Regularization and Variance-Weighted Regression Achieves Minimax Optimality in Linear MDPs: Theory and Practice

Toshinori Kitamura [1] **Tadashi Kozuno** [2] **Yunhao Tang** [3] **Nino Vieillard** [4] **Michal Valko** [3] **Wenhao Yang** [5] **Jincheng Mei** [4] **Pierre Ménard** [6] **Mohammad Gheshlaghi Azar** [3] **Rémi Munos** [3] **Olivier Pietquin** [4] **Matthieu Geist** [4] **Csaba Szepesvári** [7 3] **Wataru Kumagai** [1] **Yutaka Matsuo** [1]

Abstract

Mirror descent value iteration (MDVI), an abstraction of Kullback–Leibler (KL) and entropy-regularized reinforcement learning (RL), has served as the basis for recent high-performing practical RL algorithms. However, despite the use of function approximation in practice, the theoretical understanding of MDVI has been limited to tabular Markov decision processes (MDPs). We study MDVI with linear function approximation through its sample complexity required to identify an ε-optimal policy with probability $1 - \delta$ under the settings of an infinite-horizon linear MDP, generative model, and G-optimal design. We demonstrate that least-squares regression weighted by the variance of an estimated optimal value function of the next state is crucial to achieving minimax optimality. Based on this observation, we present Variance-Weighted Least-Squares MDVI (VWLS-MDVI), the first theoretical algorithm that achieves nearly minimax optimal sample complexity for infinite-horizon linear MDPs. Furthermore, we propose a practical VWLS algorithm for value-based deep RL, Deep Variance Weighting (DVW). Our experiments demonstrate that DVW improves the performance of popular value-based deep RL algorithms on a set of MinAtar benchmarks.

[1]The University of Tokyo, Japan [2]OMRON SINIC X, Japan [3]DeepMind [4]Google Research, Brain team [5]Peking University [6]Otto von Guericke University Magdeburg [7]University of Alberta. Correspondence to: Toshinori Kitamura <toshinori-k@weblab.t.u-tokyo.ac.jp>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

1. Introduction

Kullback–Leibler (KL) divergence and entropy regularization play an important role in recent reinforcement learning (RL) algorithms. These regularizations are often introduced to promote exploration (Haarnoja et al., 2017; 2018), make algorithms more robust to errors (Husain et al., 2021; Bellemare et al., 2016), and ensure that performance improves over time (Schulman et al., 2015). The behavior of RL algorithms under these regularizations can be studied using mirror descent value iteration (MDVI; Geist et al. (2019)), a value iteration algorithm that incorporates KL and entropy regularization in its value and policy updates. Notably, when both regularizations are combined, MDVI is proven to achieve nearly minimax optimal sample complexity[1] with the generative model (simulator) in infinite-horizon MDPs, which indicates that it can exhibit good performance with relatively few samples (Kozuno et al., 2022). This analysis supports the state-of-the-art performance of the recent Munchausen DQN (M-DQN, Vieillard et al. (2020b)), which is a natural extension of MDVI to a value-based deep RL algorithm.

However, the minimax optimality of MDVI has only been proven for tabular Markov decision processes (MDPs), and does not consider the challenge of generalization in RL. As practical RL algorithms often use function approximators to obtain generalizability, this leads to a natural question: *Is MDVI minimax optimal with function approximation?* The answer to this question should reveal room for improvement in existing practical MDVI-based algorithms such as M-DQN. This study addresses the question by investigating the sample complexity of a model-free infinite-horizon (ε, δ)-PAC RL algorithm, i.e., the expected number of calls to the generative model to identify an ε-optimal policy with a failure probability less than δ, under the assumptions of linear MDP (Jin et al., 2020), access to all the state-action pairs with a generative model, and a G-optimal design (Lattimore et al., 2020). Intuitively, these assumptions allow

[1]We only study the sample complexity (number of calls to a generative model) and ignore the computational complexity (total number of logical and arithmetic operations that the agent uses).

us to focus on the value update rule, which is the core of RL algorithms, based on the following mechanisms; the access to all the state-action pairs with the generative model removes difficulties of exploration, the linear MDP provides a good representation, and the G-optimal design provides access to an effective dataset. We explain in Section 2 why the study of infinite-horizon RL is of value.

In Section 4, we provide positive and negative answers to the aforementioned question. We demonstrate that a popular method for extending tabular algorithms to function approximation, i.e., regressing the target value with least-squares (Bellman et al., 1963; Munos, 2005), can result in sub-optimal sample complexity in MDVI. This suggests that in the case of function approximation, algorithms such as M-DQN, which rely mainly on the power of regularization, may exhibit a sub-optimal performance in terms of sample complexity. However, we confirm that MDVI achieves nearly minimax optimal sample complexity when the least-squares regression is weighted by the variance of the optimal value function of the next state. We prove these scenarios using our novel proof tool, the *weighted Kiefer–Wolfowitz* (KW) theorem, which allows us to use the total variance (TV) technique (Azar et al., 2013) to provide a $\sqrt{(1-\gamma)^{-1}}$ tighter performance bound than the vanilla KW theorem (Kiefer & Wolfowitz, 1960; Lattimore et al., 2020), where γ denotes the discount factor.

Based on the theoretical observations, we propose both theoretical and practical algorithms; a minimax optimal extension of MDVI to infinite-horizon linear MDPs, called Variance-Weighted Least-Squares MDVI (`VWLS-MDVI`, Section 5), and a practical weighted regression algorithm for value-based deep RL, called Deep Variance Weighting (`DVW`, Section 6). `VWLS-MDVI` is the first-ever algorithm with nearly minimax sample complexity under the setting of both model-based and model-free infinite-horizon linear MDPs. `DVW` is also the first algorithm that extends the minimax optimal theory of function approximation to deep RL. Our experiments demonstrate the effectiveness of `DVW` to value-based deep RL through an environment where we can compute oracle values (Section 7.2.1) and a set of MinAtar benchmarks (Young & Tian (2019), Section 7.2.2).

2. Related Work

Minimax Infinite-Horizon RL with Linear Function Approximation. The development of minimax optimal RL with linear function approximation has significantly advanced in recent years owing to the study of Zhou et al. (2021). Zhou et al. (2021) proposed the Bernstein-type *self-normalized concentration inequality* (Abbasi-Yadkori et al., 2011) and combined it with variance-weighted regression (VWR) to achieve minimax optimal regret bound for linear mixture MDPs. Then, Hu et al. (2022) and He et al. (2022) built upon the VWR technique for linear MDPs to achieve minimax optimality. VWR has also been used for tight analyses in offline RL (Yin et al., 2022b; Xiong et al., 2022), off-policy policy evaluation (Min et al., 2021), and RL with nonlinear function approximation (Yin et al., 2022c; Agarwal et al., 2022).

Table 1. Sample complexity comparison to find an ε-optimal policy under infinite-horizon Linear MDP. In the table, d denotes the dimension of a linear MDP and γ denotes the discount factor .

Algorithm (Publication)	Complexity
G-Sampling-and-Stop (Taupin et al., 2022)	$\widetilde{\mathcal{O}}\left(\frac{d^2}{\varepsilon^2(1-\gamma)^4}\right)$
`VWLS-MDVI` (proposed in this study)	$\widetilde{\mathcal{O}}\left(\frac{d^2}{\varepsilon^2(1-\gamma)^3}\right)$
Lower Bound (Weisz et al., 2022)	$\Omega\left(\frac{d^2}{\varepsilon^2(1-\gamma)^3}\right)$

Despite the development of minimax optimal RL with linear function approximation, their results are limited to the setting of finite-horizon episodic MDPs. However, in practical RL applications, it is not uncommon to encounter infinite horizons, as can be observed in robotics (Miki et al., 2022), recommendation (Maystre et al., 2023), and industrial automation (Zhan et al., 2022). Additionally, many practical deep RL algorithms, such as DQN (Mnih et al., 2015) and SAC (Haarnoja et al., 2018), are designed as model-free algorithms for the infinite-horizon discounted MDPs. Despite the practical importance of this topic, the minimax optimal algorithm for infinite-horizon discounted linear MDPs was unknown until this study. Our study not only developed the first minimax optimal algorithm but also became the first study to naturally extend it to a practical deep RL algorithm.

Generative Model Assumption. In the infinite-horizon setting, the assumption of a generative model is not uncommon because, in contrast to the finite-horizon episodic setting, the environment cannot be reset, rendering exploration difficult (Azar et al., 2013; Sidford et al., 2018; Agarwal et al., 2020). In fact, efficient learning in the infinite-horizon setting without the generative model is believed to be achievable only when an MDP has a finite diameter (Jaksch et al., 2010).

The problem setting of our theory, where the generative model can be queried for any state-action pair, is known as *random access* generative model setting. For this setting, Lattimore et al. (2020) and Taupin et al. (2022) provided infinite-horizon sample-efficient algorithms with a G-optimal design; however, their sample complexity is not minimax optimal. Yang & Wang (2019) proposed an algorithm with minimax optimal sample complexity for infinite-horizon MDPs; however, their algorithm relies on the special MDP structure, called anchor state-action pairs, as input to the algorithm. In contrast, the proposed VWLS-MDVI algorithm can be executed as long as we have access to all state-action pairs. Comparison of sample complexity with that of previous algorithms for infinite-horizon Linear MDPs is summarized in Table 1.

Computational Complexity. Unfortunately, the computational complexity of algorithms using a G-optimal design, including our theoretical algorithm, can be inefficient (Lattimore et al., 2020). This issue is addressed by extending the problem setting to more practical scenarios, e.g., *local access*, where the agent can query to the generative model only previously visited state-action pairs (Yin et al., 2022a; Weisz et al., 2022), or online RL. We empirically address the issue by proposing the practical VWR algorithm, i.e., DVW, and demonstrate its effectiveness in an online RL setting. Unlike previous practical algorithms that utilize weighted regression (Schaul et al., 2015; Kumar et al., 2020; Lee et al., 2021), the proposed DVW possesses a theoretical background of statistical efficiency. We leave theoretical extensions to wider problem settings as future works.

3. Preliminaries

For a set $\mathcal{S}$, we denote its complement and its size by $\mathcal{S}^c$ and $|\mathcal{S}|$, respectively. For $N \in \mathbb{N}$, let $[N] := \{1 \dots N\}$. For a measurable space, say $(\mathcal{S}, \mathfrak{F})$, the set of probability measures over $(\mathcal{S}, \mathfrak{F})$ is denoted by $\Delta(\mathcal{S}, \mathfrak{F})$ or $\Delta(\mathcal{S})$ when the σ-algebra is clear from the context. $\mathbb{E}[X]$ and $\mathbb{V}[X]$ denotes the expectation and variance of a random variable X, respectively. The empty sum is defined to be 0, e.g., $\sum_{i=j}^{k} c_i = 0$ if $j > k$.

We consider an infinite-horizon discounted MDP defined by $(\mathcal{X}, \mathcal{A}, \gamma, r, P)$, where $\mathcal{X}$ denotes the state space, $\mathcal{A}$ denotes finite action space with size A, $\gamma \in [0, 1)$ denotes the discount factor, $r : \mathcal{X} \times \mathcal{A} \to [-1, 1]$ denotes the reward function, and $P : \mathcal{X} \times \mathcal{A} \to \Delta(\mathcal{X})$ denotes the state-transition probability kernel. We denote the sets of all bounded Borel-measurable functions over $\mathcal{X}$ and $\mathcal{X} \times \mathcal{A}$ by $\mathcal{F}_v$ and $\mathcal{F}_q$, respectively. Let H be the (effective) time horizon $(1-\gamma)^{-1}$. For both $\mathcal{F}_v$ and $\mathcal{F}_q$, let $\mathbf{0}$ and $\mathbf{1}$ denote functions that output zero and one everywhere, respectively. Whether $\mathbf{0}$ and $\mathbf{1}$ are defined in $\mathcal{F}_v$ or $\mathcal{F}_q$ shall be clear from the context. All the scalar operators and inequalities applied to $\mathcal{F}_v$ and $\mathcal{F}_q$ should be understood point-wise.

With an abuse of notation, let P be an operator from $\mathcal{F}_q$ to $\mathcal{F}_v$ such that $(Pv)(x, a) = \int v(y) P(dy|x, a)$ for any $v \in \mathcal{F}_v$. A policy is a probability kernel over $\mathcal{A}$ conditioned on $\mathcal{X}$. For any policy π and $q \in \mathcal{F}_q$, let π be an operator from $\mathcal{F}_v$ to $\mathcal{F}_q$ such that $(\pi q)(x) = \sum_{a \in \mathcal{A}} \pi(a|x) q(x, a)$. We adopt a shorthand notation, i.e., $P_\pi := P\pi$. We define the Bellman operator T_π for a policy π as $T_\pi q := r + \gamma P_\pi q$, which has the unique fixed point, i.e., q_π. The state-value function v_π is defined as πq_π. An optimal policy π_* is a policy such that $v_* := v_{\pi_*} \geq v_\pi$ for any policy π, where the inequality is point-wise.

3.1. Tabular MDVI

To better understand the motivation of our theorems for function approximation, we provide a background on Tabular MDVI of Kozuno et al. (2022).

3.1.1. Tabular MDVI Algorithm

For any policies π and μ, let $\mathcal{H}(\pi) := -\pi \log \pi \in \mathcal{F}_v$ be the entropy of π and $\mathrm{KL}(\pi \| \mu) := \pi \log \frac{\pi}{\mu} \in \mathcal{F}_v$ be the KL divergence of π and μ. For all $(x, a) \in \mathcal{X} \times \mathcal{A}$, the update rule of Tabular MDVI is written as follows:

$$\begin{aligned} & q_{k+1} = r + \gamma \widehat{P}_k(M) v_k \ , \\ \text{where} \ \ & v_k = \pi_k q_k - \tau \mathrm{KL}(\pi_k \| \pi_{k-1}) + \kappa \mathcal{H}(\pi_k) \ , \\ & \pi_k(a|x) \propto \pi_{k-1}(a|x)^\alpha \exp(\beta q_k(x, a)) \ . \end{aligned} \tag{1}$$

Here, we define $\alpha := \tau/(\tau + \kappa)$ and $\beta := 1/(\tau + \kappa)$. Furthermore, let $\widehat{P}_k(M) v_k : (x, a) \mapsto \frac{1}{M} \sum_{m=1}^{M} v_k(y_{k,m,x,a})$ where $(y_{k,m,x,a})_{m=1}^{M}$ are $M \in \mathbb{N}$ samples obtained from the generative model $P(\cdot|x, a)$ at the k th iteration.

Similar to Kozuno et al. (2022), we use the idea of the non-stationary policy (Scherrer & Lesner, 2012) to provide a tight analysis. For a sequence of policies $(\pi_k)_{k \in \mathbb{Z}}$, let $P_j^i := P_{\pi_i} P_{\pi_{i-1}} \cdots P_{\pi_{j+1}} P_{\pi_j}$ for $i \geq j$, otherwise let $P_j^i := I$. As a special case with $\pi_k = \pi_*$ for all k, let $P_*^i := (P_{\pi_*})^i$. Moreover, for a sequence of policies $(\pi_k)_{k=0}^{K}$, let π'_k be the non-stationary policy that follows π_{k-t} at the t-th time step until $t = k$, after which π_0 is followed.[2] The value function of such a non-stationary policy is given by $v_{\pi'_k} = \pi_k T_{\pi_{k-1}} \cdots T_{\pi_1} q_{\pi_0}$. While not covered in this work, we anticipate that our main results remain valid for the last policy case, at the expense of the range of valid ε, by extending the analysis of Kozuno et al. (2022).

3.1.2. Techniques to Minimax Optimality

The key to achieving the minimax optimality of Tabular MDVI is combining the *averaging property* (Vieillard et al., 2020a) and *TV* technique (Azar et al., 2013).

Averaging Property. Let $s_k := \sum_{j=0}^{k-1} \alpha^j q_{k-j}$ be the moving average of past q-functions and w_k be the function $x \mapsto \beta^{-1} \log \sum_{a \in \mathcal{A}} \exp(\beta s_k(x, a))$ over $\mathcal{X}$. Then, the update (1) can be rewritten as (derivation in Appendix B):

$$q_{k+1} = r + \gamma \widehat{P}_k(M) v_k \ , \tag{2}$$

where $v_k = w_k - \alpha w_{k-1}$, and $\pi_k(a|x) \propto \exp(\beta s_k(x, a))$. To simplify the analysis, we consider the limit of $\tau, \kappa \to 0$

[2] The time step index t starts from 0.

while keeping $\tau/(\tau+\kappa)$ constant. This limit corresponds to letting $\beta \to \infty$, letting $w_k : x \mapsto \max_{a\in\mathcal{A}} s_k(x,a)$ over $\mathcal{X}$, and having π_k be greedy with respect to s_k[3].

Intuitively, s_k, i.e., the moving average of past q-values, averages past errors caused during the update. Kozuno et al. (2022) confirmed that this allows Azuma–Hoeffding inequality (Lemma D.1) to provide a tighter upper bound of $\|v_* - v_{\pi'_k}\|_\infty$ than that in the absence of averaging, where errors appear as a sum of the norms (Vieillard et al., 2020a). We provide the pseudocode of Tabular MDVI with (2) in Appendix K.

Total Variance Technique. The TV technique is a common theoretical technique used to sharpen the upper bound of $\|v_* - v_{\pi'_k}\|_\infty$ (referred to as the performance bound in this study). For any $v \in \mathcal{F}_v$, let $\mathrm{Var}(v)$ be the "variance" function.

$$\mathrm{Var}(v) : (x,a) \mapsto (Pv^2)(x,a) - (Pv)^2(x,a)\ .$$

We often write $\sqrt{\mathrm{Var}(v)}$ as $\sigma(v)$. For a discounted sum of variances of policy values, the TV technique provides the following bound (the corollary follows from Lemma E.2):

Corollary 3.1. *Let* $\heartsuit_k^{TV} := \sum_{j=0}^{k-1} \gamma^j \pi_k P_{k-j}^{k-1} \sigma(v_{\pi_{k-j}})$ *and* $\clubsuit_k^{TV} := \sum_{j=0}^{k-1} \gamma^j \pi_* P_*^j \sigma(v_*)$. *For any* $k \in [K]$ *in Tabular MDVI,* $\heartsuit_k^{TV} \le \sqrt{2H^3}\mathbf{1}$ *and* $\clubsuit_k^{TV} \le \sqrt{2H^3}\mathbf{1}$.

Kozuno et al. (2022) used this TV technique to improve the performance bound of Tabular MDVI. As $\sigma(v_{\pi_{k-j}}) \le H$ and $\sigma(v_*) \le H$ due to Lemma D.5, the TV technique provides approximately $\sqrt{H}$ tighter bound than the naive bounds of $\heartsuit_k^{\mathrm{TV}} \le H^2\mathbf{1}$ and $\clubsuit_k^{\mathrm{TV}} \le H^2\mathbf{1}$. This leads to $\sqrt{H}$ better performance bound.

3.2. Linear MDP and G-Optimal Design

We assume access to a good feature representation with which an MDP is linear (Jin et al., 2020).

