Why Should I Trust You, Bellman? The Bellman Error is a Poor Replacement for Value Error

Scott Fujimoto, David Meger, Doina Precup, Ofir Nachum, Shixiang Shane Gu

May 19, 2025

Introduction

In common RL settings, the Bellman error is used as a proxy for the value error. Prominent use cases:

- Approximate Dynamic Programming and Reinforcement Learning: Training a neural network to approximate the value function.

However, this paper argues that the Bellman error is a poor replacement for the value error.

Fix an approximate Q-function $\hat{q} : \mathsf{X} \times \mathsf{A} \to \mathbb{R}$ of a $bm(\mathsf{X} \times \mathsf{A})$ map q, and an action $a' = \sigma(x')$.

Fix a policy $\sigma \in \Sigma$, we want to find $\hat{q}$ such that:

$$
\begin{aligned}
\hat{q}(x, a) &= r(x, a) + \beta \int_{x' \in \mathsf{X}} \hat{q}(x', \sigma(x')) P(x, a, dx') \\
&= r(x, a) + \beta \mathbb{E}_{x'} \left[\hat{q}(x', \sigma(x')) \right] \\
&= T_\sigma \hat{q}(x, a)
\end{aligned}
\tag{1}
$$

Bellman Error vs Value Error

To achieve that, we can define the Bellman error and value error as follows:

- Bellman error:

$$\begin{aligned} \varepsilon(x, a) &:= \hat{q}(x, a) - \left\{ r(x, a) + \beta \int_{x' \in \mathsf{X}} \hat{q}(x', \sigma(x')) P(x, a, dx') \right\} \\ &= \hat{q}(x, a) - T_\sigma \hat{q}(x, a) \end{aligned} \tag{2}$$

  which is the difference between the two sides of the Bellman equation.

- Value error:

$$\delta(x, a) := \hat{q}(x, a) - q(x, a) \tag{3}$$

  where q is the Q-function.

To simplify notation, in the following slides, we shorten $a' := \sigma(x')$.

Bellman Error vs Value Error

Equivalently, we can use value function instead of Q-function.

Fix an approximate value function $\hat{v} : \mathsf{X} \to \mathbb{R}$, we can define the Bellman error and value error as follows:

$$\varepsilon(x) \coloneqq \hat{v}_\sigma(x) - \max_{a \in \mathsf{A}} \left[r(x, a) + \beta \mathbb{E}_{x'} v_\sigma(x') \right] \tag{4}$$

and

$$\delta(x) \coloneqq \hat{v}_\sigma(x) - v_\sigma(x) \tag{5}$$

where $\hat{v}$ is the approximate value function and v_σ is the true σ-value function.

In most RL settings, function $r(x, a)$ is unknown, so we can replace it with a sample r.

Issues with the Bellman error:

- The magnitude of the Bellman error hides bias (i.e., the metric is only a weak proxy for the value error).
- In the "finite data regime", the Bellman error can be minimized by infinitely many suboptimal solutions.

Theoretically, it works

Let $d_\sigma(x', a' \mid x, a)$ be conditional distribution of (x', a') given (x, a) following policy σ. Then

Theorem 1: Value Error and Bellman Error

For any $(x, a) \in \mathsf{X} \times \mathsf{A}$,

$$\delta(x, a) = \frac{1}{1-\beta} \; \mathbb{E}_{x', a' \sim d_\sigma(x', a' \mid x, a)} \left[\varepsilon(x', a')\right] \tag{6}$$

Proposition 1: Guarantee over $\mathsf{X} \times \mathsf{A}$

$$\varepsilon(x, a) = 0 \quad \forall (x, a) \in \mathsf{X} \times \mathsf{A} \iff \delta(x, a) = 0 \quad \forall (x, a) \in \mathsf{X} \times \mathsf{A}$$

Proof.

First by definition:

$$\delta(x,a) \coloneqq \hat{q}(x,a) - q(x,a)$$
$$\Rightarrow q(x,a) = \hat{q}(x,a) - \delta(x,a)$$

Then we can decompose value error:

$$\begin{aligned}
\delta(x,a) &= \hat{q}(x,a) - q(x,a) \\
&= \hat{q}(x,a) - (r(x,a) + \beta\mathbb{E}[q(x',a')]) \\
&= \hat{q}(x,a) - (r(x,a) + \beta\mathbb{E}[\hat{q}(x',a') - \delta(x',a')]) \\
&= \hat{q}(x,a) - (r(x,a) + \beta\mathbb{E}[\hat{q}(x',a')]) + \beta\mathbb{E}[\delta(x',a')] \\
&= \varepsilon(x,a) + \beta\mathbb{E}[\delta(x',a')]
\end{aligned}$$

By treating $\delta(x,a)$ as a value function and $\varepsilon(x',a')$ as the reward, we can see that:

$$\delta(x,a) = \frac{1}{1-\beta}\mathbb{E}_{(x',a')\sim d_\sigma(\cdot|x,a)}[\varepsilon(x',a')] \tag{7}$$

□

Proof of Proposition 1

Proof.

