

# Value Function Geometry and Gradient TD

Ashwin Rao

ICME, Stanford University

February 22, 2021

# Overview

- 1 Motivation and Notation
- 2 Vector/Geometric Representation of Value Functions
- 3 BE-minimization with a Linear System Formulation
- 4 Residual Gradient TD
- 5 PBE-minimization with a Linear System Formulation
- 6 Gradient TD

# Motivation for understanding Value Function Geometry

- Helps us better understand transformations of Value Functions (VFs)
- Across the various DP and RL algorithms
- Particularly helps when VFs are approximated, esp. with linear approx
- Provides insights into stability and convergence
- Particularly when dealing with the “Deadly Triad”
- Deadly Triad := [Bootstrapping, Func Approx, Off-Policy]
- **Leads us to Gradient TD**

- Assume state space  $\mathcal{S}$  consists of  $n$  states:  $\{s_1, s_2, \dots, s_n\}$
- Action space  $\mathcal{A}$  consisting of finite number of actions
- This exposition extends easily to continuous state/action spaces too
- This exposition is for a fixed (often stochastic) policy denoted  $\pi(s, a)$
- VF for a policy  $\pi$  is denoted as  $\mathbf{V}^\pi : \mathcal{S} \rightarrow \mathbb{R}$
- $m$  feature functions  $\phi_1, \phi_2, \dots, \phi_m : \mathcal{S} \rightarrow \mathbb{R}$
- Feature vector for a state  $s \in \mathcal{S}$  denoted as  $\phi(s) \in \mathbb{R}^m$
- For linear function approximation of VF with weights  $\mathbf{w} = (w_1, w_2, \dots, w_m)$ , VF  $\mathbf{V}_{\mathbf{w}} : \mathcal{S} \rightarrow \mathbb{R}$  is defined as:

$$\mathbf{V}_{\mathbf{w}}(s) = \phi(s)^T \cdot \mathbf{w} = \sum_{j=1}^m \phi_j(s) \cdot w_j \text{ for any } s \in \mathcal{S}$$

- $\mu_\pi : \mathcal{S} \rightarrow [0, 1]$  denotes the states' probability distribution under  $\pi$

# VF Geometry and VF Linear Approximations

- Consider  $n$ -dim space  $\mathbb{R}^n$ , with each dim corresponding to a state in  $\mathcal{S}$
- Think of a VF (typically denoted  $\mathbf{V}$ ):  $\mathcal{S} \rightarrow \mathbb{R}$  as a vector in this space
- Each dimension's coordinate is the VF for that dimension's state
- Coordinates of vector  $\mathbf{V}^\pi$  for policy  $\pi$  are:  $[\mathbf{V}^\pi(s_1), \dots, \mathbf{V}^\pi(s_n)]$
- Consider  $m$  independent vectors with  $j^{th}$  vector:  $[\phi_j(s_1), \dots, \phi_j(s_n)]$
- These  $m$  vectors are the  $m$  columns of  $n \times m$  matrix  $\Phi = [\phi_j(s_i)]$
- Their span represents  $m$ -dim subspace within this  $n$ -dim space
- Spanned by the set of all  $\mathbf{w} = [w_1, w_2, \dots, w_m] \in \mathbb{R}^m$
- Vector  $\mathbf{V}_{\mathbf{w}} = \Phi \cdot \mathbf{w}$  in this subspace has coordinates  $[\mathbf{V}_{\mathbf{w}}(s_1), \dots, \mathbf{V}_{\mathbf{w}}(s_n)]$
- Vector  $\mathbf{V}_{\mathbf{w}}$  is fully specified by  $\mathbf{w}$  (so we often say  $\mathbf{w}$  to mean  $\mathbf{V}_{\mathbf{w}}$ )

# Some more notation

- Denote  $\mathcal{R}(s, a)$  as the Expected Reward upon action  $a$  in state  $s$
- Denote  $\mathcal{P}(s, a, s')$  as the probability of transition  $s \rightarrow s'$  upon action  $a$
- Define

$$\mathcal{R}^\pi(s) = \sum_{a \in \mathcal{A}} \pi(s, a) \cdot \mathcal{R}(s, a)$$

$$\mathcal{P}^\pi(s, s') = \sum_{a \in \mathcal{A}} \pi(s, a) \cdot \mathcal{P}(s, a, s')$$

