Lecture Note 13

1 Temporal-Difference Learning

We now consider the problem of computing an appropriate parameter $\tilde{r}$, so that, given an approximation architecture $\tilde{J}(x,r)$, $\tilde{J}(\cdot,\tilde{r}) \approx J^*(\cdot)$.

A class of iterative methods are the so-called *temporal-difference learning* algorithms, which generates a series of approximations $\tilde{J}_k = \tilde{J}(\cdot, r_k)$ as follows. Consider generating a trajectory $(x_1, u_1, \ldots, x_k, u_k)$, where u_k is the greedy policy with respect to $\tilde{J}_k$. We then have the error/temporal differences

$$d_k = g_{u_k}(x_k) + \alpha \tilde{J}_k(x_{k+1}, r_k) - \tilde{J}_k(x_k, r_k),$$

which represent an approximation to the Bellman error $(T\tilde{J}_k)(x_k) - \tilde{J}_k(x_k)$ at state x_k. Based on the temporal differences, an intuitive way of updating the parameters r_k is to make updates proportional to the observed Bellman error/temporal difference:

$$r_{k+1} = r_k + \gamma_k d_k z_k,$$

where γ_k is the step size and z_k is called an *eligibility vector* — it measures how much updates to each component of the vector r_k would affect the Bellman error.

To gather more intuition about how to choose the eligibility vector, we will consider the case of autonomous systems, i.e., systems that do not involve control. In this case, we can estimate the cost-to-go function via sampling as follows. Suppose that we have a trajectory $x_1, \ldots, x_n$. Then we have

$$\begin{aligned} J^*(x_1) &\approx \sum_{t=1}^{n} \alpha^{n-1} g(x_t) \\ J^*(x_2) &\approx \sum_{t=2}^{n} \alpha^{n-2} g(x_t) \\ &\vdots \end{aligned}$$

In other words, from a trajectory $x_1, \ldots, x_n$, we can derive pairs $(x_i, \hat{J}(x_i))$, where $\hat{J}(x_i)$ is a noisy and biased estimate of $J^*(x_i)$. Therefore we may consider fitting the approximation $\tilde{J}(x,r)$ by minimizing the empirical squared error:

$$\min_r \sum_{t=1}^{n} \left(\hat{J}_n(x_t) - \tilde{J}(x_t, r) \right)^2 \tag{1}$$

We derive an incremental, approximate version of (1). First note that $\hat{J}_n(x_t)$ could be updated incrementally as follows:

$$\hat{J}_{n+1}(x_t) = \hat{J}_n(x_t) + \alpha^{n+1-t} g(x_{n+1}) \tag{2}$$

Alternatively, we may use a small-step update of the form

$$\hat{J}_{n+1}(x_t) = \hat{J}_n(x_t) + \gamma \left(\sum_{j=t}^{n+1} \alpha^{j-t} g(x_j) - \hat{J}_n(x_t) \right), \tag{3}$$

which makes $\hat{J}_{n+1}(x_t)$ an average of the "old estimate" $\hat{J}_n(x_t)$ and the "new estimate" (2). Finally, we may approximate (3) to have $\hat{J}_n(x_t)$ function $d_1, d_2, \ldots, d_n$:

$$\begin{aligned}\sum_{j=t}^{n} \alpha^{j-t} g(x_j) - \hat{J}_n(x_t) &= g(x_t) + \alpha \hat{J}_n(x_{t+1}) - \hat{J}_n(x_t) + \alpha(g(x_{t+1}) + \alpha \hat{J}_n(x_{t+2}) - \hat{J}_n(x_{t+1})) + \ldots \\ &\quad + \alpha^{n-t}(g(x_n) + \alpha \hat{J}_n(x_{n+1}) - \hat{J}_n(x_n)) - \alpha^{n+1-t} \hat{J}_n(x_{n+1} \\ &\approx \sum_{j=t}^{n} \alpha^{j-t} d_t.\end{aligned}$$

Hence

$$\hat{J}_{n+1}(x_t) = \hat{J}_n(x_t) + \gamma \sum_{j=t}^{n+1} \alpha^{j-t} d_j. \tag{4}$$

