History-Dependent Off-Policy Evalation for POMDPs: Exponential-Free Error Bound via Future-Dependent Functional Characterization

π

1 Introduction

Off-policy evaluation (OPE) concerns the problem of estimating the expected return of a target policy using data collected under a different behavior policy [Jiang and Li, 2016, Hao et al., 2021, Zhang et al., 2022]. This problem is fundamental in reinforcement learning (RL) and critical for applications where deploying the target policy directly is impractical or unsafe, such as in healthcare, finance, and autonomous systems [Tang and Wiens, 2021, Fu et al., 2021]. In fully observable Markov decision processes (MDPs), a rich body of work has developed estimators with strong theoretical guarantees that avoid exponential dependence on the planning horizon. However, many real-world environments exhibit partial observability and complex, high-dimensional observations, which significantly complicate the evaluation task. In such partially observable Markov decision processes (POMDPs), a straightforward reduction to a history-based MDP leads to estimation errors that depend on ratios over entire observation histories, resulting in exponential blowups with respect to the horizon. This exponential dependence — often referred to as the "curse of history" — poses a severe challenge to scalable and reliable OPE.

Recent work has introduced the concept of future-dependent value functions as a promising framework to mitigate this issue [Uehara et al., 2023]. Unlike traditional value functions defined solely over past histories, future-dependent value functions incorporate future observation proxies that can capture latent state information without relying on full history sequences. This framework potentially enables estimators whose performance guarantees depend on densities over latent states rather than on combinatorially large history distributions, suggesting a path toward polynomial sample complexity. Nevertheless, existing analyses reveal that the boundedness and related properties of future-dependent value functions may still grow exponentially with the horizon under standard assumptions, thereby erasing the intended advantage. **As a reproduction for Zhang and Jiang [2024], we aim to provide a rigorous and positive response to the following question:**

Is it theoretically possible to achieve exponential-free error bounds, in terms of off-policy evaluation for history-dependent policies in POMDPs?

2 Setup

We consider an episodic learning scenario with timesteps indexed by $h \in \{0, \dots, H-1\}$. At each step, the observation space has cardinality $|\mathcal{O}_h| \equiv O$, the state space satisfies $|\mathcal{S}_h| = S_h$, the action space has cardinality $|\mathcal{A}| = A$, and the feedback space has cardinality $|\mathcal{F}_h| = F_h$. For convenience, we encode each state as an integer so that $\mathcal{S}_h = [S_h]$. The full notation is summarized in Table 1. We now formalize the assumptions used throughout this work.

Assumption 1. We consider an undiscounted POMDP M with a finite horizon H. Each i.i.d. rollout of M generates a complete trajectory

$$(o_0, a_0, r_0, \dots, o_{H-1}, a_{H-1}, r_{H-1}, o_H),$$

Disclaimer: This report contains π's personal academic project submitted for *CS542 Statistical Reinforcement Learning, 2025 Fall.* It is shared for educational and portfolio purposes only.

containing the sequence of observations, actions, and rewards encountered along the episode. The performance measure of interest is the undiscounted cumulative return

$$G = \sum_{t=0}^{H-1} r_t,$$

where we assume, without loss of generality, that each reward is bounded in $[0, 1]$. This assumption standardizes the scale of returns and simplifies the subsequent analysis.

Assumption 2. (Coverage for History-Dependent Policies) To enable off-policy evaluation and learning, we assume access to a behavior policy $\pi_{b,h} : \mathcal{T}_h \times \mathcal{O}_h \to \Delta(\mathcal{A})$. For any target history-dependent policy $\pi_h : \mathcal{T}_h \to \Delta(\mathcal{A})$, we require a uniform coverage condition ensuring that the behavior policy sufficiently explores the action space. Specifically, the density ratio

$$\rho(a_h \mid \tau_h) = \frac{\pi_h(a_h \mid \tau_h)}{\pi_{b,h}(a_h \mid \tau_h)}$$

is assumed to be bounded by a constant C_{cvg} for all timesteps $h \in [H]$, actions $a_h \in \mathcal{A}$, and histories $\tau_h \in \mathcal{T}_h$. This prevents degenerate data-collection policies and ensures that importance-weighted estimators remain well-defined.

Assumption 3. (Realizability) At least one $V_{\mathcal{F}}^* \in \mathcal{V}$ can minimize the objective in Eq. 12.

Definition 2.1. (Projected Bellman Error) Given a future-dependent value function $V_{\mathcal{F}}$ that satisfies 2, we define its expected Bellman error conditioned on the history τ_h by

$$\left(\mathcal{B}_{\mathcal{H}}^{\pi} V_{\mathcal{F}}\right)(\tau_h) = \left\langle b(\tau_h), \mathcal{B}_{\mathcal{S}}^{\pi} V_{\mathcal{S}} \right\rangle, \tag{1}$$

where $\mathcal{B}_{\mathcal{S}}^{\pi} V_{\mathcal{S}} \in \mathbb{R}^{\mathcal{S}_h}$ ensembles Bellman error across $s_h \in \mathcal{S}_h$.

Assumption 4. (Cross Bellman Completeness) The history-dependent function class $\mathcal{G}$ is finite, and $\mathcal{B}_{\mathcal{H}}^{\pi} V_{\mathcal{F}} \in \mathcal{G}, \forall V_{\mathcal{F}} \in \mathcal{V}$,

3 Future-Dependent Value Functions: OPE Error, Bellman Residual Optimization, and Theoretical Guarantees

Value functions for classical MDPs are typically constructed by $V_{\mathcal{S}}^{\pi}(s_h) = \mathbb{E}_{\pi}\left[\sum_{t=h}^{H-1} r_t \,\middle|\, s_h\right]$ as an expectation conditioned on a given state s_h at step h, yet this definition is not fully tractable in POMDPs due to the learner's limited exposure to the underlying system states. Alternatively, we can propose a *future-dependent value function (FDVF)* $V_{\mathcal{F},h} : \mathcal{F}_h \to \mathbb{R}$ that maps a given future $f_h \in \mathcal{F}_h$ to a real value estimate, satisfying

$$\mathbb{E}_{\pi_b}\left[V_{\mathcal{F}}(f_h) \,|\, s_h\right] = V_{\mathcal{S}}(s_h) \tag{2}$$

as a parallel modeling route, where we denote $V_{\mathcal{S},h}(s_h) = V_{\mathcal{S}}(s_h)$ and $V_{\mathcal{F},h}(f_h) = V_{\mathcal{F}}(f_h)$ for simplicity of the notation. This formulation automatically enjoys a matrix form

$$U_{b,h}^{T} V_{\mathcal{F}} = V_{\mathcal{S}}$$

for vectorized $V_{\mathcal{F}} \in \mathbb{R}^{\mathcal{F}_h}$ and $V_{\mathcal{S}} \in \mathbb{R}^{\mathcal{S}_h}$, leveraging the outcome matrix $U_{b,h}$. Ultimately, the (undiscounted) state-space Bellman error for a specific choice of $V_{\mathcal{S}}$ can be represented as

$$\begin{aligned}
\left(\mathcal{B}_{\mathcal{S}}^{\pi} V_{\mathcal{S}}\right)(s_h) &= \mathbb{E}_{a_h \sim \pi(\cdot|s_h), s_{h+1} \sim P(\cdot|s_h, a_h)}\left[r_h + V_{\mathcal{S}}(s_{h+1})\right] - V_{\mathcal{S}}(s_h) & (3)\\
&= \mathbb{E}_{a_h \sim \pi(\cdot|s_h), s_{h+1} \sim P(\cdot|s_h, a_h)}\left[r_h + \mathbb{E}_{\pi_b}\left[V_{\mathcal{F}}(f_{h+1}) \,|\, s_{h+1}\right]\right] - \mathbb{E}_{\pi_b}\left[V_{\mathcal{F}}(f_h) \,|\, s_h\right] & (4)\\
&= \mathbb{E}_{a_h \sim \pi(\cdot|s_h), a_{h+1:H-1} \sim \pi_b}\left[r_h + V_{\mathcal{F}}(f_{h+1})\right] - \mathbb{E}_{\pi_b}\left[V_{\mathcal{F}}(f_h) \,|\, s_h\right] & (5)
\end{aligned}$$

by invoking the definition of FDVF. The primary result for FDVF-based OPE is verified in Theorem 3.1.

