Bellman Unbiasedness: Toward Provably Efficient Distributional Reinforcement Learning with General Value Function Approximation

Taehyun Cho[1] Seungyub Han[1] Seokhun Ju[1]
Dohyeong Kim[1] Kyungjae Lee[2] Jungwoo Lee[1]

[1]Seoul National University, [2]Korea University

ICML
June 14, 2025

Seoul National University
Cognitive Machine Learning Lab.

KOREA UNIVERSITY

Motivation & Challenges

Why Distributional RL?

- **Distributional RL (DistRL)** models the entire distribution of returns, not just the expectation.

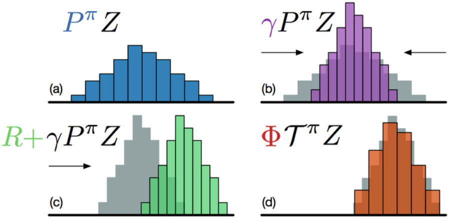


Figure: Distributional Bellman Update

Bellemare, Marc G., Will Dabney, and Rémi Munos. "A distributional perspective on reinforcement learning." International conference on machine learning. PMLR, 2017.

Why Distributional RL?

- **Distributional RL (DistRL)** models the entire distribution of returns, not just the expectation.
- Offers richer insight into **uncertainty**, such as variance, skewness, and quantiles.

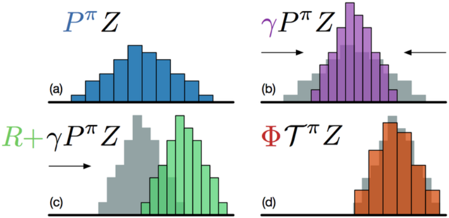


Figure: Distributional Bellman Update

Bellemare, Marc G., Will Dabney, and Rémi Munos. "A distributional perspective on reinforcement learning." International conference on machine learning. PMLR, 2017.

Why Distributional RL?

- **Distributional RL (DistRL)** models the entire distribution of returns, not just the expectation.
- Offers richer insight into **uncertainty**, such as variance, skewness, and quantiles.
- Facilitates safer and more effective decision-making by explicitly considering **risk**.

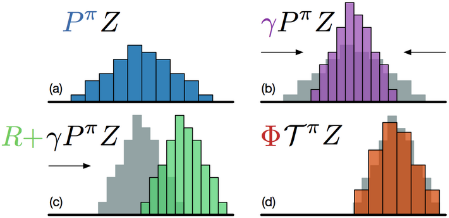


Figure: Distributional Bellman Update

Bellemare, Marc G., Will Dabney, and Rémi Munos. "A distributional perspective on reinforcement learning." International conference on machine learning. PMLR, 2017.

Two challenges in distRL

1. **Infinite-dimensionality**

2. **Online distributional update**

Two challenges in distRL

1. **Infinite-dimensionality**
   - Return distributions contain an infinite amount of information.
   - We must approximate it using a finite number of parameters or statistical functionals.
   - However, not all statistical functionals can be exactly learned through the Bellman operator, as the meaning is not preserved.
2. **Online distributional update**

Two challenges in distRL

1. **Infinite-dimensionality**
   - Return distributions contain an infinite amount of information.
   - We must approximate it using a finite number of parameters or statistical functionals.
   - However, not all statistical functionals can be exactly learned through the Bellman operator, as the meaning is not preserved.
2. **Online distributional update**
   - Decoupling the policy update and the distribution estimation via additional rollouts is sample-inefficient.
   - Limited rollouts inevitably introduce approximation errors into the estimated distribution.

Two challenges in distRL

1. **Infinite-dimensionality**
   - Return distributions contain an infinite amount of information.
   - We must approximate it using a finite number of parameters or statistical functionals.
   - However, not all statistical functionals can be exactly learned through the Bellman operator, as the meaning is not preserved.
2. **Online distributional update**
   - Decoupling the policy update and the distribution estimation via additional rollouts is sample-inefficient.
   - Limited rollouts inevitably introduce approximation errors into the estimated distribution.

"Can we design a representation that is both exactly learned and provably sample-efficient?"

