Generalizing the Projected Bellman Error Objective for Nonlinear Value Estimation

Martha White
Associate Professor
University of Alberta

amii

UNIVERSITY OF ALBERTA

The Value Estimation Problem

- Find approximate values v that minimizes the value error objective:

$$\|v - v_\pi\|_d = \sum_s d(s) \ \left(v(s) - v_\pi(s)\right)^2$$

- We **cannot** directly optimize this objective

Motivation and History

- Sound off-policy value estimation was an open problem for some time
- Significant **progress** since the introduction of the mean squared projected Bellman Error ($\overline{\text{PBE}}$) and resulting gradient TD algorithms
- $\overline{\text{PBE}}$ primarily for the **linear** setting
    - nonlinear $\overline{\text{PBE}}$ relatively complex, with Hessian-vector products
- $\overline{\text{BE}}$ **difficult to optimize** due to the double-sampling problem
    - plus, it has **identifiability** issues
    - though recent **positive** developments using conjugate form

What is the right objective for value estimation under nonlinear function approximation?

My Answer:

- The **Generalized $\overline{\text{PBE}}$**
  - which uses a more general projection on the Bellman Error
- With a potentially different weighting over states d in the objective
  - than the weighting d_{ideal} in the $\overline{\text{VE}}$

Outline

- Derive the Generalized $\overline{\text{PBE}}$
- Explain the role of the state-weighting in the objective
- Highlight two possible gradient estimates to optimize the Generalized $\overline{\text{PBE}}$
- [Maybe] Show positive empirical results for an algorithm using these insights

- Slides and working paper on website: marthawhite.ca
- Paper title: “Investigating Objectives for Off-policy Value Estimation in Reinforcement Learning”

Let’s start by deriving the Generalized $\overline{\text{PBE}}$

A Conjugate Form of the Bellman Error

- Beautiful result from Bo Dai and others: “Learning from Conditional Distributions via Dual Embeddings”
- Reformulate $\overline{\text{BE}}$ as a saddlepoint problem (min-max form)
  - Auxiliary variable h learned to estimate a part of the objective
  - Non-parametric approaches for h provide a close estimate for the $\overline{\text{BE}}$
- **Key Insight (for us):**
  - Now have some practical algorithms to (nearly) optimize the $\overline{\text{BE}}$

We build on this work to derive a generalized $\overline{\text{PBE}}$

- Let's understand the steps for the finite state case
- Some notation:
- $\hat{v}(s, \mathbf{w})$ is the parameterized value function, with function space $\mathcal{F}$
- $\delta(\mathbf{w}) = R + \gamma \hat{v}(S', \mathbf{w}) - \hat{v}(S, \mathbf{w})$ is the TD-error
- $\mathbb{E}_\pi[\delta(\mathbf{w}) \,|\, S = s] = T\hat{v}(\,\cdot\,, \mathbf{w})(s) - \hat{v}(S, \mathbf{w})$ for Bellman operator T
- Let $\mathcal{F}_{\text{all}}$ be the space of all functions

Deriving a Conjugate Form for the Bellman Error

$$\overline{\mathrm{BE}}(\boldsymbol{w}) = \sum_{s\in\mathcal{S}} d(s)\mathbb{E}_\pi[\delta(\boldsymbol{w}) \mid S=s]^2$$

$$y^2 = \max_{h\in\mathbb{R}} 2yh - h^2$$

$$= \sum_{s\in\mathcal{S}} d(s) \max_{h\in\mathbb{R}} \left(2\mathbb{E}_\pi[\delta(\boldsymbol{w}) \mid S=s]\, h - h^2\right)$$

$$= \max_{h\in\mathcal{F}_{\mathrm{all}}} \sum_{s\in\mathcal{S}} d(s) \left(2\mathbb{E}_\pi[\delta(\boldsymbol{w}) \mid S=s]\, h(s) - h(s)^2\right)$$

The function $h^*(s) = \mathbb{E}_\pi[\delta \mid S=s]$ provides the minimal error of zero.

Why is this useful?

- Computing a gradient update for the weights is now straightforward

$$h(s)(\nabla_{\boldsymbol{w}}\hat{v}(s,\boldsymbol{w}) - \gamma\nabla_{\boldsymbol{w}}\hat{v}(S',\boldsymbol{w}))$$

- h(s) needs to estimate $\mathbb{E}_\pi[\delta(\mathbf{w})\,|\,S=s]$
  - This estimator can be updated simultaneously with w

$$\left(2\mathbb{E}_\pi[\delta(\boldsymbol{w}) \mid S=s]\,h(s) - h(s)^2\right)$$

$$\delta(\mathbf{w}) = R + \gamma\hat{v}(S',\mathbf{w}) - \hat{v}(S,\mathbf{w})$$

Why is this useful?