Theorem 3.1. *(Performance Difference Lemma)*

$$J(\pi) - \mathbb{E}_{\pi_b}\left[V_{\mathcal{F}}(f_0)\right] = \sum_{t=0}^{H-1} \mathbb{E}_{\pi}\left[\left(\mathcal{B}_{\mathcal{S}}^{\pi} V_{\mathcal{S}}\right)(s_t)\right] \tag{6}$$

Proof. We employ classical telescoping techniques for estimating $J(\pi)$, which is detailed by

$$\begin{aligned} J(\pi) &= \mathbb{E}_{\pi}\left[\sum_{t=0}^{H-1} r_t\right] && (7)\\ &= \mathbb{E}_{\pi_b}\left[V_{\mathcal{F}}(f_0)\right] + \mathbb{E}_{\pi}\left[\sum_{t=0}^{H-1} - \mathbb{E}_{f_t \sim \pi_b}\left[V_{\mathcal{F}}(f_t) \mid s_t\right] + r_t + \mathbb{E}_{f_{t+1} \sim \pi_b}\left[V_{\mathcal{F}}(f_{t+1}) \mid s_{t+1}\right]\right] && (8)\\ &= \mathbb{E}_{\pi_b}\left[V_{\mathcal{F}}(f_0)\right] + \sum_{t=0}^{H-1} \mathbb{E}_{(s_t,a_t)\sim d_t^{\pi}}\left[-\mathbb{E}_{\pi_b}\left[V_{\mathcal{F}}(f_t) \mid s_t\right] + r_t + \mathbb{E}_{\pi_b}\left[V_{\mathcal{F}}(f_{t+1}) \mid s_{t+1}\right]\right] && (9)\\ &= \mathbb{E}_{\pi_b}\left[V_{\mathcal{F}}(f_0)\right] + \sum_{t=0}^{H-1} \mathbb{E}_{\pi}\left[\left(\mathcal{B}_{\mathcal{S}}^{\pi} V_{\mathcal{S}}\right)(s_t)\right]. && (10) \end{aligned}$$

With this telescoping result, we can further establish the aforementioned PD lemma in theorem 3.1. □

The PD lemma motivates us to search for a function approximator $V_{\mathcal{F}}$ with low Bellman error $\left(\mathcal{B}_{\mathcal{S}}^{\pi} V_{\mathcal{S}}\right)(s) \approx 0$, such that it approaches the groundtruth $J(\pi) \approx \mathbb{E}_{\pi_b}\left[V_{\mathcal{F}}(f_0)\right]$ simply by accessing the offline data $f_0 \sim \mathcal{D}_b$. However, minimizing the squared Bellman error will introduce double-sampling issues, making it highly intractable to unbiasedly estimate the inner expectation with only single-sample observations of the random variables involved. To provide an statistically unbiased estimate for $\mathbb{E}\left[\left(\mathcal{B}_{\mathcal{S}}^{\pi} V_{\mathcal{S}}\right)(s_t)\right]^2$, Zhang and Jiang [2024] suggests a minimax loss construction

$$\begin{aligned} &\mathbb{E}_{s_t \sim d_t^{\pi_b}}\left[\left(\mathcal{B}_{\mathcal{S}}^{\pi} V_{\mathcal{S}}\right)(s_t)\right]^2 && (11)\\ &= \mathbb{E}_{\tau_t \sim \mu_t^{\pi_b}, s_t \sim b(\cdot|\tau_t)}\left[\mathbb{E}_{f_t \sim u^{\pi_b}(\cdot|s_t)}\left[V_{\mathcal{F}}(f_t)\right] - \mathbb{E}_{o_t \sim \mathbb{O}(\cdot|s_t), a_t \sim \pi(\cdot|\tau_t), s_{t+1} \sim P(\cdot|s_t,a_t), f_{t+1} \sim u^{\pi_b}(\cdot|s_{t+1})}\left[r_h + V_{\mathcal{F}}(f_{t+1})\right]\right]^2\\ &= \mathbb{E}_{\tau_t \sim \mu_t^{\pi_b}, s_t \sim b(\cdot|\tau_t)}\left[\mathbb{E}_{f_t \sim u^{\pi_b}(\cdot|s_t)}\left[V_{\mathcal{F}}(f_t)\right] - \mathbb{E}_{\substack{o_t \sim \mathbb{O}(\cdot|s_t), a_t \sim \pi(\cdot|\tau_t),\\ s_{t+1} \sim P(\cdot|s_t,a_t), f_{t+1} \sim u^{\pi_b}(\cdot|s_{t+1})}}\left[r_h + V_{\mathcal{F}}(f_{t+1})\right]\right]^2\\ &= \mathbb{E}_{\tau_t \sim \mu_t^{\pi_b}, s_t \sim b(\cdot|\tau_t)}\left[\max_{g \in \mathbb{R}}\left[-g^2 \pm 2\left[\mathbb{E}_{f_t \sim u^{\pi_b}(\cdot|s_t)}\left[V_{\mathcal{F}}(f_t)\right] - \mathbb{E}_{\substack{o_t \sim \mathbb{O}(\cdot|s_t), a_t \sim \pi(\cdot|\tau_t),\\ s_{t+1} \sim P(\cdot|s_t,a_t), f_{t+1} \sim u^{\pi_b}(\cdot|s_{t+1})}}\left[r_h + V_{\mathcal{F}}(f_{t+1})\right]\right] g\right]\right]\\ &= \max_{V_{\mathcal{H}}:\tau_t \to \mathbb{R}} \mathbb{E}_{\pi_b}\left[-V_{\mathcal{H}}^2(\tau_t) \pm 2\left[V_{\mathcal{F}}(f_t) - \frac{\pi(a_t|\tau_t)}{\pi_b(a_t|\tau_t)}\left[r_h - V_{\mathcal{F}}(f_{t+1})\right]\right] V_{\mathcal{H}}(\tau_t)\right]\\ &= \max_{V_{\mathcal{H}}:\tau_t \to \mathbb{R}} \mathbb{E}_{\pi_b}\left[-V_{\mathcal{H}}^2(\tau_t) + 2\left[\frac{\pi(a_t|\tau_t)}{\pi_b(a_t|\tau_t)}\left[r_h - V_{\mathcal{F}}(f_{t+1})\right] - V_{\mathcal{F}}(f_t)\right] V_{\mathcal{H}}(\tau_t)\right]\\ &= \max_{V_{\mathcal{H}}:\mathcal{T}_t \to \mathbb{R}} \mathcal{L}_t(V_{\mathcal{F}}, V_{\mathcal{H}}) \end{aligned}$$

by a functional re-characterization, such that the future-dependent value function can be satisfactorily approximated via

$$\hat{V_{\mathcal{F}}} \in \arg\min_{V_{\mathcal{F}} \in \mathcal{V}} \max_{V_{\mathcal{H}} \in \mathcal{G}} \sum_{t=0}^{H-1} \hat{\mathcal{L}}_t(V_{\mathcal{F}}, V_{\mathcal{H}}), \tag{12}$$

where the empirical loss $\hat{\mathcal{L}}_t$ substitutes $\mathbb{E}_{\pi_b}$ with $\mathbb{E}_{\mathcal{D}_b}$. As a substantial result for FDVF, the future-dependent value function fitted by 12 will enjoy the suboptimality illustrated below.