Backgrounds

1. Bellman Closedness
2. Bellman Unbiasedness
3. Statistical Functional Bellman Completeness (SFBC)
4. Statistical Functional Least Square Value Iteration (SF-LSVI)

Statistical Functional, Sketch

Statistical Functional, Sketch; [Bellemare, 2023]

A **statistical functional** is a mapping from a probability distribution to a real value $\psi : \mathscr{P}(\mathbb{R}) \to \mathbb{R}$. A **sketch** is a vector-valued function $\psi_{1:N} : \mathscr{P}(\mathbb{R}) \to \mathbb{R}^N$ specified by an N-tuple where each component is a statistical functional,

$$\psi_{1:N}(\cdot) = (\psi_1(\cdot), \cdots, \psi_N(\cdot)).$$

Bellman Closedness (BC)

Bellman Closedness; [Rowland, 2019]

A sketch $\psi_{1:N}$ is **Bellman closed** if there exists an operator $\mathcal{T}_{\psi_{1:N}} : I_{\psi_{1:N}}^{\mathcal{S}} \to I_{\psi_{1:N}}^{\mathcal{S}}$ such that

$$\psi_{1:N}(\mathcal{T}\bar{\eta}) = \mathcal{T}_{\psi_{1:N}}\psi_{1:N}(\bar{\eta}) \quad \text{for all } \bar{\eta} \in \mathscr{P}(\mathbb{R})^{\mathcal{S}}$$

which is closed under a distributional Bellman operator $\mathcal{T} : \mathscr{P}(\mathbb{R})^{\mathcal{S}} \to \mathscr{P}(\mathbb{R})^{\mathcal{S}}$.

Bellman Closedness (BC)

Theorem ([Rowland, 2019])

The only finite sets of statistics of the form $\psi(\bar{\eta}) = \mathbb{E}_{Z\sim\bar{\eta}}[h(Z)]$ that are Bellman closed are given by the collections of $\psi_1, \dots, \psi_N$ where its linear span $\{\sum_{n=0}^{N} \alpha_n \psi_n | \ \alpha_n \in \mathbb{R}\ , \forall N\}$ is equal to the set of exponential polynomial functionals $\{\eta \to \mathbb{E}_{Z\sim\eta}[Z^l \exp(\lambda Z)] | \ l = 0, 1, \dots, L, \lambda \in \mathbb{R}\}$, where ψ_0 is the constant functional equal to 1.
In discount setting, it is equal to the linear span of the set of moment functionals $\{\eta \to \mathbb{E}_{Z\sim\eta}[Z^l] | \ l = 0, 1, \dots, L\}$ for some $L \leq N$.

Bellman Closedness (BC)

Theorem ([Rowland, 2019])

The only finite sets of statistics of the form $\psi(\bar{\eta}) = \mathbb{E}_{Z\sim\bar{\eta}}[h(Z)]$ that are Bellman closed are given by the collections of $\psi_1, \ldots, \psi_N$ where its linear span $\{\sum_{n=0}^{N} \alpha_n \psi_n | \ \alpha_n \in \mathbb{R}\ , \forall N\}$ is equal to the set of exponential polynomial functionals $\{\eta \to \mathbb{E}_{Z\sim\eta}[Z^l \exp(\lambda Z)] | \ l = 0, 1, \ldots, L, \lambda \in \mathbb{R}\}$, where ψ_0 is the constant functional equal to 1.
In discount setting, it is equal to the linear span of the set of moment functionals $\{\eta \to \mathbb{E}_{Z\sim\eta}[Z^l] | \ l = 0, 1, \ldots, L\}$ for some $L \leq N$.

Although both the first and second moments are Bellman closed, the variance is **nonlinear**.
As a result, its Bellman closedness cannot be determined by the existing theory, which only applies to **linear** statistical functionals.

Bellman Closedness (BC)

Theorem

Quantile functional cannot be Bellman closed under any additional sketch.

Along with a technical clarification of the proof in [Rowland, 2019], we provide an improved version of the proof.

Bellman Closedness (BC)

Theorem

Quantile functional cannot be Bellman closed under any additional sketch.

