Notes on the Derivation of Least Squares Policy Iteration

- "LSPI is a model-free, off-policy method which can use efficiently (and reuse in each iteration) sample experiences collected in any manner."

The state-action value function $Q^\pi(s,a)$ of any policy π, including a randomized policy, can be found by solving the Bellman equations:

$$Q^\pi(s,a) = \mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s,a,s') \sum_{a' \in \mathcal{A}} \pi(a';s') Q^\pi(s',a').$$

$\pi(a;s)$ is the probability that policy π chooses action a in state s. We can write this in matrix form:

$$Q^\pi = \mathcal{R} + \gamma \mathbf{P} \mathbf{\Pi}_\pi Q^\pi.$$

- Q^π and $\mathcal{R}$ are vectors of size $|\mathcal{S}|\,|\mathcal{A}|$.
- $\mathbf{P}$ is a stochastic matrix of size $|\mathcal{S}|\,|\mathcal{A}| \times \mathcal{S}$ where

$$\mathbf{P}\big((s,a),s'\big) = \mathcal{P}(s,a,s').$$

- $\mathbf{\Pi}_\pi$ is a stochastic matrix of size $\mathcal{S} \times |\mathcal{S}|\,|\mathcal{A}|$ that describes π:

$$\mathbf{\Pi}_\pi\big(s',(s',a')\big) = \pi(a';s')$$

Then we can find Q^π by solving

$$(\mathbf{I} - \gamma \mathbf{P} \mathbf{\Pi}_\pi) Q^\pi = \mathcal{R}.$$

We can also think of this as a fixed point of the Bellman operator T_π:

$$(T_\pi Q)(s,a) = \mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s,a,s') \sum_{a' \in \mathcal{A}} \pi(a';s') Q(s',a').$$

Example

Recall Puterman's favorite 2-state Markov chain on Page 34 of *Markov Decision Processes*. Two states, s_1 and s_2, and two actions, a_1 and a_2. Then:

$$Q^\pi = [Q^\pi(s_1,a_1),\ Q^\pi(s_1,a_2),\ Q^\pi(s_2,a_1),\ Q^\pi(s_2,a_2)]^\intercal$$

$$\mathcal{R} = [5,\ 10,\ -1,\ -\infty]^\intercal\,,$$

and

$$\mathbf{P} = \begin{matrix} & s_1 & s_2 \\ (s_1,a_1) & .5 & .5 \\ (s_1,a_2) & 0 & 1 \\ (s_2,a_1) & 0 & 1 \\ (s_2,a_2) & 0 & 1 \end{matrix}$$

We construct the following policy:

$$\mathbf{\Pi}_\pi = \begin{matrix} & (s_1,a_1) & (s_1,a_2) & (s_2,a_1) & (s_2,a_2) \\ s_1 & .5 & .5 & 0 & 0 \\ s_2 & 0 & 0 & 1 & 0 \end{matrix}.$$

This results in

$$\mathbf{I} - \gamma \mathbf{P}\Pi_\pi = \begin{pmatrix} 1-0.25\gamma & -0.25\gamma & -0.5\gamma & 0. \\ 0. & 1 & -\gamma & 0. \\ 0. & 0. & 1-\gamma & 0. \\ 0. & 0. & -\gamma & 1 \end{pmatrix}$$

and we can solve $(\mathbf{I} - \gamma \mathbf{P}\Pi_\pi)Q^\pi = \mathcal{R}$ for any value of γ to obtain Q^π.