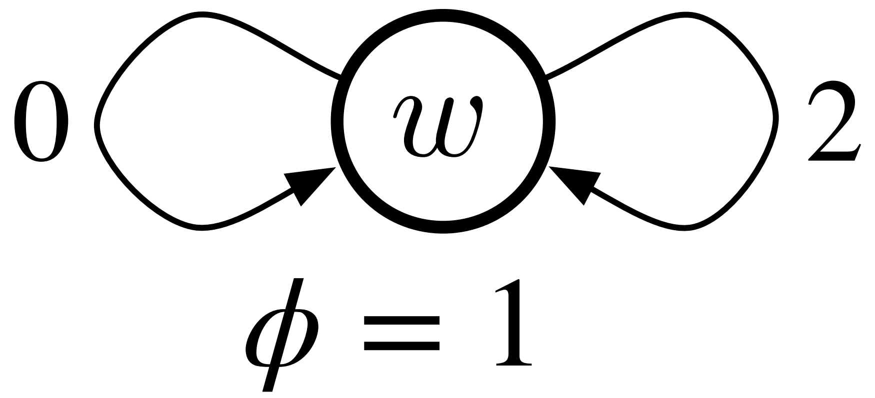
- Computing a gradient update for the weights is now straightforward

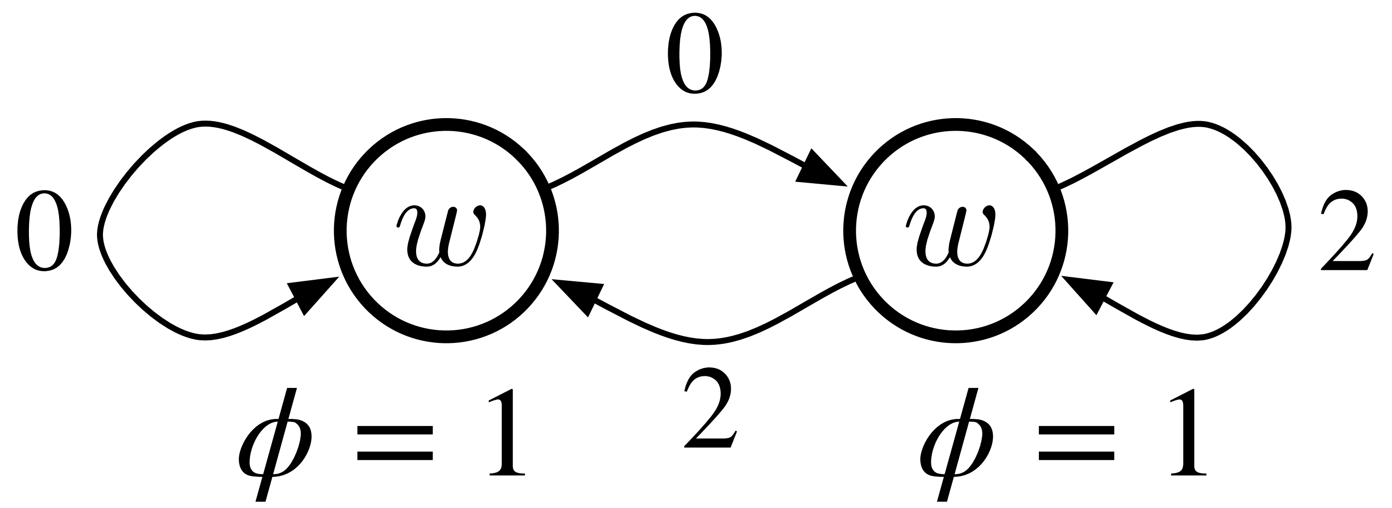
$$h(s)(\nabla_{\boldsymbol{w}} \hat{v}(s, \boldsymbol{w}) - \gamma \nabla_{\boldsymbol{w}} \hat{v}(S', \boldsymbol{w}))$$

- h(s) needs to estimate $\mathbb{E}_\pi[\delta(\mathbf{w}) \,|\, S = s]$
- But, wait! Isn't the $\overline{\text{BE}}$ non-identifiable (or non-learnable)?
  - This reformulation helps us solve that problem too