Along with a technical clarification of the proof in [Rowland, 2019], we provide an improved version of the proof.

Theorem

Maximum and minimum functional are nonlinear and Bellman closed.

- $\mathcal{T}_{\psi_{\max}}\Big(\psi_{\max}(\bar{\eta}(s))\Big) = \max_{s' \sim \mathbb{P}(\cdot|s,a)}\Big(r + \psi_{\max}\big(\bar{\eta}(s')\big)\Big)$.
- $\mathcal{T}_{\psi_{\min}}\Big(\psi_{\min}(\bar{\eta}(s))\Big) = \min_{s' \sim \mathbb{P}(\cdot|s,a)}\Big(r + \psi_{\min}\big(\bar{\eta}(s')\big)\Big)$.

Key Concepts

1. Bellman Closedness
2. Bellman Unbiasedness
3. Statistical Functional Bellman Completeness (SFBC)
4. Statistical Functional Least Square Value Iteration (SF-LSVI)

Bellman Unbiasedness (BU)

$$\begin{tikzcd}
\eta \arrow[r] \arrow[d] & \mathcal{T}\eta \arrow[r] & \psi(\mathcal{T}\eta) \\
\psi(\eta) \arrow[r] & \mathcal{T}_\psi\psi(\eta) \arrow[ur] &
\end{tikzcd}$$

Figure 3. Bellman Closedness

$$\begin{tikzcd}
\eta \arrow[r] \arrow[d] & \mathcal{T}\eta \arrow[r] & \psi(\mathcal{T}\eta) \\
\psi(\eta) \arrow[r, dashed, "\text{Sample next state}"'] & \hat{\mathcal{T}}_\psi\psi(\eta) \arrow[r, dashed, "\text{Unbiasedly estimate}"'] & \mathbb{E}_{\mathbb{P}}[\phi_\psi \circ \hat{\mathcal{T}}_\psi\psi(\eta)] \arrow[u]
\end{tikzcd}$$

Figure 4. Bellman Unbiasedness

$$
\begin{tikzcd}
\eta \arrow[r] \arrow[d] & \mathcal{T}\eta \arrow[r] & \psi(\mathcal{T}\eta) \\
\psi(\eta) \arrow[r] & \mathcal{T}_\psi\psi(\eta) \arrow[ur] &
\end{tikzcd}
$$

Figure 3. Bellman Closedness

$$
\begin{tikzcd}
\eta \arrow[r] \arrow[d] & \mathcal{T}\eta \arrow[r] & \psi(\mathcal{T}\eta) \\
\psi(\eta) \arrow[r, dashed, "\text{Sample next state}"'] & \hat{\mathcal{T}}_\psi\psi(\eta) \arrow[r, dashed, "\text{Unbiasedly estimate}"'] & \mathbb{E}_{\mathbb{P}}[\phi_\psi \circ \hat{\mathcal{T}}_\psi\psi(\eta)] \arrow[u]
\end{tikzcd}
$$

Figure 4. Bellman Unbiasedness

Bellman Closedness

- Exact learnability
- Exact update in finite dimensional space.

Bellman Unbiasedness (BU)

$$\begin{tikzcd}
\eta \arrow[r] \arrow[d] & \mathcal{T}\eta \arrow[r] & \psi(\mathcal{T}\eta) \\
\psi(\eta) \arrow[r] & \mathcal{T}_\psi \psi(\eta) \arrow[ur] &
\end{tikzcd}$$

Figure 3. Bellman Closedness

$$\begin{tikzcd}
\eta \arrow[r] \arrow[d] & \mathcal{T}\eta \arrow[r] & \psi(\mathcal{T}\eta) \\
\psi(\eta) \arrow[r, dashed, "\text{Sample}"', "\text{next state}"' {yshift=-12pt}] & \hat{\mathcal{T}}_\psi \psi(\eta) \arrow[r, dashed, "\text{Unbiasedly}"', "\text{estimate}"' {yshift=-12pt}] & \mathbb{E}_{\mathbb{P}}[\phi_\psi \circ \hat{\mathcal{T}}_\psi \psi(\eta)] \arrow[u]
\end{tikzcd}$$

Figure 4. Bellman Unbiasedness

Bellman Closedness

- Exact learnability
- Exact update in finite dimensional space.