- Notation  $\mathcal{R}^\pi$  refers to vector  $[\mathcal{R}^\pi(s_1), \mathcal{R}^\pi(s_2), \dots, \mathcal{R}^\pi(s_n)]$
- Notation  $\mathcal{P}^\pi$  refers to matrix  $[\mathcal{P}^\pi(s_i, s_{i'})], 1 \leq i, i' \leq n$
- Denote  $\gamma$  as the MDP discount factor

# Bellman operator $B^\pi$

- Bellman Policy Operator  $B^\pi$  for policy  $\pi$  operating on VF vector  $\mathbf{V}$ :

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

- $B^\pi$  is a linear operator in vector space  $\mathbb{R}^n$
- So we denote and treat  $B^\pi$  as a  $n \times n$  matrix
- Note that  $\mathbf{V}^\pi$  is the fixed point of  $B^\pi$ , i.e.,

$$B^\pi \cdot \mathbf{V}^\pi = \mathbf{V}^\pi$$

- If we start with an arbitrary VF vector  $\mathbf{V}$  and repeatedly apply  $B^\pi$ , by Fixed-Point Theorem, we will reach the fixed point  $\mathbf{V}^\pi$
- This is the Dynamic Programming Policy Evaluation algorithm
- Monte Carlo without func approx also converges to  $\mathbf{V}^\pi$  (albeit slowly)

# Projection operator $\Pi_\Phi$

- First we define “distance”  $d(\mathbf{V}_1, \mathbf{V}_2)$  between VF vectors  $\mathbf{V}_1, \mathbf{V}_2$
- Weighted by  $\mu_\pi$  across the  $n$  dimensions of  $\mathbf{V}_1, \mathbf{V}_2$

$$d(\mathbf{V}_1, \mathbf{V}_2) = \sum_{i=1}^n \mu_\pi(s_i) \cdot (\mathbf{V}_1(s_i) - \mathbf{V}_2(s_i))^2 = (\mathbf{V}_1 - \mathbf{V}_2)^T \cdot \mathbf{D} \cdot (\mathbf{V}_1 - \mathbf{V}_2)$$

where  $\mathbf{D}$  is the square diagonal matrix consisting of  $\mu_\pi(s_i), 1 \leq i \leq n$

- Projection operator for subspace spanned by  $\Phi$  is denoted as  $\Pi_\Phi$
- $\Pi_\Phi$  performs an orthogonal projection of VF vector  $\mathbf{V}$  on subspace  $\Phi$
- So,  $\Pi_\Phi(\mathbf{V})$  is the VF in subspace  $\Phi$  defined by  $\arg \min_{\mathbf{w}} d(\mathbf{V}, \mathbf{V}_w)$
- This is a weighted least squares regression with solution:

$$\mathbf{w} = (\Phi^T \cdot \mathbf{D} \cdot \Phi)^{-1} \cdot \Phi^T \cdot \mathbf{D} \cdot \mathbf{V}$$

- So, we denote and treat Projection operator  $\Pi_\Phi$  as a  $n \times n$  matrix:

$$\Pi_\Phi = \Phi \cdot (\Phi^T \cdot \mathbf{D} \cdot \Phi)^{-1} \cdot \Phi^T \cdot \mathbf{D}$$



## 4 VF vectors of interest in the $\Phi$ subspace

Note: We will refer to the  $\Phi$ -subspace VF vectors by their weights  $\mathbf{w}$