Theorem 3.2. *Under Assumptions 3 and 4, w.p.* $\geq 1 - \delta$,

$$\left|J(\pi) - \mathbb{E}_{\mathcal{D}_b}\left[\hat{V_{\mathcal{F}}}(f_0)\right]\right| \leq cH \cdot C_{cvg} \cdot \max\left\{D_{\mathcal{V}} + 1, D_{\mathcal{G}}\right\} \cdot IV(\mathcal{V}) \cdot Dr_{\mathcal{V}}\left(\pi \| \pi_b\right) \cdot \sqrt{\frac{\log \frac{|\mathcal{V}||\mathcal{G}|}{\delta}}{n}}, \tag{13}$$

where

- $D_{\mathcal{V}} = \max_{V_{\mathcal{F}} \in \mathcal{V}} \|V_{\mathcal{F}}\|_{\infty}$ *and* $D_{\mathcal{G}} = \max_{V_{\mathcal{H}} \in \mathcal{G}} \|V_{\mathcal{H}}\|_{\infty}$ *represents the upper bounds of the function classes,*

- $$\mathit{IV}(\mathcal{V}) = \max_h \sup_{V_\mathcal{F} \in \mathcal{V}} \sqrt{\frac{\mathbb{E}_{s_h \sim d_h^{\pi_b}} \left[\left(\mathcal{B}_\mathcal{S}^\pi V_\mathcal{S}\right)(s_h)\right]^2}{\mathbb{E}_{\tau_h \sim \mu_h^{\pi_b}} \left[\left(\mathcal{B}_\mathcal{H}^\pi V_\mathcal{F}\right)(\tau_h)\right]^2}}$$

  measures the worst-case conversion ratio between $\mathcal{B}_\mathcal{S}^\pi V_\mathcal{S}$ and its proxy $\mathcal{B}_\mathcal{H}^\pi V_\mathcal{F}$, where $\left(\mathcal{B}_\mathcal{S}^\pi V_\mathcal{S}\right)(s_h)$ can be represented by $V_\mathcal{F}$ via FDVF definition 5,

- $$\mathit{Dr}(\pi \| \pi_b) = \max_h \sup_{V_\mathcal{F} \in \mathcal{V}} \sqrt{\frac{\mathbb{E}_{s_h \sim d_h^{\pi}} \left[\left(\mathcal{B}_\mathcal{S}^\pi V_\mathcal{S}\right)(s_h)\right]^2}{\mathbb{E}_{s_h \sim d_h^{\pi_b}} \left[\left(\mathcal{B}_\mathcal{S}^\pi V_\mathcal{S}\right)(s_h)\right]^2}}$$

  corresponds to the coverage measure given π_b and π.

See our proof in Appendix A.2. At a high level, the proof shows that our two-level optimization framework can reliably recover the policy value by separating the problem into an inner signal-estimation step and an outer value-approximation step. The inner step learns a projected Bellman-error signal that aligns with what the outer estimator needs, while the outer step uses this learned signal to construct a stable and statistically controlled estimate of the target policy's return. By ensuring that each stage introduces only a small amount of statistical deviation, the analysis demonstrates that the overall procedure remains accurate even under distribution mismatch between the behavior and target policies. The result is a principled off-policy evaluation method whose reliability follows from the systematic control of errors at both levels.

4 Other Notes

Our main result in Appendix A.2 does not fully match the order of coverage reported in the original work (we obtain an $O(C_{\text{cvg}})$ bound, whereas they claimed $O(\sqrt{C_{\text{cvg}}})$). This discrepancy may be attributed to a potential error in their proof. Moreover, in certain instances, our constants are tighter than those in the original paper, which can be explained by a more refined exploitation of Bellman completeness.

References

Justin Fu, Mohammad Norouzi, Ofir Nachum, George Tucker, Ziyu Wang, Alexander Novikov, Mengjiao Yang, Michael R. Zhang, Yutian Chen, Aviral Kumar, Cosmin Paduraru, Sergey Levine, and Tom Le Paine. Benchmarks for deep off-policy evaluation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=kWSeGEeHvF8.

Botao Hao, Xiang Ji, Yaqi Duan, Hao Lu, Csaba Szepesvári, and Mengdi Wang. Bootstrapping fitted q-evaluation for off-policy inference. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4074–4084. PMLR, 2021. URL http://proceedings.mlr.press/v139/hao21b.html.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In Maria-Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 652–661. JMLR.org, 2016. URL http://proceedings.mlr.press/v48/jiang16.html.

Shengpu Tang and Jenna Wiens. Model selection for offline reinforcement learning: Practical considerations for healthcare settings. In Ken Jung, Serena Yeung, Mark P. Sendak, Michael W. Sjoding, and Rajesh Ranganath, editors, *Proceedings of the Machine Learning for Healthcare Conference, MLHC 2021, 6-7 August 2021, Virtual Event*, volume 149 of *Proceedings of Machine Learning Research*, pages 2–35. PMLR, 2021. URL https://proceedings.mlr.press/v149/tang21a.html.

Masatoshi Uehara, Haruka Kiyohara, Andrew Bennett, Victor Chernozhukov, Nan Jiang, Nathan Kallus, Chengchun Shi, and Wen Sun. Future-dependent value-based off-policy evaluation in pomdps. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 15991–16008. Curran Associates, Inc., 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/3380e8116452e0efbf36f35d95e88c94-Paper-Conference.pdf`.

Ruiqi Zhang, Xuezhou Zhang, Chengzhuo Ni, and Mengdi Wang. Off-policy fitted q-evaluation with differentiable function approximators: Z-estimation and inference theory. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 26713–26749. PMLR, 2022. URL `https://proceedings.mlr.press/v162/zhang22al.html`.

Yuheng Zhang and Nan Jiang. On the curses of future and history in future-dependent value functions for off-policy evaluation, 2024. URL `https://arxiv.org/abs/2402.14703`.

A Appendix

A.1 Notations

Table 1: A summary wrapping up the notations we employ in throughout the text.