Bellman Unbiasedness

- Provable efficiency
- Unbiased update using sampled sketches

Bellman Unbiasedness (BU)

Bellman Unbiasedness

A sketch $\psi(=\psi_{1:N})$ is **Bellman unbiased** if a vector-valued estimator $\phi_\psi = \phi_\psi(\psi(\cdot), \cdots, \psi(\cdot)) : (I_\psi^{\mathcal{S}})^k \to I_\psi^{\mathcal{S}}$ exists where the sketch of expected distribution can be unbiasedly estimated by ϕ_ψ using the k sampled sketches from the sample distribution , i.e.,

$$\mathbb{E}_{s_i' \sim \mathbb{P}}\left[\phi_\psi\left(\underbrace{\psi\Big((\mathcal{B}_r)_\# \bar{\eta}(s_1')\Big), \cdots, \psi\Big((\mathcal{B}_r)_\# \bar{\eta}(s_k')\Big)}_{k \text{ sampled sketches from sample distribution } \hat{\mathcal{T}}_\psi \psi(\bar{\eta}(s))} \right)\right] = \psi\Big((\mathcal{B}_r)_\# \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)}[\bar{\eta}(s')]\Big).$$

Example) Mean-variance sketch

$$\begin{aligned}(\bar{\mu}, \bar{\sigma}^2) &= \phi_{(\mu,\sigma^2)}\Big((\hat{\mu}_1, \hat{\sigma}_1^2), \cdots, (\hat{\mu}_k, \hat{\sigma}_k^2)\Big) \\ &= \Big(\frac{1}{k}\sum_{i=1}^{k} \hat{\mu}_i,\ \frac{1}{k}\sum_{i=1}^{k}(\hat{\mu}_i - \frac{1}{k}\sum_{j=1}^{k}\hat{\mu}_j)^2 + \hat{\sigma}_i^2\Big)\end{aligned}$$

Bellman Unbiasedness (BU)

Lemma

*Let $F_{\bar{\eta}}$ be a CDF of the probability distribution $\bar{\eta} \in \mathscr{P}_\psi(\mathbb{R})^{\mathcal{S}}$. Then a sketch is Bellman unbiased if and only if the sketch is **homogeneous** over $\mathscr{P}_\psi(\mathbb{R})^{\mathcal{S}}$ of degree k, i.e., there exists some vector-valued function $h = h(x_1, \cdots, x_k) : \mathcal{X}^k \to \mathbb{R}^N$ such that*

$$\psi(\bar{\eta}) = \int \cdots \int h(x_1, \cdots, x_k) dF_{\bar{\eta}}(x_1) \cdots dF_{\bar{\eta}}(x_k).$$

Example) Variance is nonlinear but homogeneous of degree 2.

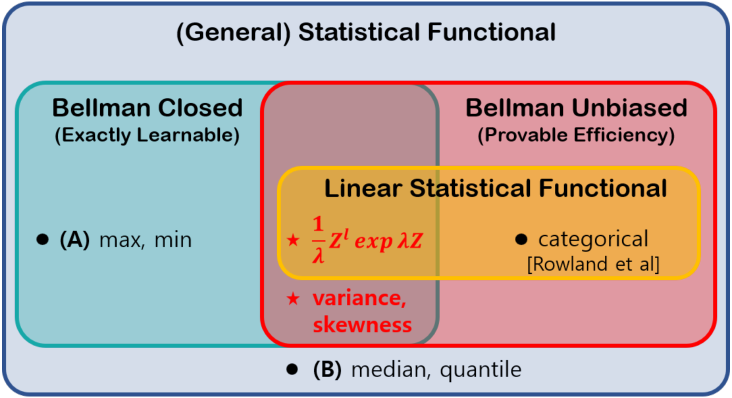
$$\begin{aligned}\mathrm{Var}(\bar{\eta}) &= \mathbb{E}_{Z_1 \sim \bar{\eta}}[(Z_1 - \mathbb{E}_{Z_2 \sim \bar{\eta}}[Z_2])^2] \\ &= \mathbb{E}_{Z_1, Z_2 \sim \bar{\eta}}[Z_1^2 - 2Z_1 Z_2 + Z_2^2] = \mathbb{E}_{Z_1, Z_2 \sim \bar{\eta}}[h(Z_1, Z_2)]\end{aligned}$$

