Lab Meeting Notes: 19th April 2024

Dhawal Gupta
dgupta@cs.umass.edu

Overview of the Material

In this document, we will cover the material in the following order:

- TD learning and its algorithm.
- Convergence TD in the on-policy setting.
- Condition of divergence in the case of off-policy learning and the reason behind it (with example).
- Objective for TD, i.e., mean squared TD error, followed by residual gradient methods, how MSTDE is a bad objective for learning value functions.
- Mean squared Bellman error, and how it is not representable.
- Projection operator for a linear span of the feature span, and how to project vectors back into the desirable space.
- Mean squared projected Bellman error (MSPBE), and how the minimizer of that coincides with TD learning fixed point.
- Gradient update for the MSPBE and how GTD2 and TDC methods come about.

Notation

Markov decision process (MDP), i.e., $(\mathcal{S}, \mathcal{A}, p, r, \gamma, d_0)$, S_t, A_t, R_t random variables, $t \in \{1, 2, 3, \dots\}$,where $|\mathcal{S}| = n, \pi : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$, $\boldsymbol{\phi} : \mathcal{S} \rightarrow \mathbb{R}^d$, $\boldsymbol{\phi}(S_t) \equiv \boldsymbol{\phi}_t, p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$, $p(s, a, s') = \Pr(S_{t+1}{=}s' | S_t = s, A_t{=}a)$, $r : \mathcal{S} \times \mathcal{A} \rightarrow [-R_{\max}, R_{\max}], R_{\max} \in \mathbb{R}^+$, $r(s, a) = \mathbf{E}[R_t | S_t{=}s, A_t{=}a]$ $p^\pi(s, s') = \sum_a \pi(s, a) p(s, a, s')$, $d^\pi : \mathcal{S} \rightarrow [0, 1]$, where $d^\pi(s) = \sum_{s'} d^\pi(s') p^\pi(s', s)$.

Bellman Operator:

Matrix Rules

$$\frac{\partial \mathbf{x}^\intercal \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^\intercal)\mathbf{x}, \frac{\partial \mathbf{x}^\intercal \mathbf{A}}{\partial \mathbf{x}} = \mathbf{A}, \frac{\partial \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}^\intercal.$$

Vector Notation

Let n denote the total number of states, i.e., $n = |\mathcal{S}|$, and let $\Phi \in \mathbb{R}^{n \times d}$, Let $\mathbf{P}^\pi \in \mathbb{R}^{n \times n}, \mathbf{P}^\pi_{i,j} = p^\pi(s_i, s_j) \doteq \Pr(S_{t+1} = s_j | S_t = s_i; \pi) = \sum_a \pi(s_i, a) p(s_i, a, s_j)$. Let $\mathbf{d}^\pi \in [0, 1]^n$, where $\mathbf{d}^\pi_i \doteq d^\pi(s_i)$. Similarly, let $\mathbf{D}^\pi = \texttt{diag}(\mathbf{d}^\pi) \in \mathbb{R}^{n \times n}$, be the diagonal matrix of $\mathbf{d}^\pi$. Note that as $\mathbf{d}^\pi$ is the stationary distribution, we have that, or:

$$\mathbf{d}^{\pi\intercal} = \mathbf{d}^{\pi\intercal} \mathbf{P}^\pi.$$

Preprint. Under review.

Let, $\mathbf{r}^\pi \in \mathbb{R}^n$ such that, $\mathbf{r}^\pi_i = \mathbf{E}[R_t | S_t = s_i; \pi] = \sum_a \pi(s_i, a) r(s_i, a)$. Similarly, we can define the value function vector, i.e., $\mathbf{v}^\pi \in \mathbb{R}^n$, where $\mathbf{v}^\pi_i = v^\pi(s_i)$, and hence the approximated version as $\mathbf{v}_{\boldsymbol{\theta}} \doteq \Phi\boldsymbol{\theta}, \mathbf{v}_{\boldsymbol{\theta}} \in \mathbb{R}^n$.The Bellman equation in case of the vector notation becomes:

$$\begin{aligned}\mathbf{v}^\pi &= \mathbf{r}^\pi + \gamma \mathbf{P}^\pi \mathbf{v}^\pi \\ \mathbf{v}^\pi &= (\mathbf{I} - \gamma \mathbf{P}^\pi)^{-1} \mathbf{r}^\pi.\end{aligned}$$