Symbol	Description
H	The finite-horizon
$d_0(\cdot) \in \Delta(\mathcal{S}_0)$	Initial state distribution
$\mathbb{O}(\cdot\|s_h) \in \Delta(\mathcal{S}_h)$	Emission probabilities
$d_t^\pi(\cdot) : \Delta(\mathcal{S}_t)$	t-step state occupancy
$z_t^\pi(\cdot)$	t-step observation occupancy
τ_h	Historical transitions $\{o_0, a_0, ..., o_{h-1}, a_{h-1}\}$ terminating at a_{h-1}
$b(\cdot\|\tau_h) \in \Delta(\mathcal{S}_h)$	Posterior for s_h given history τ_h
$b(\tau_h)$	A column vector $[b(0\|\tau_h), ..., b(S_h - 1\|\tau_h)]^T$
$\mathcal{T}_h$	$\ni \{o_0, a_0, ..., o_{h-1}, a_{h-1}\}$
π^{hd}	A sequence of history-dependent policies $\left\{\pi_h^{\text{hd}}\right\}_{[H]}$
$\pi_h^{\text{hd}} : \mathcal{T}_h \times \mathcal{O}_h \to \Delta(\mathcal{A})$	The history-dependent policy at step h given τ_h and o_h
π^{ml}	A sequence of memoryless policies $\left\{\pi^{\text{ml}}\right\}_{[H]}$
$\pi_h^{\text{ml}} : \mathcal{O}_h \to \mathcal{A}$	The memoryless policy at step h given o_h
$R : \cup_{h\in[H]}\mathcal{O}_h \to \mathbb{R}$	A time-invariant reward function that does not depend on the action
$d_h^\pi(s, a)$	State-action occupancy at step h
$\mathbb{E}_\pi[X]$	The expectation for a random variable X when performing rollouts with π
$\mathbb{P}^\pi(X)$	The probability for an incident X when performing rollouts with π
f_h	A future trajectory $\{o_h, a_h, ..., o_{H-2}, a_{H-2}, o_{H-1}\}$ starting at o_h
$P(\cdot\|s,a)$	The transition dynamics
$\mathcal{F}_h$	$\ni \{o_h, a_h, ..., o_{H-2}, a_{H-2}, o_{H-1}\}$
$u^\pi(\cdot\|s_h) \in \Delta(\mathcal{F}_h)$	The posterior for f_h given s_h and a policy π
$u^\pi(f_h)$	A column vector $[u^\pi(f_h\|0), ..., u^\pi(f_h\|S_h - 1)]^T$
$U_{b,h} \in \mathbb{R}^{F_h \times S_h}$	The outcome matrix $\left[u^{\pi_b}(f^{(0)}), ..., u^{\pi_b}(f^{(F_h-1)})\right]^T$
$\mathcal{D}_b = \left\{\tau_H^{(i)}\right\}_{[D]}$	The off-policy dataset collected by π_b that contains D many i.i.d. trajectories $\tau_H^{(i)}$
$\mathcal{B}^{\mathcal{S}}$	State-level Bellman error
$\mathcal{B}^{\mathcal{H}}$	History-dependent Bellman error
$\mu^\pi(\cdot) \in \Delta(\mathcal{T}_h)$	Probability for τ_h induced by a behavior policy π_b
$\mathcal{V}$	The future-dependent value function class
$\mathcal{G}$	The history-dependent value function class
$V_{\mathcal{H}}$	The history-dependent value function

A.2 Proof of Main Theorem

Proof. For shorthand we denote the random Bellman error as

$$\mathcal{E}_{\mathcal{F},h} = \rho(a_h|\tau_h)\,[r_h + V_{\mathcal{F}}(f_{h+1})] - V_{\mathcal{F}}(f_h),$$

with importance weight $\rho(a_h|\tau_h) = \frac{\pi(a_h|\tau_h)}{\pi_b(a_h|\tau_h)}$, then it suffices that the conditional expectation

$$\mathbb{E}_{\pi_b}\left[\mathcal{E}_{\mathcal{F},h} \mid \tau_h\right] = \left(\mathcal{B}_{\mathcal{H}}^\pi V_{\mathcal{F}}\right)(\tau_h)$$

well measures the true Bellman error projected on τ_h. Additionally, the re-characterization can be further represented by

$$\begin{aligned}\mathcal{L}_t(V_\mathcal{F}, V_\mathcal{H}) &= \mathbb{E}_{\pi_b}\left[-V^2_\mathcal{H}(\tau_t) + 2\mathcal{E}_{\mathcal{F},t}V_\mathcal{H}(\tau_t)\right] \\ &= \mathbb{E}_{\pi_b}\left[\mathcal{E}^2_{\mathcal{F},t} - \left[V_\mathcal{H}(\tau_t) - \mathcal{E}_{\mathcal{F},t}\right]^2\right]\end{aligned} \tag{14}$$

Step I. Inner Maximizer. For a fixed $V_\mathcal{F} \in \mathcal{V}$, we denote $\hat{V_\mathcal{H}}$ as the solution that empirically maximizes $\sum_{t=0}^{H-1} \hat{\mathcal{L}}_t(V_\mathcal{F}, V_\mathcal{H})$ from the dataset $\mathcal{D}_b$. Now we want to bound the empirical estimator $\hat{V_\mathcal{H}}$ from $\mathbb{E}_{\pi_b}\left[\mathcal{E}_{\mathcal{F},t} \,|\, \tau_t\right] = \left(\mathcal{B}^\pi_\mathcal{H} V_\mathcal{F}\right)(\tau_h)$, the target function defined on τ_t that desirably re-characterizes the objective in 11. Given realizability and the orthogonal decomposition [1]

$$\mathbb{E}_{\mathcal{D}_b}\left[V_\mathcal{H}(\tau_t) - \mathcal{E}_{\mathcal{F},t}\right]^2 = \mathbb{E}_{\mathcal{D}_b}\left[V_\mathcal{H}(\tau_t) - \left(\mathcal{B}^\pi_\mathcal{H} V_\mathcal{F}\right)(\tau_t)\right]^2 + \mathbb{E}_{\mathcal{D}_b}\left[\left(\mathcal{B}^\pi_\mathcal{H} V_\mathcal{F}\right)(\tau_t) - \mathcal{E}_{\mathcal{F},t}\right]^2,$$

we know that

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_b}\left[\sum_{t=0}^{H-1}\mathcal{E}^2_{\mathcal{F},t} - \left[\hat{V_\mathcal{H}(\tau_t)} - \mathcal{E}_{\mathcal{F},t}\right]^2\right] &= \mathbb{E}_{\mathcal{D}_b}\left[\sum_{t=0}^{H-1}\mathcal{E}^2_{\mathcal{F},t}\right] - \mathbb{E}_{\mathcal{D}_b}\left[\sum_{t=0}^{H-1}\left[\hat{V_\mathcal{H}}(\tau_t) - \left(\mathcal{B}^\pi_\mathcal{H} V_\mathcal{F}\right)(\tau_t)\right]^2\right] \\ &\quad - \mathbb{E}_{\mathcal{D}_b}\left[\sum_{t=0}^{H-1}\left[\left(\mathcal{B}^\pi_\mathcal{H} V_\mathcal{F}\right)(\tau_t) - \mathcal{E}_{\mathcal{F},t}\right]^2\right] \\ &\geq \mathbb{E}_{\mathcal{D}_b}\left[\sum_{t=0}^{H-1}\mathcal{E}^2_{\mathcal{F},t} - \left[\left(\mathcal{B}^\pi_\mathcal{H} V_\mathcal{F}\right)(\tau_t) - \mathcal{E}_{\mathcal{F},t}\right]^2\right],\end{aligned} \tag{15}$$

which implicates $\mathbb{E}_{\mathcal{D}_b}\left[\sum_{t=0}^{H-1}\left[\hat{V_\mathcal{H}}(\tau_t) - \left(\mathcal{B}^\pi_\mathcal{H} V_\mathcal{F}\right)(\tau_t)\right]^2\right] = 0$ (see footnote [2]). Leveraging this result, we can further derive the generalization loss

$$\sum_{t=0}^{H-1}\mathbb{E}_{\pi_b}\left[\hat{V_\mathcal{H}}(\tau_t) - \left(\mathcal{B}^\pi_\mathcal{H} V_\mathcal{F}\right)(\tau_t)\right]^2 \tag{16}$$