Finally, we may consider having the sum in (1) implemented incrementally, so that the previous temporal differences do not have to be stored:

$$\hat{J}_{n+1}(x_t) = \hat{J}_n(x_t) + \gamma \alpha^{n+1-t} d_{n+1}.$$

Hence, in each time stage, we would like to find r_n minimizing

$$\min_r \sum_{t=1}^{n} \left(\hat{J}_n(x_t) + \gamma \alpha^{n-t} d_n - \tilde{J}(x_t, r) \right)^2. \tag{5}$$

Starting from the solution r_n to the problem at stage n, we can approximate the solution of the problem at stage $n+1$ by updating r_{n+1} along the gradient of (5). This leads to

$$r_{n+1} = r_n + \gamma \left(\sum_{t=1}^{n} \alpha^{t-n} \nabla_r \tilde{J}(r_n, x_t) \right) d_{n+1}.$$

We can also have an incremental version, given by

$$\begin{aligned} r_{k+1} &= r_k + \gamma_k z_k d_k \\ z_k &= \alpha z_{k-1} + \nabla_r \tilde{J}(x_k, r_k) \end{aligned}$$

The algorithm above is known as $TD(1)$. We have the generalization $TD(\lambda)$, $\lambda \in [0, 1]$.

$$\boxed{\begin{aligned} r_{k+1} &= r_k + \gamma_k z_k d_k \\ z_k &= \alpha \lambda z_{k-1} + \nabla_r \tilde{J}(x_k, r_k) \end{aligned}} \qquad \text{TD}(\lambda)$$

Before analyzing the behavior of $TD(\lambda)$, we are going to study a related, deterministic algorithm — approximate value iteration. The analysis of $TD(\lambda)$ will be based on interpreting it as a stochastic approximation version of approximate value iteration.

2 Approximate Value Iteration

Define the operator T_λ

$$\begin{aligned} T_\lambda J &= (1-\lambda) \sum_{m=0}^{\infty} \lambda^m T^{m+1} J, \quad && \text{for } \lambda \in [0, 1) \\ T_\lambda J &= J^*, && \text{for } \lambda = 1. \end{aligned}$$

We can show that T_λ satisfies the same properties as T:

Lemma 1

$$\|T_\lambda J - T_\lambda \bar{J}\|_\infty \leq \frac{\alpha(1-\lambda)}{1-\alpha\lambda}\|J - \bar{J}\|_\infty, \quad \forall J, \bar{J}$$
$$J^* = T_\lambda J^*$$

The motivation for T_λ is as follows. Recall that, in value iteration, we have $J_{k+1} = TJ_k$. However, we could also implement value iteration with $J_{k+1} = T^L J_k$, which implies L steps look ahead. Finally, we can have an update that is a weighted average over all possible values of L; $J_{k+1} = T_\lambda J_k$ gives one such update.

In what follows, we are going to restrict attention to linear approximation architectures. Let

$$\begin{aligned}
\tilde{J}(x,r) &= \sum_{i=1}^{P} \phi_i(x) r_i, \qquad \text{and} \\
\Phi &= \begin{bmatrix} \phi_1(1) & \phi_2(1) & \dots & \phi_P(1) \\ \phi_1(2) & \phi_2(2) & \dots & \phi_P(2) \\ \vdots & \vdots & \dots & \vdots \\ \phi_1(n) & \phi_2(n) & \dots & \phi_P(n) \end{bmatrix} \\
\tilde{J} &= \Phi r
\end{aligned}$$

Moreover, we are going to consider only autonomous systems. We denote by P the transition matrix associated with the system.

Let us introduce some notation. First, we have

$$D = \begin{bmatrix} d(1) & 0 & \dots & 0 \\ 0 & d(2) & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & d(n) \end{bmatrix}$$

where $d : \mathcal{S} \to (0,1)^{\mathcal{S}}$ is a probability distribution over states. Define the weighted Euclidean norms

$$\begin{aligned}
\|J\|_{2,D} &= J^T D J = \sum_{x \in \mathcal{S}} d(x) J^2(x) \\
< J\ \bar{J} >_D &= J^T D \bar{J} = \sum_{x \in \mathcal{S}} d(x) J(x) \bar{J}(x)
\end{aligned}$$