Temporal Difference Update

$$\begin{aligned}\boldsymbol{\theta}_{t+1} &\doteq \boldsymbol{\theta}_t + \alpha(R_t + \gamma v_{\theta_t}(S_{t+1}) - v_{\theta_t}(S_t)) \frac{\partial v_{\boldsymbol{\theta}}(S_t)}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_t}, \\ &= \boldsymbol{\theta}_t + \alpha(R_t + \gamma \boldsymbol{\phi}_{t+1}^\intercal \boldsymbol{\theta}_t - \boldsymbol{\phi}_t^\intercal \boldsymbol{\theta}_t)\boldsymbol{\phi}_t, \\ &= \boldsymbol{\theta}_t + \alpha(R_t \boldsymbol{\phi}_t - \boldsymbol{\phi}_t(\boldsymbol{\phi}_t - \gamma \boldsymbol{\phi}_{t+1})^\intercal \boldsymbol{\theta}_t), \\ \mathbf{E}[\boldsymbol{\theta}_{t+1} | \boldsymbol{\theta}_t] &= \boldsymbol{\theta}_t + \alpha(\mathbf{b} - \mathbf{A}\boldsymbol{\theta}_t).i\end{aligned}$$

Where, $\mathbf{b} \doteq \mathbf{E}[R_t \boldsymbol{\phi}_t]$, $\mathbf{A} \doteq \mathbf{E}\big[\boldsymbol{\phi}_t(\boldsymbol{\phi}_t - \gamma \boldsymbol{\phi}_{t+1})^\intercal\big]$, and hence this is minimized when $\mathbf{b} = \mathbf{A}\boldsymbol{\theta}_{\texttt{TD}}$. When TD converges, it converges to the fixed point of :

$$\mathbf{A}^{-1}\mathbf{b} = \boldsymbol{\theta}_{\texttt{TD}}.$$

Note: I was confused as to why this might be the case and if it will exist or not, but it will always exist because if you look at it $\mathbf{b} = \mathbf{E}[R_t \boldsymbol{\phi}_t]$, which is a linear combination of the features, and also when we look at $\mathbf{A}$ its span is also in the span of a combination of $\boldsymbol{\phi}$ (later we will see $\mathbf{A} = \Phi^\intercal \mathbf{D}^\pi (\mathbf{I} - \gamma \mathbf{P}^\pi)\Phi$), so hence $\mathbf{b}$ and $\mathbf{A}\boldsymbol{\theta}$ lie in the span of $\boldsymbol{\phi}$ and hence can be solved.

Proof of Convergence

We have all seen in the RL class the convergence proof for the tabular case, where we show that the Bellman operator is a contraction mapping under the $\inf$ norm and hence converges to a unique value function on repeated application, now we will look at the convergence of TD learning under linear function approximation and conditions where it might end up diverging.

Reordering the terms in the update, we can see that :

$$\mathbf{E}[\boldsymbol{\theta}_{t+1} | \boldsymbol{\theta}_t] = (\mathbf{I} - \alpha \mathbf{A})\boldsymbol{\theta}_t + \alpha \mathbf{b}.$$

We see that $(\mathbf{I} - \alpha \mathbf{A})$ multiplies the weight vector and hence this should contract to shrink the effect of $\boldsymbol{\theta}_t$, consider a diagonal matrix of $\mathbf{A}$ if all diagonal entries are negative, then at least one term in the diagonal > 1, which will amplify the weights, and if $\mathbf{A}$ has all positive elements, then we can set α such that all the values $\in [0, 1]$. Hence, if $\mathbf{A}$ is *positive definite* (PD) matrix, then we can ensure that the system converges.

Vector notation for $\mathbf{A}$:

$$\begin{aligned}\mathbf{A} &= \sum_s d^\pi(s) \sum_{s'} p^\pi(s, s') \boldsymbol{\phi}(s)(\boldsymbol{\phi}(s) - \gamma \boldsymbol{\phi}(s'))^\intercal, \\ &= \sum_s d^\pi(s) \boldsymbol{\phi}(s)(\boldsymbol{\phi}(s) - \gamma \sum_{s'} p^\pi(s, s') \boldsymbol{\phi}(s'))^\intercal \\ &= \Phi^\intercal \mathbf{D}^\pi (\mathbf{I} - \gamma \mathbf{P}^\pi)\Phi.\end{aligned}$$

A matrix $\mathbf{C}$ is positive definite if $\forall \mathbf{x}, \mathbf{x}^\intercal \mathbf{C} \mathbf{x} > 0$, and hence $\mathbf{A}$ is PD if $\forall \mathbf{x}, \mathbf{x}^\intercal \mathbf{A} \mathbf{x} \geq 0$.

$$\begin{aligned}\mathbf{x}^\intercal \mathbf{A} \mathbf{x} &= \mathbf{x}^\intercal \Phi^\intercal \mathbf{D}^\pi (\mathbf{I} - \gamma \mathbf{P}^\pi)\Phi \mathbf{x} \\ &= \mathbf{z}^\intercal \mathbf{D}^\pi (\mathbf{I} - \gamma \mathbf{P}^\pi)\mathbf{z}.\end{aligned}$$

If Φ is full-rank, then $\mathbf{z}$ are in the column span of Φ, and hence, if it is full rank, $\mathbf{z} \neq 0$. Hence it is important to determine the PD'ness of $\mathbf{D}^\pi(\mathbf{I} - \gamma \mathbf{P}^\pi)$.