$$\leq \underbrace{\mathbb{E}_{\mathcal{D}_b}\left[\sum_{t=0}^{H-1}\left[\hat{V_\mathcal{H}}(\tau_t) - \left(\mathcal{B}^\pi_\mathcal{H} V_\mathcal{F}\right)(\tau_t)\right]^2\right]}_{I_1=0} + \left|\left[\mathbb{E}_{\pi_b} - \mathbb{E}_{\mathcal{D}_b}\right]\sum_{t=0}^{H-1}\left[\hat{V_\mathcal{H}}(\tau_t) - \left(\mathcal{B}^\pi_\mathcal{H} V_\mathcal{F}\right)(\tau_t)\right]^2\right| \tag{17}$$

$$= \underbrace{\left|\left[\mathbb{E}_{\pi_b} - \mathbb{E}_{\mathcal{D}_b}\right]\sum_{t=0}^{H-1}\left[\hat{V_\mathcal{H}}(\tau_t) - \left(\mathcal{B}^\pi_\mathcal{H} V_\mathcal{F}\right)(\tau_t)\right]\left[\hat{V_\mathcal{H}}(\tau_t) - 2\mathcal{E}_{\mathcal{F},t} + \left(\mathcal{B}^\pi_\mathcal{H} V_\mathcal{F}\right)(\tau_t)\right]\right|}_{I_2} \tag{18}$$

where we let $X_{H,t} = \left[\hat{V_\mathcal{H}}(\tau_t) - \left(\mathcal{B}^\pi_\mathcal{H} V_\mathcal{F}\right)(\tau_t)\right]\left[\hat{V_\mathcal{H}}(\tau_t) - 2\mathcal{E}_{\mathcal{F},t} + \left(\mathcal{B}^\pi_\mathcal{H} V_\mathcal{F}\right)(\tau_t)\right]$ in the second term $I_2 = \left|\left[\mathbb{E}_{\pi_b} - \mathbb{E}_{\mathcal{D}_b}\right]X_H\right| = \left|\left[\mathbb{E}_{\pi_b} - \mathbb{E}_{\mathcal{D}_b}\right]\sum_{t=0}^{H-1} X_{H,t}\right|$. With $\bar{D} = \max\left\{D_\mathcal{V}+1, D_\mathcal{G}\right\}$ and $|r_t| \leq 1$, we know that

$$\begin{aligned}&|\mathcal{E}_{\mathcal{F},t}| \leq |\rho(a_t|\tau_t)| \cdot |r_h + V_\mathcal{F}(f_{h+1})| + |V_\mathcal{F}(f_h)| \leq C_{\text{cvg}}R_{max} + C_{\text{cvg}}D_\mathcal{V} + D_\mathcal{V} \leq 3C_{\text{cvg}}\bar{D}, \\ &|V_\mathcal{H}(\tau_t)| \leq D_\mathcal{V} \leq \bar{D} \leq C_{\text{cvg}}\bar{D}, \\ &\left|\left(\mathcal{B}^\pi_\mathcal{H} V_\mathcal{F}\right)(\tau_t)\right| \leq D_\mathcal{G} \leq \bar{D} \leq C_{\text{cvg}}\bar{D}, \quad \text{(Completeness in 4)}\end{aligned}$$

[1] $\mathbb{E}\left[X-Y\right]^2 = \mathbb{E}\left[X - \mathbb{E}\left[Y|X\right]\right]^2 + \mathbb{E}\left[\mathbb{E}\left[Y|X\right] - Y\right]^2$.

[2] Realizability suggests that, for any randomly generated dataset $\mathcal{D}_b$, the function class can at least overfit to $\mathcal{D}_b$ with 0 empirical loss. Despite the possibility of $\hat{V_\mathcal{H}} \neq \mathcal{B}^\pi_\mathcal{H} V_\mathcal{F}$, the generalization error for $\hat{V_\mathcal{H}}$ can asymptotically approach 0 when the effective data size $|\mathcal{D}_b|$ is sufficiently large.

which leads to $|X_{H,t}| \leq 2\bar{D} \cdot 8C_{\text{cvg}}\bar{D} = 16C_{\text{cvg}}\bar{D}^2$, and $|X_H| \leq 16HC_{\text{cvg}}\bar{D}^2$. Furthermore, we can theoretically validate that

$$\begin{aligned}\mathbb{E}_{\pi_b}\left[X_{H,t}^2\right] &= \mathbb{E}_{\pi_b}\left[\hat{V_{\mathcal{H}}}(\tau_t) - \left(\mathcal{B}_{\mathcal{H}}^{\pi}V_{\mathcal{F}}\right)(\tau_t)\right]^2\left[\hat{V_{\mathcal{H}}}(\tau_t) - 2\mathcal{E}_{\mathcal{F},t} + \left(\mathcal{B}_{\mathcal{H}}^{\pi}V_{\mathcal{F}}\right)(\tau_t)\right]^2 \\ &\leq \mathbb{E}_{\pi_b}\left[\hat{V_{\mathcal{H}}}(\tau_t) - \left(\mathcal{B}_{\mathcal{H}}^{\pi}V_{\mathcal{F}}\right)(\tau_t)\right]^2\left[C_{\text{cvg}}\bar{D} + 6C_{\text{cvg}}\bar{D} + C_{\text{cvg}}\bar{D}\right]^2 \\ &\leq 64C_{\text{cvg}}^2\bar{D}^2\mathbb{E}_{\pi_b}\left[\hat{V_{\mathcal{H}}}(\tau_t) - \left(\mathcal{B}_{\mathcal{H}}^{\pi}V_{\mathcal{F}}\right)(\tau_t)\right]^2\end{aligned} \tag{19}$$

such that, the total variance for X_H can be bounded through

$$\begin{aligned}\text{Var}\left[X_H\right] &\leq \mathbb{E}_{\pi_b}\left[X_H^2\right] \\ &\leq H\mathbb{E}_{\pi_b}\left[\sum_{t=0}^{H-1} X_{H,t}^2\right] \quad \text{(Cauchy's)} \\ &\leq 64HC_{\text{cvg}}^2\bar{D}^2\sum_{t=0}^{H-1}\mathbb{E}_{\pi_b}\left[\hat{V_{\mathcal{H}}}(\tau_t) - \left(\mathcal{B}_{\mathcal{H}}^{\pi}V_{\mathcal{F}}\right)(\tau_t)\right]^2,\end{aligned} \tag{20}$$

and

$$\begin{aligned}\left|X_H - \mathbb{E}_{\pi_b}\left[X_H\right]\right| &\leq \sum_{t=0}^{H-1}\left|X_{H,t} - \mathbb{E}_{\pi_b}\left[X_{H,t}\right]\right| \\ &\leq 2\sum_{t=0}^{H-1}\sup\left\{X_{H,t}\right\} \\ &\leq 32HC_{\text{cvg}}\bar{D}\end{aligned} \tag{21}$$