Linear Architecture

We now consider approximating Q^π by a $\hat{Q}^\pi$, a linear combination of basis functions. Suppose we have $\phi_j : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ for $j = 1, 2, \ldots, k$. Define $\phi(s,a)$, column vector of size k, and

$$\phi(s,a) = \begin{pmatrix} \phi_1(s,a) \\ \cdots \\ \phi_1(s,a) \\ \cdots \\ \phi_k(s,a) \end{pmatrix}.$$

Define $\mathbf{\Phi}$ is a $(|\mathcal{S}|\,|\mathcal{A}| \times k)$ matrix of the form

$$\mathbf{\Phi} = \begin{pmatrix} \phi(s_1,a_1)^\intercal \\ \cdots \\ \phi(s,a)^\intercal \\ \cdots \\ \phi(s_{|\mathcal{S}|}, a_{|\mathcal{A}|})^\intercal \end{pmatrix}.$$

If w_j^π is the weight for each function, we can write

$$\hat{Q}^\pi = \mathbf{\Phi} w^\pi.$$

Least-Squares Fixed-Point Approximation

Recall that the Q-values of a policy π are a fixed point of the Bellman operator: $T_\pi Q^\pi = Q^\pi$. We could approximate the value function by finding a fixed point in the space of the linear approximations:

$$T_\pi \hat{Q}^\pi = \hat{Q}^\pi.$$

However, this approximation space is not guaranteed to contain a fixed point.

Recall that if $\mathcal{M}$ is an r-dimensional subspace of $\mathbb{R}^n$, $\mathbf{M}_{n\times r}$ is a basis for $\mathcal{M}$, and

$$\mathbf{P}_{\mathcal{M}} = \mathbf{M}(\mathbf{M}^\intercal \mathbf{M})^{-1}\mathbf{M}^\intercal,$$

then $\mathbf{P}_{\mathcal{M}}$ is the unique orthogonal projector onto $\mathcal{M}$. That is, for any $v = m + n \in \mathbb{R}^n$, where $m \in \mathcal{M}$ and $n \in \mathcal{M}^\perp$, $\mathbf{P}_{\mathcal{M}} v = m$.

Also, for vector $b \in \mathbb{R}^n$,

$$\min_{m\in\mathcal{M}} \|b - m\|_2 = \|b - \mathbf{P}_{\mathcal{M}} b\|_2 .$$

That is, the vector in $\mathcal{M}$ closest to b is the projection of b onto $\mathcal{M}$, $\mathbf{P}_{\mathcal{M}} b$.

We might hope to find a pseudo-fixed point of the Bellman operator on an approximation of the value function. Find the weights for a value function approximation $\hat{Q}^\pi$ so that if we apply the Bellman operator (which may be outside the approximation space), then *project* this into the approximation space, we get the original approximation function. That is, we want weights w^π so that

$$\begin{aligned}\hat{Q}^\pi &= \mathbf{\Phi}(\mathbf{\Phi}^\intercal\mathbf{\Phi})^{-1}\mathbf{\Phi}^\intercal(T_\pi \hat{Q}^\pi) \\ &= \mathbf{\Phi}(\mathbf{\Phi}^\intercal\mathbf{\Phi})^{-1}\mathbf{\Phi}^\intercal(\mathcal{R} + \gamma \mathbf{P}\mathbf{\Pi}_\pi \hat{Q}^\pi). \qquad (1)\end{aligned}$$

In effect, the weights make the Bellman operator perpendicular to the approximation space.

We can manipulate (1) into solving a linear system for the weights:

$$\begin{aligned}
\mathbf{\Phi}(\mathbf{\Phi}^\intercal\mathbf{\Phi})^{-1}\mathbf{\Phi}^\intercal(\mathcal{R}+\gamma\mathbf{P}\mathbf{\Pi}_\pi\hat{Q}^\pi) &= \hat{Q}^\pi \\
\mathbf{\Phi}(\mathbf{\Phi}^\intercal\mathbf{\Phi})^{-1}\mathbf{\Phi}^\intercal(\mathcal{R}+\gamma\mathbf{P}\mathbf{\Pi}_\pi\mathbf{\Phi}w^\pi) &= \mathbf{\Phi}w^\pi \\
\mathbf{\Phi}(\mathbf{\Phi}^\intercal\mathbf{\Phi})^{-1}\mathbf{\Phi}^\intercal(\mathcal{R}+\gamma\mathbf{P}\mathbf{\Pi}_\pi\mathbf{\Phi}w^\pi) - \mathbf{\Phi}w^\pi &= 0 \\
\mathbf{\Phi}\left((\mathbf{\Phi}^\intercal\mathbf{\Phi})^{-1}\mathbf{\Phi}^\intercal(\mathcal{R}+\gamma\mathbf{P}\mathbf{\Pi}_\pi\mathbf{\Phi}w^\pi) - w^\pi\right) &= 0
\end{aligned}$$