Assumption 3.2 (Linear MDP)**.** Suppose an MDP $\mathcal{M}$ with the state-action space $\mathcal{X}\times\mathcal{A}$. We have access to a known feature map $\phi : \mathcal{X}\times\mathcal{A} \to \mathbb{R}^d$ that satisfies the following condition: there exist a vector $\psi \in \mathbb{R}^d$ and d (signed) measures $\mu := (\mu_1, \dots, \mu_d)$ on $\mathcal{X}$ such that $P(\cdot|x,a) = \phi(x,a)^\top \mu$ for any $(x,a) \in \mathcal{X}\times\mathcal{A}$, and $r = \phi^\top \psi$. Let $\Phi := \{\phi(x,a) : (x,a) \in \mathcal{X}\times\mathcal{A}\} \subset \mathbb{R}^d$ be the set of all feature vectors. We assume that Φ is compact and spans $\mathbb{R}^d$.

A crucial property of the linear MDP is that, for any policy π, q_π is always linear in the feature map ϕ (Jin et al., 2020). The compactness and span assumptions of Φ are made for the purpose of constructing a G-optimal design later on.

[3] Even if β is finite, the minimax optimality holds as long as β is sufficiently large (**Remark 1** in Kozuno et al. (2022)).

Furthermore, we assume access to a good finite subset of $\mathcal{X}\times\mathcal{A}$ called a core set $\mathcal{C}$. The key properties of the core set are that it has a few elements while $\{\phi(y,b) : (y,b) \in \mathcal{C}\}$ provides a "good coverage" of the feature space in the sense that we describe now. For a distribution ρ over $\mathcal{X}\times\mathcal{A}$, let $G \in \mathbb{R}^{d\times d}$ and $g(\rho) \in \mathbb{R}$ be defined by

$$\begin{aligned} G &:= \sum_{(y,b)\in\mathcal{C}} \rho(y,b)\phi(y,b)\phi(y,b)^\top \\ \text{and}\ \ g(\rho) &:= \max_{(x,a)\in\mathcal{X}\times\mathcal{A}} \phi(x,a)^\top G^{-1} \phi(x,a)\ , \end{aligned} \tag{3}$$

respectively. We denote ρ as the design, G as the design matrix underlying ρ, and $\mathcal{C} := \mathrm{Supp}(\rho)$ as the support of ρ, which we denote as the core set of ρ. The problem of finding a design that minimizes g is known as the *G-optimal design* problem. The Kiefer–Wolfowitz (KW) theorem (Kiefer & Wolfowitz, 1960) states the optimal design ρ_* must satisfy $g(\rho_*) = d$. Furthermore, the following theorem shows that there exists a near-optimal design with a small core set for Φ. The proof is provided in Appendix F.

Theorem 3.3. *Let* $u_\mathcal{C} := 4d\log\log(d+4) + 28$. *For* Φ *satisfying Assumption 3.2, there exists a design* ρ *such that* $g(\rho) \le 2d$ *and the core set of* ρ *has size at most* $u_\mathcal{C}$.

4. MDVI with Linear Function Approximation

In this section, we provide essential components to extend MDVI from tabular to linear with minimax optimality. To illustrate how linear MDVI fails or succeeds in attaining minimax optimality, we begin by introducing the general algorithm, called Weighted Least-Squares MDVI (`WLS-MDVI`).

4.1. Weighted Least-Squares MDVI Algorithm

Let $q_k(x,a) := \phi^\top(x,a)\theta_k$ be the linearly parameterized value function using the basis function $\theta_k \in \mathbb{R}^d$. For this q_k, the moving average of past q-values can be implemented as

$$s_k := \phi^\top \overline{\theta}_k \text{ where } \overline{\theta}_k = \theta_k + \alpha\overline{\theta}_{k-1}\ .$$

Using these q_k and s_k, let w_k, v_k, and the policy π_k be the same as those of Section 3.1.2. Given a bounded positive weighting function $f : \mathcal{X}\times\mathcal{A} \mapsto (0,\infty)$, we learn θ_k based on weighted least-squares regression.

$$\begin{aligned} \theta_k &= \operatorname*{arg\,min}_{\theta\in\mathbb{R}^d} \sum_{(y,b)\in\mathcal{C}_f} \frac{\rho_f(y,b)}{f^2(y,b)} \left(\phi^\top(y,b)\theta - \widehat{q}_k(y,b)\right)^2\ , \\ &\text{where}\ \ \widehat{q}_k(y,b) = r(y,b) + \gamma\widehat{P}_{k-1}(M)v_{k-1}(y,b)\ . \end{aligned} \tag{4}$$

Here, ρ_f is a design over $\mathcal{X}\times\mathcal{A}$ and $\mathcal{C}_f := \mathrm{Supp}(\rho_f)$ is a core set of ρ_f. When $f = \mathbf{1}$, we recover the vanilla least-squares regression (Bellman et al., 1963; Munos, 2005), which is a common strategy in practice. We call this algorithm `WLS-MDVI`. The next section presents our novel theoretical tool to provide minimax sample complexity.

4.2. Weighted Kiefer–Wolfowitz Theorem

Let $\theta_k^* \in \mathbb{R}^d$ be the oracle parameter satisfying $\phi^\top \theta_k^* = r + \gamma P v_{k-1}$. θ_k^* is ensured to exist by the property of linear MDPs. To derive the sample complexity, we need a bound of the regression errors outside the core set $\mathcal{C}_f$, i.e., $\|\phi^\top(\theta_k - \theta_k^*)\|_\infty$. Lattimore et al. (2020) derived such a bound using Theorem 3.3.

Instead of the vanilla G-optimal design, we consider the following *weighted* design with a bounded positive function $f : \mathcal{X} \times \mathcal{A} \mapsto (0, \infty)$. For a design ρ over $\mathcal{X} \times \mathcal{A}$, let $G_f \in \mathbb{R}^{d\times d}$ and $g_f(\rho) \in \mathbb{R}$ be defined by

$$G_f := \sum_{(y,b)\in\mathcal{C}_f} \rho(y,b)\frac{\phi(y,b)\phi(y,b)^\top}{f(y,b)^2} \,, \quad \text{and } \; g_f(\rho) := \max_{(y,b)\in\mathcal{X}\times\mathcal{A}} \frac{\phi(y,b)^\top G_f^{-1}\phi(y,b)}{f(y,b)^2} \,, \tag{5}$$

respectively. Equation (5) is the weighted generalization of Equation (3) with ϕ scaled by $1/f$. For this weighted optimal design, we derived the weighted KW theorem, which almost immediately follows from Theorem 3.3 by considering a weighted feature map $\phi_f : (x,a) \mapsto \phi(x,a)/f(x,a)$.

Theorem 4.1 (Weighted KW Theorem)**.** *For Φ satisfying Assumption 3.2, there exists a design ρ_f such that $g_f(\rho_f) \leq 2d$ and the core set of ρ_f has size at most $u_\mathcal{C}$.*

Such the design under Assumption 3.2 with finite $\mathcal{X}$ can be obtained using the Frank-Wolfe algorithm of **Lemma 3.9** mentioned in Todd (2016). We provide the pseudocode of Frank-Wolfe algorithm in Appendix K. We assume that we have access to the weighted optimal design in constructing our theory:

Assumption 4.2 (Weighted Optimal design)**.** There is an oracle called `ComputeOptimalDesign` that accepts a bounded positive function $f : \mathcal{X} \times \mathcal{A} \mapsto (0, \infty)$ and returns ρ_f, $\mathcal{C}_f$, and G_f as in Theorem 4.1.

Combined with this `ComputeOptimalDesign`, we provide the pseudocode of `WLS-MDVI` in Algorithm 1. The weighted KW theorem yields the following bound on the optimal design. The proof can be found in Appendix G.

Lemma 4.3 (Weighted KW Bound)**.** *Let $f : \mathcal{X} \times \mathcal{A} \mapsto (0, \infty)$ be a positive function and z be a function defined over $\mathcal{C}_f$. Then, there exists $\rho_f \in \Delta(\mathcal{X} \times \mathcal{A})$ with a finite support $\mathcal{C}_f := \mathrm{Supp}(\rho_f)$ of size less than or equal to $u_\mathcal{C}$ such that*

$$|\phi^\top W(f,z)| \leq \sqrt{2d} f \max_{(y',b')\in\mathcal{C}_f} \left| \frac{z(y',b')}{f(y',b')} \right| \,,$$

where $W(f,z) := G_f^{-1} \sum_{(y,b)\in\mathcal{C}_f} \frac{\rho_f(y,b)\phi(y,b)z(y,b)}{f^2(y,b)}$.

4.3. Sample Complexity of `WLS-MDVI`

Lemma 4.3 helps derive the sample complexity of `WLS-MDVI`. Let ε_k be the sampling error for v_{k-1} and E_k be its moving average:

$$\varepsilon_k : (x,a) \mapsto \gamma\left(\widehat{P}_{k-1}(M)v_{k-1} - Pv_{k-1}\right)(x,a)$$
$$\text{and } \; E_k : (x,a) \mapsto \sum_{j=1}^{k} \alpha^{k-j}\varepsilon_j(x,a) \,.$$

Furthermore, for any non-negative integer k, let $A_{\gamma,k} := \sum_{j=0}^{k-1}\gamma^{k-j}\alpha^j$, $A_k := \sum_{j=0}^{k-1}\alpha^j$, and $A_\infty := 1/(1-\alpha)$. Then, the performance bound of `WLS-MDVI` is derived as

$$|v_* - v_{\pi'_k}| \leq \frac{\sqrt{2d}}{A_\infty}\left(\heartsuit_k^{\text{wls}} + \clubsuit_k^{\text{wls}}\right) + \diamondsuit_k \,, \tag{6}$$
$$\text{where } \; \heartsuit_k^{\text{wls}} := \sum_{j=0}^{k-1}\gamma^j \pi_k P_{k-j}^{k-1}\left|\max_{(y,b)\in\mathcal{C}_f}\frac{E_{k-j}(y,b)}{f(y,b)}\right| f$$
$$\text{and } \; \clubsuit_k^{\text{wls}} := \sum_{j=0}^{k-1}\gamma^j \pi_* P_*^j\left|\max_{(y,b)\in\mathcal{C}_f}\frac{E_{k-j}(y,b)}{f(y,b)}\right| f \,.$$

Here, $\diamondsuit_k := 2H\left(\alpha^k + A_{\gamma,k}/A_\infty\right)\mathbf{1}$. The formal lemma can be found in Lemma H.3. This performance bound provides the negative and positive answers to our main question: *Is MDVI minimax optimal with function approximation?*

4.3.1. Negative Result of $f = \mathbf{1}$

When $f = \mathbf{1}$, the performance bound becomes incompatible with the TV technique (Corollary 3.1), which is necessary for minimax optimality. In this case, $\heartsuit_k^{\text{wls}} = \clubsuit_k^{\text{wls}} = \sum_{j=0}^{k-1}\gamma^j|\max_{(y,b)\in\mathcal{C}} E_{k-j}(y,b)|\mathbf{1}$. Therefore, even when we relate E_{k-j} to $\sigma(v_{\pi_{k-j}}) \leq H\mathbf{1}$ or $\sigma(v_*) \leq H\mathbf{1}$ using a Bernstein-type inequality, we only obtain a H^2 bound inside the first term of the inequality (6). This implies that the sample complexity can be sub-optimal, as we need more samples by $\sqrt{H}$ than using the TV technique to obtain a near-optimal policy.

4.3.2. Positive Result of $f \approx \sigma(v_*)$

When we carefully select the weighting function f, the performance bound becomes compatible with the TV technique. For example, when $f = \sigma(v_*)$ and E_{k-j} is related to $\sigma(v_*)$ using a Bernstein-type inequality, we obtain $\sum_{j=0}^{k-1}\gamma^j\pi_*P_*^j\sigma(v_*) \leq H\sqrt{H}\mathbf{1}$ inside $\clubsuit_k^{\text{wls}}$ owing to the TV technique. This helps achieve a performance bound that is approximately $\sqrt{H}$ tighter than the bound of $f = \mathbf{1}$.

Indeed, when $f \approx \sigma(v_*)$, we obtain the following minimax optimal sample complexity of `WLS-MDVI`:

Theorem 4.4 (Sample complexity of `WLS-MDVI` with $f \approx \sigma(v_*)$, informally)**.** *When $\varepsilon \in (0, 1/H]$, $\alpha = \gamma$,*

Algorithm 1 WLS-MDVI (α, f, K, M)

Input: $\alpha \in [0,1)$, $f : \mathcal{X} \times \mathcal{A} \to (0, \infty)$, $K \in \mathbb{N}$, $M \in \mathbb{N}$.
Initialize $\theta_0 = \overline{\theta}_0 = \mathbf{0} \in \mathbb{R}^d$, $s_0 = \mathbf{0} \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$, and $w_0 = w_{-1} = \mathbf{0} \in \mathbb{R}^{\mathcal{X}}$.
$\rho_f, \mathcal{C}_f, G_f \coloneqq$ `ComputeOptimalDesign`(f).
for $k = 0$ **to** $K-1$ **do**
  $v_k = w_k - \alpha w_{k-1}$.
  for each state-action pair $(y, b) \in \mathcal{C}_f$ **do**
    Compute $\widehat{q}_{k+1}(y, b)$ by Equation (4).
  end for
  Compute θ_{k+1} by Equation (4).
  $\overline{\theta}_{k+1} = \theta_{k+1} + \alpha\overline{\theta}_k$ and $s_{k+1} = \phi^\top \overline{\theta}_{k+1}$.
  $w_{k+1}(x) = \max_{a\in\mathcal{A}} s_{k+1}(x, a)$ for each $x \in \mathcal{X}$.
end for
Return: v_K and $(\pi_k)_{k=0}^K$, where π_k is greedy policy with respect to s_k

and $\sigma(v_*) \le f \le \sigma(v_*) + 2\sqrt{H}\mathbf{1}$, WLS-MDVI outputs a sequence of policies $(\pi_k)_{k=0}^K$ such that $\|v_* - v_{\pi'_K}\|_\infty \le \varepsilon$ with probability at least $1-\delta$, using $\widetilde{\mathcal{O}}\left(d^2H^3\varepsilon^{-2}\log(1/\delta)\right)$ samples from the generative model.

The formal theorem and proof are provided in Appendix H. The sample complexity matches the lower bound by Weisz et al. (2022) up to logarithmic factors. This means that `WLS-MDVI` is nearly minimax optimal as long as $f \approx \sigma(v_*)$ and ε is sufficiently small. The remaining challenge is to learn such weighting function.

5. Variance Weighted Least-Squares MDVI

In this section, we present a simple algorithm for learning the weighting function and introduce our `VWLS-MDVI`, which combines the weight learning algorithm with `WLS-MDVI` to achieve minimax optimal sample complexity.

5.1. Learning the Weighting Function

As stated in Theorem 4.4, the weighting function should be close to $\sigma(v_*)$ by a factor of $\sqrt{H}$. We accomplish this by learning the weighting function in two steps: learning a $\sqrt{H}$-optimal value function (Section 5.1.1) and learning the variance of the value function (Section 5.1.2).

5.1.1. Learning the $\sqrt{H}$-optimal value function

Theorem 5.1 shows that `WLS-MDVI` with $f = \mathbf{1}$ yields a $\sqrt{H}$-optimal value function with sample complexity that is $1/\varepsilon$ smaller than that of Theorem 4.4.

Theorem 5.1 (Sample complexity of `WLS-MDVI` with $f = \mathbf{1}$, informally)**.** *When $\varepsilon \in (0, 1/H]$, $\alpha = \gamma$, and $f = \mathbf{1}$, WLS-MDVI outputs v_K satisfying $\|v_* - v_K\|_\infty \le \frac{1}{2}\sqrt{H}$ with probability at least $1-\delta$, using $\widetilde{\mathcal{O}}\left(d^2H^3\varepsilon^{-1}\log(1/\delta)\right)$ samples from the generative model.*

The formal theorem and proof are provided in Appendix H.

5.1.2. Learning the Variance Function

Algorithm 2 VarianceEstimation (v_σ, M_σ)

Input: $v_\sigma \in \mathbb{R}^{\mathcal{X}}$, $M_\sigma \in \mathbb{N}$.
$\rho, \mathcal{C}, G \coloneqq$ `ComputeOptimalDesign`$(\mathbf{1})$.
for each state-action pair $(x, a) \in \mathcal{C}$ **do**
  Compute $\widehat{\mathrm{Var}}(x, a)$ by Equation (7).
end for
$\omega = G^{-1}\sum_{(x,a)\in\mathcal{C}} \rho(x, a)\phi(x, a)\widehat{\mathrm{Var}}(x, a)$.
Return: ω.

Given a $\sqrt{H}$-optimal value function v_σ by Theorem 5.1, we linearly approximate the variance function as $\mathrm{Var}_\omega(x, a) \coloneqq \phi^T(x, a)\omega$ with $\omega \in \mathbb{R}^d$. Using ρ, $\mathcal{C}$, and G of the vanilla optimal design, ω is learned using least-squares estimation.

$$\omega = G^{-1}\sum_{(x,a)\in\mathcal{C}} \rho(x, a)\phi(x, a)\widehat{\mathrm{Var}}(x, a)\ , \quad \text{where}$$

$$\widehat{\mathrm{Var}}(x, a) = \frac{1}{2M_\sigma}\sum_{m=1}^{M_\sigma}\Big(v_\sigma(y_{m,x,a}) - v_\sigma(z_{m,x,a})\Big)^2 . \tag{7}$$

Here, $(y_{m,x,a})_{m=1}^{M_\sigma}$ and $(z_{m,x,a})_{m=1}^{M_\sigma}$ denote M_σ independent samples from $P(\cdot|x, a)$.

The pseudocode of the algorithm is shown in Algorithm 2. Theorem 5.2 shows that with a small number of samples, the learned ω estimates $\sigma(v_*)$ with $\sqrt{H}$ accuracy.

Theorem 5.2 (Accuracy of `VarianceEstimation`, informally)**.** *When v_σ satisfies $\|v_* - v_\sigma\|_\infty \le \frac{1}{2}\sqrt{H}$, VarianceEstimation outputs ω such that $\sigma(v_*) \le \sqrt{\max(\phi^\top\omega, \mathbf{0})} + \sqrt{H}\mathbf{1} \le \sigma(v_*) + 2\sqrt{H}\mathbf{1}$ with probability at least $1-\delta$, using $\widetilde{\mathcal{O}}\left(d^2H^2\log(1/\delta)\right)$ samples from the generative model.*

The formal theorem and proof are provided in Appendix I.

5.2. Put Everything Together

The proposed `VWLS-MDVI` algorithm consists of three steps: (1) executing `WLS-MDVI` with $f = \mathbf{1}$, (2) performing `VarianceEstimation`, and (3) executing `WLS-MDVI` again with the output from (2). The technical novelty of our theory lies in the ingenuity to run `WLS-MDVI` twice to use the TV technique, which was not seen in previous studies such as Lattimore et al. (2020) and Kozuno et al. (2022). By combining these three steps, the `VWLS-MDVI` obtains an ϵ-optimal policy within minimax optimal sample complexity.