Now we subsequently apply Bernstein's inequality [3] for term I_2, which verifies that w.p. $\geq 1-\delta$,

$$I_2 \leq \frac{64HC_{\text{cvg}}\bar{D}\log\frac{2|\mathcal{V}||\mathcal{G}|}{\delta}}{3n} + \sqrt{\frac{128HC_{\text{cvg}}^2\bar{D}^2\log\frac{2|\mathcal{V}||\mathcal{G}|}{\delta}}{n}}\sqrt{\mathbb{E}_{\pi_b}\left[\sum_{t=0}^{H-1}\hat{V_{\mathcal{H}}}(\tau_t) - \left(\mathcal{B}_{\mathcal{H}}^{\pi}V_{\mathcal{F}}\right)(\tau_t)\right]^2}, \tag{22}$$

where n represents the number of trajectories contained in the dataset $\mathcal{D}_b$. And finally combining 18 with 26, we yield that

$$\begin{aligned}&\mathbb{E}_{\pi_b}\left[\sum_{t=0}^{H-1}\hat{V_{\mathcal{H}}}(\tau_t) - \left(\mathcal{B}_{\mathcal{H}}^{\pi}V_{\mathcal{F}}\right)(\tau_t)\right]^2 \\ &\leq \frac{64HC_{\text{cvg}}\bar{D}\log\frac{2|\mathcal{V}||\mathcal{G}|}{\delta}}{3n} + \sqrt{\frac{128HC_{\text{cvg}}^2\bar{D}^2\log\frac{2|\mathcal{V}||\mathcal{G}|}{\delta}}{n}}\sqrt{\mathbb{E}_{\pi_b}\left[\sum_{t=0}^{H-1}\hat{V_{\mathcal{H}}}(\tau_t) - \left(\mathcal{B}_{\mathcal{H}}^{\pi}V_{\mathcal{F}}\right)(\tau_t)\right]^2} \\ &\leq \frac{64HC_{\text{cvg}}^2\bar{D}^2\log\frac{2|\mathcal{V}||\mathcal{G}|}{\delta}}{3n} + \sqrt{\frac{128HC_{\text{cvg}}^2\bar{D}^2\log\frac{2|\mathcal{V}||\mathcal{G}|}{\delta}}{n}}\sqrt{\mathbb{E}_{\pi_b}\left[\sum_{t=0}^{H-1}\hat{V_{\mathcal{H}}}(\tau_t) - \left(\mathcal{B}_{\mathcal{H}}^{\pi}V_{\mathcal{F}}\right)(\tau_t)\right]^2}\end{aligned} \tag{23}$$

[3] Bernstein Inequality: $\mathbb{P}(\left|\frac{1}{n}\sum_{i=1}^{n}X_i - \mathbb{E}\left[X\right]\right| \geq \epsilon) \leq 2\exp\left(-\frac{3n\epsilon^2}{6\text{Var}[X]+2M\epsilon}\right)$, where $X_1, \cdots, X_n$ are i.i.d. samples and $|X - \mathbb{E}\left[X\right]| \leq M$. This yields w.p. $\geq 1-\delta$,

$$\left|\frac{1}{n}\sum_{i=1}^{n}X_i - \mathbb{E}\left[X\right]\right| \leq \frac{M\log\frac{2}{\delta}}{3n} + \frac{\sqrt{4M^2\log^2\frac{2}{\delta} + 72n\text{Var}\left[X\right]\log\frac{2}{\delta}}}{6n} \leq \frac{2M\log\frac{2}{\delta}}{3n} + \sqrt{\frac{2\text{Var}\left[X\right]\log\frac{2}{\delta}}{n}}$$

Solving the inequality in 23 derives

$$
\begin{aligned}
\mathbb{E}_{\pi_b}\left[\sum_{t=0}^{H-1} \hat{V}_{\mathcal{H}}(\tau_t) - \left(\mathcal{B}^{\pi}_{\mathcal{H}} V_{\mathcal{F}}\right)(\tau_t)\right]^2 &\leq \frac{64\left(4+\sqrt{15}\right) H C^2_{\text{cvg}} \bar{D}^2 \log \frac{2|\mathcal{V}||\mathcal{G}|}{\delta}}{3n} \\
&\leq \frac{512 H C^2_{\text{cvg}} \bar{D}^2 \log \frac{2|\mathcal{V}||\mathcal{G}|}{\delta}}{3n},
\end{aligned}
\tag{24}
$$

where the upper bound is denoted as $\epsilon_{\text{stat}} = \frac{512 H C^2_{\text{cvg}} \bar{D}^2 \log \frac{2|\mathcal{V}||\mathcal{G}|}{\delta}}{3n}$, such that

$$
\begin{aligned}
I_2 &\leq \frac{64 H C_{\text{cvg}} \bar{D} \log \frac{2|\mathcal{V}||\mathcal{G}|}{\delta}}{3n} + \sqrt{\frac{128 H C^2_{\text{cvg}} \bar{D}^2 \log \frac{2|\mathcal{V}||\mathcal{G}|}{\delta}}{n} \cdot \frac{512 H C^2_{\text{cvg}} \bar{D}^2 \log \frac{2|\mathcal{V}||\mathcal{G}|}{\delta}}{3n}} \\
&\leq \frac{64 H C^2_{\text{cvg}} \bar{D}^2 \log \frac{2|\mathcal{V}||\mathcal{G}|}{\delta}}{3n} + \frac{256 H C^2_{\text{cvg}} \bar{D}^2 \log \frac{2|\mathcal{V}||\mathcal{G}|}{\delta}}{\sqrt{3}n} \\
&= \frac{(64 + 256\sqrt{3}) H C^2_{\text{cvg}} \bar{D}^2 \log \frac{2|\mathcal{V}||\mathcal{G}|}{\delta}}{3n} \\
&\leq \frac{512 H C^2_{\text{cvg}} \bar{D}^2 \log \frac{2|\mathcal{V}||\mathcal{G}|}{\delta}}{3n} \\
&\leq \epsilon_{\text{stat}}
\end{aligned}
\tag{25}
$$

and we can conclude that

$$
\sum_{t=0}^{H-1} \mathbb{E}_{\pi_b}\left[\hat{V}_{\mathcal{H}}(\tau_t) - \left(\mathcal{B}^{\pi}_{\mathcal{H}} V_{\mathcal{F}}\right)(\tau_t)\right]^2 \tag{26}
$$

$$
= \left|\left[\mathbb{E}_{\pi_b} - \mathbb{E}_{\mathcal{D}_b}\right] \sum_{t=0}^{H-1} \left[\hat{V}_{\mathcal{H}}(\tau_t) - \left(\mathcal{B}^{\pi}_{\mathcal{H}} V_{\mathcal{F}}\right)(\tau_t)\right]\left[\hat{V}_{\mathcal{H}}(\tau_t) - 2\mathcal{E}_{\mathcal{F},t} + \left(\mathcal{B}^{\pi}_{\mathcal{H}} V_{\mathcal{F}}\right)(\tau_t)\right]\right| \tag{27}
$$

$$
\leq \epsilon_{\text{stat}}. \tag{28}
$$

Step II. Outer Analysis. Previous result in 26 verifies that for any fixed $V_{\mathcal{F}} \in \mathcal{V}$, the **error between $\hat{V_{\mathcal{H}}}$ and true $(\mathcal{B}^{\pi}_{\mathcal{H}} V_{\mathcal{F}})$ (which is given realizability, equivalent to the statistical error on the empirical mean estimation)**, can be bounded by a constant ϵ_{stat}. We will further leverage this analysis, and extend it to the final error bound in theorem 3.2 where $V_{\mathcal{F}}$ is tailored to $\hat{V_{\mathcal{F}}}$ that minimizes the target loss.