An Identifiable $\overline{\text{BE}}$

- The counterexample involves partial observability in the data





* from Sutton and Barto, 2018, Chapter 11.6

- **Issue**: $\overline{\text{BE}}$ defined on quantities not available in the data

An Identifiable $\overline{\text{BE}}$

- **Issue**: $\overline{\text{BE}}$ defined on quantities not available in the data
- **Solution**:

$$\mathcal{H}_{\text{all}} \stackrel{\text{def}}{=} \{h = f \circ \phi \mid \text{ where } f \text{ is any function on the space produced by } \phi\}.$$

$$\text{Identifiable } \overline{\text{BE}}(\boldsymbol{w}) \stackrel{\text{def}}{=} \max_{h \in \mathcal{H}_{\text{all}}} \mathbb{E}\left[2\mathbb{E}_\pi[\delta(\boldsymbol{w}) \mid S]\, h(S) - h(S)^2\right].$$

Restricting the Function Space for h Corresponds to a Projection on the Bellman Error

$$\Pi_{\mathcal{H},d} u = \arg\min_{h \in \mathcal{H}} \|u - h\|_d$$

$$\overline{\mathrm{PBE}}(\boldsymbol{w}) \stackrel{\text{def}}{=} \max_{h \in \mathcal{H}} \sum_{s \in \mathcal{S}} d(s) \left(2\mathbb{E}_\pi[\delta \mid S = s]\, h(s) - h(s)^2 \right)$$

...

$$= \|\Pi_{\mathcal{H},d}(\mathcal{T}\hat{v}(\cdot, \boldsymbol{w}) - \hat{v}(\cdot, \boldsymbol{w}))\|_d^2$$

$$\|v\|_d^2 = \sum_s d(s) v(s)^2$$

The Generalized $\overline{\text{PBE}}$

$$\overline{\text{PBE}}(\boldsymbol{w}) \stackrel{\text{def}}{=} \max_{h \in \mathcal{H}} \sum_{s \in \mathcal{S}} d(s) \left(2\mathbb{E}_\pi[\delta \mid S = s]\, h(s) - h(s)^2\right)$$

- For $\mathscr{H} = \mathscr{F} =$ a linear function space, this equals the linear $\overline{\text{PBE}}$
- For $\mathscr{H} = \mathscr{F} =$ a nonlinear function space, we get a natural extension of the linear $\overline{\text{PBE}}$ to the nonlinear setting
- For $\mathscr{H} = \mathscr{H}_{\text{all}}$, this equals the Identifiable $\overline{\text{BE}}$
- For $\mathscr{F} \subset \mathscr{H} \subset \mathscr{H}_{\text{all}}$, this provides a new Projected Bellman Error

Let’s move now to the Role of the Weighting in the Generalized $\overline{\text{PBE}}$

Upper Bound on the Value Error

Theorem 1 *If* $\mathcal{H} \supseteq \mathcal{F}$, *then the solution* $v_{\boldsymbol{w}_{\mathcal{H},d}}$ *to the generalized* $\overline{PBE}$ *satisfies*

$$\|v_\pi - v_{\boldsymbol{w}_{\mathcal{H},d}}\|_d \leq \|\Pi_{\mathcal{F},H}\|_d \|v_\pi - \Pi_{\mathcal{F},d} v_\pi\|_d.$$

H is a (non-diagonal) matrix, where the projection to $\mathscr{F}$ is weighted by H

Impact of the Weighting

- Kolter's counterexample a two-state MDP with small approximation error
- Shows that with d corresponding to off-policy stationary distribution d_b, the solution to the linear $\overline{\text{PBE}}$ can have arbitrarily bad $\overline{\text{VE}}$
- Using an emphatic weighting for d prevents this, and gives

$$\|v_\pi - v_{\boldsymbol{w}_{\mathcal{H},d}}\|_d \leq C(P_\pi, \gamma, d) \|v_\pi - \Pi_{\mathcal{F},d} v_\pi\|_d.$$

$$\|v_\pi - v_{\boldsymbol{w}_{\mathcal{H},d}}\|_{d_b} \leq C(P_\pi, \gamma, d, d_b) \|v_\pi - \Pi_{\mathcal{F},d} v_\pi\|_d.$$