Theorem 5.3 (Sample complexity of `VWLS-MDVI`, informally)**.** *When $\varepsilon \in (0, 1/H]$ and $\alpha = \gamma$, VWLS-MDVI out-*

Algorithm 3 VWLS-MDVI $(\alpha, K, M, \widetilde{K}, \widetilde{M}, M_\sigma)$

Input: $\alpha \in [0,1)$, $f : \mathcal{X} \times \mathcal{A} \mapsto (0, \infty)$, $K, \widetilde{K} \in \mathbb{N}$, $M, \widetilde{M} \in \mathbb{N}$, $M_\sigma \in \mathbb{N}$.
$v_K, \_ = \texttt{WLS-MDVI}(\alpha, \mathbf{1}, K, M)$.
$\omega = \texttt{VarianceEstimation}(v_K, M_\sigma)$.
$\widetilde{\sigma} = \min\left(\sqrt{\max(\phi^T\omega, \mathbf{0})} + \sqrt{H}\mathbf{1}, H\mathbf{1}\right)$.
$\_, \pi' = \texttt{WLS-MDVI}(\alpha, \widetilde{\sigma}, \widetilde{K}, \widetilde{M})$.
Return: π'

puts a sequence of policies π' such that $\|v_* - v_{\pi'}\|_\infty \leq \varepsilon$ with probability at least $1-\delta$, using $\widetilde{\mathcal{O}}\left(d^2H^3\varepsilon^{-2}\log(1/\delta)\right)$ samples from the generative model.

The formal theorem and proof are provided in Appendix J, and the pseudocode of the algorithm is provided in Algorithm 3. The sample complexity of VWLS-MDVI matches the lower bound described by Weisz et al. (2022) up to logarithmic factors as long as ε is sufficiently small. This is the first algorithm that achieves nearly minimax sample complexity under inifinite-horizon linear MDPs.

6. Deep Variance Weighting

Motivated on the theoretical observations, we propose a practical algorithm to re-weight the least-squares loss of value-based deep RL algorithms, called Deep Variance Weighting (DVW).

6.1. Weighted Loss Function for the Q-Network

As Munchausen DQN (M-DQN, Vieillard et al. (2020b)) is the effective deep extension of MDVI, we use it as our base algorithm to apply DVW. However, the proposed DVW can be potentially applied to any DQN-like algorithms[4]. We provide the pseudocode for the general case in Algorithm 4 and for online RL in Appendix K.

Similar to M-DQN, let q_θ be the q-network and $q_{\overline{\theta}}$ be its target q-network with parameters θ and $\overline{\theta}$, respectively. In this section, x' denotes the next state sampled from $P(\cdot|x,a)$. $\widehat{\mathbb{E}}_{\mathcal{B}}$ denotes the expectation over using samples $(x,a,r,x') \in \mathcal{X} \times \mathcal{A} \times \mathbb{R} \times \mathcal{X}$ from some dataset $\mathcal{B}$. With a weighting function $f : \mathcal{X} \times \mathcal{A} \to (0,\infty)$, we consider the following weighted version of M-DQN's loss function:

$$\mathcal{L}(\theta) = \widehat{\mathbb{E}}_{\mathcal{B}}\left[\left(\frac{r_{\overline{\theta}}(x,a) + \gamma v_{\overline{\theta}}(x') - q_\theta(x,a)}{f(x,a)}\right)^2\right], \qquad (8)$$

[4]Van Hasselt et al. (2019) stated that DQN may not be a completely model-free algorithm, which could potentially conflict with the model-free structure of VWLS-MDVI. Nevertheless, we do not consider such discrepancies from our theory to be problematic, as the primary aim of DVW is to improve the popular algorithms rather than to validate the theoretical analysis.

Algorithm 4 DVW for (Munchausen-)DQN

Input: $\theta, \omega, K \in \mathbb{N}, F \in \mathbb{N}, \mathcal{B}$
Set $\overline{\theta} = \widehat{\theta} = \theta$ and $\overline{\omega} = \omega$. $\eta = 1$.
for $k = 0$ **to** $K-1$ **do**
  Sample a random batch of transition $B_k \in \mathcal{B}$.
  On B_k, update ω by $\mathcal{L}(\omega)$ of (9).
  On B_k, update η by $\mathcal{L}(\eta)$ of (11).
  On B_k and f^{DVW} of (10), update θ by $\mathcal{L}(\theta)$ of (8).
  if $k \mod F = 0$ **then**
    $\widehat{\theta} \leftarrow \overline{\theta}, \overline{\theta} \leftarrow \theta, \overline{\omega} \leftarrow \omega$.
  end if
end for
Return: A greedy policy with respect to q_θ

where $r_{\overline{\theta}} = r + \tau\log\pi_{\overline{\theta}}$, $\pi_{\overline{\theta}}(a|x) \propto \exp\left(\beta q_{\overline{\theta}}(x,a)\right)$, and $v_{\overline{\theta}}(x') = \sum_{a'\in\mathcal{A}}\pi_{\overline{\theta}}(a'|x')\left(q_{\overline{\theta}}(x',a') - \frac{1}{\beta}\log\pi_{\overline{\theta}}(a'|x')\right)$. Equation (8) is equivalent to M-DQN when $f = \mathbf{1}$. Furthermore, when $\tau = \kappa = 0$, we assume that $\tau\log\pi_{\overline{\theta}} = \frac{1}{\beta}\log\pi_{\overline{\theta}} = \mathbf{0}$ and $\sum_{a'\in\mathcal{A}}\pi_{\overline{\theta}}(a'|x')q_{\overline{\theta}}(x',a') = \max_{a'\in\mathcal{A}} q_{\overline{\theta}}(x',a')$. This allows us to generalize Equation (8) to DQN's loss when $f = \mathbf{1}$ and $\tau = \kappa = 0$.

We update θ by stochastic gradient descent (SGD) with respect to $\mathcal{L}(\theta)$. We replace $\overline{\theta}$ with θ for every F iteration.

6.2. Loss Function for the Variance Network

Let Var_ω be the variance network with parameter ω. We also define $q_{\widehat{\theta}}$ as the preceding target q-network to $q_{\overline{\theta}}$. The parameter $\widehat{\theta}$ of $q_{\widehat{\theta}}$ is replaced with $\overline{\theta}$ for every F iteration.

For sufficiently large F, we expect that $q_{\overline{\theta}}$ well approximates $q_{\overline{\theta}} \approx r_{\widehat{\theta}} + \gamma P v_{\widehat{\theta}}$. Using this approximation and based on VarianceEstimation, we construct the loss function for the variance network as

$$\mathcal{L}(\omega) = \widehat{\mathbb{E}}_{\mathcal{B}}\left[h\left(y^2 - \mathrm{Var}_\omega(x,a)\right)\right], \qquad (9)$$

where $y = r_{\widehat{\theta}}(x,a) + \gamma v_{\widehat{\theta}}(x') - q_{\overline{\theta}}(x,a)$. Here, we use the Huber loss function h: for $x \in \mathbb{R}$, $h(x) = x^2$ if $x < 1$; otherwise, $h(x) = |x|$. This is to alleviate the issue with large y^2 in contrast to the L2 loss. We update ω by iterating SGD with respect to $\mathcal{L}(\omega)$.

6.3. Weighting Function Design

According to VWLS-MDVI, the weighting function f should be inversely proportional to the learned variance function with lower and upper thresholds. Moreover, uniformly scaling f with some constant variables does not affect the solution of weighted regression. Therefore, we design the weighting function f^{DVW} such that

$$\frac{1}{f^{\text{DVW}}(x,a)^2} = \max\left(\frac{\eta}{\mathrm{Var}_{\overline{\omega}}(x,a) + \overline{c}_f}, \underline{c}_f\right), \qquad (10)$$

where $\eta \in (0, \infty)$ denotes a scaling constant, $\underline{c}_f$ and $\overline{c}_f \in (0, \infty)$ denote constants for the lower and upper thresholds, respectively. Here, we use the frozen parameter $\overline{\omega}$, which is replaced with ω for every $F \in \mathbb{N}$ iteration, as we should use the weight learned via Equation (9).

To further stabilize training, we automatically adjust η so that $\widehat{\mathbb{E}}_{\mathcal{B}}[f^{\text{DVW}}(x,a)^{-2}] \approx 1$. We adjust η by SGD with respect to the following loss function:

$$\mathcal{L}(\eta) = \left(\widehat{\mathbb{E}}_{\mathcal{B}} \left[\frac{\eta}{\text{Var}_{\overline{\omega}}(x,a) + \overline{c}_f} \right] - 1 \right)^2 , \tag{11}$$

where the term $\eta / \left(\text{Var}_{\overline{\omega}}(x,a) + \overline{c}_f\right)$ is the value inside the max of Equation (10). The max is removed to avoid zero gradient. While the target value can be set to a value other than 1, doing so would be equivalent to adjusting the learning rate in the standard SGD. To avoid introducing an unnecessary hyperparameter, we have fixed the target value to 1.

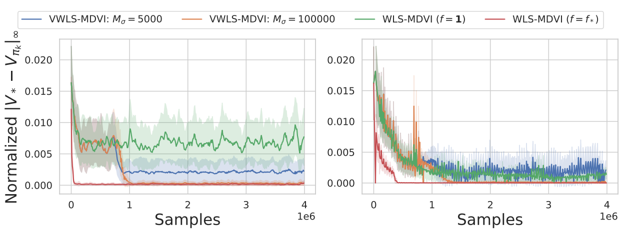

Figure 1. Comparison of the normalized gaps. `VWLS-MDVI` switches to the second run of `WLS-MDVI` around 10^6 samples. **Left:** $M = \widetilde{M} = 100$ and **Right:** $M = \widetilde{M} = 1000$.

7. Experiments

This section reports the experimental sample efficiency of the proposed `VWLS-MDVI` and deep RL with `DVW`.

7.1. Linear MDP Experiments

To empirically validate the negative and positive claims made in Section 4.3 and demonstrate the sample efficiency of `VWLS-MDVI`, we compare `VWLS-MDVI` to `WLS-MDVI` with two different weighting functions: $f = \mathbf{1}$ and $f = f_*$, where $f_* := \min(\sigma(v_*) + \sqrt{H}\mathbf{1}, H\mathbf{1})$ is the oracle weighting function from Theorem 4.4. The evaluation is conducted on randomly generated hard linear MDPs that are based on **Theorem H.3** in Weisz et al. (2022). For simplicity, all algorithms use the last policy for evaluation. Specifically, for the $k \in [K]$ th iteration to update the parameter θ, we report the normalized optimality gap $\|v_* - v_{\pi_k}\|_\infty / \|v_*\|_\infty$ in terms of the total number of samples used so far. We normalize the gap by $\|v_*\|_\infty$ as the maximum gap can vary depending on the MDPs.

Figure 1 compares algorithms under $M = \widetilde{M} = 100$ (Left) and $M = \widetilde{M} = 1000$ (Right). The results are averaged over 300 random MDPs. For WLS-MDVI ($f = 1$), increasing M from 100 to 1000 results in a smaller optimality gap, which is expected due to the increase in the number of samples. On the other hand, WLS-MDVI ($f = f_*$) achieves a gap very close to 0 even with $M = 100$, demonstrating the effectiveness of variance-weighted regression in improving sample efficiency, as claimed in Section 4.3. Similarly, it is observed that the VWLS-MDVI ($M_\sigma = 100000$) achieves a smaller gap with much fewer samples than that of WLS-MDVI. However, the gap of VWLS-MDVI ($M_\sigma = 5000$) does not reach that of $f = f_*$. This suggests that the accuracy of the `VarianceEstimation` is important for guaranteeing good performance. Further experimental details are provided in Appendix L.1.

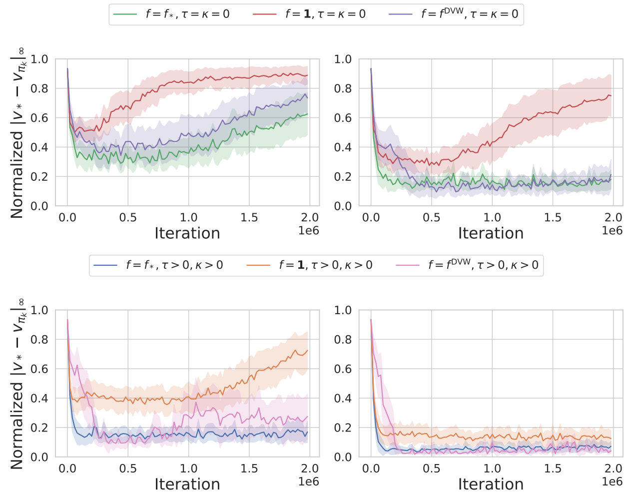

Figure 2. Comparison of the normalized gaps. **Top Row:** $\tau = \kappa = 0$ and **Bottom Row:** $\tau > 0, \kappa > 0$. **Left Column:** $M = 3$ and **Right Column:** $M = 10$.

7.2. Deep RL Experiments

We perform two deep RL experiments to evaluate the effectiveness of `DVW`: one to compare `DVW` with the oracle weighting function of Theorem 4.4, and another to demonstrate the effectiveness of `DVW` to online deep RL. The details of the experiments are provided in Section 7.2.1.

7.2.1. Comparison of $f = f^{\text{DVW}}$ with $f \approx \sigma(v_*)$

To investigate the effectiveness of `DVW`, we evaluate the behavior of M-DQN with weighted regression (8) under three weighting functions: the oracle weighting ($f = f_*$), the uniform weighting ($f = \mathbf{1}$), and the `DVW` weighting ($f = f^{\text{DVW}}$). Furthermore, for the purpose of ablation study, we compare the algorithms with and without regularization ($\tau > 0, \kappa > 0$ vs $\tau = 0, \kappa = 0$). To remove the challenge

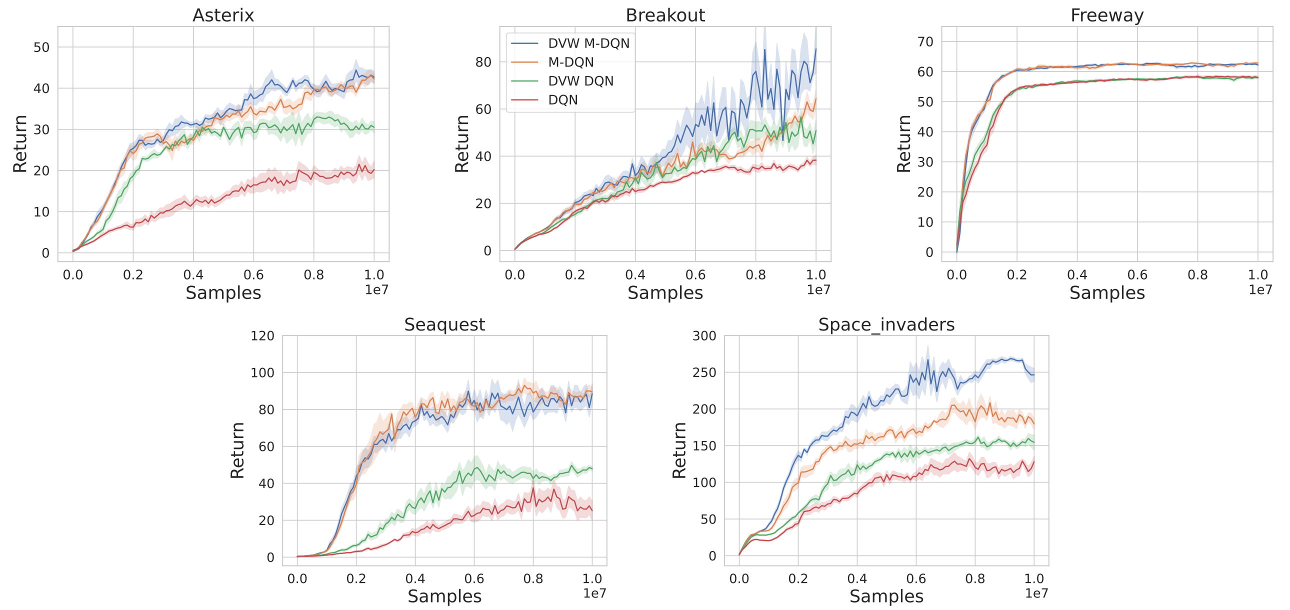

Figure 3. Comparison of returns on MinAtar benchmarks. We report the return of the greedy policy with respect to q_θ for each algorithm.

of exploration for didactic purposes, we use a dataset $\mathcal{B}$, which is constructed by pairs of (x, a, r, x') for the entire state-action space with M next-state samples. In other words, $\mathcal{B}$ is a dataset of size MXA.

We evaluate them in randomly generated environments where we can compute oracle values. Specifically, we use a modified version of the gridworld environment described by Fu et al. (2019). For the k th iteration to update the q-networks, we evaluate the normalized optimality gap averaged over 20 environments and 3 random seeds for each.

Figure 2 compares algorithms under $M = 3$ (Left Column) and $M = 10$ (Right Column). In both cases, `DVW` consistently achieves a smaller gap compared to $f = \mathbf{1}$, and moreover, the gap of `DVW` is comparable to that of the oracle weighting $f = f_*$. In addition, the gap is smaller when $\tau > 0, \kappa > 0$ compared to when $\tau = \kappa = 0$. It can be inferred that `DVW` weighting and KL-entropy regularization contribute to improving sample efficiency, and that performance is significantly improved when both are present.

7.2.2. DVW for Online RL

We evaluate the effectiveness of DVW using a set of the challenging benchmarks for online RL. Similar to Section 7.2.1, we evaluate four algorithms that varied with and without `DVW` (DVW vs N/A), and with and without regularization (M-DQN vs DQN). We compare their performance on the MinAtar environment (Young & Tian, 2019), which possesses high-dimensional features and more challenging exploration than Section 7.2.1, while facilitating fast training. For a fair comparison, the algorithms use the same network architecture and same epsilon-greedy exploration strategy. Each algorithm is executed five times with different random seeds for each environment.

Figure 3 shows the average returns of the algorithms. We observe that `DVW` improves the performance of M-DQN and DQN in almost all the environments. Although our theory does not cover online RL, this experiment suggests that the extension of `DVW` to wider problem settings is effective.

8. Conclusion

In this study, we proposed both a theoretical algorithm, i.e., `VWLS-MDVI`, and a practical algorithm, i.e., `DVW`. `VWLS-MDVI` achieved the first-ever nearly minimax optimal sample complexity in infinite-horizon Linear MDPs by utilizing the combination of KL-entropy regularization and variance-weighted regression. We extended our theoretical observations and developed the `DVW` algorithm, which reweights the least-squares loss of value-based RL algorithms using the estimated variance of the value function. Our experiments demonstrated that `DVW` effectively helps improve the performance of value-based deep RL algorithms.

Acknowledgements

Csaba Szepesvári gratefully acknowledges the funding from Natural Sciences and Engineering Research Council (NSERC) of Canada and the Canada CIFAR AI Chairs Program for Amii.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved Algorithms for Linear Stochastic Bandits. In *Advances in Neural Information Processing Systems*, 2011.

Agarwal, A., Kakade, S., and Yang, L. F. Model-Based Reinforcement Learning with a Generative Model is Minimax Optimal. In *Conference on Learning Theory*, 2020.