For simplicity, we assume that $\phi_i, i = 1, \dots, p$ is an orthonormal basis to the subspace $J = \Phi r$, i.e.,

$$\|\phi_i\|_{2,D} = 1 \text{ and } < \phi_i, \phi_j >= 0, \forall i \neq j$$

In matrix notation, we have

$$\Phi^T D \Phi = I.$$

We are going to use the following *projection operator* Π:

$$\Pi J = \Phi r_J, \quad \text{where } r_J = \arg\min_r \|\Phi r - J\|_{2,D}$$

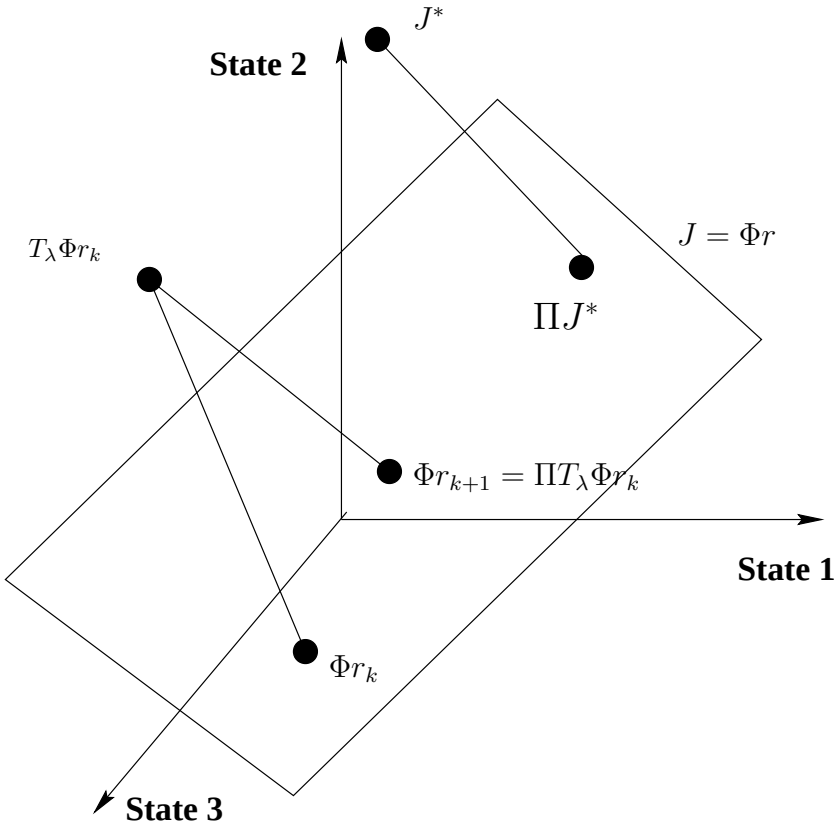

Figure 1: Approximate Value Iteration

We can characterize Π explicitly by solving the associated minimizing problem. We have

$$\begin{aligned} r_J &= \arg\min_r \|\Phi r - J\|_{2,D}^2 \\ &= \arg\min_r (\Phi r - J)^T D(\Phi r - J) \\ &= \left(\Phi^T D\Phi\right)^{-1} \Phi^T DJ \\ &= < \Phi, J >_D \end{aligned}$$

Hence, we have $\Pi J = \Phi < \Phi, J >_D$.

Lemma 2 *For all J,*

$$\Pi J = \Phi < \Phi, J >_D \tag{6}$$

$$< \Pi J, J - \Pi J >_D = 0 \tag{7}$$

$$\|J\|_{2,D}^2 = \|\Pi J\|_{2,D}^2 + \|J - \Pi J\|_{2,D}^2 \tag{8}$$

Note that $\Phi r_{k+1} = \Pi T_\lambda \Phi r_k$. We know that the projection Π is a nonexpansion from

$$\|\Pi J - \Pi \bar{J}\|_{2,D} = \|\Pi(J - \bar{J})\|_{2,D} \leq \|J - \bar{J}\|_{2,D}.$$

Moreover, T_λ is a contraction:

$$\|T_\lambda J - T_\lambda \bar{J}\|_\infty \leq K\|J - \bar{J}\|_\infty.$$

However, the fact that Π and T_λ are a non-expansion and a contraction with respect to different norms implies that convergence of approximate value iteration cannot be guaranteed by a contraction argument, as was the case with exact value iteration. Indeed, as illustrated in Figure 2, ΠT_λ is not necessarily a contraction with respect to any norm, and one can find counterexamples where $TD(\lambda)$ fails to converge.

