad \text{and successor states } x_{t+1} \sim \mathbb{P}_{u_t}(x_{t+1} \mid x_t). \tag{9.1}$$

We frequently adopt the shorthand f_θ for this quantity.

We define the discounted reward over this trajectory as

$$S_\gamma(\tau) = \sum_{t=0}^{\infty} \gamma^t r(x_t, u_t).$$

Proposition 9.1 (Importance sampling for policy gradients) *Let f be any probability density over the trajectory space $\mathcal{T}$ that takes strictly positive values on the support of*

f_θ. Then we have the stochastic gradient representation

$$\nabla_\theta V^{\pi_\theta}(\xi_{in}) = \mathbb{E}_{\tau\sim f}\left[\frac{S_\gamma(\tau)\nabla_\theta f_\theta(\tau)}{f(\tau)}\right] \tag{9.2a}$$

In particular, with the choice $f = f_\theta$, we have

$$\nabla_\theta V^{\pi_\theta}(\xi_{in}) = \mathbb{E}_{\tau\sim f_\theta}\left[S_\gamma(\tau)\nabla_\theta \log f_\theta(\tau)\right] \tag{9.2b}$$

Recall that

$$\log f_\theta(\tau) = \xi_{\text{in}}(x_0) + \sum_{t=0}^{\infty} \log \pi_\theta(u_t \mid x_t).$$

It is useful to observe the alternative expression

$$\mathbb{E}_{\tau\sim f_\theta}\left[S_\gamma(\tau)\nabla_\theta \log f_\theta(\tau)\right] = \mathbb{E}_{\tau\sim f_\theta}\left[S_\gamma(\tau)\sum_{t=0}^{\infty}\nabla_\theta \log_\theta \pi_\theta(u_t \mid x_t)\right].$$

Consider the particular case of soft-max policies

$$\log \pi_\theta(u \mid x) = \langle \theta,\, \varphi(x,u)\rangle - \log \sum_a \exp\left(\langle \theta,\, \varphi(x,a)\rangle\right),$$

and hence

$$\nabla_\theta \log \pi_\theta(u \mid x) = \varphi(x,u) - \mathbb{E}_\theta[\varphi(x,U)]$$

where $\mathbb{E}_\theta$ denotes expectation over the action space under the (stochastic) policy indexed by θ.

Proof By definition, we have

$$\begin{aligned}
\nabla_\theta V^{\pi_\theta}(\xi_{\text{in}}) = \nabla_\theta \mathbb{E}_{\tau\sim f_{\pi_\theta,\xi_{\text{in}}}}[S_\gamma(\tau)] &= \int S_\gamma(\tau)\nabla_\theta f_{\pi_\theta,\xi_{\text{in}}}(\tau)d\tau \\
&= \int \frac{S_\gamma(\tau)\nabla_\theta f_{\pi_\theta,\xi_{\text{in}}}(\tau)}{g(\tau)} g(\tau)d\tau \\
&= \mathbb{E}_{\tau\sim g}\left[\frac{S_\gamma(\tau)\nabla_\theta f_{\pi_\theta,\xi_{\text{in}}}(\tau)}{g(\tau)}\right],
\end{aligned}$$

which establishes the first claim (9.2a).

Finally, by chain rule, we can write

$$\nabla_\theta \log f_{\pi_\theta,\xi_{\text{in}}}(\tau) = \frac{\nabla_\theta f_{\pi_\theta,\xi_{\text{in}}}(\tau)}{f_{\pi_\theta,\xi_{\text{in}}}(\tau)}, \tag{9.3}$$

from which the claim (9.2b) follows. □

9.3 Value-based policy gradients

For each time $t = 0, 1, 2, \ldots$ and some initial state distribution ξ_{in}, let $f^t_{\pi_\theta,\xi_{\text{in}}}$ denote the marginal probability density over states at time t over trajectories τ generated according

to the joint density $f_{\pi_\theta,\xi_{\text{in}}}$. We then define the *discounted state occupation* density

$$d^{\pi_\theta}_{\xi_{\text{in}}}(x) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t f^t_{\pi_\theta,\xi_{\text{in}}}(x) \tag{9.4}$$