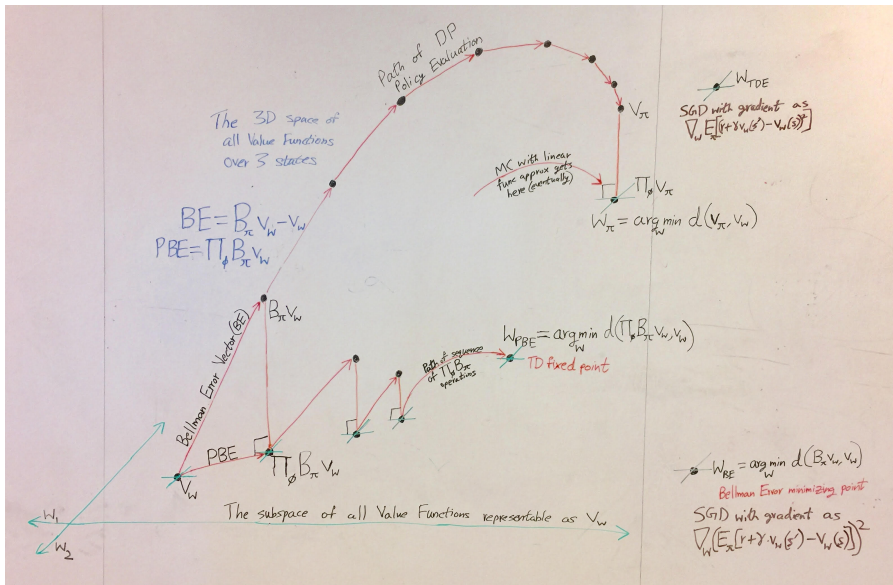
- ① Projection  $\Pi_{\Phi} \cdot \mathbf{V}^{\pi}$  yields  $\mathbf{w}_{\pi} = \arg \min_{\mathbf{w}} d(\mathbf{V}^{\pi}, \mathbf{V}_{\mathbf{w}})$ 
  - This is the VF we seek when doing linear function approximation
  - Because it is the VF vector “closest” to  $\mathbf{V}^{\pi}$  in the  $\Phi$  subspace
  - Monte-Carlo with linear func approx will (slowly) converge to  $\mathbf{w}_{\pi}$
- ② Bellman Error (BE)-minimizing:  $\mathbf{w}_{BE} = \arg \min_{\mathbf{w}} d(\mathbf{B}^{\pi} \cdot \mathbf{V}_{\mathbf{w}}, \mathbf{V}_{\mathbf{w}})$ 
  - This can be expressed as the solution of a linear system  $\mathbf{A} \cdot \mathbf{w} = \mathbf{b}$
  - Matrix  $\mathbf{A}$  and Vector  $\mathbf{b}$  comprises of  $\mathcal{R}^{\pi}, \mathcal{P}^{\pi}, \Phi, \mu_{\pi}$
  - In model-free setting,  $\mathbf{A}$  and  $\mathbf{b}$  can be estimated with batch data
  - For non-linear approx or off-policy, Residual Gradient TD Algorithm
  - Based on observation:  $\mathbf{w}_{BE} = \arg \min_{\mathbf{w}} (\mathbb{E}_{\pi}[\delta])^2$ , where  $\delta$  is TD Error
  - Cannot learn if we can only access features, and not underlying states
- ③ Temporal Difference Error (TDE)-minimizing:  
 $\mathbf{w}_{TDE} = \arg \min_{\mathbf{w}} \mathbb{E}_{\pi}[\delta^2]$ 
  - Naive Residual Gradient TD Algorithm

## 4 VF vectors of interest in the $\Phi$ subspace (continued)

### ④ Projected Bellman Error (PBE)-minimizing:

$$\mathbf{w}_{PBE} = \arg \min_{\mathbf{w}} d((\Pi_{\Phi} \cdot \mathbf{B}^{\pi}) \cdot \mathbf{V}_{\mathbf{w}}, \mathbf{V}_{\mathbf{w}})$$

- The minimum is 0, i.e.,  $\Phi \cdot \mathbf{w}_{PBE}$  is the fixed point of operator  $\Pi_{\Phi} \cdot \mathbf{B}^{\pi}$
- Starting with an arbitrary VF vector  $\mathbf{V}$  and repeatedly applying  $\mathbf{B}^{\pi}$  (potentially taking it out of the subspace) followed by  $\Pi_{\Phi}$  (projecting it back to the subspace), we will reach the fixed point  $\Phi \cdot \mathbf{w}_{PBE}$
- $\mathbf{w}_{PBE}$  can be expressed as the solution of a linear system  $\mathbf{A} \cdot \mathbf{w} = \mathbf{b}$
- In model-free setting,  $\mathbf{A}$  and  $\mathbf{b}$  can be estimated with batch data
- This yields the *Least Squares Temporal Difference (LSTD)* algorithm
- For non-linear approx or off-policy, Gradient TD Algorithms