- for some constants dependent on the problem

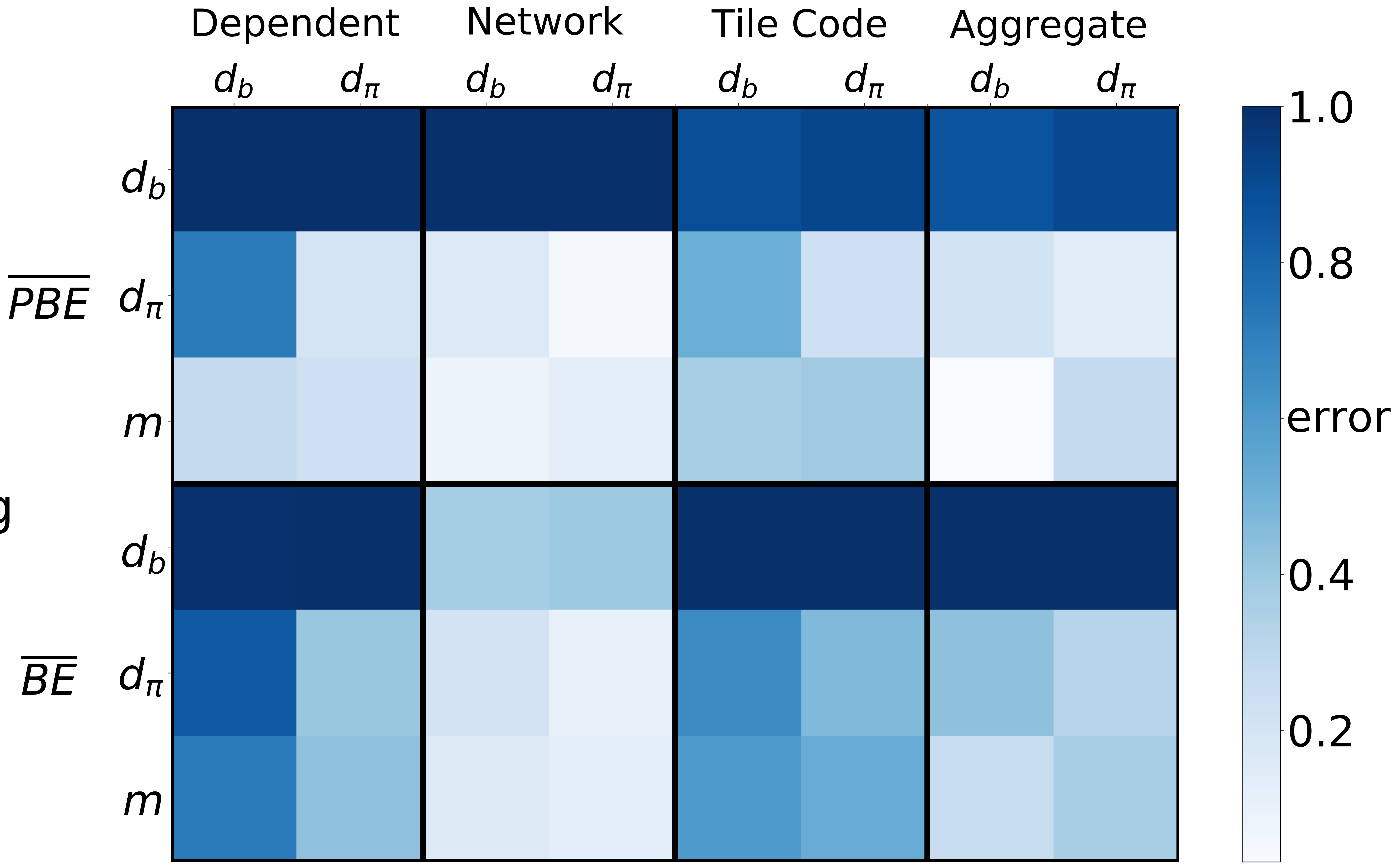
Empirical Results for Solution Quality

Key Conclusion:

Weighting with d_b bad

Weighting with m good

…even when measuring performance with d_b



The final step to obtaining a practical algorithm using the generalized $\overline{\text{PBE}}$:
Reducing reliance on our estimate h

Sampling the Gradient

- The saddlepoint update

$$\Delta \boldsymbol{w} \leftarrow h(s)(\nabla_{\boldsymbol{w}} \hat{v}(s, \boldsymbol{w}) - \gamma \nabla_{\boldsymbol{w}} \hat{v}(S', \boldsymbol{w}))$$