Agarwal, A., Jin, Y., and Zhang, T. VO Q L: Towards Optimal Regret in Model-free RL with Nonlinear Function Approximation. *arXiv preprint arXiv:2212.06069*, 2022.

Azar, M., Munos, R., and Kappen, H. J. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning*, 91(3): 325–349, 2013.

Bellemare, M. G., Ostrovski, G., Guez, A., Thomas, P., and Munos, R. Increasing the Action Gap: New Operators for Reinforcement Learning. In *AAAI Conference on Artificial Intelligence*, 2016.

Bellman, R., Kalaba, R., and Kotkin, B. Polynomial Approximation–A New Computational Technique in Dynamic Programming: Allocation Processes. *Mathematics of Computation*, 17(82):155–161, 1963.

Bernstein, S. N. *The Theory of Probabilities*. Gastehizdat Publishing House, 1946.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration Inequalities - A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Fu, J., Kumar, A., Soh, M., and Levine, S. Diagnosing Bottlenecks in Deep Q-Learning Algorithms. In *International Conference on Machine Learning*, 2019.

Geist, M., Scherrer, B., and Pietquin, O. A Theory of Regularized Markov Decision Processes. In *International Conference on Machine Learning*, 2019.

Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. Reinforcement Learning with Deep Energy-Based Policies. In *International Conference on Machine Learning*, 2017.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *International Conference on Machine Learning*, 2018.

He, J., Zhao, H., Zhou, D., and Gu, Q. Nearly minimax optimal reinforcement learning for linear markov decision processes. *arXiv preprint arXiv:2212.06132*, 2022.

Hu, P., Chen, Y., and Huang, L. Nearly Minimax Optimal Reinforcement Learning with Linear Function Approximation. In *International Conference on Machine Learning*, 2022.

Huang, S., Dossa, R. F. J., Ye, C., Braga, J., Chakraborty, D., Mehta, K., and Araújo, J. G. CleanRL: High-quality Single-file Implementations of Deep Reinforcement Learning Algorithms. *Journal of Machine Learning Research*, 23(274):1–18, 2022.

Husain, H., Ciosek, K., and Tomioka, R. Regularized Policies are Reward Robust. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal Regret Bounds for Reinforcement Learning. *Journal of Machine Learning Research*, 11(51):1563–1600, 2010.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably Efficient Reinforcement Learning with Linear Function Approximation. In *Conference on Learning Theory*, 2020.

Kiefer, J. and Wolfowitz, J. The Equivalence of Two Extremum Problems. *Canadian Journal of Mathematics*, 12:363–366, 1960.

Kitamura, T. and Yonetani, R. ShinRL: A Library for Evaluating RL Algorithms from Theoretical and Practical Perspectives. *arXiv preprint arXiv:2112.04123*, 2021.

Kozuno, T., Uchibe, E., and Doya, K. Theoretical Analysis of Efficiency and Robustness of Softmax and Gap-Increasing Operators in Reinforcement Learning. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Kozuno, T., Yang, W., Vieillard, N., Kitamura, T., Tang, Y., Mei, J., Ménard, P., Azar, M. G., Valko, M., Munos, R., et al. KL-Entropy-Regularized RL with a Generative Model is Minimax Optimal. *arXiv preprint arXiv:2205.14211*, 2022.

Kumar, A., Gupta, A., and Levine, S. Discor: Corrective Feedback in Reinforcement Learning via Distribution Correction. *Advances in Neural Information Processing Systems*, 2020.

Lattimore, T. and Szepesvari, C. *Bandit Algorithms*. Cambridge University Press, 1st edition, 2020.

Lattimore, T., Szepesvari, C., and Weisz, G. Learning with Good Feature Representations in Bandits and in RL with a Generative Model. In *International Conference on Machine Learning*, 2020.

Lee, K., Laskin, M., Srinivas, A., and Abbeel, P. Sunrise: A simple unified framework for ensemble learning in deep reinforcement learning. In *International Conference on Machine Learning*, 2021.

Maystre, L., Russo, D., and Zhao, Y. Optimizing Audio Recommendations for the Long-Term: A Reinforcement Learning Perspective. *arXiv preprint arXiv:2302.03561*, 2023.

Miki, T., Lee, J., Hwangbo, J., Wellhausen, L., Koltun, V., and Hutter, M. Learning robust perceptive locomotion for quadrupedal robots in the wild. *Science Robotics*, 7 (62):eabk2822, 2022.

Min, Y., Wang, T., Zhou, D., and Gu, Q. Variance-Aware Off-Policy Evaluation with Linear Function Approximation. In *Advances in neural information processing systems*, 2021.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529–533, 2015.

Munos, R. Error bounds for approximate value iteration. In *AAAI Conference on Artificial Intelligence*, 2005.

Schaul, T., Quan, J., Antonoglou, I., and Silver, D. Prioritized Experience Replay. In *International Conference on Learning Representations*, 2015.

Scherrer, B. and Lesner, B. On the Use of Non-Stationary Policies for Stationary Infinite-Horizon Markov Decision Processes. In *Advances in Neural Information Processing Systems*, 2012.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust Region Policy Optimization. In *International Conference on Machine Learning*, 2015.

Sidford, A., Wang, M., Wu, X., Yang, L., and Ye, Y. Near-Optimal Time and Sample Complexities for Solving Markov Decision Processes with a Generative Model. In *Advances in Neural Information Processing Systems*, 2018.

Taupin, J., Jedra, Y., and Proutiere, A. Best Policy Identification in Linear MDPs. *arXiv preprint arXiv:2208.05633*, 2022.

Todd, M. J. *Minimum-Volume Ellipsoids: Theory and Algorithms*. Society for Industrial and Applied Mathematics, 2016.

Van Hasselt, H. P., Hessel, M., and Aslanides, J. When to use parametric models in reinforcement learning? In *Advances in Neural Information Processing Systems*, 2019.

Vieillard, N., Kozuno, T., Scherrer, B., Pietquin, O., Munos, R., and Geist, M. Leverage the Average: an Analysis of KL Regularization in Reinforcement Learning. In *Advances in Neural Information Processing Systems*, 2020a.

Vieillard, N., Pietquin, O., and Geist, M. Munchausen Reinforcement Learning. In *Advances in Neural Information Processing Systems*, 2020b.

Weisz, G., György, A., Kozuno, T., and Szepesvári, C. Confident Approximate Policy Iteration for Efficient Local Planning in q^π-realizable MDPs. *arXiv preprint arXiv:2210.15755*, 2022.

Xiong, W., Zhong, H., Shi, C., Shen, C., Wang, L., and Zhang, T. Nearly Minimax Optimal Offline Reinforcement Learning with Linear Function Approximation: Single-Agent MDP and Markov Game. *arXiv preprint arXiv:2205.15512*, 2022.

Yang, L. and Wang, M. Sample-Optimal Parametric Q-learning Using Linearly Additive Features. In *International Conference on Machine Learning*, 2019.

Yin, D., Hao, B., Abbasi-Yadkori, Y., Lazić, N., and Szepesvári, C. Efficient Local Planning with Linear Function Approximation. In *International Conference on Algorithmic Learning Theory*, 2022a.

Yin, M., Duan, Y., Wang, M., and Wang, Y.-X. Near-optimal Offline Reinforcement Learning with Linear Representation: Leveraging Variance Information with Pessimism. *arXiv preprint arXiv:2203.05804*, 2022b.

Yin, M., Wang, M., and Wang, Y.-X. Offline Reinforcement Learning with Differentiable Function Approximation is Provably Efficient. In *International Conference on Learning Representations*, 2022c.

Young, K. and Tian, T. Minatar: An atari-inspired testbed for thorough and reproducible reinforcement learning experiments. *arXiv preprint arXiv:1903.03176*, 2019.

Zhan, X., Xu, H., Zhang, Y., Zhu, X., Yin, H., and Zheng, Y. Deepthermal: Combustion optimization for thermal power generating units using offline reinforcement learning. In *AAAI Conference on Artificial Intelligence*, 2022.

Zhang, Z., Zhou, Y., and Ji, X. Model-free reinforcement learning: from clipped pseudo-regret to sample complexity. In *International Conference on Machine Learning*, 2021.

Zhou, D., Gu, Q., and Szepesvari, C. Nearly Minimax Optimal Reinforcement Learning for Linear Mixture Markov Decision Processes. In *Conference on Learning Theory*, 2021.

Contents

- Appendix A lists notations for the theoretical analysis and their meaning;
- Appendix B proves the MDVI transformation stated in Section 3.1.2;
- Appendix C provides auxiliary lemmas necessary for proofs;
- Appendix E provides the formal theorem and the proofs of the total variance technique;
- Appendix F provides the proof of the existence of a small core set for a compact set (Theorem 3.3);
- Appendix G provides the proof of the weighted Kiefer Wolfowitz bound (Lemma 4.3);
- Appendix H provides the formal theorems and the proofs for sample complexity of `WLS-MDVI` (Theorem 4.4 and Theorem 5.1)
- Appendix I provides the formal theorem and the proof for sample complexity of `VarianceEstimation` (Theorem 5.2)
- Appendix J provides the formal theorem and the proof for sample complexity of `VWLS-MDVI` (Theorem 5.3);
- Appendix K provides the pseudocode of algorithms missed in the main pages;
- Appendix L provides the details of experiments stated in Section 7.

A. Notations for Theoretical Analysis

Table 2. Table of Notations for Theoretical Analysis

Notation	Meaning
$\mathcal{A}$, $\mathcal{X}$	action space of size A, state space
γ, H	discount factor in $[0,1)$ and effective horizon $H \coloneqq 1/(1-\gamma)$
ϕ, d	feature map of a linear MDP and its dimension (Assumption 3.2)
r	reward function bounded by 1
P, P_π	transition kernel, $P_\pi \coloneqq P\pi$
$\mathcal{F}_v$, $\mathcal{F}_q$	the sets of all bounded Borel-measurable functions over $\mathcal{X}$ and $\mathcal{X} \times \mathcal{A}$, respectively
π'_k	a non-stationary policy that follows $\pi_k, \pi_{k-1}, \dots$ sequentially (Section 3.1)
P^i_j, P^i_*	$P^i_j \coloneqq P_{\pi_i} P_{\pi_{i-1}} \cdots P_{\pi_{j+1}} P_{\pi_j}$ and $P^i_* \coloneqq (P_{\pi_*})^i$ (Section 3.1)
T_π, T^i_j	Bellman operator for a policy π, $T^i_j \coloneqq T_{\pi_i} T_{\pi_{i-1}} \cdots T_{\pi_{j+1}} T_{\pi_j}$ (Section 3.1)
$v_{\pi'_k}$	value function of π'_k; $v_{\pi'_k} = \pi_k T_{\pi_{k-1}} \cdots T_{\pi_1} q_{\pi_0}$.
ε, δ	admissible suboptimality, admissible failure probability
ε_k	$\varepsilon_k : (x,a) \mapsto \gamma \widehat{P}_{k-1}(M) v_{k-1}(x,a) - \gamma P v_{k-1}(x,a)$ in `WLS-MDVI`
E_k	$E_k : (x,a) \mapsto \sum_{j=1}^{k} \alpha^{k-j} \varepsilon_j(x,a)$
f	a bounded positive weighting function over $\mathcal{X} \times \mathcal{A}$
ρ_f	a design over $\mathcal{X} \times \mathcal{A}$
$\mathcal{C}_f$, $u_\mathcal{C}$	core set, $u_\mathcal{C} \coloneqq 4d \log\log(d+4) + 28$ (Section 3.2)
G_f	design matrix with respect to f, ϕ, and ρ_f (Theorem 4.1)
u_f, l_f	$u_f \coloneqq \max_{(x,a)\in\mathcal{X}\times\mathcal{A}} f(x,a)$, $l_f \coloneqq \min_{(x,a)\in\mathcal{X}\times\mathcal{A}} f(x,a)$ (Appendix H)
$W(f_1, f_2)$	solution of a weighted least-squares estimation (Lemma 4.3)
$\widehat{P}_k$	$\widehat{P}_k(M) v_k : (x,a) \mapsto \frac{1}{M} \sum_{m=1}^{M} v_k(y_{k,m,x,a})$ (Section 3.1)
$\theta_k, \overline{\theta}_k$	parameter of q_k in `WLS-MDVI` ($q_k = \phi^\top \theta_k$), $\overline{\theta}_k = \theta_k + \alpha \overline{\theta}_{k-1} = \sum_{j=0}^{k} \alpha^{k-j} \theta_j$
$\theta^*_k, \overline{\theta}^*_k$	parameter that satisfies $\phi^\top \theta^*_k = r + \gamma P v_{k-1}$, $\overline{\theta}^*_k = \sum_{j=1}^{k} \alpha^{k-j} \theta^*_j$ (Appendix H)
s_k, v_k, w_k	$s_k \coloneqq \phi^\top(x,a)\overline{\theta}_k$, $v_k \coloneqq w_k - \alpha w_{k-1}$, $w_k(x) \coloneqq \max_{a\in\mathcal{A}} s_k(x,a)$ (`WLS-MDVI`)
α, β	weights for MDVI updates $\alpha \coloneqq \tau/(\tau+\kappa)$, $\beta \coloneqq 1/(\tau+\kappa)$ (Section 3.1)
K, M	the number of iterations and the number of samples from the generative model in `WLS-MDVI`
$\widehat{\mathrm{Var}}$	$\widehat{\mathrm{Var}}(x,a) = \frac{1}{2M} \sum_{m=1}^{M} \Big(v_\sigma(y_{m,x,a}) - v_\sigma(z_{m,x,a}) \Big)^2$ (Section 5.1)
M_σ	number of samples from the generative model in `VarianceEstimation`
v_σ	the input value function to `VarianceEstimation`
ω	parameter for `VarianceEstimation`
$A_k, A_\infty, A_{\gamma,k}$	$\sum_{j=0}^{k-1} \alpha^j$, $\sum_{j=0}^{\infty} \alpha^j$, $\sum_{j=0}^{k-1} \alpha^j \gamma^{k-j}$
$\mathbf{F}_{k,m}$	σ-algebra in the filtration for `WLS-MDVI` (Appendix H)
$\mathbf{F}_m$	σ-algebra in the filtration for `VarianceEstimation` (Appendix I)
ι_1, $\iota_{2,n}$	$\iota_1 = \log(2c_0 u_\mathcal{C} K\delta)$, $\iota_{2,n} = \log(2c_0^2 u_\mathcal{C} K/(c_0 - n)\delta)$ for $n \in \mathbb{N}$ (Appendix H)
$\xi_{2,n}$	$\xi_{2,n} = \iota_{2,n} + \log\log_2(16KH^2)$ (Appendix H)
$\square$	an indefinite constant independent of H, X, A, ε, and δ (Appendix H)
$\mathcal{E}_1$	event of f close to $\sigma(v_*)$
$\mathcal{E}_2$	event of v_k bound for all k
$\mathcal{E}_3$	event of small E_k for all k (not variance-aware)
$\mathcal{E}_4$	event of small ε_k for all k (not variance-aware)
$\mathcal{E}_5$	event of small E_k for all k (variance-aware)
$\mathcal{E}_6$	event of v_σ close to v_* (Appendix I)
$\mathcal{E}_7$	event of learned $\phi^\top \omega$ close to $\sigma(v_*)$ (Appendix I)

B. Equivalence of MDVI Update Rules (Kozuno et al., 2022)

We show the equivalence of MDVI's updates (1) to those used in `Tabular MDVI`. The following transformation is identical to that of Kozuno et al. (2022) but is included here for completeness. We first recall MDVI's updates (1):

$$q_{k+1} = r + \gamma \widehat{P}_k \pi_k \left(q_k - \tau \log \frac{\pi_k}{\pi_{k-1}} - \kappa \log \pi_k \right),$$

$$\text{where } \pi_k(\cdot|x) = \operatorname*{arg\,max}_{p \in \Delta(\mathcal{A})} \sum_{a \in \mathcal{A}} p(a) \left(q_k(x,a) - \tau \log \frac{p(a)}{\pi_{k-1}(a|x)} - \kappa \log p(a) \right) \text{ for all } x \in \mathcal{X},$$

The policy update can be rewritten in a closed-form solution as follows (e.g., Equation (5) of Kozuno et al. (2019)):

$$\pi_k(a|x) = \frac{\pi_{k-1}(a|x)^\alpha \exp(\beta q_k(x,a))}{\sum_{b \in \mathcal{A}} \pi_{k-1}(b|x)^\alpha \exp(\beta q_k(x,b))},$$

where $\alpha := \tau/(\tau+\kappa)$, and $\beta := 1/(\tau+\kappa)$. It can be further rewritten as, defining $s_k = q_k + \alpha s_{k-1}$,

$$\pi_k(a|x) = \frac{\exp(\beta s_k(x,a))}{\sum_{b \in \mathcal{A}} \exp(\beta s_k(x,b))}.$$

Plugging in this policy expression to v_k, we deduce that

$$\begin{aligned} v_k(x) &= \frac{1}{\beta} \log \sum_{a \in \mathcal{A}} \exp(\beta q_k(x,a) + \alpha \log \pi_{k-1}(a|x)) \\ &= \frac{1}{\beta} \log \sum_{a \in \mathcal{A}} \exp(\beta s_k(x,a)) - \frac{\alpha}{\beta} \log \sum_{a \in \mathcal{A}} \exp(\beta s_{k-1}(x,a)). \end{aligned}$$

Kozuno et al. (2019, Appendix B) show that when $\beta \to \infty$, $v_k(x) = w_k(x) - \alpha w_{k-1}(x)$. Furthermore, the Boltzmann policy becomes a greedy policy. Accordingly, the update rules used in `Tabular MDVI` is a limit case of the original MDVI updates.

C. Auxiliary Lemmas

In this appendix, we prove some auxiliary lemmas used in the proof. Some of the lemmas are identical to those of Kozuno et al. (2022) but are included here for completeness.