We will make use of two results:
Lemma: (Sutton (1988), pg. 27): If a real symmetric matrix $\mathbf{C}$ is PD if all its diagonal elements are

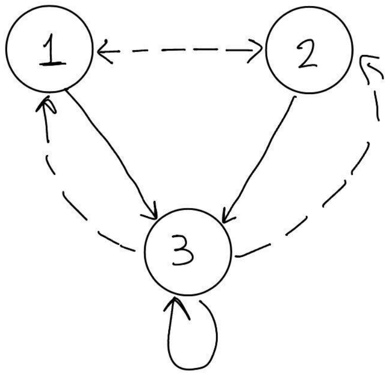

Figure 1: A small MDP with 3 states.

positive and is diagonally dominant, i.e., $\forall i \in [n], \mathbf{C}_{ii} \geq \sum_{j \neq i} |\mathbf{C}_{i,j}|$.
But $\mathbf{D}^{\pi}(\mathbf{I} - \gamma \mathbf{P}^{\pi})$ need not be symmetric, hence we make use of another property,
Theorem: (Varga (1962), pg. 23): $\mathbf{M}$ is PD iff $\mathbf{S} = \mathbf{M} + \mathbf{M}^{\intercal}$ is PD.

Hence for this we simply need to show that for $\mathbf{A}$ to be PD, we need to show that $\mathbf{D}^{\pi}(\mathbf{I} - \gamma \mathbf{P}^{\pi}) + (\mathbf{D}^{\pi}(\mathbf{I} - \gamma \mathbf{P}^{\pi}))^{\intercal}$ is diagonall dominant and has positive diagonal elements. Which means we have to talk about the diagonal entries in $\mathbf{D}^{\pi}(\mathbf{I} - \gamma \mathbf{P}^{\pi})$, being larger than sum of all values in a given row and a given column.

row: This is true by default as the sum of rows in $\forall i \mathbf{P}^{\pi}_{:,i} = 1$ as it is a stocashci matrix, and the hence $\gamma < 1$ and the diaognal is $1 - \gamma \geq 0$.

cols: Consider $\mathbf{1}_{\intercal} \mathbf{M}$ will give a row vector where each element in the row vector is the sum of the col's of $\mathbf{M}$, hence:

$$\begin{aligned}
\mathbf{1}^{\intercal}\mathbf{D}^{\pi}(\mathbf{I} - \gamma \mathbf{P}^{\pi}) &= \mathbf{d}^{\pi\intercal}(\mathbf{I} - \gamma \mathbf{P}^{\pi}) \\
&= \mathbf{d}^{\pi\intercal} - \gamma \mathbf{d}^{\pi\intercal}\mathbf{P}^{\pi} \\
&= \mathbf{d}^{\pi} - \gamma \mathbf{d}^{\pi} \\
&= (1 - \gamma)\mathbf{d}^{\pi} \geq 0.
\end{aligned}$$

Hence, $\mathbf{A}$ is PD, as all elements are positive, and we can say that the system convergencs.

So, for convergence under function approximation and bootstrapping we needed, on-policy distribution and this is where actually under the disrtibution mistmatch is when the system can diverge, i.e., lets say for a behavior policy β, when :

$$\mathbf{d}^{\beta\intercal}(\mathbf{I} - \gamma \mathbf{P}^{\pi}) = \mathbf{d}^{\beta\intercal} - \gamma \mathbf{d}^{\beta\intercal}\mathbf{P}^{\pi}.$$

as $\mathbf{d}^{\beta}$ is not the stationary distribution, and hence $\mathbf{d}^{\beta}\mathbf{P}^{\pi} \neq \mathbf{d}^{\beta}$,we can have a system where we might get a negative value which can potentially lead the system to diverge.