$\Longrightarrow$ is trivial from Theorem **??**.

$\Longleftarrow$:

$$\delta(x,a) = 0 \quad \forall (x,a) \in \mathsf{X} \times \mathsf{A} \Rightarrow \hat{q}(x,a) = q(x,a) \quad \forall (x,a) \in \mathsf{X} \times \mathsf{A}$$

From the Bellman equation, it follows that:

$$\begin{aligned} \hat{q}(x,a) &= q(x,a) \\ &= r(x,a) + \beta \mathbb{E}[q(x',a')] =: Tq \\ &= r(x,a) + \beta \mathbb{E}[\hat{q}(x',a')] =: T\hat{q} \end{aligned}$$

Therefore:

$$\begin{aligned} \varepsilon(x,a) &= \hat{q}(x,a) - \{r(x,a) + \beta \mathbb{E}[\hat{q}(x',a')]\} \\ &= \hat{q}(x,a) - \hat{q}(x,a) \\ &= 0 \end{aligned}$$

□

In reality, it doesn't work

- The Bellman error is a poor proxy for the value error.
- In the finite data regime, the Bellman error can be minimized by infinitely many suboptimal solutions.

To show this, first we observe that

$$\varepsilon(x, a) = \delta(x, a) - \beta \mathbb{E}_{x'}[\delta(x', a')] \tag{8}$$

This is easy to verify since

$$\begin{aligned}
\varepsilon(x, a) &= \hat{q}(x, a) - T_\sigma \hat{q}(x, a) \\
&= \hat{q}(x, a) - \left\{ r(x, a) + \beta \mathbb{E}_{x'}[\hat{q}(x', a')] \right\} \\
&= \hat{q}(x, a) - \left\{ r(x, a) + \beta \mathbb{E}_{x'}[q(x', a')] + \beta \mathbb{E}_{x'}[\delta(x', a')] \right\} \\
&= \delta(x, a) - \beta \mathbb{E}_{x'}[\delta(x', a')]
\end{aligned}$$

We shorthand $\mathbb{E}_{x'}$ as $\mathbb{E}$ in the following slides.

Problem 1: The Magnitude of Bellman Error Hides Bias

- The Bellman error is not guaranteed to correspond with the magnitude of the value error.
- Fundamental issue: cancellation between error terms.

Example 1: Same Value Error, Different Bellman Error

Let q be the true value function for some MDP and policy σ.

We define two approximate value functions:

$$\hat{q}_1 = q + 1 \quad \text{(correlated error)}$$

$$\hat{q}_2 = q \pm 1 \quad \text{(uncorrelated error, randomly + or -)}$$

Here, ± 1 is a random variable that is either 1 or -1 with equal probability.

In both cases, the absolute value error $|\delta|$ is 1 for all state-action pairs.

- For $\hat{q}_1$:

$$\begin{aligned}\varepsilon_1(x,a) &= \delta_1(x,a) - \beta\mathbb{E}[\delta_1(x',a')]\\ &= 1 - \beta\mathbb{E}[1]\\ &= 1 - \beta\cdot 1 = 1-\beta\end{aligned}$$

- For $\hat{q}_2$:

$$\begin{aligned}\varepsilon_2(x,a) &= \delta_2(x,a) - \beta\mathbb{E}[\delta_2(x',a')]\\ &= \pm 1 - \beta\mathbb{E}[\pm 1]\\ &= \pm 1 - \beta\cdot 0 = \pm 1\end{aligned}$$

So $|\varepsilon_2| = 1 > |\varepsilon_1| = 1-\beta$

Example 2: Same Bellman Error, Different Value Error

Define two more approximate value functions:

$$\hat{q}_1 = q + \frac{1}{1-\beta}$$

$$\hat{q}_2 = q \pm 1$$

For $\hat{q}_1$:

$$\begin{aligned}\varepsilon_1(x,a) &= \delta_1(x,a) - \beta\mathbb{E}[\delta_1(x',a')] \\ &= \frac{1}{1-\beta} - \beta\mathbb{E}[\frac{1}{1-\beta}] \\ &= \frac{1}{1-\beta} - \beta\cdot\frac{1}{1-\beta} \\ &= \frac{1}{1-\beta}\cdot(1-\beta) = 1\end{aligned}$$