Note if ρ is the stationary density for the Markov chain induced by π_θ, then we have $d^{\pi_\theta}_{\rho} = \rho$ by construction. Otherwise, the discounted state occupation accounts for the potential non-stationarity of the input distribution ξ_{in}. In terms of this object, we have the following representation:

Proposition 9.2 *For any integrable function $\varphi : \mathcal{X} \to \mathbb{R}$, we have*

$$\nabla_\theta V^{\pi_\theta}(\xi_{in}) = \frac{1}{1-\gamma}\mathbb{E}_{X\sim d^{\pi_\theta}_{\xi_{in}}}\mathbb{E}_{U\sim\pi_\theta(\cdot|X)}\left[(Q^{\pi_\theta}(X,U)-\varphi(X))\nabla_\theta\log\pi_\theta(U\mid X)\right]. \tag{9.5}$$

Proof For any integrable φ, we have

$$\begin{aligned}\mathbb{E}_{U\sim\pi_\theta(\cdot|x)}\left[\varphi(x)\nabla_\theta\log\pi_\theta(U\mid x)\right] &= \varphi(x)\mathbb{E}_{U\sim\pi_\theta(\cdot|x)}\left[\frac{\nabla_\theta\pi_\theta(U\mid x)}{\pi_\theta(U\mid x)}\right]\\ &= \varphi(x)\int_{\mathcal{U}}\nabla_\theta\pi_\theta(u\mid x)du\\ &= \varphi(x)\nabla_\theta\int_{\mathcal{U}}\pi_\theta(u\mid x)du\\ &= 0,\end{aligned}$$

where the final inequality follows from the fact that $\int_{\mathcal{U}}\pi_\theta(u\mid x)du = 1$ is independent of θ. Consequently, it suffices to prove the claim (9.5) for $\varphi \equiv 0$.

By the relation between the value and Q-value functions, we have

$$\begin{aligned}\nabla_\theta V^{\pi_\theta}(\xi_{\text{in}}) &= \nabla_\theta\mathbb{E}_{X\sim\xi_{\text{in}}}\left[\int_{\mathcal{U}}Q^{\pi_\theta}(X,u)\pi_\theta(u\mid X)du\right]\\ &\overset{(i)}{=} \underbrace{\mathbb{E}_{X\sim\xi_{\text{in}}}\Big[\int_{\mathcal{U}}\{\nabla_\theta Q^{\pi_\theta}(X,u)\}\,\pi_\theta(u\mid X)du\Big]}_{T_1} + \underbrace{\mathbb{E}_{X\sim\xi_{\text{in}}}\Big[\int_{\mathcal{U}}Q^{\pi_\theta}(X,u))\,\nabla_\theta\pi_\theta(u\mid X)du\Big]}_{T_2},\end{aligned}$$

where, in step (i), we have exchanged the order of integration and differentiation, and applied the chain rule. By the elementary relation (9.3), the second term T_2 can be written as

$$T_2 = \mathbb{E}_{X\sim\xi_{\text{in}}}\mathbb{E}_{U\sim\pi_\theta(\cdot|X)}[Q^{\pi_\theta}(X,U)\nabla_\theta\log\pi_\theta(U\mid X)].$$

As for the first term T_1, we have

$$\begin{aligned}\nabla_\theta Q^{\pi_\theta}(x,u) &= \nabla_\theta\Big\{r(x,u)+\gamma\mathbb{E}_{X'\sim\mathbb{P}_u(x|X')}\big[V^{\pi_\theta}(X')\big]\Big\}\\ &= \gamma\mathbb{E}_{X'\sim\mathbb{P}_u(x|X')}\big[\nabla_\theta V^{\pi_\theta}(X')\big].\end{aligned}$$

By iterating this argument, we find that

$$\nabla_\theta V^{\pi_\theta}(\xi_{\text{in}}) = \mathbb{E}_{\tau\sim f_{\pi_\theta,\xi_{\text{in}}}}\Big[\sum_{t=0}^{\infty}\gamma^t Q^{\pi_\theta}(x_t,u_t)\nabla_\theta\log\pi_\theta(u_t\mid x_t)\Big].$$

From the definition of the discounted occupation measure, this expression is equivalent to the claim (9.5). □