Consider the following target policy $\pi(\cdot, \texttt{solid}) = 1$, and the following behaviour policy $\beta(\cdot, \texttt{dashed}) = 2/3$ and $\beta(\cdot, \texttt{solid}) = 1/3$ and $\gamma = 0$ and all rewards $= 0$. We can see that the system can diverge in this case, as the distribution mismatch can lead to a negative value in the update. The stationary distribution for the MDP in Figure 1 is given by

$$\mathbf{d}^{\pi} = [0, 0, 1]^{\intercal}, \mathbf{d}^{\beta} = [1/3, 1/3, 1/3]^{\intercal}, \mathbf{P}^{\pi} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \mathbf{P}^{\beta} = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}.$$

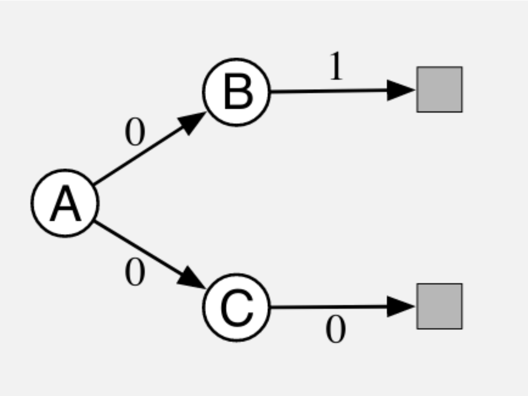

Figure 2: A small MDP with 3 states(Sutton and Barto, 2018).

We see that

$$\begin{aligned}
\mathbf{d}^{\beta\intercal}(\mathbf{I}-\gamma\mathbf{P}^{\pi}) &= \begin{bmatrix}1/3\\1/3\\1/3\end{bmatrix} - \gamma\,[1/3 \quad 1/3 \quad 1/3]\begin{bmatrix}0&0&1\\0&0&1\\0&0&1\end{bmatrix}\\
&= \begin{bmatrix}1/3\\1/3\\1/3\end{bmatrix} - \gamma\begin{bmatrix}0\\0\\1\end{bmatrix}\\
&= \begin{bmatrix}1/3\\1/3\\(1/3-\gamma)\end{bmatrix}.
\end{aligned}$$

Here the last term can be negative if $\gamma > 1/3$, and hence the system can diverge.

TD Error

What objective does the TD error minimize? The Mean Squared TD error is defined as:

$$\begin{aligned}
\text{MSTDE}(\boldsymbol{\theta}) &= \mathbf{E}\big[(R_t+\gamma\boldsymbol{\phi}_{t+1}^{\intercal}\boldsymbol{\theta}-\boldsymbol{\phi}_t^{\intercal}\boldsymbol{\theta})^2\big]\\
&= \mathbf{E}\big[\delta_t^2(\boldsymbol{\theta})\big]\\
&= \sum_s d^{\pi}(s)\mathbf{E}\big[\delta_t^2(\boldsymbol{\theta})|S_t=s\big]\\
\nabla\text{MSTDE} &= \sum_s d^{\pi}(s)\nabla\mathbf{E}\big[\delta_t^2(\boldsymbol{\theta})|S_t=s\big]\\
&= \sum_s d^{\pi}(s)\nabla\mathbf{E}\big[2\delta_t(\boldsymbol{\theta})(-\boldsymbol{\phi}_t+\gamma\boldsymbol{\phi}_{t+1})|S_t=s\big]
\end{aligned}$$

When implementing TD learning, the update we implement is the following:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha\mathbf{E}[\delta_t\boldsymbol{\phi}_t]\,,$$

i.e., we ignore the residual term $\gamma\boldsymbol{\phi}_{t+1}$, which makes it a semi gradient method, and we plug this back , we get what is called as the residual gradient methods, which minimizes MSTDE. There exists example which show that minimzing the MSTDE might not be the best thing to do, as often solitions based on MSVE have higher MSTDE. For example, consider the following MDP, where we have 3 states, and the transitions as showin in Figure 2.

In this case going from B or C is probability $1/2$ and hence $v^{\pi}(A) = 1/2, v^{\pi}(B) = 1, v^{\pi}(C) = 0$. The value function which minimizes MSTDE is $v_{\texttt{MSTDE}} = [1/2, 3/4, 1/4]^{\intercal}$, which is not a desirable property we want from our method.

$$\begin{aligned}
\text{MSTDE}(\boldsymbol{\theta}_{\text{MSTDE}}) &= (1/16) + (1/16) = 1/8\\
\text{MSTDE}(\boldsymbol{\theta}_{\text{MSVE}}) &= (1/4) + (1/4) = 1/2.
\end{aligned}$$

Bellman Error

The Bellman operator is defined as:

$$\mathcal{T}^{\pi} v(s) = \mathbf{r}^{\pi}(s) + \gamma \sum_{s'} p^{\pi}(s, s') v(s'),$$

for a given policy π and value at state s for some approximation v. The vector notation to write the same is :

$$\mathcal{T}^{\pi} \mathbf{v} = \mathbf{r}^{\pi} + \gamma \mathbf{P}^{\pi} \mathbf{v}.$$