Bounding $\sum_{t=0}^{H-1} \mathbb{E}_{\pi_b}\left[\left(\mathcal{B}^{\pi}_{\mathcal{H}} \hat{V}_{\mathcal{F}}\right)(\tau_t)\right]^2$. Adhering to the PD Lemma 3.1, our analysis follows that

$$
\begin{aligned}
&\left|J(\pi)-\mathbb{E}_{\pi_b}\left[\hat{V}_{\mathcal{F}}(f_0)\right]\right| = \left|\sum_{t=0}^{H-1}\mathbb{E}_{\pi}\left[\left(\mathcal{B}^{\pi}_{\mathcal{S}}\hat{V}_{\mathcal{S}}\right)(s_t)\right]\right| \\
&\le \sqrt{H\sum_{t=0}^{H-1}\mathbb{E}_{\pi_b}\left[\left(\mathcal{B}^{\pi}_{\mathcal{H}}\hat{V}_{\mathcal{F}}\right)(\tau_t)\right]^2}\sqrt{\frac{\sum_{t=0}^{H-1}\mathbb{E}_{\pi}\left[\left(\mathcal{B}^{\pi}_{\mathcal{S}}\hat{V}_{\mathcal{S}}\right)(s_t)\right]^2}{\sum_{t=0}^{H-1}\mathbb{E}_{\pi_b}\left[\left(\mathcal{B}^{\pi}_{\mathcal{H}}\hat{V}_{\mathcal{F}}\right)(\tau_t)\right]^2}} \\
&\le \sqrt{H\sum_{t=0}^{H-1}\mathbb{E}_{\pi_b}\left[\left(\mathcal{B}^{\pi}_{\mathcal{H}}\hat{V}_{\mathcal{F}}\right)(\tau_t)\right]^2}\sqrt{\frac{\sum_{t=0}^{H-1}\mathbb{E}_{\pi}\left[\left(\mathcal{B}^{\pi}_{\mathcal{S}}\hat{V}_{\mathcal{S}}\right)(s_t)\right]^2}{\sum_{t=0}^{H-1}\mathbb{E}_{\pi_b}\left[\left(\mathcal{B}^{\pi}_{\mathcal{S}}\hat{V}_{\mathcal{S}}\right)(s_t)\right]^2}}\sqrt{\frac{\sum_{t=0}^{H-1}\mathbb{E}_{\pi_b}\left[\left(\mathcal{B}^{\pi}_{\mathcal{S}}\hat{V}_{\mathcal{S}}\right)(s_t)\right]^2}{\sum_{t=0}^{H-1}\mathbb{E}_{\pi_b}\left[\left(\mathcal{B}^{\pi}_{\mathcal{H}}\hat{V}_{\mathcal{F}}\right)(\tau_t)\right]^2}} \\
&\le \sqrt{H}\cdot\mathrm{Dr}(\pi\|\pi_b)\cdot\mathrm{IV}(\mathcal{V})\cdot\sqrt{\sum_{t=0}^{H-1}\mathbb{E}_{\pi_b}\left[\left(\mathcal{B}^{\pi}_{\mathcal{H}}\hat{V}_{\mathcal{F}}\right)(\tau_t)\right]^2}
\end{aligned}
\tag{29}
$$

which demonstrates that the OPE error is determined by $\sum_{t=0}^{H-1}\mathbb{E}_{\pi_b}\left[\left(\mathcal{B}^{\pi}_{\mathcal{H}}\hat{V}_{\mathcal{F}}\right)(\tau_t)\right]^2$. And our objective is to finialize it via

$$
\begin{aligned}
\sum_{t\ =0}^{H-1}\mathbb{E}_{\pi_b}\left[\left(\mathcal{B}^{\pi}_{\mathcal{H}}\hat{V}_{\mathcal{F}}\right)(\tau_t)\right]^2 &\le \sum_{t=0}^{H-1}\mathbb{E}_{\pi_b}\left[\hat{\mathcal{E}}^2_{\mathcal{F},t}-\left(\hat{\mathcal{E}}_{\mathcal{F},t}-\left(\mathcal{B}^{\pi}_{\mathcal{H}}\hat{V}_{\mathcal{F}}\right)(\tau_t)\right)^2\right] \\
&\le \left|\sum_{t=0}^{H-1}\left(\mathbb{E}_{\pi_b}-\mathbb{E}_{\mathcal{D}_b}\right)\left[\hat{\mathcal{E}}^2_{\mathcal{F},t}-\left(\hat{\mathcal{E}}_{\mathcal{F},t}-\left(\mathcal{B}^{\pi}_{\mathcal{H}}\hat{V}_{\mathcal{F}}\right)(\tau_t)\right)^2\right]\right| \\
&\quad+\left|\sum_{t=0}^{H-1}\mathbb{E}_{\mathcal{D}_b}\left[\hat{\mathcal{E}}^2_{\mathcal{F},t}-\left(\hat{\mathcal{E}}_{\mathcal{F},t}-\left(\mathcal{B}^{\pi}_{\mathcal{H}}\hat{V}_{\mathcal{F}}\right)(\tau_t)\right)^2\right]\right| \\
&= I_3+I_4.
\end{aligned}
\tag{30}
$$

Bounding I_3. Let $Y_H=\sum_{t=0}^{H-1}\hat{\mathcal{E}}^2_{\mathcal{F},t}-\left(\hat{\mathcal{E}}_{\mathcal{F},t}-\left(\mathcal{B}^{\pi}_{\mathcal{H}}\hat{V}_{\mathcal{F}}\right)(\tau_t)\right)^2=\sum_{t=0}^{H-1}Y_t$. We know that

$$
\begin{aligned}
|Y_H| &\le \sum_{t=0}^{H-1}2\left|\hat{\mathcal{E}}_{\mathcal{F},t}\right|\left|\left(\mathcal{B}^{\pi}_{\mathcal{H}}\hat{V}_{\mathcal{F}}\right)(\tau_t)\right|+\left|\left(\mathcal{B}^{\pi}_{\mathcal{H}}\hat{V}_{\mathcal{F}}\right)(\tau_t)\right|^2 \\
&\le H(2\cdot 3C_{\text{cvg}}\bar{D}\cdot C_{\text{cvg}}\bar{D}+C^2_{\text{cvg}}\bar{D}^2) \\
&= 7HC^2_{\text{cvg}}\bar{D}^2
\end{aligned}
\tag{31}
$$

and

$$
\begin{aligned}
\mathbb{E}_{\pi_b}\left[Y_t^2\right] &= \mathbb{E}_{\pi_b}\left[\left[2\hat{\mathcal{E}}_{\mathcal{F},t}-\left(\mathcal{B}^{\pi}_{\mathcal{H}}\hat{V}_{\mathcal{F}}\right)(\tau_t)\right]^2\left(\mathcal{B}^{\pi}_{\mathcal{H}}\hat{V}_{\mathcal{F}}\right)(\tau_t)^2\right] \\
&\le \left[4\cdot 9C^2_{\text{cvg}}\bar{D}^2+C^2_{\text{cvg}}\bar{D}^2+4\cdot 3C_{\text{cvg}}\bar{D}\cdot C_{\text{cvg}}\bar{D}\right]\mathbb{E}_{\pi_b}\left[\left(\mathcal{B}^{\pi}_{\mathcal{H}}\hat{V}_{\mathcal{F}}\right)(\tau_t)\right]^2 \\
&= 49C^2_{\text{cvg}}\bar{D}^2\cdot\mathbb{E}_{\pi_b}\left[\left(\mathcal{B}^{\pi}_{\mathcal{H}}\hat{V}_{\mathcal{F}}\right)(\tau_t)\right]^2
\end{aligned}
\tag{32}
$$

such that

$$
\mathrm{Var}_{\pi_b}\left[Y_H\right]\le\mathbb{E}_{\pi_b}\left[Y_H^2\right]\le H\sum_{t=0}^{H-1}\mathbb{E}_{\pi_b}\left[Y_t^2\right]=49HC^2_{\text{cvg}}\bar{D}^2\cdot\mathbb{E}_{\pi_b}\left[\left(\mathcal{B}^{\pi}_{\mathcal{H}}\hat{V}_{\mathcal{F}}\right)(\tau_t)\right]^2.
$$