# Solution of $\mathbf{w}_{BE}$ with a Linear System Formulation

$$\begin{aligned}\mathbf{w}_{BE} &= \arg \min_{\mathbf{w}} d(\mathbf{V}_{\mathbf{w}}, \mathcal{R}^{\pi} + \gamma \mathcal{P}^{\pi} \cdot \mathbf{V}_{\mathbf{w}}) \\ &= \arg \min_{\mathbf{w}} d(\Phi \cdot \mathbf{w}, \mathcal{R}^{\pi} + \gamma \mathcal{P}^{\pi} \cdot \Phi \cdot \mathbf{w}) \\ &= \arg \min_{\mathbf{w}} d(\Phi \cdot \mathbf{w} - \gamma \mathcal{P}^{\pi} \cdot \Phi \cdot \mathbf{w}, \mathcal{R}^{\pi}) \\ &= \arg \min_{\mathbf{w}} d((\Phi - \gamma \mathcal{P}^{\pi} \cdot \Phi) \cdot \mathbf{w}, \mathcal{R}^{\pi})\end{aligned}$$

This is a weighted least-squares linear regression of  $\mathcal{R}^{\pi}$  versus  $\Phi - \gamma \mathcal{P}^{\pi} \cdot \Phi$  with weights  $\mu_{\pi}$ , whose solution is:

$$\mathbf{w}_{BE} = ((\Phi - \gamma \mathcal{P}^{\pi} \cdot \Phi)^T \cdot \mathbf{D} \cdot (\Phi - \gamma \mathcal{P}^{\pi} \cdot \Phi))^{-1} \cdot (\Phi - \gamma \mathcal{P}^{\pi} \cdot \Phi)^T \cdot \mathbf{D} \cdot \mathcal{R}^{\pi}$$

# Model-Free Learning of $\mathbf{w}_{BE}$

- Let us refer to  $(\Phi - \gamma \mathcal{P}^\pi \cdot \Phi)^T \cdot \mathbf{D} \cdot (\Phi - \gamma \mathcal{P}^\pi \cdot \Phi)$  as  $\mathbf{A}$
- Let us refer to  $(\Phi - \gamma \mathcal{P}^\pi \cdot \Phi)^T \cdot \mathbf{D} \cdot \mathcal{R}^\pi$  as  $\mathbf{b}$
- So that  $\mathbf{w}_{BE} = \mathbf{A}^{-1} \cdot \mathbf{b}$
- Following policy  $\pi$ , each time we perform a model-free transition from  $s$  to  $s'$  getting reward  $r$ , we get a sample estimate of  $\mathbf{A}$  and  $\mathbf{b}$
- Estimate of  $\mathbf{A}$  is the outer-product of vector  $\phi(s) - \gamma \cdot \phi(s')$  with itself
- Estimate of  $\mathbf{b}$  is scalar  $r$  times vector  $\phi(s) - \gamma \cdot \phi(s')$
- Average these estimates across many such model-free transitions
- However, this requires  $m$  (number of features) to not be too large

# Residual Gradient Algorithm to solve for $\mathbf{w}_{BE}$

- We defined  $\mathbf{w}_{BE}$  as the vector in the  $\Phi$  subspace that minimizes BE
- But BE for a state is the expected TD error  $\delta$  in that state when following policy  $\pi$
- So we want to do SGD with gradient of square of expected TD error

$$\begin{aligned}\Delta \mathbf{w} &= -\frac{1}{2}\alpha \cdot \nabla_{\mathbf{w}}(\mathbb{E}_{\pi}[\delta])^2 \\ &= -\alpha \cdot \mathbb{E}_{\pi}[r + \gamma \cdot \phi(s')^T \cdot \mathbf{w} - \phi(s)^T \cdot \mathbf{w}] \cdot \nabla_{\mathbf{w}}\mathbb{E}_{\pi}[\delta] \\ &= \alpha \cdot (\mathbb{E}_{\pi}[r + \gamma \cdot \phi(s')^T \cdot \mathbf{w}] - \phi(s)^T \cdot \mathbf{w}) \cdot (\phi(s) - \gamma \cdot \mathbb{E}_{\pi}[\phi(s')])\end{aligned}$$

- This is called the *Residual Gradient* algorithm
- Requires two independent samples of  $s'$  transitioning from  $s$
- In that case, converges to  $\mathbf{w}_{BE}$  robustly (even for non-linear approx)
- But it is slow, and doesn't converge to a desirable place
- Cannot learn if we can only access features, and not underlying states

# Naive Residual Gradient Algorithm to solve for $\mathbf{w}_{TDE}$

- We defined  $\mathbf{w}_{TDE}$  as the vector in the  $\Phi$  subspace that minimizes the expected square of the TD error  $\delta$  when following policy  $\pi$

$$\mathbf{w}_{TDE} = \arg \min_{\mathbf{w}} \sum_{s \in \mathcal{S}} \mu_{\pi}(s) \sum_{r, s'} \mathbb{P}_{\pi}(r, s' | s) \cdot (r + \gamma \cdot \phi(s')^T \cdot \mathbf{w} - \phi(s)^T \cdot \mathbf{w})^2$$