Similarly we can define the Bellman error as the expected TD error at a given state, i.e.,

$$\begin{aligned} \bar{\boldsymbol{\delta}}(\boldsymbol{\theta}) &\doteq \mathcal{T}^{\pi} \Phi \boldsymbol{\theta} - \Phi \boldsymbol{\theta}, \\ \bar{\delta}(s, \boldsymbol{\theta}) &\doteq \mathbf{r}^{\pi}(s) + \gamma \sum_{s'} p^{\pi}(s, s') \boldsymbol{\phi}(s')^{\intercal} \boldsymbol{\theta} - \boldsymbol{\phi}(s)^{\intercal} \boldsymbol{\theta}. \end{aligned}$$

Hence the Mean Squared Bellman Error is defined as:

$$\begin{aligned} \text{MSBE}(\boldsymbol{\theta}) &\doteq \|\bar{\boldsymbol{\delta}}(\boldsymbol{\theta})\|_{d^{\pi}}^{2} \\ &= \sum_{s} d^{\pi}(s) \bar{\delta}(s, \boldsymbol{\theta})^{2}. \end{aligned}$$

Note, that it is often not possible to minimize the Bellman Error directly, as the Bellman Error might lie outside the representable space of our function approximator. To see this, lets look at the Bellman operator:

$$\mathcal{T}^{\pi} \mathbf{v} = \mathbf{r}^{\pi} + \gamma \mathbf{P}^{\pi} \Phi.$$

We see that the second term, $\mathbf{P}^{\pi} \Phi$ might not lie in the column space of Φ, as well as $\mathbf{r}^{\pi}$ might also lie outside the column span and hence we might not be able to represent the Bellman operator in the space of Φ. To be able to work with this error, we shouild be able to project this back into our representable function space and then minimize the error.

Project Back to Function Space

Whenever we have vector lets say $\mathbf{y} \in \mathbb{R}^{n}$, which might not lie in coloumn span of Φ, i.e., not representable as $\Phi \boldsymbol{\theta}$, we can project it back into the coloumn span of Φ by finding the closest point in the coloumn span of Φ to $\mathbf{y}$, i.e., $\Phi \boldsymbol{\theta}$. This is done by finding the $\boldsymbol{\theta}$ which minimizes the distance between $\mathbf{y}$ and $\Phi \boldsymbol{\theta}$, i.e., $\boldsymbol{\theta}_{\mathbf{y}} = \arg\min_{\boldsymbol{\theta}} \|\mathbf{y} - \Phi \boldsymbol{\theta}\|_{d^{\pi}}$.

$$\begin{aligned} \|\mathbf{y} - \Phi \boldsymbol{\theta}\|_{d^{\pi}} &= (\mathbf{y} - \Phi \boldsymbol{\theta})^{\intercal} \mathbf{D}^{\pi} (\mathbf{y} - \Phi \boldsymbol{\theta}) \\ &= \mathbf{y}^{\intercal} \mathbf{D}^{\pi} \mathbf{y} - 2 \boldsymbol{\theta}^{\intercal} \Phi^{\intercal} \mathbf{D}^{\pi} \mathbf{y} + \boldsymbol{\theta}^{\intercal} \Phi^{\intercal} \mathbf{D}^{\pi} \Phi \boldsymbol{\theta} \\ \nabla \|\mathbf{y} - \Phi \boldsymbol{\theta}\|_{d^{\pi}} &= -2 \Phi^{\intercal} \mathbf{D}^{\pi} \mathbf{y} + 2 \Phi^{\intercal} \mathbf{D}^{\pi} \Phi \boldsymbol{\theta}. \end{aligned}$$

The above term is minimized when we put the gradient to 0, i.e., when:

$$\begin{aligned} \Phi^{\intercal} \mathbf{D}^{\pi} \Phi \boldsymbol{\theta} &= \Phi^{\intercal} \mathbf{D}^{\pi} \mathbf{y}, \\ \boldsymbol{\theta}_{\mathbf{y}} &= (\Phi^{\intercal} \mathbf{D}^{\pi} \Phi)^{-1} \Phi^{\intercal} \mathbf{D}^{\pi} \mathbf{y}. \end{aligned}$$

Hence if we want to define a projection operator, that project $\mathbf{y}$ back into the column span of Φ, we can define it as:

$$\begin{aligned} \Pi \mathbf{y} &= \Phi \boldsymbol{\theta}_{\mathbf{y}} \\ &= \Phi (\Phi^{\intercal} \mathbf{D}^{\pi} \Phi)^{-1} \Phi^{\intercal} \mathbf{D}^{\pi} \mathbf{y}. \end{aligned}$$