Stochastic gradients

This gives us another mechanism for computing stochastic gradients based on observed trajectories. Let $\tau = \{(x_t, u_t, r_t)_{t\geq 0}\}$ be an observed trajectory, and for each $h = 0, 1, 2, \ldots$, define the estimate

$$\widehat{Q}^{\pi_\theta}(x_h, u_h) = \sum_{t=h}^{\infty} \gamma^{t-h} \underbrace{r_t(x_t, u_t)}_{r_t}. \tag{9.6}$$

Note that from the construction of the trajectory and the definition of Q-functions we have

$$\mathbb{E}_\tau[\widehat{Q}^{\pi_\theta}(x_h, u_h) \mid x_h, u_h] = \widehat{Q}^{\pi_\theta}(x_h, u_h),$$

so that we have formed unbiased estimates of the Q-function at each step. We can thus form the stochastic gradient estimate

$$G^{\pi_\theta} = \sum_{t=0}^{\infty} \gamma^t \widehat{Q}^{\pi_\theta}(x_t, u_t) \nabla_\theta \log \pi_\theta(x_t \mid u_t),$$

and by the tower property and linearity of expectation, we have an unbiased estimate of the gradient. This can be used for a stochastic gradient algorithm.

9.3.1 Advantage functions

The function φ can be chosen for variance reduction purposes. One choice is to set

$$\varphi(x) = V^{\pi_\theta}(x) \equiv \mathbb{E}_{U\sim\pi_\theta(\cdot|x)} Q^{\pi_\theta}(x, U). \tag{9.7}$$

In this case, we obtain a representation in terms of the so-called *advantage function*

$$A^{\pi_\theta}(x, u) = Q^{\pi_\theta}(x, u) - V^{\pi_\theta}(x). \tag{9.8}$$

This advantage function plays an important role in the sequel.

9.4 Gradient methods for the soft-max policy classes

Given a feature mapping $\varphi : \mathcal{X} \times \mathcal{U} \to \mathbb{R}^d$, let us consider a general soft-max policy class of the form

$$\pi_\theta(x \mid u) = \frac{e^{\langle \theta, \varphi(x,u)\rangle}}{\sum_{u'} e^{\langle \theta, \varphi(x,u')\rangle}}.$$

Here $\theta \in \mathbb{R}^d$ parameterizes the policy.

The following result shows that the gradient $\nabla_\theta V^{\pi_\theta}(\rho) \in \mathbb{R}^d$ has a relatively simple representation in terms of the feature vector and the advantage function:

Lemma 9.3 *For any distribution ρ, the soft-max policy class has gradients*

$$\nabla_\theta V^{\pi_\theta}(\rho) = \mathbb{E}_{\pi_\theta \rho \pi_\theta}[\varphi(X,U) A^{\pi_\theta}(X,U)]. \tag{9.9}$$

Proof Introduce the shorthand $\Phi_x(\theta) = \log\big(\sum_{u'} e^{\langle \theta, \varphi(x,u')\rangle}\big)$, so that we have

$$\log \pi_\theta(u \mid x) = \langle \theta, \varphi(x,u)\rangle - \Phi_x(\theta).$$

For each fixed x, we see that the distribution $\pi_\theta(\cdot \mid x)$ is an exponential family with feature vector $\varphi(\cdot, x)$, and Φ_x is the associated cumulant generating function. Thus, by standard properties of exponential families, we have $\nabla_\theta \Phi_x(\theta) = \mathbb{E}_{U \sim \pi_\theta(\cdot|x)}[\varphi(x,U)]$, and hence

$$\nabla_\theta \log \pi_\theta(u \mid x) = \varphi(x,u) - \underbrace{\mathbb{E}_{U' \sim \pi_\theta(\cdot|x)}[\varphi(x,U')]}_{:= f(x)}$$