Because $\mathbf{\Phi}$ has linearly independent columns:

$$\begin{aligned}
(\mathbf{\Phi}^\intercal\mathbf{\Phi})^{-1}\mathbf{\Phi}^\intercal(\mathcal{R}+\gamma\mathbf{P}\mathbf{\Pi}_\pi\mathbf{\Phi}w^\pi) - w^\pi &= 0 \\
(\mathbf{\Phi}^\intercal\mathbf{\Phi})^{-1}\mathbf{\Phi}^\intercal(\mathcal{R}+\gamma\mathbf{P}\mathbf{\Pi}_\pi\mathbf{\Phi}w^\pi) &= w^\pi \\
\mathbf{\Phi}^\intercal(\mathcal{R}+\gamma\mathbf{P}\mathbf{\Pi}_\pi\mathbf{\Phi}w^\pi) &= (\mathbf{\Phi}^\intercal\mathbf{\Phi})w^\pi \\
\mathbf{\Phi}^\intercal\mathcal{R} + \mathbf{\Phi}^\intercal(\gamma\mathbf{P}\mathbf{\Pi}_\pi\mathbf{\Phi}w^\pi) - (\mathbf{\Phi}^\intercal\mathbf{\Phi})w^\pi &= 0 \\
\mathbf{\Phi}^\intercal(\gamma\mathbf{P}\mathbf{\Pi}_\pi\mathbf{\Phi}w^\pi - \mathbf{\Phi})w^\pi &= -\mathbf{\Phi}^\intercal\mathcal{R} \\
\underbrace{\mathbf{\Phi}^\intercal\left(\mathbf{\Phi}-\gamma\mathbf{P}\mathbf{\Pi}_\pi\mathbf{\Phi}\right)}_{(k\times k)} w^\pi &= \underbrace{\mathbf{\Phi}^\intercal\mathcal{R}}_{(k\times 1)}
\end{aligned}$$

Thus for a policy matrix $\mathbf{\Pi}_\pi$, we can find the least-squares weights minimizing the L_2 distance between $\hat{Q}$ and the projection of $T_\pi\hat{Q}$ onto the $\mathbf{\Phi}$ plane:

$$w^\pi = \left(\mathbf{\Phi}^\intercal\left(\mathbf{\Phi}-\gamma\mathbf{P}\mathbf{\Pi}_\pi\mathbf{\Phi}\right)\right)^{-1}\mathbf{\Phi}^\intercal\mathcal{R},$$

assuming the inverse exists. Koller and Parr (2000) showed this inverse exists for all but finitely many values of γ. The proof follows from the determinant of $\mathbf{\Phi}^\intercal\left(\mathbf{\Phi}-\gamma\mathbf{P}\mathbf{\Pi}_\pi\mathbf{\Phi}\right)$ being a polynomial of γ, and the polynomial only having a finite number of roots.

We can also use the $|\mathcal{S}|\,|\mathcal{A}|\times|\mathcal{S}|\,|\mathcal{A}|$ diagonal matrix Δ_μ weight the projection matrix according to $\mu(s,a)$:

$$w^\pi = \left(\mathbf{\Phi}^\intercal\Delta_\mu\left(\mathbf{\Phi}-\gamma\mathbf{P}\mathbf{\Pi}_\pi\mathbf{\Phi}\right)\right)^{-1}\mathbf{\Phi}^\intercal\Delta_\mu\mathcal{R}.$$

This is analogous to weighted regressions. Letting $\mathbf{A} = \mathbf{\Phi}^\intercal\Delta_\mu\left(\mathbf{\Phi}-\gamma\mathbf{P}\mathbf{\Pi}_\pi\mathbf{\Phi}\right)$ and $b = \mathbf{\Phi}^\intercal\Delta_\mu\mathcal{R}$, we can solve $w^\intercal$ by solving $k\times k$ linear system:

$$\mathbf{A}w^\pi = b.$$

If A and b were known, this linear system would be tractable for a reasonable number of features; because, however, $\mathbf{P}$ and $\mathcal{R}$ are likely either unknown or too large, A and b cannot be directly computed.