Lemma C.1. *For any events A and B, $\mathbb{P}(A \cap B) \geq \mathbb{P}(B) - \mathbb{P}(A^c|B)$.*

Proof. $\mathbb{P}(A \cap B) = \mathbb{P}((A \cup B^c) \cap B) \geq 1 - \mathbb{P}(A^c \cap B) - \mathbb{P}(B^c) = \mathbb{P}(B) - \mathbb{P}(A^c \cap B)$. The claim holds by $\mathbb{P}(A^c \cap B) = \mathbb{P}(A^c|B)\mathbb{P}(B) \leq \mathbb{P}(A^c|B)$. □

Lemma C.2. *For any positive real values a and b, $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$.*

Proof. Indeed, $a + b \leq a + 2\sqrt{ab} + b = (\sqrt{a} + \sqrt{b})^2$. □

Lemma C.3. *Let a, b, and c are positive real values. If $|a^2 - b^2| \leq c^2$, then $|a - b| \leq c$.*

Proof. Without loss of generality, assume that $a \geq b$. Then, $c^2 \geq (a^2 - b^2) = (a+b)(a-b) \geq (a-b)^2$ and thus $|a - b| \leq c$. □

Lemma C.4. *For any real values $(a_n)_{n=1}^N$, $(\sum_{n=1}^N a_n)^2 \leq N \sum_{n=1}^N a_n^2$.*

Proof. Indeed, from the Cauchy–Schwarz inequality,

$$\left(\sum_{n=1}^N a_n \cdot 1\right)^2 \le \left(\sum_{n=1}^N 1\right)\left(\sum_{n=1}^N a_n^2\right) = N \sum_{n=1}^N a_n^2 \,,$$

which is the desired result. □

Lemma C.5. *For any $k \in [K]$,*

$$A_{\gamma,k} = \begin{cases} \gamma \dfrac{\alpha^k - \gamma^k}{\alpha - \gamma} & \text{if } \alpha \neq \gamma \\ k\gamma^k & \text{otherwise} \end{cases} .$$

Proof. Indeed, if $\alpha \neq \gamma$

$$A_{\gamma,k} = \sum_{j=0}^{k-1} \alpha^j \gamma^{k-j} = \gamma^k \frac{(\alpha/\gamma)^k - 1}{(\alpha/\gamma) - 1} = \gamma \frac{\alpha^k - \gamma^k}{\alpha - \gamma} \,.$$

If $\alpha = \gamma$, $A_{\gamma,k} = k\gamma^k$ by definition. □

Lemma C.6. *For any real value $x \in (0, 1]$, $1 - x \le \log(1/x)$.*

Proof. Since $\log(1/x)$ is convex and differentiable, $\log(1/x) \ge \log(1/y) - (x - y)/y$. Choosing $y = 1$, we concludes the proof. □

Lemma C.7. *Suppose $\alpha, \gamma \in [0, 1)$, $\varepsilon \in (0, 1]$, $c \in [1, \infty)$, $m \in \mathbb{N}$, and $n \in [0, \infty)$. Let $K := \frac{m}{1-\alpha} \log \frac{cH}{\varepsilon}$. Then,*

$$K^n \alpha^K \le \left(\frac{mn}{(1-\alpha)e}\right)^n \left(\frac{\varepsilon}{cH}\right)^{m-1} .$$

Proof. Using Lemma C.6 for $\alpha \in [0, 1)$,

$$K = \frac{m}{1-\alpha} \log \frac{cH}{\varepsilon} \ge \log_\alpha \left(\frac{\varepsilon}{cH}\right)^m .$$

Therefore,

$$K^n \alpha^K \le \left(\frac{m}{1-\alpha} \log \frac{cH}{\varepsilon}\right)^n \left(\frac{\varepsilon}{cH}\right)^m = \frac{m^n}{(1-\alpha)^n} \left(\frac{\varepsilon}{cH}\right)^m \left(\log \frac{cH}{\varepsilon}\right)^n .$$

Since $x\left(\log \frac{1}{x}\right)^n \le \left(\frac{n}{e}\right)^n$ for any $x \in (0, 1]$ as shown later,

$$K^n \alpha^K \le \left(\frac{mn}{(1-\alpha)e}\right)^n \left(\frac{\varepsilon}{cH}\right)^{m-1} .$$

Now it remains to show $f(x) := x\left(\log \frac{1}{x}\right)^n \le \left(\frac{n}{e}\right)^n$ for $x < 1$. We have that

$$f'(x) = (-\log x)^n - n(-\log x)^{n-1} \implies f'(x) = 0 \text{ at } x = e^{-n}.$$

Therefore, f takes its maximum $\left(\frac{n}{e}\right)^n$ at e^{-n} when $x \in (0, 1)$. □

The following lemma is a special case of a well-known inequality that for any increasing function f

$$\sum_{k=1}^K f(k) \le \int_1^{K+1} f(x) dx \,.$$

Lemma C.8. *For any $K \in \mathbb{N}$ and $n \in [0, \infty)$, $\sum_{k=1}^K k^n \le \frac{1}{n+1}(K+1)^{n+1}$.*

D. Tools from Probability Theory

We extensively use the following concentration inequality, which is derived based on a proof idea of Bernstein's inequality (Bernstein, 1946; Boucheron et al., 2013) for a martingale (Lattimore & Szepesvari, 2020, Excercises 5.14 (f)). For a real-valued stochastic process $(X_n)_{n=1}^N$ adapted to a filtration $(\mathcal{F}_n)_{n=1}^N$, we let $\mathbb{E}_n[X_n] \coloneqq \mathbb{E}[X_n|\mathcal{F}_{n-1}]$ for $n \geq 1$, and $\mathbb{E}_1[X_1] \coloneqq \mathbb{E}[X_1]$.

Lemma D.1 (Azuma-Hoeffding Inequality)**.** *Consider a real-valued stochastic process $(X_n)_{n=1}^N$ adapted to a filtration $(\mathcal{F}_n)_{n=1}^N$. Assume that $X_n \in [l_n, u_n]$ and $\mathbb{E}_n[X_n] = 0$ almost surely, for all n. Then,*

$$\mathbb{P}\left(\sum_{n=1}^N X_n \geq \sqrt{\sum_{n=1}^N \frac{(u_n - l_n)^2}{2} \log \frac{1}{\delta}}\right) \leq \delta$$

for any $\delta \in (0, 1)$.

Lemma D.2 (Conditional Azuma-Hoeffding's Inequality)**.** *Consider a real-valued stochastic process $(X_n)_{n=1}^N$ adapted to a filtration $(\mathcal{F}_n)_{n=1}^N$. Assume that $\mathbb{E}_n[X_n] = 0$ almost surely, for all n. Furthermore, let $\mathcal{E}$ be an event that implies $X_n \in [l_n, u_n]$ with $\mathbb{P}(\mathcal{E}) \geq 1 - \delta'$ for all n and for some $\delta' \in (0, 1)$. Then,*

$$\mathbb{P}\left(\sum_{n=1}^N X_n \geq \sqrt{\sum_{n=1}^N \frac{(u_n - l_n)^2}{2} \log \frac{1}{\delta(1-\delta')}} \middle| \mathcal{E}\right) \leq \delta$$

for any $\delta \in (0, 1)$.

Proof. Let A denote the events of

$$\sum_{n=1}^N X_n \geq \sqrt{\sum_{n=1}^N \frac{(u_n - l_n)^2}{2} \log \frac{1}{\delta(1-\delta')}} .$$

Accordingly,

$$\mathbb{P}(A|\mathcal{E}) = \frac{\mathbb{P}(A \cap \mathcal{E})}{\mathbb{P}(\mathcal{E})} \overset{(a)}{\leq} \frac{\delta(1-\delta')}{\mathbb{P}(\mathcal{E})} \overset{(b)}{\leq} \delta ,$$

where (a) follows from the Azuma-Hoeffding inequality (Lemma D.1), and (b) follows from $\mathbb{P}(\mathcal{E}) \geq 1 - \delta'$. □

Lemma D.3 (Lemma 13 in Zhang et al. (2021))**.** *Consider a real-valued stochastic process $(X_n)_{n=1}^N$ adapted to a filtration $(\mathcal{F}_n)_{n=1}^N$. Suppose that $|X_n| \leq U$ and $\mathbb{E}_n[X_n] = 0$ almost surely, for all n and for some $U \in [0, \infty)$. Then, letting $V_N \coloneqq \sum_{n=1}^N \mathbb{E}_n[X_n^2]$,*

$$\mathbb{P}\left(\left|\sum_{n=1}^N X_n\right| \geq 2\sqrt{2}\sqrt{V_N \log\left(\frac{1}{\delta}\right)} + 2\sqrt{\epsilon \log\left(\frac{1}{\delta}\right)} + 2U \log\left(\frac{1}{\delta}\right)\right) \leq 2\left(\log_2\left(\frac{NU^2}{\epsilon} + 1\right)\right)\delta ,$$

for any $\epsilon, \delta > 0$.

In our analysis, we use the following corollary of this inequality.

Lemma D.4 (Conditional Bernstein-type Inequality)**.** *Consider a real-valued stochastic process $(X_n)_{n=1}^N$ adapted to a filtration $(\mathcal{F}_n)_{n=1}^N$. Suppose that $\mathbb{E}_n[X_n] = 0$ almost surely, for all n. Furthermore, let $\mathcal{E}$ be an event that implies $|X_n| \leq U$ with $\mathbb{P}(\mathcal{E}) \geq 1 - \delta'$ for all n, for some $\delta' \in (0, 1)$ and $U \in [0, \infty)$. Then, letting $V_N \coloneqq \sum_{n=1}^N \mathbb{E}_n[X_n^2]$,*

$$\mathbb{P}\left(\left|\sum_{n=1}^N X_n\right| \geq 2\sqrt{2}\sqrt{(1 + V_N) \log\left(\frac{2\log_2(NU^2)}{\delta(1-\delta')}\right)} + 2U \log\left(\frac{2\log_2(NU^2)}{\delta(1-\delta')}\right) \middle| \mathcal{E}\right) \leq \delta ,$$

for any $\delta > 0$.

Proof. Let A and B denote the events of

$$\left|\sum_{n=1}^{N} X_n\right| \geq 2\sqrt{2}\sqrt{(1+V_N)\log\left(\frac{2\log_2(NU^2)}{\delta(1-\delta')}\right)} + 2U\log\left(\frac{2\log_2(NU^2)}{\delta(1-\delta')}\right)$$

and $|X_n| \leq U$ for all n, respectively. Since $\mathcal{E} \subset B$, it follows that $A \cap \mathcal{E} \subset A \cap B$, and $\mathbb{P}(A \cap \mathcal{E}) \leq \mathbb{P}(A \cap B)$. Accordingly,

$$\mathbb{P}(A|\mathcal{E}) = \frac{\mathbb{P}(A \cap \mathcal{E})}{\mathbb{P}(\mathcal{E})} \leq \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(\mathcal{E})} \overset{(a)}{\leq} \frac{\delta(1-\delta')}{\mathbb{P}(\mathcal{E})} \overset{(b)}{\leq} \delta\,,$$

where (a) follows from Lemma D.3 with $1 + \sqrt{V_N} \geq \sqrt{1 + V_N}$ due to Lemma C.2 and $\epsilon = 2$. Then, (b) follows from $\mathbb{P}(\mathcal{E}) \geq 1 - \delta'$. □

Lemma D.5 (Popoviciu's Inequality for Variances)**.** *The variance of any random variable bounded by x is bounded by x^2.*

E. Total Variance Technique (Kozuno et al., 2022)

This section introduces the total variance technique for non-stationary policy. The proof is identical to that of Kozuno et al. (2022) but is included here for completeness.

The following lemma is due to Azar et al. (2013).

Lemma E.1. *Suppose two real-valued random variables X, Y whose variances, $\mathbb{V}X$ and $\mathbb{V}Y$, exist and are finite. Then, $\sqrt{\mathbb{V}X} \leq \sqrt{\mathbb{V}[X-Y]} + \sqrt{\mathbb{V}Y}$.*

For completeness, we prove Lemma E.1.

Proof. Indeed, from Cauchy-Schwartz inequality,

$$\begin{aligned}\mathbb{V}X &= \mathbb{V}[X - Y + Y] \\ &= \mathbb{V}[X-Y] + \mathbb{V}Y + 2\mathbb{E}\left[(X - Y - \mathbb{E}[X-Y])(Y - \mathbb{E}Y)\right] \\ &\leq \mathbb{V}[X-Y] + \mathbb{V}Y + 2\sqrt{\mathbb{V}[X-Y]\mathbb{V}Y} = \left(\sqrt{\mathbb{V}[X-Y]} + \sqrt{\mathbb{V}Y}\right)^2 .\end{aligned}$$

This is the desired result. □

The following lemma is an extension of Lemma 7 by Azar et al. (2013) and its refined version by Agarwal et al. (2020).

Lemma E.2. *Suppose a sequence of deterministic policies $(\pi_k)_{k=0}^K$ and let*

$$q_{\pi'_k} := \begin{cases} r + \gamma P v_{\pi'_{k-1}} & \text{for } k \in [K] \\ q_{\pi_0} & \text{for } k = 0 \end{cases}.$$

Furthermore, let σ_k^2 and Σ_k^2 be non-negative functions over $\mathcal{X} \times \mathcal{A}$ defined by

$$\sigma_k^2(x,a) := \begin{cases} P(v_{\pi'_{k-1}})^2(x,a) - (Pv_{\pi'_{k-1}})^2(x,a) & \text{for } k \in [K] \\ P(v_{\pi_0})^2(x,a) - (Pv_{\pi_0})^2(x,a) & \text{for } k = 0 \end{cases}$$

and

$$\Sigma_k^2(x,a) := \mathbb{E}_k\left[\left(\sum_{t=0}^{\infty}\gamma^t r(X_t,A_t) - q_{\pi'_k}(X_0,A_0)\right)^2 \middle| X_0 = x, A_0 = a\right] \tag{12}$$

for $k \in \{0\} \cup [K]$, where $\mathbb{E}_k$ is the expectation over $(X_t, A_t)_{t=0}^{\infty}$ wherein $A_t \sim \pi_{k-t}(\cdot|X_t)$ until $t = k$, and $A_t \sim \pi_0(\cdot|X_t)$ thereafter. Then,

$$\sum_{j=0}^{k-1}\gamma^{j+1}P_{k-j}^{k-1}\sigma_{k-j} \leq \sqrt{2H^3}\mathbf{1}$$

for any $k \in [K]$.

For its proof, we need the following lemma.

Lemma E.3. *Suppose a sequence of deterministic policies $(\pi_k)_{k=0}^K$ and notations in Lemma E.2. Then, for any $k \in [K]$, we have that*

$$\Sigma_k^2 = \gamma^2 \sigma_k^2 + \gamma^2 P_{\pi_{k-1}} \Sigma_{k-1}^2 \,.$$

Proof. Let $R_s^u \coloneqq \sum_{t=s}^u \gamma^{t-s} r(X_t, A_t)$ and $\mathbb{E}_k[\cdot|x,a] \coloneqq \mathbb{E}_k[\cdot|X_0 = x, A_0 = a]$. We have that

$$\Sigma_k^2(x,a) = \mathbb{E}_k\left[\left(R_0^\infty - q_{\pi'_k}(X_0,A_0)\right)^2 \middle| x,a\right] \coloneqq \mathbb{E}_k\left[(I_1 + \gamma I_2)^2 \middle| x,a\right] ,$$

where $I_1 \coloneqq r(X_0,A_0) + \gamma q_{\pi'_{k-1}}(X_1,A_1) - q_{\pi'_k}(X_0,A_0)$, and $I_2 \coloneqq R_1^\infty - q_{\pi'_{k-1}}(X_1,A_1)$. With these notations, we see that

$$\begin{aligned}
\Sigma_k^2(x,a) &= \mathbb{E}_k\left[I_1^2 + \gamma^2 I_2^2 + 2\gamma I_1 I_2 \middle| x,a\right] \\
&= \mathbb{E}_k\left[I_1^2 + \gamma^2 I_2^2 + 2\gamma I_1 \mathbb{E}_{k-1}\left[I_2 | X_1, A_1\right] \middle| x,a\right] \\
&= \mathbb{E}_k\left[I_1^2 \middle| x,a\right] + \gamma^2 \mathbb{E}_k\left[I_2^2 \middle| x,a\right] \\
&= \mathbb{E}_k\left[I_1^2 \middle| x,a\right] + \gamma^2 P^{\pi_{k-1}} \Sigma_{k-1}^2(x,a) \,,
\end{aligned}$$

where the second line follows from the law of total expectation, and the third line follows since $\mathbb{E}_{k-1}[I_2|X_1,A_1] = 0$ due to the Markov property. The first term in the last line is $\gamma^2\sigma_k^2(x,a)$ because

$$\begin{aligned}
\mathbb{E}_k\left[I_1^2 \middle| x,a\right] &\overset{(a)}{=} \gamma^2 \mathbb{E}_k\left[\left(\underbrace{q_{\pi'_{k-1}}(X_1,A_1)}_{v_{\pi'_{k-1}}(X_1) \text{ from (b)}} - (P v_{\pi'_{k-1}})(X_0,A_0)\right)^2 \middle| x,a\right] \\
&= \gamma^2 \left(P\left(v_{\pi'_{k-1}}\right)^2\right)(x,a) + \gamma^2 (P v_{\pi'_{k-1}})^2(x,a) - 2(P v_{\pi'_{k-1}})^2(x,a) \\
&= \gamma^2 \left(P\left(v_{\pi'_{k-1}}\right)^2\right)(x,a) - \gamma^2 (P v_{\pi'_{k-1}})^2(x,a) \,,
\end{aligned}$$

where (a) follows from the definition that $q_{\pi'_k} = r + \gamma P v_{\pi'_{k-1}}$, and (b) follows since the policies are deterministic. From this argument, it is clear that $\Sigma_k^2 = \gamma^2\sigma_k^2 + \gamma^2 P_{\pi_{k-1}} \Sigma_{k-1}^2$, which is the desired result. ☐

Now, we are ready to prove Lemma E.2.

Proof of Lemma E.2. Let $H_k \coloneqq \sum_{j=0}^{k-1} \gamma^j$. Using Jensen's inequality twice,

$$\begin{aligned}
\sum_{j=0}^{k-1} \gamma^{j+1} P_{k-j}^{k-1} \sigma_{k-j} &\le \sum_{j=0}^{k-1} \gamma^{j+1} \sqrt{P_{k-j}^{k-1} \sigma_{k-j}^2} \\
&\le \gamma H_k \sum_{j=0}^{k-1} \frac{\gamma^{j+1}}{H_k} \sqrt{P_{k-j}^{k-1} \sigma_{k-j}^2} \\
&\le \sqrt{H_k \sum_{j=0}^{k-1} \gamma^{j+2} P_{k-j}^{k-1} \sigma_{k-j}^2} \le \sqrt{H \sum_{j=0}^{k-1} \gamma^{j+2} P_{k-j}^{k-1} \sigma_{k-j}^2} \,.
\end{aligned}$$

From Lemma E.3, we have that

$$\begin{aligned}\sum_{j=0}^{k-1} \gamma^{j+2} P_{k-j}^{k-1} \sigma_{k-j}^2 &= \sum_{j=0}^{k-1} \gamma^j P_{k-j}^{k-1} \left(\Sigma_{k-j}^2 - \gamma^2 P^{\pi_{k-1-j}} \Sigma_{k-1-j}^2 \right) \\ &= \sum_{j=0}^{k-1} \gamma^j P_{k-j}^{k-1} \left(\Sigma_{k-j}^2 - \gamma P^{\pi_{k-1-j}} \Sigma_{k-1-j}^2 + \gamma(1-\gamma) P^{\pi_{k-1-j}} \Sigma_{k-1-j}^2 \right) \\ &= \sum_{j=0}^{k-1} \gamma^j P_{k-j}^{k-1} \Sigma_{k-j}^2 - \sum_{j=1}^{k} \gamma^j P_{k-j}^{k-1} \Sigma_{k-j}^2 + \gamma(1-\gamma) \sum_{j=0}^{k-1} \gamma^j P_{k-1-j}^{k-1} \Sigma_{k-1-j}^2 \,.\end{aligned}$$

The final line is equal to $\Sigma_k^2 - \gamma^k P_0^{k-1} \Sigma_0^2 + \gamma(1-\gamma) \sum_{j=0}^{k-1} \gamma^j P_{k-1-j}^{k-1} \Sigma_{k-1-j}^2$. Finally, from the monotonicity of stochastic matrices and that $\mathbf{0} \le \Sigma_j^2 \le H^2 \mathbf{1}$ for any j,

$$\sum_{j=0}^{k-1} \gamma^{j+1} P_{k-j}^{k-1} \sigma_{k-j} \le \sqrt{2H^3} \mathbf{1} \,.$$

This concludes the proof. □

F. Proof of Theorem 3.3

As a reminder, let $\Phi \coloneqq \{\phi(x,a) : (x,a) \in \mathcal{X} \times \mathcal{A}\} \subset \mathbb{R}^d$. For $G \in \mathbb{R}^{d \times d}$ and $\phi \in \mathbb{R}^d$, we use the notation $\|\phi\|_G^2 \coloneqq \phi^\top G \phi$. Additionally, we use the operator norm of a matrix G and denote it as $\|G\| = \sup_{\phi^\top \phi = 1} \sqrt{(G\phi)^\top G\phi}$.