Applying Bernstein's we obtain that

$$
\begin{aligned}
I_3 &\leq \sqrt{2 \cdot 49 C_{\text{cvg}}^2 \bar{D}^2 \cdot \mathbb{E}_{\pi_b}\left[\left(\mathcal{B}_{\mathcal{H}}^{\pi} \hat{V}_{\mathcal{F}}\right)(\tau_t)\right]^2 \cdot \frac{\log \frac{2|\mathcal{V}||\mathcal{G}|}{\delta}}{n}} + \frac{2}{3} \cdot 14 H C_{\text{cvg}}^2 \bar{D}^2 \frac{\log \frac{2|\mathcal{V}||\mathcal{G}|}{\delta}}{n} \\
&\leq \sqrt{2 \cdot 49 H C_{\text{cvg}}^2 \bar{D}^2 \cdot \mathbb{E}_{\pi_b}\left[\left(\mathcal{B}_{\mathcal{H}}^{\pi} \hat{V}_{\mathcal{F}}\right)(\tau_t)\right]^2 \cdot \frac{\log \frac{2|\mathcal{V}||\mathcal{G}|}{\delta}}{n}} + \frac{2}{3} \cdot 14 H C_{\text{cvg}}^2 \bar{D}^2 \frac{\log \frac{2|\mathcal{V}||\mathcal{G}|}{\delta}}{n}
\end{aligned} \tag{33}
$$

Bounding I_4. From 26 it can be verified that

$$
\begin{aligned}
&\left|\sum_{t=0}^{H-1} \mathbb{E}_{\mathcal{D}_b}\left[\hat{V}_{\mathcal{F}}(\tau_t) - \hat{\mathcal{E}}_{\mathcal{F},t}\right]^2 - \mathbb{E}_{\mathcal{D}_b}\left[\left(\mathcal{B}_{\mathcal{H}}^{\pi} \hat{V}_{\mathcal{F}}\right)(\tau_t) - \hat{\mathcal{E}}_{\mathcal{F},t}\right]^2\right| \\
&\leq \left|\sum_{t=0}^{H-1} \mathbb{E}_{\pi_b}\left[\hat{V}_{\mathcal{F}}(\tau_t) - \hat{\mathcal{E}}_{\mathcal{F},t}\right]^2 - \mathbb{E}_{\pi_b}\left[\left(\mathcal{B}_{\mathcal{H}}^{\pi} \hat{V}_{\mathcal{F}}\right)(\tau_t) - \hat{\mathcal{E}}_{\mathcal{F},t}\right]^2\right| + 2\epsilon_{\text{stat}} \\
&= \sum_{t=0}^{H-1} \mathbb{E}_{\pi_b}\left[\left(\mathcal{B}_{\mathcal{H}}^{\pi} \hat{V}_{\mathcal{F}}\right)(\tau_t) - \hat{V}_{\mathcal{F}}(\tau_t)\right]^2 + 2\epsilon_{\text{stat}} \\
&\leq 3\epsilon_{\text{stat}}.
\end{aligned} \tag{34}
$$

Consequently, we can bound I_4 via

$$
\begin{aligned}
&\sum_{t\ =0}^{H-1} \mathbb{E}_{\mathcal{D}_b}\left[\hat{\mathcal{E}}_{\mathcal{F},t}^2 - \left(\hat{\mathcal{E}}_{\mathcal{F},t} - \left(\mathcal{B}_{\mathcal{H}}^{\pi} \hat{V}_{\mathcal{F}}\right)(\tau_t)\right)^2\right] \\
&\leq \sum_{t=0}^{H-1} \mathbb{E}_{\mathcal{D}_b}\left[\hat{\mathcal{E}}_{\mathcal{F},t}^2 - \left(\hat{\mathcal{E}}_{\mathcal{F},t} - \hat{V}_{\mathcal{F}}(\tau_t)\right)^2\right] + 3\epsilon_{\text{stat}} \\
&\leq \sum_{t=0}^{H-1} \mathbb{E}_{\mathcal{D}_b}\left[\hat{\mathcal{E}}_{\mathcal{F},t}^2 - \left(\hat{\mathcal{E}}_{\mathcal{F},t} - \left(\mathcal{B}_{\mathcal{H}}^{\pi} \hat{V}_{\mathcal{F}}\right)(\tau_t)\right)^2\right] + 6\epsilon_{\text{stat}} \\
&= 6\epsilon_{\text{stat}}
\end{aligned} \tag{35}
$$

Final Guarantee. Leveraging the aforementioned results, it can be derived that

$$
\begin{aligned}
\sum_{t\ =0}^{H-1} \mathbb{E}_{\pi_b}\left[\left(\mathcal{B}_{\mathcal{H}}^{\pi} \hat{V}_{\mathcal{F}}\right)(\tau_t)\right]^2 &\leq \sqrt{2 \cdot 49 H C_{\text{cvg}}^2 \bar{D}^2 \cdot \mathbb{E}_{\pi_b}\left[\left(\mathcal{B}_{\mathcal{H}}^{\pi} \hat{V}_{\mathcal{F}}\right)(\tau_t)\right]^2 \cdot \frac{\log \frac{2|\mathcal{V}||\mathcal{G}|}{\delta}}{n}} + \frac{2}{3} \\
&\quad \cdot 14 H C_{\text{cvg}}^2 \bar{D}^2 \frac{\log \frac{2|\mathcal{V}||\mathcal{G}|}{\delta}}{n} + \frac{1024 H C_{\text{cvg}}^2 \bar{D}^2 \log \frac{2|\mathcal{V}||\mathcal{G}|}{\delta}}{n} \\
&\leq \sqrt{2 \cdot 49 H C_{\text{cvg}}^2 \bar{D}^2 \cdot \mathbb{E}_{\pi_b}\left[\left(\mathcal{B}_{\mathcal{H}}^{\pi} \hat{V}_{\mathcal{F}}\right)(\tau_t)\right]^2 \cdot \frac{\log \frac{2|\mathcal{V}||\mathcal{G}|}{\delta}}{n}} \\
&\quad + \frac{1034 H C_{\text{cvg}}^2 \bar{D}^2 \log \frac{2|\mathcal{V}||\mathcal{G}|}{\delta}}{n}
\end{aligned} \tag{36}
$$

Solving this inequality we can finalize the OPE error via

$$
\begin{aligned}
\left|J(\pi) - \mathbb{E}_{\pi_b}\left[\hat{V}_{\mathcal{F}}(f_0)\right]\right| &\leq \sqrt{H} \cdot \mathrm{Dr}(\pi \| \pi_b) \cdot \mathrm{IV}(\mathcal{V}) \cdot \sqrt{\sum_{t=0}^{H-1} \mathbb{E}_{\pi_b}\left[\left(\mathcal{B}_{\mathcal{H}}^{\pi} \hat{V}_{\mathcal{F}}\right)(\tau_t)\right]^2} \\
&\leq \sqrt{H} \cdot \mathrm{Dr}(\pi \| \pi_b) \cdot \mathrm{IV}(\mathcal{V}) \sqrt{\frac{833 H C_{\text{cvg}}^2 \bar{D}^2 \log \frac{2|\mathcal{V}||\mathcal{G}|}{\delta}}{n}} \\
&\leq 29 H \cdot \mathrm{Dr}(\pi \| \pi_b) \cdot \mathrm{IV}(\mathcal{V}) \cdot C_{\text{cvg}} \bar{D} \cdot \frac{\log \frac{2|\mathcal{V}||\mathcal{G}|}{\delta}}{n}.
\end{aligned} \tag{37}
$$

□