- To perform SGD, we have to estimate the gradient of the expected square of TD error by sampling
- The weight update for each sample in the SGD will be:

$$\begin{aligned} \Delta \mathbf{w} &= -\frac{1}{2} \alpha \cdot \nabla_{\mathbf{w}} (r + \gamma \cdot \phi(s')^T \cdot \mathbf{w} - \phi(s)^T \cdot \mathbf{w})^2 \\ &= \alpha \cdot (r + \gamma \cdot \phi(s')^T \cdot \mathbf{w} - \phi(s)^T \cdot \mathbf{w}) \cdot (\phi(s) - \gamma \cdot \phi(s')) \end{aligned}$$

- This algorithm (named *Naive Residual Gradient*) converges robustly, but not to a desirable place

# Solution of $\mathbf{w}_{PBE}$ with a Linear System Formulation

$\Phi \cdot \mathbf{w}_{PBE}$  is the fixed point of operator  $\Pi_\Phi \cdot \mathbf{B}^\pi$ . We know:

$$\Pi_\Phi = \Phi \cdot (\Phi^T \cdot \mathbf{D} \cdot \Phi)^{-1} \cdot \Phi^T \cdot \mathbf{D}$$

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

Therefore,

$$\Phi \cdot (\Phi^T \cdot \mathbf{D} \cdot \Phi)^{-1} \cdot \Phi^T \cdot \mathbf{D} \cdot (\mathcal{R}^\pi + \gamma \mathcal{P}^\pi \cdot \Phi \cdot \mathbf{w}_{PBE}) = \Phi \cdot \mathbf{w}_{PBE}$$

Since columns of  $\Phi$  are assumed to be independent (full rank),

$$(\Phi^T \cdot \mathbf{D} \cdot \Phi)^{-1} \cdot \Phi^T \cdot \mathbf{D} \cdot (\mathcal{R}^\pi + \gamma \mathcal{P}^\pi \cdot \Phi \cdot \mathbf{w}_{PBE}) = \mathbf{w}_{PBE}$$

$$\Phi^T \cdot \mathbf{D} \cdot (\mathcal{R}^\pi + \gamma \mathcal{P}^\pi \cdot \Phi \cdot \mathbf{w}_{PBE}) = \Phi^T \cdot \mathbf{D} \cdot \Phi \cdot \mathbf{w}_{PBE}$$

$$\Phi^T \cdot \mathbf{D} \cdot (\Phi - \gamma \mathcal{P}^\pi \cdot \Phi) \cdot \mathbf{w}_{PBE} = \Phi^T \cdot \mathbf{D} \cdot \mathcal{R}^\pi$$

This is a square linear system of the form  $\mathbf{A} \cdot \mathbf{w}_{PBE} = \mathbf{b}$  whose solution is:

$$\mathbf{w}_{PBE} = \mathbf{A}^{-1} \cdot \mathbf{b} = (\Phi^T \cdot \mathbf{D} \cdot (\Phi - \gamma \mathcal{P}^\pi \cdot \Phi))^{-1} \cdot \Phi^T \cdot \mathbf{D} \cdot \mathcal{R}^\pi$$



# Model-Free Learning of $\mathbf{w}_{PBE}$

- How do we construct matrix  $\mathbf{A} = \Phi^T \cdot \mathbf{D} \cdot (\Phi - \gamma \mathcal{P}^\pi \cdot \Phi)$  and vector  $\mathbf{b} = \Phi^T \cdot \mathbf{D} \cdot \mathcal{R}^\pi$  without a model?
- Following policy  $\pi$ , each time we perform a model-free transition from  $s$  to  $s'$  getting reward  $r$ , we get a sample estimate of  $\mathbf{A}$  and  $\mathbf{b}$
- Estimate of  $\mathbf{A}$  is outer-product of vectors  $\phi(s)$  and  $\phi(s) - \gamma \cdot \phi(s')$
- Estimate of  $\mathbf{b}$  is scalar  $r$  times vector  $\phi(s)$
- Average these estimates across many such model-free transitions
- This algorithm is called Least Squares Temporal Difference (LSTD)
- Alternative: Our usual Semi-Gradient TD descent with updates:

$$\Delta \mathbf{w} = \alpha \cdot (r + \gamma \cdot \phi(s')^T \cdot \mathbf{w} - \phi(s)^T \cdot \mathbf{w}) \cdot \phi(s)$$