We first introduce an algorithm for computing the G-optimal design for finite $\mathcal{X}$, called the Frank-Wolfe algorithm from Todd (2016). The pseudocode is provided in Algorithm 7. The following theorem shows that Algorithm 7 outputs a near-optimal design with a small core set.

Theorem F.1 (**Proposition 3.17**, Todd (2016)). *Let $u_{\mathcal{C}} \coloneqq 4d \log\log(d+4) + 28$. For Φ satisfying Assumption 3.2 and if Φ is finite, Algorithm 7 with $f : \mathcal{X} \times \mathcal{A} \to (0, \infty)$ and $\varepsilon^{\mathrm{FW}} = d$ outputs a design ρ such that $g(\rho) \le 2d$ and the core set C with size at most $u_{\mathcal{C}}$.*

We extend the theorem to a compact Φ by passing to the limit. The proof of Theorem 3.3 is a modification of **Exercise 21.3** in Lattimore & Szepesvari (2020).

Proof of Theorem 3.3. Suppose that Φ satisfies Assumption 3.2 such that Φ is a compact subset of $\mathbb{R}^d$ and spans $\mathbb{R}^d$. Let $(\Phi_n)_n$ be a sequence of finite subsets with $\Phi_n \subset \Phi_{n+1}$. We suppose that Φ_n spans $\mathbb{R}^d$ and $\lim_{n\to\infty} D(\Phi, \Phi_n) = 0$ where D is the Hausdorff metric. Then let ρ_n be a G-optimal design for Φ_n with support of size at most $u_{\mathcal{C}}$ and $G_n \coloneqq \sum_{(x,a)\in\mathcal{X}\times\mathcal{A}} \rho_n(x,a)\phi(x,a)\phi(x,a)^\top$. Such the design is ensured to exist by Theorem F.1. Given any $\phi \in \Phi$, we have

$$\|\phi\|_{G_n^{-1}} \le \min_{b \in \Phi_n} \left(\|\phi - b\|_{G_n^{-1}} + \|b\|_{G_n^{-1}} \right) \le \sqrt{2d} + \min_{b\in\Phi_n} \|\phi - b\|_{G_n^{-1}} \,, \tag{13}$$

where the first inequality is due to the triangle inequality and the second inequality is due to Theorem F.1. Let $W \in \mathbb{R}^{d\times d}$ be an invertible matrix and $w_i \in \mathbb{R}^d$ be its $i \in [d]$ th column. We suppose that $w_i \in \Phi$ for any $i \in [d]$. Such W can be constructed due to the assumption that Φ spans $\mathbb{R}^d$. Then, the operator norm of $G_n^{-1/2}$ is bounded by

$$\left\|G_n^{-1/2}\right\| = \left\|W^{-1} W G_n^{-1/2}\right\| \le \left\|W^{-1}\right\| \left\|G_n^{-1/2} W\right\| = \left\|W^{-1}\right\| \sup_{\phi^\top\phi=1} \|W\phi\|_{G_n^{-1}} \,, \tag{14}$$

where the last equality is due to $\left\|G_n^{-1/2} W\right\| = \sup_{\phi^\top\phi=1} \sqrt{(G_n^{-1/2} W\phi)^\top G_n^{-1/2} W\phi} = \sup_{\phi^\top\phi=1} \|W\phi\|_{G_n^{-1}}$. Let ϕ_i be the i th element of $\phi \in \mathbb{R}^d$. Equation (14) is further bounded by

$$\sup_{\phi^\top \phi = 1} \|W\phi\|_{G_n^{-1}} \leq \sup_{\phi^\top \phi = 1} \sum_{i=1}^{d} |\phi_i| \underbrace{\|w_i\|_{G_n^{-1}}}_{\leq \sqrt{2d}} \leq 2d \, .$$

Therefore, we have $\left\|G_n^{-1/2}\right\| \leq 2d \left\|W^{-1}\right\|$. Taking the limit $n \to \infty$ shows that

$$\begin{aligned}\limsup_{n\to\infty} \|\phi\|_{G_n^{-1}} &\overset{(a)}{\leq} \sqrt{2d} + \limsup_{n\to\infty} \min_{b \in \Phi} \|\phi - b\|_{G_n^{-1}} \\ &\overset{(b)}{\leq} \sqrt{2d} + 2d \left\|W^{-1}\right\| \limsup_{n\to\infty} \min_{b\in\Phi} \sqrt{(\phi - b)^\top (\phi - b)} = \sqrt{2d} \, ,\end{aligned}$$

where (a) is due to (13) and (b) uses $\left\|G_n^{-1/2}\right\| \leq 2d \left\|W^{-1}\right\|$.

Since $\|\cdot\|_{G_n^{-1}} : \Phi \to \mathbb{R}$ is continuous and Φ is compact, it follows that

$$\limsup_{n\to\infty} \sup_{\phi\in\Phi} \|\phi\|^2_{G_n^{-1}} \leq 2d \, . \tag{15}$$

Notice that ρ_n may be represented as a tuple of vector/probability pairs with at most $u_{\mathcal{C}}$ entries and where the vectors lie in Φ. Since the set of all such tuples with the obvious topology forms a compact set, it follows that (ρ_n) has a cluster point ρ^*, which represents a distribution on Φ with support at most $u_{\mathcal{C}}$. Then, Equation (15) shows that $g(\rho^*) \leq 2d$. This concludes the proof. □

G. Proof of Weighted KW Bound (Lemma 4.3)

Proof. $\left|\phi^\top(x,a) W(f,z)\right|$ can be rewritten as

$$\begin{aligned}\left|\phi^\top(x,a) W(f,z)\right| &= \left|\phi^\top(x,a) G_f^{-1} \sum_{(y,b)\in\mathcal{C}_f} \rho_f(y,b) \frac{\phi(y,b)}{f(y,b)} \frac{z(y,b)}{f(y,b)}\right| \\ &\overset{(a)}{\leq} \left|\sum_{(y,b)\in\mathcal{C}_f} \rho_f(y,b) \phi^\top(x,a) G_f^{-1} \frac{\phi(y,b)}{f(y,b)}\right| \max_{(y',b')\in\mathcal{C}_f} \left|\frac{z(y',b')}{f(y',b')}\right| \\ &\overset{(b)}{\leq} \sum_{(y,b)\in\mathcal{C}_f} \left|\rho_f(y,b) \phi^\top(x,a) G_f^{-1} \frac{\phi(y,b)}{f(y,b)}\right| \max_{(y',b')\in\mathcal{C}_f} \left|\frac{z(y',b')}{f(y',b')}\right| \, ,\end{aligned} \tag{16}$$

where (a) is due to Hölder's inequality and (b) is due to the triangle inequality.

Next, for any $(x,a) \in \mathcal{X} \times \mathcal{A}$, we have

$$\begin{aligned}\left(\sum_{(y,b)\in\mathcal{C}_f} \left|\rho_f(y,b) \phi(x,a)^\top G_f^{-1} \frac{\phi(y,b)}{f(y,b)}\right|\right)^2 &\overset{(a)}{\leq} \sum_{(y,b)\in\mathcal{C}_f} \rho_f(y,b) \left|\phi(x,a)^\top G_f^{-1} \frac{\phi(y,b)}{f(y,b)}\right|^2 \\ &\overset{(b)}{=} f^2(x,a) \underbrace{\frac{\phi(x,a)^\top}{f(x,a)} G_f^{-1} \frac{\phi(x,a)}{f(x,a)}}_{\leq 2d \text{ from Theorem 4.1}}\end{aligned} \tag{17}$$

where (a) is due to Jensen's inequality, (b) is due to the definition of G_f. The claim holds by taking the square root for both sides of the inequality (17) and applying the result to the inequality (16). □

H. Formal Theorems and Proofs of Theorem 4.4 and Theorem 5.1

This section provides the concrete proofs of Theorem 4.4 and Theorem 5.1. Instead of the informal theorems of Theorem 4.4 and Theorem 5.1, we are going to prove the formal theorems below, Theorem H.1 and Theorem H.2, respectively.

Theorem H.1 (Sample complexity of `WLS-MDVI` with $f \approx \sigma(v_*)$). *Let c_0 be a positive constant such that $8 \geq c_0 \geq 6$ and $\widetilde{\sigma} \in \mathcal{F}_q$ be a random variable. Assume that $\varepsilon \in (0, 1/H]$ and an event*

$$\sigma(v_*) \leq \widetilde{\sigma} \leq \sigma(v_*) + 2\sqrt{H}\mathbf{1}$$

occurs with probability at least $1 - 4\delta/c_0$. Define

$$f^{wls} := \max\left(\min(\widetilde{\sigma}, H\mathbf{1}), \sqrt{H}\mathbf{1}\right),$$
$$K^{wls} := \left\lceil \frac{3}{1-\alpha} \log c_1 H + 1 \right\rceil,$$
$$\text{and } M^{wls} := \left\lceil \frac{c_2 d H^2}{\varepsilon^2} \log\left(\frac{2c_0^2 u_{\mathcal{C}} K^{wls}}{(c_0 - 5)\delta} \log_2 \frac{16 K^{wls} H^2}{(c_0 - 5)\delta} \right) \right\rceil$$

where $c_1, c_2 \geq 1$ are positive constants and $u_{\mathcal{C}} = 4d \log\log(d+4) + 28$. Then, there exist $c_1, c_2 \geq 1$ independent of d, H, X, A, ε, and δ such that `WLS-MDVI` is run with the settings $\alpha = \gamma$, $f = f^{wls}$, $K = K^{wls}$, $M = M^{wls}$ it outputs a sequence of policies $(\pi_k)_{k=0}^K$ such that $\|v_* - v_{\pi'_K}\|_\infty \leq \varepsilon$ with probability at least $1 - \delta$, using $\widetilde{\mathcal{O}}\left(u_{\mathcal{C}} K M\right) = \widetilde{\mathcal{O}}\left(d^2 H^3 / \varepsilon^2\right)$ samples from the generative model.

Theorem H.2 (Sapmle complexity of `WLS-MDVI` with $f = \mathbf{1}$). *Assume that $\varepsilon \in (0, 1/H]$. Let c_0 be a positive constant such that $8 \geq c_0 \geq 6$. Define*

$$K^{ls} := \left\lceil \frac{3}{1-\alpha} \log c_3 H + 1 \right\rceil$$
$$\text{and } M^{ls} := \left\lceil \frac{c_4 d H^2}{\varepsilon} \log \frac{2c_0^2 u_{\mathcal{C}} K^{ls}}{(c_0 - 5)\delta} \right\rceil$$

where $c_3, c_4 \geq 1$ are positive constants and $u_{\mathcal{C}} = 4d \log\log(d+4) + 28$. Then, there exist $c_3, c_4 \geq 1$ independent of d, H, X, A, ε and δ such that when `WLS-MDVI` is run with the settings $\alpha = \gamma$, $f = \mathbf{1}$, $K = K^{ls}$, and $M = M^{ls}$, it outputs v_K such that $\|v_* - v_K\|_\infty \leq \frac{1}{2}\sqrt{H}$ with probability at least $1 - 3\delta/c_0$, using $\widetilde{\mathcal{O}}\left(u_{\mathcal{C}} K M\right) = \widetilde{\mathcal{O}}\left(d^2 H^3 / \varepsilon\right)$ samples from the generative model.

The proof sketch is provided in Appendix H.2.

H.1. Notation and Frequently Used Facts for Proofs

Before moving on to the proofs, we introduce some notations and frequently used facts for theoretical analysis.

Notation for proofs. $\square$ denotes an indefinite constant that changes throughout the proof and is independent of d, H, X, A, ε, and δ.

For a sequence of policies $(\pi_k)_{k \in \mathbb{Z}}$, we let $T_j^i := T_{\pi_i} T_{\pi_{i-1}} \cdots T_{\pi_{j+1}} T_{\pi_j}$ for $i \geq j$, and $T_j^i := I$ otherwise.

For $k \in \{1, \ldots, N\}$, we write $\theta_k^* \in \mathbb{R}^d$ as the underlying unknown parameter vector satisfying $\phi^\top \theta_k^* = r + \gamma P v_{k-1}$. θ_k^* is ensured to exist by the property of linear MDPs. We also write $\overline{\theta}_k^*$ as its past moving average, i.e., $\overline{\theta}_k^* = \sum_{j=1}^k \alpha^{k-j} \theta_j^*$.

For Theorem H.2, $\mathbf{F}_{k,m}$ denotes the σ-algebra generated by random variables $\{y_{j,n,x,a} | (j, n, x, a) \in [k-2] \times [M] \times \mathcal{X} \times \mathcal{A}\} \cup \{y_{j,n,x,a} | (j, n, x, a) \in \{k-1\} \times [m-1] \times \mathcal{X} \times \mathcal{A}\}$. With an abuse of notation, for Theorem H.1, $\mathbf{F}_{k,m}$ denotes the σ-algebra generated by random variables $\{\widetilde{\sigma}\} \cup \{y_{j,n,x,a} | (j, n, x, a) \in [k-2] \times [M] \times \mathcal{X} \times \mathcal{A}\} \cup \{y_{j,n,x,a} | (j, n, x, a) \in \{k-1\} \times [m-1] \times \mathcal{X} \times \mathcal{A}\}$. Whether $\mathbf{F}_{k,m}$ is for Theorem H.2 or Theorem H.1 shall be clear from the context.

For the bounded positive function f used in `WLS-MDVI`, we introduce the shorthand $u_f := \max_{(x,a) \in \mathcal{X} \times \mathcal{A}} f(x, a)$ and $l_f := \min_{(x,a) \in \mathcal{X} \times \mathcal{A}} f(x, a)$.

Finally, throughout the proof, for $8 \geq c_0 > n > 0$, we write $\iota_1 := \log(2c_0 u_{\mathcal{C}} K / \delta)$, $\iota_{2,n} := \iota_1 + \log(c_0/(c_0 - n)) = \log(2c_0^2 u_{\mathcal{C}} K / (c_0 - n)\delta)$, and $\xi_{2,n} := \iota_{2,n} + \log\log_2(16 K H^2)$. Note that for any $8 \geq c_0 > n > 0$,

$$\xi_{2,n} \geq \iota_{2,n} \geq \iota_1 \tag{18}$$

due to $8 \geq c_0 - n > 0$ and $16 K H^2 / \delta \geq 16$. Whether K is from Theorem H.1 or Theorem H.2 shall be clear from the context.

Frequently Used Facts. Recall that $A_{\gamma,k} \coloneqq \sum_{j=0}^{k-1} \gamma^{k-j}\alpha^j$ and $A_k \coloneqq \sum_{j=0}^{k-1} \alpha^j$ for any non-negative integer k with $A_\infty \coloneqq 1/(1-\alpha)$. We often use $\alpha = \gamma$ due to the settings of Theorems H.1 and H.2. This indicates that $A_\infty = H$ and $A_{\gamma,k} = k\gamma^k$.

Recall that $\theta_k = \arg\min_{\theta \in \mathbb{R}^d} \sum_{(y,b)\in\mathcal{C}_f} \frac{\rho_f(y,b)}{f^2(y,b)} \left(\phi^\top(y,b)\theta - \widehat{q}_k(y,b)\right)^2$. Using the definition of W defined in Lemma 4.3 and G_f defined in Equation (5), the closed-form solution to θ_k is represented as $\theta_k = W(f, \hat{q}_k)$. In the similar manner, $\theta_k^* = W(f, \phi^\top \theta_k^*)$.

Since $\hat{q}_k - \phi^\top \theta_k^* = \varepsilon_k$, we have

$$
\theta_k - \theta_k^* = W(f, \hat{q}_k) - W(f, \phi^\top \theta_k^*) = W(f, \varepsilon_k)
$$
$$
\text{and } \overline{\theta}_k - \overline{\theta}_k^* = W\left(f, \sum_{j=1}^{k} \alpha^{k-j}\varepsilon_j\right) = W(f, E_k)\ ,
$$

Moreover, for any $k \in \{1, \dots, K\}$, we have that

$$
\begin{aligned}
s_k &= \phi^\top \overline{\theta}_k \\
&= \phi^\top \overline{\theta}_k^* + \phi^\top W(f, E_k) \\
&= \sum_{j=1}^{k} \alpha^{k-j}(r + \gamma P(w_{j-1} - \alpha w_{j-2})) + \phi^\top W(f, E_k) \\
&= A_k r + \gamma P w_{k-1} + \phi^\top W(f, E_k)\ .
\end{aligned}
\tag{19}
$$

In addition, we often mention the "monotonicity" of stochastic matrices: any stochastic matrix ρ satisfies that $\rho v \geq \rho u$ for any vectors v, u s.t. $v \geq u$. Examples of stochastic matrices in the proof are P, π, P^π, and πP. The monotonicity property is so frequently used that we do not always mention it.

H.2. Proof Sketch

This section provides proof sketches of Theorems H.1 and H.2, those are necessary to show Theorem J.1. The proofs follow the strategy of Kozuno et al. (2022) but with modifications for the linear function approximation.

Step 1: Error Propagation Analysis. The proof of Theorem H.1 is done by deriving a tight bound for $v_* - v_{\pi'_K}$. Recall that K is the number of iterations in `WLS-MDVI` and W is the operator defined in Lemma 4.3. The following lemmas provide the bound for any $k \in [K]$. We provide the proof in Appendix H.3.1.

Lemma H.3 (Error Propagation Analysis ($v_{\pi'_k}$))**.** *For any $k \in [K]$, $\mathbf{0} \leq v_* - v_{\pi'_k} \leq \Gamma_k$ where*

$$
\Gamma_k \coloneqq \frac{1}{A_\infty} \sum_{j=0}^{k-1} \gamma^j \left(\pi_k P_{k-j}^{k-1} - \pi_* P_*^j\right) \phi^\top W(f, E_{k-j}) + 2H\left(\alpha^k + \frac{A_{\gamma,k}}{A_\infty}\right)\mathbf{1}\ .
$$

Let

$$
\heartsuit_k \coloneqq H^{-1} \sum_{j=0}^{k-1} \gamma^j \pi_k P_{k-j}^{k-1} \left|\phi^\top W(f, E_{k-j})\right|,
$$
$$
\clubsuit_k \coloneqq H^{-1} \sum_{j=0}^{k-1} \gamma^j \pi_* P_*^j \left|\phi^\top W(f, E_{k-j})\right|,
$$
$$
\text{and } \diamondsuit_k \coloneqq \square H\left(\alpha^k + \frac{A_{\gamma,k}}{A_\infty}\right).
$$

We derive the bound of $\|v_* - v_{\pi'_K}\|_\infty$ by bounding $\heartsuit_K$, $\clubsuit_K$ and $\diamondsuit_K$. Since $\diamondsuit_k$ can be easily controlled by Lemma C.7, we focus on the bounds of $\heartsuit_K$ and $\clubsuit_K$. To derive the tight bounds of $\heartsuit_K$ and $\clubsuit_K$, we need to transform them into "TV technique compatible" forms; we will transform $\heartsuit_k$ into $\sum_{j=0}^{k-1} \gamma^j \pi_k P_{k-j}^{k-1} \sigma(v_{\pi_{k-j}})$ and $\clubsuit_k$ into $\sum_{j=0}^{k-1} \gamma^j \pi_* P_*^j \sigma(v_*)$. The transformations are provided in **Step 3** and **4**.