For $\hat{q}_2$ (as before): $|\varepsilon_2| = 1$

But the value errors are vastly different:

$$|\delta_1| = \frac{1}{1-\beta} \quad \text{vs.} \quad |\delta_2| = 1$$

Inverse Relationship Between Bellman Error and Value Error

Proposition 2: Inverse relationship

For any MDP, discount factor $\beta \in (0, 1)$, and $C > 0$, we can define a q-function $\hat{q}_1$ and a stochastic q-function $\hat{q}_2$ such that for any state-action pair $(x, a) \in \mathsf{X} \times \mathsf{A}$:

$$1. |\delta_1(x, a)| - |\delta_2(x, a)| > C$$

$$2. \mathbb{E}_{\hat{q}_2}[|\varepsilon_2(x, a)|] - |\varepsilon_1(x, a)| > C$$

- This means that lower absolute Bellman error over all state-action pairs does not guarantee lower value error
- In fact, value functions with higher Bellman error might have better approximation of the true value function

Issue 2: Bellman Error can be Minimized by Infinitely Many Suboptimal Solutions

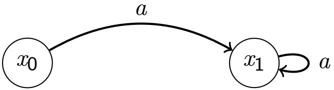
- Theorem 1 implies uniqueness of the Bellman equation over complete transition data
- But in practice, we only have finite datasets with missing transitions
- Consequence: Infinitely many suboptimal solutions can satisfy the Bellman equation

Proposition 3: Non-uniqueness of Bellman equation with incomplete data

If there exists a state-action pair (x', a') not in dataset D, where $d_\sigma(x', a'|x, a) > 0$ for some $(x, a) \in D$, then there exists a value function and $C > 0$ such that:

1. For all $(\hat{x}, \hat{a}) \in D$, the Bellman error $\varepsilon(\hat{x}, \hat{a}) = 0$
2. There exists $(x, a) \in D$, such that the value error $\delta(x, a) = C$

Two-State MDP Example



- Simple two-state MDP with reward $r(x, a) = 0 \quad \forall (x, a) \in \mathsf{X} \times \mathsf{A}$.
- True value function is $q(x, a) = 0$ for all (x, a)
- Assume dataset contains only one transition: $(x_0, a, r = 0, x_1)$
- The transition $(x_1, a, r = 0, x_1)$ is missing from our dataset

Example 1: Zero Bellman Error, Large Value Error

We can construct a value function with $C \in \mathbb{R}$ such that:

$$\hat{q}(x_0, a) = C$$
$$\hat{q}(x_1, a) = \frac{C}{\beta}$$

This gives:

$$\begin{aligned}\varepsilon(x_0, a) &= \hat{q}(x_0, a) - \{r(x_0, a) + \beta \hat{q}(x_1, a)\} \\ &= C - \{0 + \beta \cdot \frac{C}{\beta}\} \\ &= C - C = 0\end{aligned}$$

But the value error is:

$$\begin{aligned}\delta(x_0, a) &= \hat{q}(x_0, a) - q(x_0, a) \\ &= C - 0 = C\end{aligned}$$

Zero Bellman error despite arbitrarily large value error!

If the the transition is not missing, then we can derive from the self-transition of s_1 (with $r = 0$):

$$\hat{q}(s_1, a) = 0 + \beta\, \hat{q}(s_1, a) \implies (1 - \beta)\, \hat{q}(s_1, a) = 0 \implies \hat{q}(s_1, a) = 0.$$

Hence

$$\hat{q}(s_0, a) = 0 + \beta\, \hat{q}(s_1, a) = \beta \cdot 0 = 0,$$

and hence both $\delta(s_0, a)$ and $\delta(s_1, a)$ are zero.

Example 2: Large Bellman Error, Zero Value Error

Conversely, we can construct:

$$\hat{q}(x_0, a) = 0$$
$$\hat{q}(x_1, a) = -\frac{C}{\beta}$$

This gives:

$$\begin{aligned}\varepsilon(x_0, a) &= \hat{q}(x_0, a) - \{r(x_0, a) + \beta \hat{q}(x_1, a)\}\\ &= 0 - \{0 + \beta \cdot (-\frac{C}{\beta})\}\\ &= 0 - (-C) = C\end{aligned}$$

But the value error is:

$$\begin{aligned}\delta(x_0, a) &= \hat{q}(x_0, a) - q(x_0, a)\\ &= 0 - 0 = 0\end{aligned}$$

Perfect value function despite arbitrarily large Bellman error!

Implications for RL Algorithms

- Methods that minimize Bellman error over finite datasets may:
  - Converge to solutions with zero Bellman error but large value error
  - Reject solutions with near-optimal value functions due to Bellman error
- Missing transitions are inevitable in complex domains.
- Cannot rely on Bellman error alone as convergence criteria.
- This problem is structural, not just an implementation issue.