- This converges to  $\mathbf{w}_{PBE}$  because  $\mathbb{E}_\pi[\Delta \mathbf{w}] = 0$  yields

$$\begin{aligned} \Phi^T \cdot \mathbf{D} \cdot (\mathcal{R}^\pi + \gamma \mathcal{P}^\pi \cdot \Phi \cdot \mathbf{w} - \Phi \cdot \mathbf{w}) &= 0 \\ \Rightarrow \Phi^T \cdot \mathbf{D} \cdot (\Phi - \gamma \mathcal{P}^\pi \cdot \Phi) \cdot \mathbf{w} &= \Phi^T \cdot \mathbf{D} \cdot \mathcal{R}^\pi \end{aligned}$$

# Gradient TD Algorithms to solve for $\mathbf{w}_{PBE}$

- For on-policy linear func approx, semi-gradient TD works
- For non-linear func approx or off-policy, we need Gradient TD
  - GTD: The original Gradient TD algorithm
  - GTD-2: Second-generation GTD
  - TDC: TD with Gradient correction
- We need to set up the loss function whose gradient will drive SGD

$$\mathbf{w}_{PBE} = \arg \min_{\mathbf{w}} d(\Pi_{\Phi} \cdot \mathbf{B}^{\pi} \cdot \mathbf{V}_{\mathbf{w}}, \mathbf{V}_{\mathbf{w}}) = \arg \min_{\mathbf{w}} d(\Pi_{\Phi} \cdot \mathbf{B}^{\pi} \cdot \mathbf{V}_{\mathbf{w}}, \Pi_{\Phi} \cdot \mathbf{V}_{\mathbf{w}})$$

- So we define the loss function (denoting  $\mathbf{B}^{\pi} \cdot \mathbf{V}_{\mathbf{w}} - \mathbf{V}_{\mathbf{w}}$  as  $\delta_{\mathbf{w}}$ ) as:

$$\begin{aligned}\mathcal{L}(\mathbf{w}) &= (\Pi_{\Phi} \cdot \delta_{\mathbf{w}})^T \cdot \mathbf{D} \cdot (\Pi_{\Phi} \cdot \delta_{\mathbf{w}}) = \delta_{\mathbf{w}}^T \cdot \Pi_{\Phi}^T \cdot \mathbf{D} \cdot \Pi_{\Phi} \cdot \delta_{\mathbf{w}} \\ &= \delta_{\mathbf{w}}^T \cdot (\Phi \cdot (\Phi^T \cdot \mathbf{D} \cdot \Phi)^{-1} \cdot \Phi^T \cdot \mathbf{D})^T \cdot \mathbf{D} \cdot (\Phi \cdot (\Phi^T \cdot \mathbf{D} \cdot \Phi)^{-1} \cdot \Phi^T \cdot \mathbf{D}) \cdot \delta_{\mathbf{w}} \\ &= \delta_{\mathbf{w}}^T \cdot (\mathbf{D} \cdot \Phi \cdot (\Phi^T \cdot \mathbf{D} \cdot \Phi)^{-1} \cdot \Phi^T) \cdot \mathbf{D} \cdot (\Phi \cdot (\Phi^T \cdot \mathbf{D} \cdot \Phi)^{-1} \cdot \Phi^T \cdot \mathbf{D}) \cdot \delta_{\mathbf{w}} \\ &= (\delta_{\mathbf{w}}^T \cdot \mathbf{D} \cdot \Phi) \cdot (\Phi^T \cdot \mathbf{D} \cdot \Phi)^{-1} \cdot (\Phi^T \cdot \mathbf{D} \cdot \Phi) \cdot (\Phi^T \cdot \mathbf{D} \cdot \Phi)^{-1} \cdot (\Phi^T \cdot \mathbf{D} \cdot \delta_{\mathbf{w}}) \\ &= (\Phi^T \cdot \mathbf{D} \cdot \delta_{\mathbf{w}})^T \cdot (\Phi^T \cdot \mathbf{D} \cdot \Phi)^{-1} \cdot (\Phi^T \cdot \mathbf{D} \cdot \delta_{\mathbf{w}})\end{aligned}$$