On the other hand, the proof of Theorem H.2 is done by deriving a coarse bound of $v_* - v_K$. Then, the following bound (Lemma H.4) is helpful. The proof is provided in Appendix H.3.2.

Lemma H.4 (Error Propagation Analysis (v_k))**.** *For any $k \in [K]$,*

$$-2\gamma^k H\mathbf{1} - \sum_{j=0}^{k-1} \gamma^j \pi_{k-1} P_{k-j}^{k-1} \phi^\top W(f, \varepsilon_{k-j}) \leq v_* - v_k \leq \Gamma_{k-1} + 2H\gamma^k \mathbf{1} - \sum_{j=0}^{k-1} \gamma^j \pi_{k-1} P_{k-1-j}^{k-2} \phi^\top W(f, \varepsilon_{k-j}) \,.$$

We first prove Theorem H.2 in the next **Step 2** since it is straightforward compared to Theorem H.1.

Step 2: Prove Theorem H.2. Note that $f = \mathbf{1}$ in Theorem H.2. As you can see from Lemma H.4, we need the bounds of $|\phi^\top W(\mathbf{1}, \varepsilon_k)|$ and $|\phi^\top W(\mathbf{1}, E_k)|$ for the proof.

By bounding ε_k and E_k using the Azuma-Hoeffding inequality (Lemma D.1), the weighted KW bound with $f = \mathbf{1}$ (Lemma 4.3) and the settings of Theorem H.2 yild $|\phi^\top W(\mathbf{1}, \varepsilon_k)| \leq \widetilde{\mathcal{O}}(1/\sqrt{H})\mathbf{1}$ and $|\phi^\top W(\mathbf{1}, E_k)| \leq \widetilde{\mathcal{O}}(\sqrt{H})\mathbf{1}$ with high-probability. Furthermore, $\diamondsuit_K$ is bounded by $\widetilde{\mathcal{O}}(1)$ due to Lemma C.7.

Inserting these results into Lemma H.4, we obtain $\|v_* - v_K\|_\infty \leq \widetilde{\mathcal{O}}(\sqrt{H})$ with high-probability, which is the desired result of Theorem H.2.

The detailed proofs of **Step 2** are provided in Appendix H.4 and Appendix H.5.

Step 3: Refined Bound of $\clubsuit_K$ for Theorem H.1. Recall that the weighting function f satisfies $\sigma(v_*) \leq f \leq \sigma(v_*) + 2\sqrt{H}\mathbf{1}$ and $\sqrt{H}\mathbf{1} \leq f \leq H\mathbf{1}$ in Theorem H.1. The assumptions allow us to apply TV technique to $\clubsuit_K$ when the bound of $\phi^\top W(f, E_k)$ scales to f. This is where the weighted KW bound (Lemma 4.3) comes in.

Due to Lemma 4.3, we have $|\phi^\top W(f, E_k)| \leq \sqrt{2d} f \max_{(y,b)\in\mathcal{C}_f} |E_k(y,b)/f(y,b)|$. Thus, the tight bound of can be obtained by tightly bounding $\max_{(y,b)\in\mathcal{C}_f} |E_k(y,b)/f(y,b)|$.

By applying the Bernstein-type inequality (Lemma D.4) to $|E_k/f|$, discounted sum of $\sigma(v_j)/f$ from $j = 1$ to k appears inside the bound of $|E_k/f|$. We decompose it as $\sigma(v_j)/f \leq |v_* - v_j|/\sqrt{H} + \mathbf{1}$ by Lemma E.1 and Lemma D.5. Therefore, we obtain a discounted sum of $|v_* - v_j|$ in $|E_k/f|$ bound, which can be bounded in a similar way to **Step 2**.

Now we have the bound of $\phi^\top W(f, E_k)$ which scales to f. Combined with the settings of Theorem H.1, we obtain $|\phi^\top W(f, E_k)| \leq \widetilde{\mathcal{O}}(\varepsilon(\sigma(v_*)/\sqrt{H} + \mathbf{1}))$. The TV technique is therefore applicable to $\clubsuit_K$ and thus $\clubsuit_K \leq \widetilde{\mathcal{O}}(\varepsilon)$.

The detailed proofs of **Step 3** are provided in Appendix H.6.1.

Step 4: Refined Bound of $\heartsuit_K$ for Theorem H.1. We need a further transformation since TV technique in $\heartsuit_K$ requires $\sigma(v_{\pi'_k})$, not $\sigma(v_*)$. To this end, we decompose $\sigma(v_*)$ as $\sigma(v_*) \leq \sigma(v_* - v_{\pi'_k}) + \sigma(v_{\pi'_k}) \leq |v_* - v_{\pi'_k}| + \sigma(v_{\pi'_k})$ by Lemma E.1 and Lemma D.5. Thus, we need the bound of $|v_* - v_{\pi'_k}|$ which requires the coarse bound of $\|\phi^\top W(f, E_k)\|_\infty$.

By applying the Azuma-Hoeffding inequality to E_k, the settings of Theorem H.2 yields $\|\phi^\top W(f, E_k)\|_\infty \leq \widetilde{\mathcal{O}}(\sqrt{H})$. Inserting this bound to Lemma H.3, $\|v_* - v_{\pi'_k}\|_\infty \leq \widetilde{\mathcal{O}}(\sqrt{H}) + 2(H+k)\gamma^k$ (Lemma H.15).

By taking a similar procedure as **Step 3**, together with the bound of $\|v_* - v_{\pi'_k}\|_\infty$, $\heartsuit_K$ is bounded by $\widetilde{\mathcal{O}}(\varepsilon H^{-1.5}) \sum_{j=0}^{k-1} \pi_k P_{k-j}^{k-1} (\sigma(v_{\pi_{k-j}}) + \sqrt{H}\mathbf{1})$. Then, the TV technique yilds $\heartsuit_K \leq \widetilde{\mathcal{O}}(\varepsilon)$.

The detailed proofs of **Step 4** are provided in Appendix H.6.2.

Finally, we obtain the desired result of Theorem H.1 by inserting the bounds of $\heartsuit_K$ and $\clubsuit_K$ to Lemma H.15 (Appendix H.6.3.)

H.3. Proofs of Error Propagation Analysis (Step 1)

H.3.1. PROOF OF LEMMA H.3

Proof. Note that

$$\mathbf{0} \leq v_* - v_{\pi'_k} = \frac{A_k}{A_\infty}\left(v_* - v_{\pi'_k}\right) + \alpha^k \left(v_* - v_{\pi'_k}\right) \leq \frac{A_k}{A_\infty}\left(v_* - v_{\pi'_k}\right) + 2H\alpha^k \mathbf{1}$$

due to $v_* - v_{\pi'_k} \leq 2H\mathbf{1}$. Therefore, we need an upper bound for $A_k(v_* - v_{\pi'_k})$. We decompose $A_k(v_* - v_{\pi'_k})$ to $A_k v_* - w_k$ and $w_k - A_k v_{\pi'_k}$. Then, we derive upper bounds for each of them (inequalities (20) and (21), respectively). The desired result is obtained by summing up those bounds.

Upper bound for $A_k v_* - w_k$. We prove by induction that for any $k \in [K]$,

$$A_k v_* - w_k \leq H A_{\gamma,k} \mathbf{1} - \sum_{j=0}^{k-1} \gamma^j \pi_* P_*^j \phi^T W(f, E_{k-j}) \,. \tag{20}$$

We have that

$$\begin{aligned}
A_k v_* - w_k &\overset{(a)}{\leq} \pi_*(A_k q_* - s_k) \\
&\overset{(b)}{=} \pi_* \left(A_k q_* - A_k r - \gamma P w_{k-1} - \phi^T W(f, E_k)\right) \\
&\overset{(c)}{=} \pi_* \left(\gamma P(A_k v_* - w_{k-1}) - \phi^T W(f, E_k)\right) \\
&\overset{(d)}{\leq} \pi_* \left(\gamma P(A_{k-1} v_* - w_{k-1}) + \alpha^{k-1}\gamma H \mathbf{1} - \phi^T W(f, E_k)\right) \,,
\end{aligned}$$

where (a) is due to the greediness of π_k, (b) is due to the equation (19), (c) is due to the Bellman equation for q_*, and (d) is due to the fact that $(A_k - A_{k-1})v_* = \alpha^{k-1} v_* \leq \alpha^{k-1} H\mathbf{1}$. For $k = 1$, using (a), (b), and (c) with the facts that $w_0 = \mathbf{0}$ and $A_1 = 1$, we have

$$A_1 v_* - w_1 \leq \pi_* \left(\gamma P v_* - \phi^T W(f, E_1)\right) \leq \gamma H \mathbf{1} - \pi_* \phi^T W(f, E_1)$$

and thus the inequality (20) holds for $k = 1$. From the step (d) above and induction, it is straightforward to verify that the inequality (20) holds for other k.

Upper bound for $w_k - A_k v_{\pi'_k}$. We prove by induction that for any $k \in [K]$,

$$w_k - A_k v_{\pi'_k} \leq H A_{\gamma,k} \mathbf{1} + \sum_{j=0}^{k-1} \gamma^j \pi_k P_{k-j}^{k-1} \phi^T W(f, E_{k-j}) \,. \tag{21}$$

Recalling that $v_{\pi'_k} = \pi_k T_0^{k-1} q_{\pi_0}$, we deduce that

$$\begin{aligned}
w_k - A_k v_{\pi'_k} &\overset{(a)}{=} \pi_k \left(s_k - A_k T_0^{k-1} q_{\pi_0}\right) \\
&\overset{(b)}{=} \pi_k \left(A_k r + \gamma P w_{k-1} - A_k T_1^{k-1} q_{\pi_0} + \phi^T W(f, E_k)\right) \\
&\overset{(c)}{=} \pi_k \left(\gamma P \left(w_{k-1} - A_k v_{\pi'_{k-1}}\right) + \phi^T W(f, E_k)\right) \\
&\overset{(d)}{\leq} \pi_k \left(\gamma P(w_{k-1} - A_{k-1} v_{\pi'_{k-1}}) + \alpha^{k-1}\gamma H \mathbf{1} + \phi^T W(f, E_k)\right) \,,
\end{aligned}$$

where (a) follows from the definition of w_k, (b) is due to the equation (19) and $T_0^{k-1} q_{\pi_0} = T_1^{k-1} q_{\pi_0}$, (c) is due to the equation $r - T_1^{k-1} q_{\pi_0} = -P v_{\pi'_{k-1}}$ which follows from the definition of the Bellman operator, and (d) is due to the fact that $(A_k - A_{k-1}) v_{\pi'_{k-1}} = \alpha^{k-1} v_{\pi'_{k-1}} \geq -\alpha^{k-1} H\mathbf{1}$. For $k = 1$, using (a), (b), and (c) with the facts that $w_0 = \mathbf{0}$ and $A_1 = 1$, we have

$$w_1 - A_1 v_{\pi'_1} = \pi_1 \left(-\gamma P v_{\pi'_0} + \phi^T W(f, E_1)\right) \leq \gamma H \mathbf{1} + \pi_1 \phi^T W(f, E_1) \,,$$

and thus the inequality (21) holds for $k = 1$. From the step (d) above and induction, it is straightforward to verify that the inequality (21) holds for other k. □

H.3.2. Proof of Lemma H.4

We first prove an intermediate result.

Lemma H.5. *For any $k \in [K]$,*

$$v_{\pi'_{k-1}} + \sum_{j=0}^{k-1} \gamma^j \pi_{k-1} P_{k-1-j}^{k-2} \phi^\top W(f, \varepsilon_{k-j}) - \gamma^k H \mathbf{1} \leq v_k \leq v_{\pi'_k} + \sum_{j=0}^{k-1} \gamma^j \pi_k P_{k-j}^{k-1} \phi^\top W(f, \varepsilon_{k-j}) + \gamma^k H \mathbf{1}\,.$$

Proof. From the greediness of π_{k-1}, $v_k = w_k - \alpha w_{k-1} \leq \pi_k(s_k - \alpha s_{k-1}) = \pi_k(r + \gamma P v_{k-1} + \phi^\top W(f, \varepsilon_k))$. By induction on k, therefore,

$$v_k \leq \sum_{j=0}^{k-1} \gamma^j \pi_k P_{k-j}^{k-1} \left(r + \phi^\top W(f, \varepsilon_{k-j})\right) + \underbrace{\gamma^k \pi_k P_0^{k-1} v_0}_{=\mathbf{0}} \leq \sum_{j=0}^{k-1} \gamma^j \pi_k P_{k-j}^{k-1} \left(r + \phi^\top W(f, \varepsilon_{k-j})\right)\,.$$

Note that

$$T_0^{k-1} q_{\pi_0} = \sum_{j=0}^{k-1} \gamma^j P_{k-j}^{k-1} r + \gamma^k \underbrace{P_0^{k-1} q_{\pi_0}}_{\geq -H\mathbf{1}} \implies \sum_{j=0}^{k-1} \gamma^j P_{k-j}^{k-1} r \leq T_0^{k-1} q_{\pi_0} + \gamma^k H\,.$$

Accordingly, $v_k \leq \pi_k T_0^{k-1} q_{\pi_0} + \sum_{j=0}^{k-1} \gamma^j \pi_k P_{k-j}^{k-1} \phi^\top W(f, \varepsilon_{k-j}) + \gamma^k H \mathbf{1}\,.$

Similarly, from the greediness of π_k, $v_k = w_k - \alpha w_{k-1} \geq \pi_{k-1}(s_k - \alpha s_{k-1}) \geq \pi_{k-1}(r + \gamma P v_{k-1} + \phi^\top W(f, \varepsilon_k))$. By induction on k, therefore,

$$v_k \geq \sum_{j=0}^{k-1} \gamma^j \pi_{k-1} P_{k-1-j}^{k-2} \left(r + \phi^\top W(f, \varepsilon_{k-j})\right) + \underbrace{\gamma^{k-1} \pi_{k-1} P_0^{k-2} P v_0}_{=\mathbf{0}}\,.$$

Note that $T_0^{k-2} q_{\pi_0} = T_0^{k-2}(r + \gamma P v_{\pi_0})$ and

$$T_0^{k-2} q_{\pi_0} = \sum_{j=0}^{k-1} \gamma^j P_{k-1-j}^{k-2} r + \gamma^k \underbrace{P_0^{k-2} P v_{\pi_0}}_{\leq H\mathbf{1}} \implies \sum_{j=0}^{k-1} \gamma^j P_{k-1-j}^{k-2} r \geq T_0^{k-2} q_{\pi_0} - \gamma^k H\,.$$

Accordingly, $v_k \geq \pi_{k-1} T_0^{k-2} q_{\pi_0} + \sum_{j=0}^{k-1} \gamma^j \pi_{k-1} P_{k-1-j}^{k-2} \phi^\top W(f, \varepsilon_{k-j}) - \gamma^k H \mathbf{1}\,.$ □

Proof of Lemma H.4. From Lemma H.5 and $\pi_k T_{\pi_{k-1}} \cdots T_{\pi_1} q_{\pi_0} = v_{\pi'_k} \leq v_*$, we have that

$$v_{\pi'_{k-1}} + \sum_{j=0}^{k-1} \gamma^j \pi_{k-1} P_{k-1-j}^{k-2} \phi^\top W(f, \varepsilon_{k-j}) - 2\gamma^k H \mathbf{1} \leq v_k \leq v_* + \sum_{j=0}^{k-1} \gamma^j \pi_k P_{k-j}^{k-1} \phi^\top W(f, \varepsilon_{k-j}) + 2\gamma^k H \mathbf{1}\,, \tag{22}$$

where we loosened the bound by multiplying $\gamma^k H$ by 2. By simple algebra, for any $k \in [K]$,

$$v_* - v_k \geq -2\gamma^k H - \sum_{j=0}^{k-1} \gamma^j \pi_k P_{k-j}^{k-1} \phi^\top W(f, \varepsilon_{k-j}) \tag{23}$$

$$\text{and} \quad v_{\pi'_{k-1}} - v_k \leq 2\gamma^k H \mathbf{1} - \sum_{j=0}^{k-1} \gamma^j \pi_{k-1} P_{k-1-j}^{k-2} \phi^\top W(f, \varepsilon_{k-j})\,. \tag{24}$$

For the second inequality, from Lemma H.3,

$$v_{\pi'_{k-1}} \geq v_* - \frac{1}{A_\infty} \sum_{j=0}^{k-2} \gamma^j \left(\pi_{k-1} P_{k-1-j}^{k-2} - \pi_* P_*^j\right) \phi^\top W(f, E_{k-1-j}) - 2H \left(\alpha^{k-1} + \frac{A_{\gamma,k-1}}{A_\infty}\right) \mathbf{1} \tag{25}$$

for any $k \in \{2, \dots, K\}$. Since $v_{\pi_0} \geq v_* - 2H$ and the empty sum is defined to be 0, the inequality (25) holds for $k = 1$. Therefore, by applying (25) to (24), we have that

$$v_* - v_k \leq 2H\gamma^k + 2H \left(\alpha^{k-1} + \frac{A_{\gamma,k-1}}{A_\infty} \right) \mathbf{1} + \frac{1}{A_\infty} \sum_{j=0}^{k-2} \gamma^j \left(\pi_{k-1} P_{k-1-j}^{k-2} - \pi_* P_*^j \right) \phi^\top W(f, E_{k-1-j}) - \sum_{j=0}^{k-1} \gamma^j \pi_{k-1} P_{k-1-j}^{k-2} \phi^\top W(f, \varepsilon_{k-j}) \tag{26}$$

for any $k \in [K]$. Lemma H.4 holds by combining (26) and (23). □

H.4. Lemmas and Proofs of $\phi^\top W(f, \varepsilon_k)$ and $\phi^\top W(f, E_k)$ Bounds (Step 2)

This section provides formal lemmas and proofs about the high-probability bounds of $\phi^\top W(f, \varepsilon_k)$ and $\phi^\top W(f, E_k)$.

We first introduce the necessary events for the proofs.