Bellman Unbiasedness (BU)

Theorem

The only finite statistical functionals that are both Bellman unbiased and closed are given by the collections of $\psi_1, \ldots, \psi_N$ where its linear span $\{\sum_{n=0}^{N} \alpha_n \psi_n | \ \alpha_n \in \mathbb{R} \ , \forall N\}$ is equal to the set of exponential polynomial functionals $\{\eta \to \mathbb{E}_{Z \sim \eta}[Z^l \exp(\lambda Z)] | \ l = 0, 1, \ldots, L, \lambda \in \mathbb{R}\}$, where ψ_0 is the constant functional equal to 1.
In discount setting, it is equal to the linear span of the set of moment functionals $\{\eta \to \mathbb{E}_{Z \sim \eta}[Z^l] | \ l = 0, 1, \ldots, L\}$ for some $L \leq N$.

Venn-Diagram of Statistical Functional Classes



Assumption & Algorithm

1. Bellman Closedness
2. Bellman Unbiasedness
3. Statistical Functional Bellman Completeness (SFBC)
4. Statistical Functional Least Square Value Iteration (SF-LSVI)

Statistical Functional Bellman Completeness

Distributional Bellman Completeness (DistBC)

For any distribution $\bar{\eta} : \mathcal{S} \to \mathscr{P}([0, H])$ and $h \in [H]$, there exists $f_{\bar{\eta}} \in \mathcal{H}$ ($\subseteq \mathcal{F}^{\infty}$) which satisfies

$$f_{\bar{\eta}}(s, a) = (\mathcal{B}_{r_h})_{\#}[\mathbb{P}\bar{\eta}](s, a) \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}.$$

Statistical Functional Bellman Completeness

Distributional Bellman Completeness (DistBC)

For any distribution $\bar{\eta} : \mathcal{S} \to \mathscr{P}([0, H])$ and $h \in [H]$, there exists $f_{\bar{\eta}} \in \mathcal{H}$ ($\subseteq \mathcal{F}^\infty$) which satisfies

$$f_{\bar{\eta}}(s, a) = (\mathcal{B}_{r_h})_\#[\mathbb{P}\bar{\eta}](s, a) \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}.$$

$\Downarrow$ *(Relax the representation space to statistical functionals)*

Statistical Functional Bellman Completeness (SFBC)

For any distribution $\bar{\eta} : \mathcal{S} \to \mathscr{P}([0, H])$ and $h \in [H]$, there exists $f_{\bar{\eta}} \in \mathcal{F}^N$ which satisfies

$$f_{\bar{\eta}}(s, a) = \psi_{1:N}\Big((\mathcal{B}_{r_h})_\#[\mathbb{P}\bar{\eta}](s, a)\Big) \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}.$$

SF-LSVI

Algorithm 1 Statistical Functional Least Square Value Iteration (SF-LSVI(δ))

Input: failure probability $\delta \in (0,1)$ and the number of episodes K

```
for episode k = 1, 2, ..., K do
    Receive initial state s_1^k
    Initialize ψ_{1:N}(η̄_{H+1}^k(·)) ← 0^N
    for step h = H, H−1, ..., 1 do
        D_h^k ← { s_{h'}^τ, a_{h'}^τ, ψ_{1:N}( (B_{r_{h'}^τ})_# η̄_{h+1}^k(s_{h'+1}^τ) ) }_{(τ,h')∈[k−1]×[H]}   // Data collection
        f̃_{h,η̄}^k ← arg min_{f∈F^N} ||f||_{D_h^k}   // Distribution Estimation
        b_h^k(·,·) ← w^{(1)}((F^N)_h^k, ·, ·)
        Q_h^k(·,·) ← min{ (f̃_{h,η̄}^k)^{(1)}(·,·) + b_h^k(·,·), H }
        π_h^k(·) = arg max_{a∈A} Q_h^k(·, a) , V_h^k(·) = Q_h^k(·, π_h^k(·))   // Optimistic planning
        ψ_1(η_h^k(·,·)) ← Q_h^k(·,·),  ψ_{2:N}(η_h^k(·,·)) ← ( min{ (f̃_{h,η̄}^k)^{(n)}(·,·), H } )_{n∈[2:N]}
        ψ_1(η̄_h^k(·)) ← V_h^k(·),  ψ_{2:N}(η̄_h^k(·)) ← ψ_{1:N}( η_h^k(·, π_h^k(·)) )_{n∈[2:N]}
    for h = 1, 2, ..., H do
        Take action a_h^k ← π_h^k(s_h^k)
        Observe reward r_h^k(s_h^k, a_h^k) and get next state s_{h+1}^k.
```