# TDC Algorithm to solve for $\mathbf{w}_{PBE}$

We derive the TDC Algorithm based on  $\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w})$

$$\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}) = 2 \cdot (\nabla_{\mathbf{w}}(\Phi^T \cdot \mathbf{D} \cdot \delta_{\mathbf{w}})^T) \cdot (\Phi^T \cdot \mathbf{D} \cdot \Phi)^{-1} \cdot (\Phi^T \cdot \mathbf{D} \cdot \delta_{\mathbf{w}})$$

Now we express each of these 3 terms as expectations of model-free transitions  $s \xrightarrow{\pi} (r, s')$ , denoting  $r + \gamma \cdot \phi(s')^T \cdot \mathbf{w} - \phi(s)^T \cdot \mathbf{w}$  as  $\delta$

- $\Phi^T \cdot \mathbf{D} \cdot \delta_{\mathbf{w}} = \mathbb{E}[\delta \cdot \phi(s)]$
- $\nabla_{\mathbf{w}}(\Phi^T \cdot \mathbf{D} \cdot \delta_{\mathbf{w}})^T = \mathbb{E}[(\nabla_{\mathbf{w}}\delta) \cdot \phi(s)^T] = \mathbb{E}[(\gamma \cdot \phi(s') - \phi(s)) \cdot \phi(s)^T]$
- $\Phi^T \cdot \mathbf{D} \cdot \Phi = \mathbb{E}[\phi(s) \cdot \phi(s)^T]$

Substituting, we get:

$$\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}) = 2 \cdot \mathbb{E}[(\gamma \cdot \phi(s') - \phi(s)) \cdot \phi(s)^T] \cdot \mathbb{E}[\phi(s) \cdot \phi(s)^T]^{-1} \cdot \mathbb{E}[\delta \cdot \phi(s)]$$

# Weight Updates of TDC Algorithm

$$\begin{aligned}\Delta \mathbf{w} &= -\frac{1}{2}\alpha \cdot \nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}) \\ &= \alpha \cdot \mathbb{E}[(\phi(s) - \gamma \cdot \phi(s')) \cdot \phi(s)^T] \cdot \mathbb{E}[\phi(s) \cdot \phi(s)^T]^{-1} \cdot \mathbb{E}[\delta \cdot \phi(s)] \\ &= \alpha \cdot (\mathbb{E}[\phi(s) \cdot \phi(s)^T] - \gamma \cdot \mathbb{E}[\phi(s') \cdot \phi(s)^T]) \cdot \mathbb{E}[\phi(s) \cdot \phi(s)^T]^{-1} \cdot \mathbb{E}[\delta \cdot \phi(s)] \\ &= \alpha \cdot (\mathbb{E}[\delta \cdot \phi(s)] - \gamma \cdot \mathbb{E}[\phi(s') \cdot \phi(s)^T] \cdot \mathbb{E}[\phi(s) \cdot \phi(s)^T]^{-1} \cdot \mathbb{E}[\delta \cdot \phi(s)]) \\ &= \alpha \cdot (\mathbb{E}[\delta \cdot \phi(s)] - \gamma \cdot \mathbb{E}[\phi(s') \cdot \phi(s)^T] \cdot \boldsymbol{\theta})\end{aligned}$$

$\boldsymbol{\theta} = \mathbb{E}[\phi(s) \cdot \phi(s)^T]^{-1} \cdot \mathbb{E}[\delta \cdot \phi(s)]$  is the solution to weighted least-squares linear regression of  $\mathbf{B}^\pi \cdot \mathbf{V} - \mathbf{V}$  against  $\Phi$ , with weights as  $\mu_\pi$ .

**Cascade Learning: Update both  $\mathbf{w}$  and  $\boldsymbol{\theta}$  ( $\boldsymbol{\theta}$  converging faster)**

- $\Delta \mathbf{w} = \alpha \cdot \delta \cdot \phi(s) - \alpha \cdot \gamma \cdot \phi(s') \cdot (\boldsymbol{\theta}^T \cdot \phi(s))$
- $\Delta \boldsymbol{\theta} = \beta \cdot (\delta - \boldsymbol{\theta}^T \cdot \phi(s)) \cdot \phi(s)$

Note:  $\boldsymbol{\theta}^T \cdot \phi(s)$  operates as estimate of TD error  $\delta$  for current state  $s$