From our earlier advantage-based gradient representation (9.5), we have

$$\begin{aligned}
\nabla_\theta V^{\pi_\theta}(\rho) &= \mathbb{E}_{d_\rho^{\pi_\theta} \circ \pi_\theta}\Big[\nabla_\theta \log \pi_\theta(U \mid X) A^{\pi_\theta}(X,U)\Big] \\
&= \mathbb{E}_{d_\rho^{\pi_\theta} \circ \pi_\theta}\Big[\{\varphi(X,U) - f(X)\} A^{\pi_\theta}(X,U)\Big] \\
&= \mathbb{E}_{d_\rho^{\pi_\theta} \circ \pi_\theta}\Big[\varphi(X,U) A^{\pi_\theta}(X,U)\Big] - \mathbb{E}_{X \sim d_\rho^{\pi_\theta}}\Big[f(X) \mathbb{E}_{U \sim \pi_\theta(\cdot|X)}[A^{\pi_\theta}(X,U)]\Big] \\
&= \mathbb{E}_{d_\rho^{\pi_\theta} \circ \pi_\theta}\Big[\varphi(X,U) A^{\pi_\theta}(X,U)\Big],
\end{aligned}$$

where we have used the fact that $\mathbb{E}_{U \sim \pi_\theta(\cdot|X)}[A^{\pi_\theta}(X,U)] = 0$, by definition of the advantage function. □

It is worthwhile considering the special case in which we form the $d = |\mathcal{X}| \times |\mathcal{U}|$ dimensional feature vector using the binary indicator functions

$$\varphi_{s,a}(x,u) = \begin{cases} 1 & \text{if } (s,a) = (x,u), \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

In this particular case, we have

$$\frac{\partial V^{\pi_\theta}(\rho)}{\partial \theta_{s,a}} = \mathbb{E}_{d_\rho^{\pi_\theta} \circ \pi_\theta}\Big[\varphi_{s,a}(X,U) A^{\pi_\theta}(X,U)\Big] = d_\rho^{\pi_\theta}(s) \pi_\theta(a \mid x) A^{\pi_\theta}(s,a). \tag{9.10}$$

. We refer to this as the *indicator-based soft-max family.*

9.4.1 An interlude: Obtaining stationary points

9.4.2 What can be said about near stationary points?

Thus, we have seen that for a function with a Lipschitz-continuous gradient, it is possible to obtain near-stationary points in polynomial-time. For a general non-convex program, we have no guarantees on the quality of stationary points. The current setting is dramatically different, in that it is possible to directly relate the gradient to the sub-optimality gap in our function. The following result provides such a guarantee:

Proposition 9.4 *For the family of indicator-based soft-max policies, we have*

$$V^*(\xi_{in}) - V^{\pi_\theta}(\xi_{in}) \frac{1}{1-\gamma} \left\| \frac{d^{\pi^*}_{\xi_{in}} \circ \pi^*}{d^{\pi_\theta}_{\xi_{in}} \circ \pi_\theta} \right\|_2 \|\nabla_\theta V^{\pi_\theta}(\xi_{in})\|_2. \tag{9.11}$$

Consequently, suppose that we can find a policy π_θ such that

$$\|\nabla_\theta V^{\pi_\theta}(\xi_{\text{in}})\|_2 \leq \left\{ \left\| \frac{d^{\pi^*}_{\xi_{\text{in}}} \circ \pi^*}{d^{\pi_\theta}_{\xi_{\text{in}}} \circ \pi_\theta} \right\|_2 \right\}^{-1} (1-\gamma)\, \epsilon$$

for some $\epsilon > 0$. This result then guarantees that our policy π_θ is ϵ-suboptimal, in the sense that $V^*(\xi_{\text{in}}) - V^{\pi_\theta}(\xi_{\text{in}}) \leq \epsilon$. The delicacy here is that the likelihood ratio term

$$\left\| \frac{d^{\pi^*}_{\xi_{\text{in}}} \circ \pi^*}{d^{\pi_\theta}_{\xi_{\text{in}}} \circ \pi_\theta} \right\|_2 = \left\{ \sum_{x,u} \left(\frac{d^{\pi^*}_{\xi_{\text{in}}}(x)\pi^*(u \mid x)}{d^{\pi_\theta}_{\xi_{\text{in}}}(x)\pi_\theta(u \mid x)} \right)^2 \right\}^{1/2}$$

might be very large, which forces us to find a policy π_θ that is extremely close to being stationary.

Proof This proposition follows from a combination of our expression (9.10) for the gradients of soft-max policies, and the following useful result.