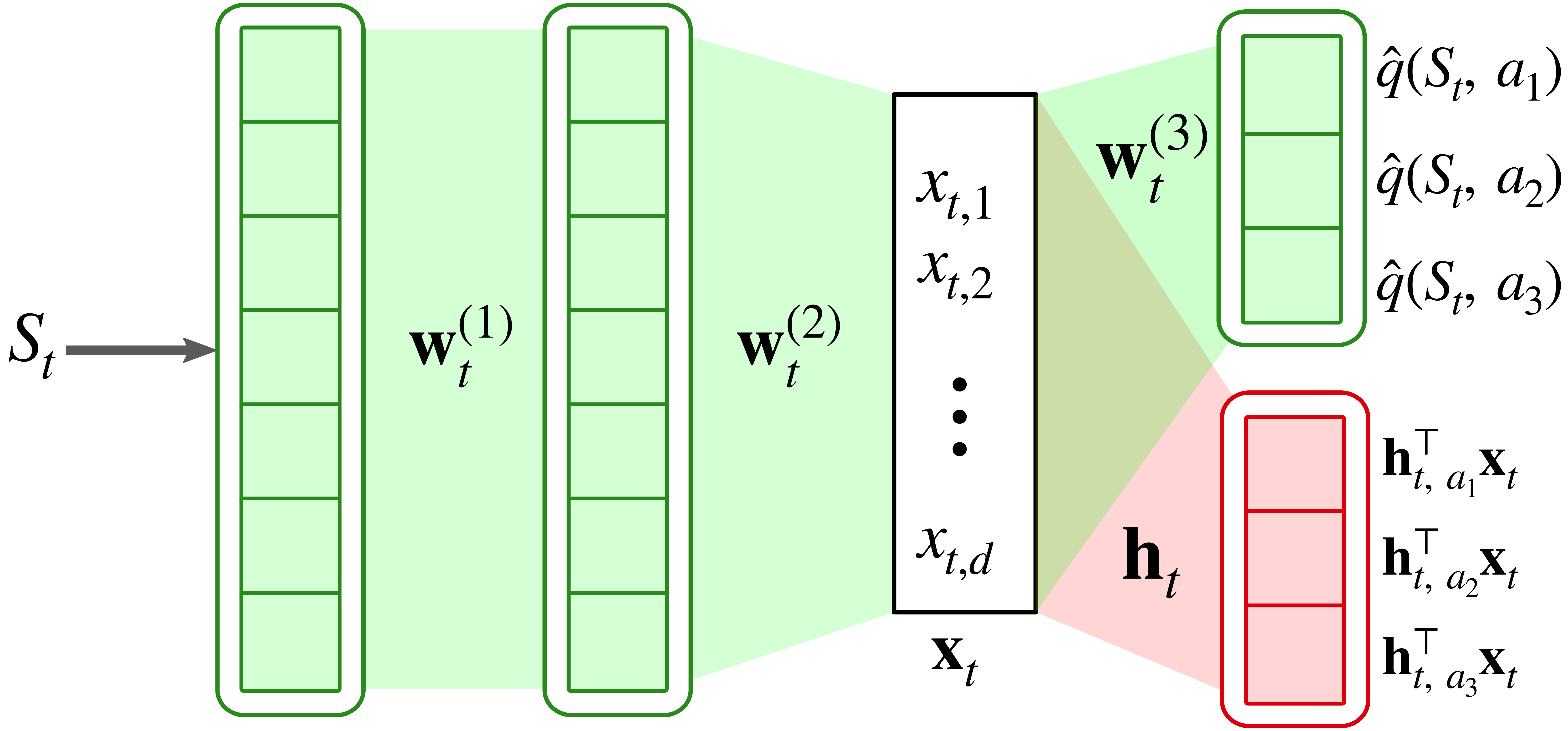
- The gradient-correction update

$$\Delta \boldsymbol{w} \leftarrow \delta(\boldsymbol{w}) \nabla_{\boldsymbol{w}} v(s, \boldsymbol{w}) - h(s)\gamma \nabla_{\boldsymbol{w}} v(S', \boldsymbol{w})$$

- To make it appropriate to use gradient-correction, analysis suggests h should be learned using the gradient of v as the features
    - the gradient vector includes the last layer of the neural network

QC and QRC (Q-learning with Corrections)

- Add head to a neural network to estimate h (gradients not passed back)