This gives us the projection operator which projects $\mathbf{y}$ back into the column span of Φ as:

$$\Pi = \Phi (\Phi^{\intercal} \mathbf{D}^{\pi} \Phi)^{-1} \Phi^{\intercal} \mathbf{D}^{\pi}.$$

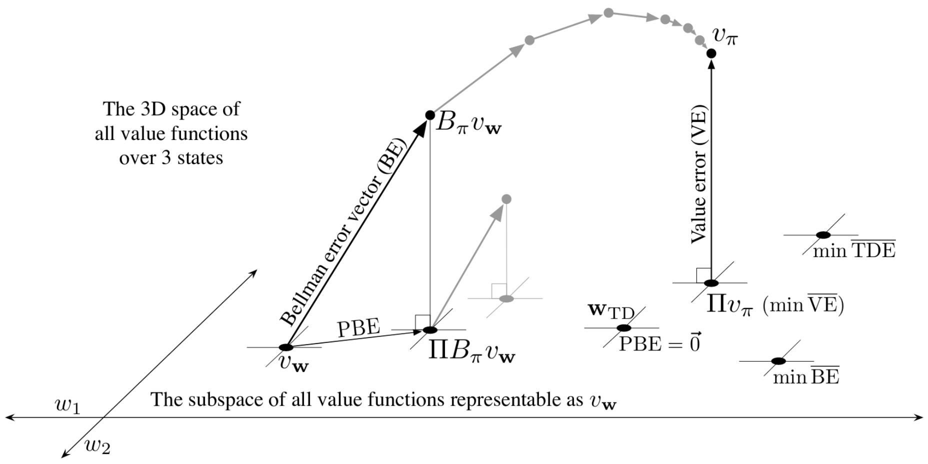
The 3D space of all value functions over 3 states
v_π
$B_\pi v_{\mathbf{w}}$
Bellman error vector (BE)
Value error (VE)
$\min \overline{\mathrm{TDE}}$
Πv_π $(\min \overline{\mathrm{VE}})$
$\mathbf{w}_{\mathrm{TD}}$
$\mathrm{PBE} = \vec{0}$
PBE
$\Pi B_\pi v_{\mathbf{w}}$
$v_{\mathbf{w}}$
$\min \overline{\mathrm{BE}}$
w_1
The subspace of all value functions representable as $v_{\mathbf{w}}$
w_2

Figure 3: The geometry of different errors (Sutton and Barto, 2018). This figure is adapted from Sutton and Barto (2018), where we have three states and two parameters or weights; hence, we can have span value functions represented by these two weight parameters.

Projected Bellman Error

The mean squared projected Bellman error is defined as the error between the projection of the Bellman operator and the current approximation of the value function, or simply the projection applied on the Bellman error, i.e.,

$$\begin{aligned}\mathrm{MSPBE} &\doteq \|\Pi\bar{\boldsymbol{\delta}}(\boldsymbol{\theta})\|^2_{d^\pi} \\ &= \sum_s d^\pi(s)\mathbf{E}\big[\Pi\bar{\delta}(S_t,\boldsymbol{\theta})^2 | S_t = s\big] \,.\end{aligned}$$

The good thing about this objective is that the minimizer of this objective coincides with $\boldsymbol{\theta}_{\mathtt{TD}}$, in the linear case.

But how does that happen, let's take a look at this:

$$\begin{aligned}\mathrm{MSPBE}(\boldsymbol{\theta}) &= (\Pi\bar{\boldsymbol{\delta}}(\boldsymbol{\theta}))^\intercal\mathbf{D}^\pi(\Pi\bar{\boldsymbol{\delta}}(\boldsymbol{\theta})) \\ &= \bar{\boldsymbol{\delta}}(\boldsymbol{\theta})^\intercal\Pi^\intercal\mathbf{D}^\pi\Pi\bar{\boldsymbol{\delta}}(\boldsymbol{\theta}) \\ &= \bar{\boldsymbol{\delta}}(\boldsymbol{\theta})^\intercal(\Phi(\Phi^\intercal\mathbf{D}^\pi\Phi)^{-1}\Phi^\intercal\mathbf{D}^\pi)^\intercal\mathbf{D}^\pi\Phi(\Phi^\intercal\mathbf{D}^\pi\Phi)^{-1}\Phi^\intercal\mathbf{D}^\pi\bar{\boldsymbol{\delta}}(\boldsymbol{\theta}) \\ &= \bar{\boldsymbol{\delta}}(\boldsymbol{\theta})^\intercal\mathbf{D}^\pi\Phi(\Phi^\intercal\mathbf{D}^\pi\Phi)^{-1^\intercal}\cancel{\Phi^\intercal\mathbf{D}^\pi\Phi}\cancel{(\Phi^\intercal\mathbf{D}^\pi\Phi)^{-1}}\Phi^\intercal\mathbf{D}^\pi\bar{\boldsymbol{\delta}}(\boldsymbol{\theta}) \\ &= \bar{\boldsymbol{\delta}}(\boldsymbol{\theta})^\intercal\mathbf{D}^\pi\Phi(\Phi^\intercal\mathbf{D}^\pi\Phi)^{-1^\intercal}\Phi^\intercal\mathbf{D}^\pi\bar{\boldsymbol{\delta}}(\boldsymbol{\theta}) \\ &= (\Phi^\intercal\mathbf{D}^\pi\bar{\boldsymbol{\delta}}(\boldsymbol{\theta}))^\intercal(\Phi^\intercal\mathbf{D}^\pi\Phi)^{-1}\Phi^\intercal\mathbf{D}^\pi\bar{\boldsymbol{\delta}}(\boldsymbol{\theta}).\end{aligned}$$