Lemma 9.5 (Performance difference) *For any policy π_θ and distribution ρ, we have*

$$V^*(\rho) - V^{\pi_\theta}(\rho) = \frac{1}{1-\gamma} \mathbb{E}_{d^{\pi^*}_\rho \circ \pi^*}[A^{\pi_\theta}(X, U)]. \tag{9.12}$$

Starting with Lemma 9.5, we can write

$$\begin{aligned}
(1-\gamma)[V^*(\xi_{\text{in}}) - V^{\pi_\theta}(\xi_{\text{in}})] &= \sum_{x,u} d^{\pi^*}_\rho(x)\pi^*(u \mid x) A^{\pi_\theta}(x,u) \\
&\overset{(i)}{=} \sum_{x,u} \frac{d^{\pi^*}_\rho(x)\pi^*(u \mid x)}{d^{\pi_\theta}_\rho(x)\pi_\theta(u \mid x)} \underbrace{d^{\pi_\theta}_\rho(x)\pi_\theta(u \mid x) A^{\pi_\theta}(x,u)}_{= \frac{\partial V^{\pi_\theta}(\rho)}{\partial \theta_{x,u}}} \\
&\overset{(ii)}{\leq} \left\| \frac{d^{\pi^*}_{\xi_{\text{in}}} \circ \pi^*}{d^{\pi_\theta}_{\xi_{\text{in}}} \circ \pi_\theta} \right\|_2 \|\nabla V^{\pi_\theta}(\xi_{\text{in}})\|_2,
\end{aligned}$$

where step (i) uses the expression (9.10) for the gradient; and step (ii) follows from the Cauchy–Schwarz inequality. □

Proof of performance difference

For each $h = 0, 1, 2, \ldots$, let $\mathbb{E}_t$ denote expectation over the marginal distribution of a state-action (x, u) at time h of the trajectory. We claim that for each $t = 0, 1, 2, \ldots$, we have

$$V^*(\rho) - V^{\pi_\theta}(\rho) = \sum_{h=0}^{t} \gamma^h \mathbb{E}_h[A^{\pi_\theta}(x,u)] + \gamma^t \mathbb{E}_t[Q^*(x,u) - Q^{\pi_\theta}(x,u)]$$

Let us verify the base case $t = 0$. We have

$$\begin{aligned} V^*(\rho) - V^{\pi_\theta}(\rho) &= \mathbb{E}_{x\sim\rho}[V^*(x) - V^{\pi_\theta}(x)] \\ &= \mathbb{E}_{x\sim\rho}\mathbb{E}_{u\sim\pi^*(\cdot|x)}\mathbb{E}[Q^*(x,u) - Q^{\pi_\theta}(x,u) + Q^{\pi_\theta}(x,u) - V^{\pi_\theta}(x)] \\ &= \mathbb{E}_{x\sim\rho}\mathbb{E}_{u\sim\pi^*(\cdot|x)}[Q^*(x,u) - Q^{\pi_\theta}(x,u)] + \mathbb{E}_{x\sim\rho}\mathbb{E}_{u\sim\pi^*(\cdot|x)}[A^{\pi_\theta}(x,u)] \\ &= \mathbb{E}_0[Q^*(x,u) - Q^{\pi_\theta}(x,u)] + \mathbb{E}_0[A^{\pi_\theta}(x,u)], \end{aligned}$$

as claimed.

Turning to the induction, let us suppose that the claim holds for some t; we prove that it holds at time $t+1$. From the form of the induction hypothesis, it suffices to show that

$$\mathbb{E}_{t+1}[Q^*(x,u) - Q^{\pi_\theta}(x,u)] = \gamma\mathbb{E}_t[Q^*(x,u) - Q^{\pi_\theta}(x,u)] + \gamma\mathbb{E}_{t+1}[A^{\pi_\theta}(x,u)].$$

Observe by the Bellman equations satisfied by Q^* and Q^{π_θ} respectively, we have

$$\begin{aligned} Q^*(x,u) &= r(x,u) + \gamma\mathbb{E}_{x'\sim\mathbb{P}_u(\cdot|x)}\mathbb{E}_{u'\sim\pi^*(\cdot|x')}Q^*(x',u'), \quad \text{and} \\ Q^{\pi_\theta}(x,u) &= r(x,u) + \gamma\mathbb{E}_{x'\sim\mathbb{P}_u(\cdot|x)}[V^{\pi_\theta}(x')]. \end{aligned}$$