Moment least square regression

$$\tilde{f}^k_{h,\bar{\eta}} \leftarrow \arg\min_{f\in\mathcal{F}} \sum_{\tau=1}^{k-1}\sum_{h'=1}^{H}\left(\sum_{n=1}^{N} f^{(n)}(s^\tau_{h'}, a^\tau_{h'}) - \psi_n\Big(\big(\mathcal{B}_{r^\tau_{h'}}\big)_{\#}\bar{\eta}^k_{h+1}(s^\tau_{h'+1})\Big)\right)^2$$

SF-LSVI

Theorem

Under SFBC assumption, with probability at least $1 - \delta$, `SF-LSVI` *achieves a regret bound of*

$$Reg(K) \leq 2H dim_E(\mathcal{F}^N, 1/T) + 4H\sqrt{KH \log(1/\delta)}.$$

SF-LSVI

Theorem

Under SFBC assumption, with probability at least $1-\delta$, SF-LSVI *achieves a regret bound of*

$$Reg(K) \le 2H dim_E(\mathcal{F}^N, 1/T) + 4H\sqrt{KH\log(1/\delta)}.$$

Table 1. Comparison for different methods under distributional RL framework. $\mathcal{H}$ represents a subspace of infinite-dimensional space $\mathcal{F}^\infty$. To bound the eluder dimesion d_E, Wang et al. (2023) and Chen et al. (2024) assumed the discretized reward MDP.

Algorithm	Regret	Eluder dimension d_E	Bellman Completeness	MDP assumption	Finite Representation	Exactly Learnable
O-DISCO (Wang et al., 2023)	$\tilde{\mathcal{O}}(\mathrm{poly}(d_E H)\sqrt{K})$	$\dim_E(\mathcal{H}, \epsilon)$	distributional BC	discretized reward, small-loss bound	✗	✗
V-EST-LSR (Chen et al., 2024)	$\tilde{\mathcal{O}}(d_E H^2 \sqrt{K})$ [2]	$\dim_E(\mathcal{H}, \epsilon)$	distributional BC	discretized reward, lipschitz continuity	✗	✗
SF-LSVI [Ours]	$\tilde{\mathcal{O}}(d_E H^{\frac{3}{2}} \sqrt{K})$	$\dim_E(\mathcal{F}^N, \epsilon)$	statistical functional BC	none	✓	✓

→ Compared to previous distRL methods, SF-LSVI achieves a **tighter** regret bound under a **weaker** structural assumption.

Conclusion

To sum up,

- **Bellman Unbiasedness** provides a foundation for designing *exactly learnable* and *provably efficient* distRL algorithm.
- We show that only **moment-based functionals** can be exactly learned—even among *nonlinear* statistical functionals.
- `SF-LSVI` achieves a *tighter* regret bound under a *weaker* assumption, **SFBC**.

Thank you!

ArXiv Link

References

Mark Rowland (2019)
Statistics and samples in distributional reinforcement learning

Marc G. Bellemare (2023)
Distributional reinforcement learning

Kaiwen Wang (2023)
The Benefits of Being Distributional: Small-Loss Bounds for Reinforcement Learning

Yu Chen (2024)
Provable Risk-Sensitive Distributional Reinforcement Learning with General Function Approximation