Lets look at $\Phi^\intercal\mathbf{D}^\pi\bar{\boldsymbol{\delta}}(\boldsymbol{\theta})$, i.e.,

$$\begin{aligned}\Phi^\intercal\mathbf{D}^\pi\bar{\boldsymbol{\delta}}(\boldsymbol{\theta}) &= \Phi^\intercal\mathbf{D}^\pi(\mathcal{T}^\pi\Phi\boldsymbol{\theta} - \Phi\boldsymbol{\theta}) \\ &= \Phi^\intercal\mathbf{D}^\pi(\mathbf{r}^\pi + \gamma\mathbf{P}^\pi\Phi\boldsymbol{\theta} - \Phi\boldsymbol{\theta}) \\ &= \Phi^\intercal\mathbf{D}^\pi(\mathbf{r}^\pi - (\mathbf{I} - \gamma\mathbf{P}^\pi)\Phi\boldsymbol{\theta}) \\ &= \Phi^\intercal\mathbf{D}^\pi\mathbf{r}^\pi - \Phi^\intercal\mathbf{D}^\pi(\mathbf{I} - \gamma\mathbf{P}^\pi)\Phi\boldsymbol{\theta} \\ &= \mathbf{b} - \mathbf{A}\boldsymbol{\theta}.\end{aligned}$$

Hence, when $\mathbf{b} = \mathbf{A}\boldsymbol{\theta}$, we have that the MSPBE is minimized, and hence the minimizer of the MSPBE is $\boldsymbol{\theta}_{\mathtt{TD}}$.

Gradient TD methods

The gradient of the MSPBE is given as:

$$\nabla \text{MSPBE}(\boldsymbol{\theta}) = 2\nabla(\Phi^\intercal \mathbf{D}^\pi \bar{\boldsymbol{\delta}}(\boldsymbol{\theta}))^\intercal((\Phi^\intercal \mathbf{D}^\pi \Phi)^{-1})(\Phi^\intercal \mathbf{D}^\pi \bar{\boldsymbol{\delta}}(\boldsymbol{\theta}))$$

Again looking at individual terms, we have:

$$\begin{aligned}
\Phi^\intercal \mathbf{D}^\pi \bar{\boldsymbol{\delta}}(\boldsymbol{\theta}) &= \mathbf{E}[\delta_t \boldsymbol{\phi}_t] \\
\Phi^\intercal \mathbf{D}^\pi \Phi &= \mathbf{E}[\boldsymbol{\phi}_t \boldsymbol{\phi}_t^\intercal] \\
\nabla \mathbf{E}[\delta_t \boldsymbol{\phi}_t]^\intercal &= \mathbf{E}[\nabla \delta_t^\intercal \boldsymbol{\phi}_t^\intercal] \\
&= \mathbf{E}\big[\nabla(R_t + \gamma \boldsymbol{\theta}^\intercal \boldsymbol{\phi}_{t+1} - \boldsymbol{\theta}^\intercal \boldsymbol{\phi}_t)^\intercal \phi_t^\intercal\big] \\
&= \mathbf{E}\big[(\gamma \boldsymbol{\phi}_{t+1} - \boldsymbol{\phi}_t)\boldsymbol{\phi}_t^\intercal\big] .
\end{aligned}$$

Hence, the gradient of the MSPBE is given as:

$$\nabla \text{MSPBE}(\boldsymbol{\theta}) = 2\mathbf{E}\big[(\gamma \boldsymbol{\phi}_{t+1} - \boldsymbol{\phi}_t)\boldsymbol{\phi}_t^\intercal\big] \, \mathbf{E}[\boldsymbol{\phi}_t \boldsymbol{\phi}_t^\intercal]^{-1} \, \mathbf{E}[\delta_t \boldsymbol{\phi}_t] .$$

In the above we have multiplications of expectation where given a state, the first and third term depend on the samples of the next state, and hence we cannot make use of a single sample of $\boldsymbol{\phi}_{t+1}$ to approximate the above term in the stochastic case, it will otherwise give us a biased update. (This is because it is a multiplication between the three expectation terms, so they need to have separate samples to estimate these quantities.)