Consequently, in terms of the shorthand $T_1(x,u) := Q^*(x,u) - Q^{\pi_\theta}(x,u)$, we have

$$\begin{aligned} T_1(x,u) &= \gamma\mathbb{E}_{x'\sim\mathbb{P}_u(\cdot|x)}\mathbb{E}_{u'\sim\pi^*(\cdot|x')}\Big[Q^*(x',u') - Q^{\pi_\theta}(x',u') + Q^{\pi_\theta}(x',u') - V^{\pi_\theta}(x')\Big] \\ &= \gamma\mathbb{E}_{x'\sim\mathbb{P}_u(\cdot|x)}\mathbb{E}_{u'\sim\pi^*(\cdot|x')}\Big[Q^*(x',u') - Q^{\pi_\theta}(x',u')\Big] + \gamma\mathbb{E}_{x'\sim\mathbb{P}_u(\cdot|x)}\mathbb{E}_{u'\sim\pi^*(\cdot|x')}\Big[A^{\pi_\theta}(x',u')\Big]. \end{aligned}$$

Thus, we have

$$\mathbb{E}_t[T_1(x,u)] = \gamma\mathbb{E}_{t+1}[Q^*(x,u) - Q^{\pi_\theta}(x,u)] + \gamma\mathbb{E}_{t+1}[A^{\pi_\theta}(x,u)],$$

as claimed.

9.4.3 Moving ahead

However, the advantage representation suggests useful ways in which to proceed. From the soft-max gradient representation, for any choice of distribution ρ, we can write

$$A^{\pi_\theta}(x,u) = \frac{1-\gamma}{d_\rho^{\pi_\theta}(x)\pi_\theta(x \mid u)}\frac{\partial V^{\pi_\theta}(\rho)}{\partial\theta_{x,u}},$$

and hence, again for any distribution ρ, we have

$$\begin{aligned} V^*(\xi_{\text{in}}) - V^{\pi_\theta}(\xi_{\text{in}}) &= \sum_{x,u}\frac{d_{\xi_{\text{in}}}^{\pi^*}(x)\pi^*(u|x)}{d_\rho^{\pi_\theta}(x)\pi_\theta(u|x)}\frac{\partial V^{\pi_\theta}(\rho)}{\partial\theta_{x,u}} \\ &\leq \left\|\frac{d_{\xi_{\text{in}}}^{\pi^*}}{d_\rho^{\pi_\theta}}\right\|_2\frac{1}{\min_{x,u}\pi(x \mid u)}\|\nabla_\theta V^{\pi_\theta}(\rho)\|_2 \\ &\leq \frac{1}{\sqrt{1-\gamma}}\left\|\frac{d_{\xi_{\text{in}}}^{\pi^*}}{\rho}\right\|_2\frac{1}{\min_{x,u}\pi(x \mid u)}\|\nabla_\theta V^{\pi_\theta}(\rho)\|_2, \end{aligned}$$

where the final line uses the fact that

$$d_\rho^{\pi_\theta}(x) \geq (1-\gamma)\rho(x) \qquad \text{for any state } x.$$

.

Two insights:

First, we see that if we have the freedom of generating trajectories that are initialized with an arbitrary starting distribution ρ, then it might in fact be beneficial to find near-stationary points of the objective $\theta \mapsto V^{\pi_\theta}(\rho)$ instead! Indeed, if

$$\left\| \frac{d^{\pi^*}_{\xi_{\text{in}}}}{\rho} \right\|_2 \ll \left\| \frac{d^{\pi^*}_{\xi_{\text{in}}}}{\xi_{\text{in}}} \right\|_2,$$

then doing so would be beneficial.

Second, we see that our guarantee will be poor unless we can ensure that $\min_{x,u} \pi(x \mid u)$ is not too small.

9.5 Bibliographic details and background

The estimator (9.2a) from Proposition 9.1 is a particular instance of an importance sampling estimator. Its specialization (9.2b) yields an instance of the likelihood ratio gradient estimate; see Glynn (1987, 1990) for an overview of such methods. In the specific context of reinforcement learning, Williams (1992) first proposed the stochastic gradient estimate (9.2b), and there is a substantial literature on its behavior (e.g.,(Sutton et al., 1999)).

9.6 Exercises

Exercise 9.1 (Inconsistency of least squares)