LSTDQ

We can leave the matrix notation to get

$$\begin{aligned}
\mathbf{A} &= \mathbf{\Phi}^{\intercal}\Delta_{\mu}\left(\mathbf{\Phi} - \gamma\mathbf{P}\mathbf{\Pi}_{\pi}\mathbf{\Phi}\right) \\
&= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\phi(s,a)\mu(s,a)\left(\phi(s,a) - \gamma\sum_{s'\in\mathcal{S}}\mathcal{P}(s,a,s')\phi(s',\pi(s'))\right)^{\intercal} \\
&= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\mu(s,a)\sum_{s'\in\mathcal{S}}\mathcal{P}(s,a,s')\left[\phi(s,a)\Big(\phi(s,a) - \gamma\phi\big(s',\pi\left(s'\right)\big)\Big)^{\intercal}\right]
\end{aligned}$$

and

$$\begin{aligned}
b &= \mathbf{\Phi}^{\intercal}\Delta_{\mu}\mathcal{R} \\
&= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\phi(s,a)\mu(s,a)\sum_{s'\in\mathcal{S}}\mathcal{P}(s,a,s')R(s,a,s') \\
&= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\mu(s,a)\sum_{s'\in\mathcal{S}}\mathcal{P}(s,a,s')\left[\phi(s,a)R(s,a,s')\right]
\end{aligned}$$

The matrix $\mathbf{A}$ is the sum of many rank one (outer product) matrices of the form

$$\phi(s,a)\Big(\phi(s,a) - \gamma\phi\big(s',\pi\left(s'\right)\big)\Big)^{\intercal}$$

and b the sum of vectors of the form

$$\phi(s,a)R(s,a,s')$$

where the sum over every (s,a,s') pair and weighted by $\mu(s,a)$ and $\mathcal{P}(s,a,s')$. We can approximate $\mathbf{A}$ and b by sampling terms from this summation. "For unbiased sampling, s and a must be drawn jointly from μ, and s' must be drawn from $\mathcal{P}(s,a,s')$." If a finite set of samples

$$D = \{(s_i, a_i, r_i, s_i') \mid i = 1, 2, \ldots, L\}$$

is sampled according to μ_D, $\mathbf{A}$ and b can be approximated according by

$$\tilde{\mathbf{A}} = \frac{1}{L}\sum_{i=1}^{L}\left[\phi(s_i, a_i)\Big(\phi(s_i, a_i) - \gamma\phi\big(s_i', \pi\left(s_i'\right)\big)\Big)^{\intercal}\right]$$
$$\tilde{b} = \frac{1}{L}\sum_{i=1}^{L}\left[\phi(s_i, a_i)r_i\right].$$

This method for approximating $\tilde{w}^{\pi}$ is what Lagoudakis and Parr called LSTDQ. They also use the Sherman-Morrison formula to provide an algorithm that maintains the inverse of A at each step which would be useful for policy improvement.

LSPI

Given some policy and basis functions, we compute the approximate policy

$$\hat{Q}(s, a;\, w) = \sum_{i=1}^{k}\phi_i(s, a)w_i = \phi(s, a)^{\intercal}w$$

by computing the weights according to LSTDQ from a set of samples D. We can then construct a greedy policy π from this by:

$$\pi(s) = \arg\max_{a\in\mathcal{A}}\hat{Q}(s, a).$$

From policy π, we can repeat the same process *reusing the same samples to compute w each time*. We repeat this process until the policy (approximately) stops changing. This is least squares policy iteration.