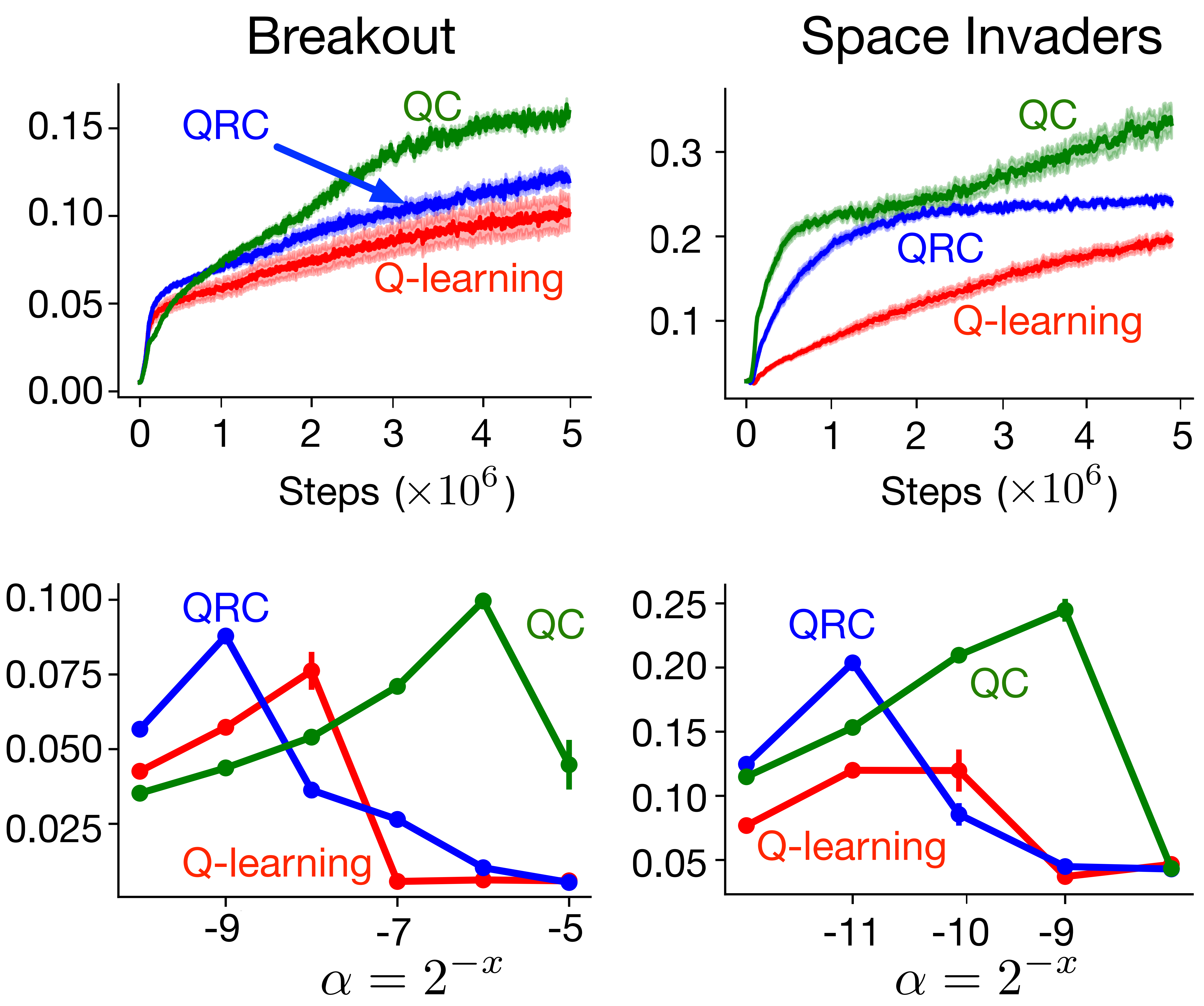
* our paper: “Gradient Temporal-Difference Learning with Regularized Corrections”, ICML, 2020

Control Results (in MinAtar)

Key Conclusion:

Gradient methods help!

Both QC and QRC



Summary of the Talk

- **Point 1:** The Generalized $\overline{\text{PBE}}$ is the natural extension of the linear $\overline{\text{PBE}}$ to the nonlinear setting

- **Point 2:** The Generalized $\overline{\text{PBE}}$ help resolve questions about the $\overline{\text{BE}}$

    - both about identifiability and connection to $\overline{\text{PBE}}$

- **Point 3:** The role of weighting should not be overlooked in the objective

Thank you! Questions?