As it turns out, there is a special choice of D that ensures convergence of $TD(\lambda)$ for all $\lambda \in [0, 1]$. Before proving that, we need the following auxiliary result. First, we present two definitions involving Markov chains.

Definition 1 *A Markov chain is called* **irreducible** *if, for every pair of states x and y, there is k such that $P^k(x, y) > 0$.*

Definition 2 *A state x is called* **periodic** *if there is m such that $P^k(x, x) > 0$ iff $k = mn$, for some $n \in \{0, 1, 2, \dots\}$. A Markov chain is called* **aperiodic** *if none of its states is periodic.*

Lemma 3 *Given a transition matrix P and assume that P is irreducible and aperiodic. Then there exists a unique π such that*

$$\pi^T P = \pi^T$$

and

$$P^n \to \begin{bmatrix} \pi^T \\ \pi^T \\ \vdots \\ \pi^T \end{bmatrix}.$$

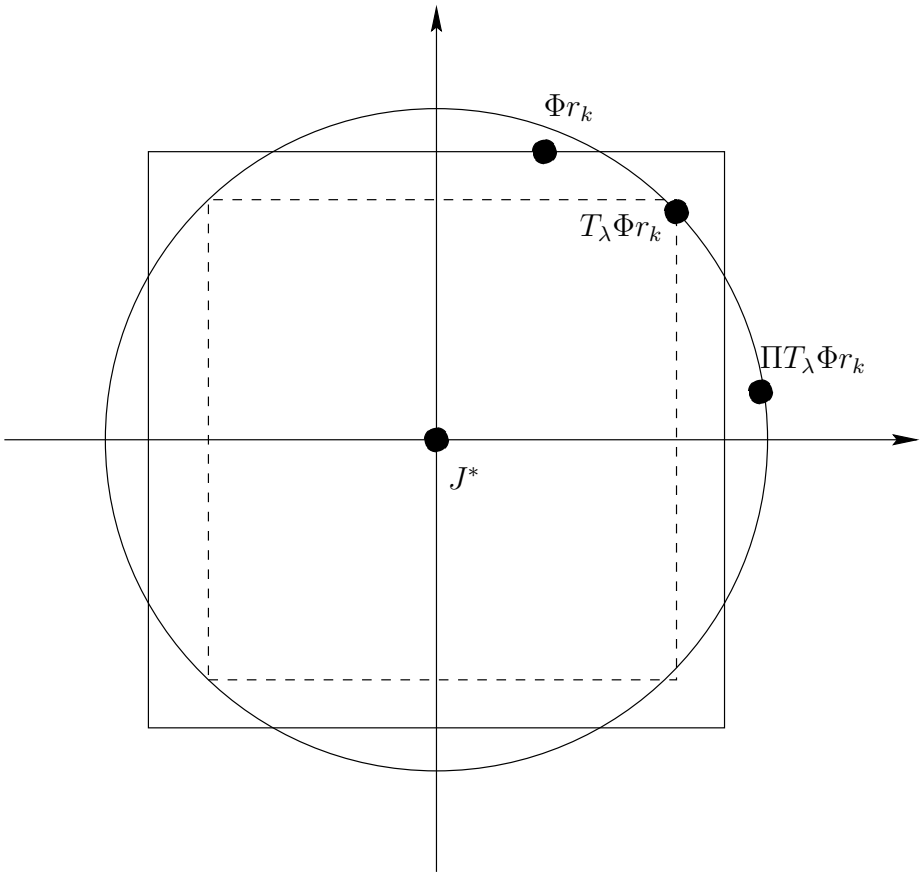

Figure 2: $T_\lambda \Phi r_k$ must be inside the smaller square and $\Pi T_\lambda \Phi r_k$ must be inside the circle, but $\Pi T_\lambda \Phi r_k$ may be outside the larger square and further away from J^* than Φr_k.

This lemma was proved in Problem Set 2, for the special case where $P(x,x) > 0$ for some x.