GTD2

The idea is to store the second and third terms and approximate them separately and make use of the stochastic samples to estimate the first term. If we look $\mathbf{E}[\boldsymbol{\phi}_t \boldsymbol{\phi}_t^\intercal]^{-1} \, \mathbf{E}[\delta_t \boldsymbol{\phi}_t]$ is the least squares solution to $\|\delta(s, \boldsymbol{\theta}) - \boldsymbol{\phi}(s)^\intercal \mathbf{w}\|_{d^\pi}$, where,

$$\mathbf{w} = \mathbf{E}[\boldsymbol{\phi}_t \boldsymbol{\phi}_t^\intercal]^{-1} \, \mathbf{E}[\boldsymbol{\phi}_t \delta_t] .$$

Hence, we can have a separate learning process, i.e.,

$$\begin{aligned}
\mathbf{w}_{t+1} &= \mathbf{w}_t - \eta(\delta_t - \boldsymbol{\phi}_t^\intercal \mathbf{w}_t)\boldsymbol{\phi}_t \\
\boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t - \frac{1}{2}\alpha \nabla \text{MSPBE}(\boldsymbol{\theta}_t) \\
&= \boldsymbol{\theta}_t + \alpha \mathbf{E}\big[(-\gamma \boldsymbol{\phi}_{t+1} + \boldsymbol{\phi}_t)\boldsymbol{\phi}_t^\intercal\big] \, \mathbf{E}[\boldsymbol{\phi}_t \boldsymbol{\phi}_t^\intercal]^{-1} \, \mathbf{E}[\delta_t \boldsymbol{\phi}_t] \\
&\approx \boldsymbol{\theta}_t + \alpha \mathbf{E}\big[(-\gamma \boldsymbol{\phi}_{t+1} + \boldsymbol{\phi}_t)\boldsymbol{\phi}_t^\intercal\big] \, \mathbf{w}_t \\
&\approx \boldsymbol{\theta}_t + \alpha(-\gamma \boldsymbol{\phi}_{t+1} + \boldsymbol{\phi}_t)\boldsymbol{\phi}_t^\intercal \mathbf{w}_t \\
&= \boldsymbol{\theta}_t + \alpha(\boldsymbol{\phi}_t^\intercal \mathbf{w}_t)(-\gamma \boldsymbol{\phi}_{t+1} + \boldsymbol{\phi}_t).
\end{aligned}$$

A slightly better method exists.

Temporal Difference with Gradient Correction (TDC)

Rearrganing the terms in the update we have:

$$\begin{aligned}
\boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t + \alpha \mathbf{E}\big[(-\gamma \boldsymbol{\phi}_{t+1} + \boldsymbol{\phi}_t)\boldsymbol{\phi}_t^\intercal\big] \, \mathbf{E}[\boldsymbol{\phi}_t \boldsymbol{\phi}_t^\intercal]^{-1} \, \mathbf{E}[\delta_t \boldsymbol{\phi}_t] \\
&= \boldsymbol{\theta}_t + \alpha(-\mathbf{E}\big[\gamma \boldsymbol{\phi}_{t+1} \boldsymbol{\phi}_t^\intercal\big] + \mathbf{E}[\boldsymbol{\phi}_t \boldsymbol{\phi}_t^\intercal])\mathbf{E}[\boldsymbol{\phi}_t \boldsymbol{\phi}_t^\intercal]^{-1} \, \mathbf{E}[\delta_t \boldsymbol{\phi}_t] \\
&\approx \boldsymbol{\theta}_t + \alpha(\mathbf{E}[\delta_t \phi_t] - \mathbf{E}\big[\gamma \boldsymbol{\phi}_{t+1} \boldsymbol{\phi}_t^\intercal\big] \, \mathbf{w}_t) \\
&\approx \boldsymbol{\theta}_t + \alpha(\delta_t \boldsymbol{\phi}_t - \gamma(\boldsymbol{\phi}_t^\intercal \mathbf{w}_t)\boldsymbol{\phi}_{t+1})
\end{aligned}$$

References

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Mach. Learn.*, 3:9–44.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Varga, R. S. (1962). Iterative analysis. *New Jersey*, 322.