We are now poised to prove the following central result used to derive a convergent version of $TD(\lambda)$:

Lemma 4 *Suppose that the transition matrix P is irreducible and aperiodic. Let*

$$D = \begin{bmatrix} \pi_1 & 0 & \dots & 0 \\ 0 & \pi_2 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & \pi_{|\mathcal{S}|} \end{bmatrix},$$

where π is the stationary distribution associated with P. Then

$$\|PJ\|_{2,D} \leq \|J\|_{2,D}.$$

Proof:

$$\begin{aligned} \|PJ\|_{2,D}^2 &= \sum_{x\in\mathcal{S}} \pi(x) \left(\sum_y P(x,y)J(y) \right)^2 \\ &\leq \sum_{x\in\mathcal{S}} \pi(x) \sum_y P(x,y)J^2(y) \\ &= \sum_y \sum_x \pi(x)P(x,y)J^2(y) \\ &= \sum_y \pi(y)J^2(y) \\ &= \|J\|_{2,D}^2 \end{aligned}$$

The first inequality follows the Jensen's inequality and the third equality holds because π is a stationary distribution. □

Based on the previous lemma, we can show that T_λ is a contraction with respect to $\|\cdot\|_{2,D_\pi}$, where

$$D_\pi = \begin{bmatrix} \pi_1 & 0 & \dots & 0 \\ 0 & \pi_2 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & \pi_{|\mathcal{S}|} \end{bmatrix}$$

and π is the stationary distribution of the transition matrix P. It follows that, if the projection Π is performed with respect to $\|\cdot\|_{2,D_\pi}$, ΠT_λ becomes a contraction with respect to the same norm, and convergence of $TD(\lambda)$ is guaranteed.

Lemma 5

$$(i) \qquad \|TJ - T\bar{J}\|_{2,D_\pi} \leq \alpha \|J - \bar{J}\|_{2,D_\pi}$$

$$(ii) \qquad \|T_\lambda J - T_\lambda \bar{J}\|_{2,D_\pi} \leq \frac{\alpha(1-\alpha)}{1-\alpha\lambda} \|J - \bar{J}\|_{2,D_\pi}$$

$$(iii) \qquad \|\Pi T_\lambda J - \Pi T_\lambda \bar{J}\|_{2,D_\pi} \leq \frac{\alpha(1-\alpha)}{1-\alpha\lambda} \|J - \bar{J}\|_{2,D_\pi}$$

Proof of (1)

$$\begin{aligned}\|TJ - T\bar{J}\|_{2,D_\pi} &= \|g + \alpha PJ - (g + \alpha \quad \bar{J})\|_{2,D_\pi} \\ &= \alpha\|PJ - P\bar{J}\|_{2,D_\pi} \\ &\leq \alpha\|J - \bar{J}\|_{2,D_\pi}\end{aligned}$$

□

Theorem 1 *Let*

$$\Phi r_{k+1} = \Pi T_\lambda \Phi r_k$$

and

$$D_\pi = \begin{bmatrix} \pi_1 & 0 & \dots & 0 \\ 0 & \pi_2 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & \pi_{|\mathcal{S}|} \end{bmatrix}.$$

Then $r_k \to r^*$ *with*

$$\|\Phi r^* J^*\|_{2,D_\pi} \leq K_{\alpha,\lambda}\|\Pi J^* - J^*\|_{2,D_\pi}.$$

Proof: Convergence follows from (*iii*). We have $\Phi r^* = \Pi T_\lambda \Phi r^*$ and $J^* - T_\lambda J^*$. Then

$$\begin{aligned}\|\Phi r^* - J^*\|^2_{2,D_\pi} &= \|\Phi r^* - \Pi J^* + \Pi J^* - J^*\|^2_{2,D_\pi} \\ &= \|\Phi r^* - \Pi J^*\|^2_{2,D_\pi} + \|\Pi J^* - J^*\|^2_{2,D_\pi} \quad \text{(orthogonal)} \\ &= \|\Pi T_\lambda \Phi r^* - \Pi T_\lambda J^*\|^2_{2,D_\pi} + \|\Pi J^* - J^*\|^2_{2,D_\pi} \\ &\leq \underbrace{\frac{\alpha^2(1-\lambda)^2}{(1-\alpha\lambda)^2}}_{\gamma} \|\Phi r^* - J^*\|^2_{2,D_\pi} + \|\Pi J^* - J^*\|^2_{2,D_\pi}\end{aligned}$$

Therefore

$$\|\Phi r^* J^*\|_{2,D_\pi} \leq \frac{1}{1-\gamma}\|\Pi J^* - J^*\|_{2,D_\pi}$$

□