Event 1 ($\mathcal{E}_1$)**.** *The input f of* `WLS-MDVI` *satisfies $\sigma(v_*)(x,a) \leq f(x,a) \leq \sigma(v_*)(x,a) + 2\sqrt{H}$, and $\sqrt{H} \leq f(x,a) \leq H$ for all $(x,a) \in \mathcal{X} \times \mathcal{A}$.*

Event 2 ($\mathcal{E}_2$)**.** *v_k is bounded by $2H$ for all $k \in [K]$.*

Event 3 ($\mathcal{E}_3$)**.** $\left|\phi^\top(x,a) W(f, E_k)\right| \leq (8Hu_f/l_f)\sqrt{dA_\infty \iota_{2,5}/M}$ *for all $(x,a,k) \in \mathcal{X} \times \mathcal{A} \times [K]$.*

Event 4 ($\mathcal{E}_4$)**.** $\left|\phi^\top(x,a) W(f, \varepsilon_k)\right| \leq (8\gamma H f(x,a)/l_f)\sqrt{d\iota_{2,5}/M}$ *for all $(x,a,k) \in \mathcal{X} \times \mathcal{A} \times [K]$.*

Event 5 ($\mathcal{E}_5$)**.** $|\phi^\top(x,a) W(f, E_k)| \leq \sqrt{2d} f(x,a) \left(8H\xi_{2,5}/(l_f M) + 2\sqrt{2\xi_{2,5}/(l_f^2 M)} + V_k \right)$ *where*

$$V_k := 2\sqrt{\frac{2\xi_{2,5}}{M} \sum_{j=1}^{k} \alpha^{2(k-j)} \max_{(y,b) \in \mathcal{C}_f} \frac{\sigma^2(v_{j-1})(y,b)}{f^2(y,b)}} \,,$$

for all $(x,a,k) \in \mathcal{X} \times \mathcal{A} \times [K]$

$\mathcal{E}_1$ is for the condition of f in Theorem H.1, and $\mathcal{E}_2$ is for the use of concentration inequalities. Our goal is to show that $\mathcal{E}_3$, $\mathcal{E}_4$, and $\mathcal{E}_5$ occur with high probability in Theorem H.2 and Theorem H.1.

H.4.1. Lemmas and Proofs of v_k Bound ($\mathcal{E}_2$)

We first show that $\mathcal{E}_2$ occurs with high probability. The following Lemma H.6 is for Theorem H.2, and Lemma H.7 is for Theorem H.1.

Lemma H.6. *With the settings of Theorem H.2, there exists $c_4 \geq 1$ independent of d, H, X, A, ε and δ such that $\mathbb{P}(\mathcal{E}_2{}^c) \leq \delta/c_0$.*

Lemma H.7. *With the settings of Theorem H.1, there exists $c_2 \geq 1$ independent of d, H, X, A, ε and δ such that $\mathbb{P}(\mathcal{E}_2{}^c | \mathcal{E}_1) \leq \delta/c_0$.*

Proof. From the greediness of the policies π_k and π_{k-1},

$$\pi_{k-1}\phi^\top\theta_k = \pi_{k-1}(s_k - \alpha s_{k-1}) \leq v_k \leq \pi_k(s_k - \alpha s_{k-1}) = \pi_k \phi^\top \theta_k. \tag{27}$$

Let $\varepsilon'_k := \varepsilon_k / \|v_{k-1}\|_\infty$ be a normalized error. We prove the claim by bounding $\phi^\top\theta_k$ as

$$\begin{aligned} \left|\phi^\top\theta_k\right| = \left|\phi^\top W(f, \hat{q}_k)\right| &\overset{(a)}{\leq} \left|\phi^\top W(f, \phi^\top\theta_k^*)\right| + \left|\phi^\top W(f, \varepsilon_k)\right| = |r + \gamma P v_{k-1}| + \left|\phi^\top W(f, \varepsilon_k)\right| \\ &\overset{(b)}{\leq} (1 + \gamma\|v_{k-1}\|_\infty)\,\mathbf{1} + \frac{u_f\sqrt{2d}}{l_f} \max_{(x,a)\in\mathcal{C}_f} |\varepsilon_k(x,a)|\,\mathbf{1} \\ &\overset{(c)}{=} (1 + \gamma\|v_{k-1}\|_\infty)\,\mathbf{1} + \frac{u_f\sqrt{2d}}{l_f} \|v_{k-1}\|_\infty \max_{(x,a)\in\mathcal{C}_f} |\varepsilon'_k(x,a)|\,\mathbf{1}\,, \end{aligned} \tag{28}$$

where (a) uses the triangle inequality, (b) is due to Lemma 4.3 and since r is bounded by 1, and (c) uses $\varepsilon'_k = \varepsilon_k / \|v_{k-1}\|_\infty$. We also used the shorthand $u_f \coloneqq \max_{(x,a)\in\mathcal{X}\times\mathcal{A}} f(x,a)$ and $l_f \coloneqq \min_{(x,a)\in\mathcal{X}\times\mathcal{A}} f(x,a)$.

We need to bound $\max_{(x,a)\in\mathcal{C}_f} |\varepsilon'_k(x,a)|$. For $(x,a)\in\mathcal{X}\times\mathcal{A}$,

$$\varepsilon'_k(x,a) = \frac{\gamma}{M}\sum_{m=1}^{M} \underbrace{\Big(v_{k-1}(y_{k-1,m,x,a}) - Pv_{k-1}(x,a)\Big)/\|v_{k-1}\|_\infty}_{\text{bounded by 2}}$$

is a sum of bounded martingale differences with respect to $(\mathbf{F}_{k,m})_{m=1}^M$. Using the Azuma-Hoeffding inequality (Lemma D.1) and taking the union bound over $(x,a)\in\mathcal{C}_f$ and $k\in[K]$, we have

$$\mathbb{P}\left(\exists(x,a,k)\in\mathcal{C}_f\times[K] \text{ s.t. } |\varepsilon'_k(x,a)| \geq \gamma\sqrt{\frac{8\iota_1}{M}}\right) \leq \frac{\delta}{c_0}\,. \tag{29}$$

We are now ready to prove Lemma H.6 and Lemma H.7 by induction. The claims hold for $k=0$ since $v_0 = \mathbf{0}$. Assume that v_{k-1} is bounded by $2H$ for some $k\geq 1$.

Lemma H.6 proof Note that $u_f/l_f = 1$ due to the settings of Theorem H.2. Therefore, the following inequality holds with probability at least $1-\delta/c_0$.

$$\|\phi^\top\theta_k\|_\infty \overset{(a)}{\leq} 1+\gamma 2H + 2H\sqrt{2d}\max_{(x,a)\in\mathcal{C}_f}|\varepsilon'_k(x,a)| \overset{(b)}{\leq} 1+\gamma 2H + 8H\gamma\sqrt{\frac{d\iota_1}{M}}\,,$$

where (a) is due to (28) with the induction hypothesis and (b) the second inequality is due to (29). Since $H=1/(1-\gamma)$, by simple algebra, some M such that $M\geq 64\gamma^2H^2d\iota_1$ satisfies $\|\phi^\top\theta_k\|_\infty\leq 2H$ with probability at least $1-\delta/c_0$.

Recall that $M=\lceil c_4 dH^2\iota_{2,5}/\varepsilon\rceil$ in Theorem H.2. Due to the assumption of $\varepsilon\leq 1/H$ and $\iota_{2,5}\geq\iota_1$ by (18), the value of M in Theorem H.2 satisfies $M\geq 64\gamma^2H^2d\iota_1$ for some c_4. Lemma H.6 hence holds by inserting the result into the inequality (27) with induction.

Lemma H.7 proof Note that $u_f/l_f\leq\sqrt{H}$ due to the condition of Lemma H.7. Therefore, the following inequality holds with probability at least $1-\delta/c_0$.

$$\|\phi^\top\theta_k\|_\infty \overset{(a)}{\leq} 1+\gamma 2H + 2H\sqrt{2dH}\max_{(x,a)\in\mathcal{C}_f}|\varepsilon'_k(x,a)| \overset{(b)}{\leq} 1+\gamma 2H + 8H\gamma\sqrt{\frac{dH\iota_1}{M}}\,,$$

where (a) is due to (28) with the induction hypothesis and (b) the second inequality is due to (29). Since $H=1/(1-\gamma)$, by simple algebra, some M such that $M\geq 64\gamma^2H^3d\iota_1$ satisfies $\|\phi^\top\theta_k\|_\infty\leq 2H$ with probability at least $1-\delta/c_0$.

Recall that $M=\lceil c_2 dH^2\xi_{2,5}/\varepsilon^2\rceil$ in Theorem H.1. Due to the assumption of $\varepsilon\leq 1/H$ and $\xi_{2,5}\geq\iota_1$ by (18), the value of M in Theorem H.1 satisfies $M\geq 64\gamma^2H^3d\iota_1$ for some c_2. Lemma H.7 hence holds by inserting the result into the inequality (27) with induction.

□

H.4.2. Lemmas and Proofs of Coarse $\phi^\top W(f, E_k)$ Bound ($\mathcal{E}_3$)

The following Lemma H.8 is for Theorem H.2, and Lemma H.9 is for Theorem H.1.

Lemma H.8. *With the settings of Theorem H.2, there exists $c_4\geq 1$ independent of d, H, X, A, ε and δ such that* $\mathbb{P}({\mathcal{E}_3}^c|\mathcal{E}_2)\leq\delta/c_0$.

Lemma H.9. *With the settings of Theorem H.1, there exists $c_2\geq 1$ independent of d, H, X, A, ε and δ such that* $\mathbb{P}({\mathcal{E}_3}^c|\mathcal{E}_1\cap\mathcal{E}_2)\leq\delta/c_0$.

Proof. For any $(x, a) \in \mathcal{X} \times \mathcal{A}$ and $k \in [K]$, we have

$$\left|\phi^\top(x,a) W(f, E_k)\right| \le \frac{\sqrt{2d} f(x,a)}{l_f} \underbrace{\max_{(y',b') \in \mathcal{C}_f} \left|\sum_{j=1}^{k} \alpha^{k-j} \varepsilon_j(y', b')\right|}_{\heartsuit_k}, \tag{30}$$

where the inequality is due to the weighted KW bound (Lemma 4.3).

We need to bound $\heartsuit_k$. Note that for a fixed $k \in [K]$ and $(x, a) \in \mathcal{C}_f$,

$$\sum_{j=1}^{k} \alpha^{k-j} \varepsilon_j(x,a) = \frac{\gamma}{M} \sum_{j=1}^{k} \alpha^{k-j} \sum_{m=1}^{M} \underbrace{\Big(v_{j-1}(y_{j-1,m,x,a}) - P v_{j-1}(x,a)\Big)}_{\text{bounded by } 4H \text{ due to } \mathcal{E}_2}$$

is a sum of bounded martingale differences with respect to $(\mathbf{F}_{j,m})_{j=1,m=1}^{k,M}$. We are now ready to prove Lemma H.8 and Lemma H.9 using the conditional Azuma-Hoeffding inequality (Lemma D.2).

Lemma H.8 proof In the settings of Theorem H.2, some c_4 satisfies that $\mathbb{P}(\mathcal{E}_2) \ge 1 - \delta/c_0$ due to Lemma H.6. Using the conditional Azuma-Hoeffding inequality (Lemma D.2) and taking the union bound over $(x, a) \in \mathcal{C}_f$ and $k \in [K]$,

$$\mathbb{P}\left(\exists (x,a,k) \in \mathcal{C}_f \times [K] \text{ s.t. } \sum_{j=1}^{k} \alpha^{k-j} \varepsilon_j(x,a) \ge \gamma H \sqrt{\frac{32 A_\infty \iota_{2,1}}{M}} \middle| \mathcal{E}_2 \right) \le \frac{\delta}{c_0}.$$

where $\iota_{2,1} = \iota_1 + \log(c_0/(c_0 - 1))$ is due to the condition by $\mathcal{E}_2$. We used $\iota_{2,1}$ since $1/(1 - \delta/c_0) \le c_0/(c_0 - 1)$.

Therefore, $\heartsuit_k \le H\sqrt{32 A_\infty \iota_{2,1}/M}$ with probability at least $1 - \delta/c_0$ for all $k \in [K]$. The claim holds by inserting $\heartsuit_k$ into the inequality (30).

Lemma H.9 proof In the settings of Theorem H.1, we have $\mathbb{P}(\mathcal{E}_1) \ge 1 - 4\delta/c_0$ and some c_2 satisfies that $\mathbb{P}\left({\mathcal{E}_2}^c | \mathcal{E}_1\right) \le \delta/c_0$ due to Lemma H.7. Therefore, $\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \ge 1 - 5\delta/c_0$ holds due to Lemma C.1.

Using Lemma D.2 and taking the union bound over $(x, a) \in \mathcal{C}_f$ and $k \in [K]$,

$$\mathbb{P}\left(\exists (x,a,k) \in \mathcal{C}_f \times [K] \text{ s.t. } \sum_{j=1}^{k} \alpha^{k-j} \varepsilon_j(x,a) \ge \gamma H \sqrt{\frac{32 A_\infty \iota_{2,5}}{M}} \middle| \mathcal{E}_1 \cap \mathcal{E}_2 \right) \le \frac{\delta}{c_0},$$

where $\iota_{2,5} = \iota_1 + \log(c_0/(c_0 - 5))$ is due to the condition by $\mathcal{E}_1 \cap \mathcal{E}_2$. We used $\iota_{2,5}$ since $1/(1 - 5\delta/c_0) \le c_0/(c_0 - 5)$. Lemma H.9 holds in the same way as the proof of Lemma H.8.

□

H.4.3. Lemmas and Proofs of Coarse $\phi^\top W(f, \varepsilon_k)$ Bound ($\mathcal{E}_4$)

The following Lemma H.10 is for Theorem H.2, and Lemma H.11 is for Theorem H.1.

Lemma H.10. *With the settings of Theorem H.2, there exists $c_4 \ge 1$ independent of d, H, X, A, ε and δ such that $\mathbb{P}({\mathcal{E}_4}^c | \mathcal{E}_2) \le \delta/c_0$.*

Lemma H.11. *With the settings of Theorem H.1, there exists $c_2 \ge 1$ independent of d, H, X, A, ε and δ such that $\mathbb{P}({\mathcal{E}_4}^c | \mathcal{E}_1 \cap \mathcal{E}_2) \le \delta/c_0$.*

Proof. For any $(x, a) \in \mathcal{X} \times \mathcal{A}$ and $k \in [K]$, we have

$$|\phi^\top(x,a) W(f, \varepsilon_k)| \le \frac{\sqrt{2d} f(x,a)}{l_f} \underbrace{\max_{(y',b') \in \mathcal{C}_f} |\varepsilon_k(y', b')|}_{\heartsuit_k}, \tag{31}$$

where the inequality is due to the weighted KW bound (Lemma 4.3).

We need to bound $\heartsuit_k$. Note that for a fixed $k \in [K]$ and $(x, a) \in \mathcal{C}_f$,

$$\varepsilon_k(x,a) = \frac{\gamma}{M} \sum_{m=1}^{M} \underbrace{\Big(v_{k-1}(y_{k-1,m,x,a}) - Pv_{k-1}(x,a)\Big)}_{\text{bounded by } 4H \text{ due to } \mathcal{E}_2}$$

is a sum of bounded martingale differences with respect to $(\mathbf{F}_{k,m})_{m=1}^{M}$. We are ready to prove Lemma H.10 and Lemma H.11 using the conditional Azuma-Hoeffding inequality (Lemma D.2).

Lemma H.10 proof Note that some c_4 satisfies that $\mathbb{P}(\mathcal{E}_2) \geq 1 - \delta/c_0$ in the settings of Theorem H.2 due to Lemma H.6. Using the conditional Azuma-Hoeffding inequality (Lemma D.2) and taking the union bound over $(x, a) \in \mathcal{C}_f$ and $k \in [K]$, we have

$$\mathbb{P}\left(\exists (x,a,k) \in \mathcal{C}_f \times [K] \text{ s.t. } |\varepsilon_k(x,a)| \geq \gamma H \sqrt{\frac{32\iota_{2,1}}{M}} \middle| \mathcal{E}_2 \right) \leq \frac{\delta}{c_0}.$$

where $\iota_{2,1} = \iota_1 + \log(c_0/(c_0 - 1))$ is due to the condition by $\mathcal{E}_2$. We used $\iota_{2,1}$ since $1/(1 - \delta/c_0) \leq c_0/(c_0 - 1)$.

Therefore, $\heartsuit_k \leq \gamma H \sqrt{32\iota_{2,1}/M}$ with probability at least $1 - \delta/c_0$ for all $k \in [K]$. The claim holds by inserting $\heartsuit_k$ into the inequality (31).

Lemma H.11 proof Due to Lemma C.1 and Lemma H.7, some c_2 satisfies that $\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \geq 1 - 5\delta/c_0$ in the settings of Theorem H.1. Therefore, using Lemma D.2 and taking the union bound over $(x, a) \in \mathcal{C}_f$ and $k \in [K]$,

$$\mathbb{P}\left(\exists (x,a,k) \in \mathcal{C}_f \times [K] \text{ s.t. } |\varepsilon_k(x,a)| \geq \gamma H \sqrt{\frac{32\iota_{2,5}}{M}} \middle| \mathcal{E}_1 \cap \mathcal{E}_2 \right) \leq \frac{\delta}{c_0}.$$

where $\iota_{2,5} = \iota_1 + \log(c_0/(c_0 - 5))$ is due to the condition by $\mathcal{E}_1 \cap \mathcal{E}_2$. We used $\iota_{2,5}$ since $1/(1 - 5\delta/c_0) \leq c_0/(c_0 - 5)$.

The claim holds in the same way as the proof of Lemma H.10. □

H.4.4. Lemma and Proof of Refined $\phi^\top W(f, E_k)$ Bound ($\mathcal{E}_5$)

The following Lemma H.12 is for Theorem H.1.

Lemma H.12. *With the settings of Theorem H.1, there exists $c_2 \geq 1$ independent of d, H, X, A, ε and δ such that $\mathbb{P}\left(\mathcal{E}_5{}^c | \mathcal{E}_1 \cap \mathcal{E}_2\right) \leq \delta/c_0$.*

Proof. For any $(x, a) \in \mathcal{X} \times \mathcal{A}$ and $k \in [K]$, we have

$$\left|\phi^\top(x,a) W(f, E_k)\right| \leq \sqrt{2d} f(x,a) \max_{(y',b') \in \mathcal{C}_f} \frac{1}{f(y',b')} \left| \sum_{j=1}^{k} \alpha^{k-j} \varepsilon_j(y',b') \right|. \tag{32}$$

where the inequality is due to the weighted KW bound (Lemma 4.3).

We further bound Equation (32) by bounding $\left|\sum_{j=1}^{k} \alpha^{k-j} \varepsilon_j(x,y)\right|$ over $(x, y) \in \mathcal{C}_f$. For a fixed $k \in [K]$ and $(x, a) \in \mathcal{C}_f$,

$$\sum_{j=1}^{k} \alpha^{k-j} \varepsilon_j(x,a) = \gamma \sum_{j=1}^{k} \alpha^{k-j} \frac{1}{M} \sum_{m=1}^{M} \underbrace{\Big(v_{j-1}(y_{j-1,m,x,a}) - Pv_{j-1}(x,a)\Big)}_{\text{bounded by } 4H \text{ due to } \mathcal{E}_2}$$

is a sum of bounded martingale differences with respect to $(\mathbf{F}_{j,m})_{j=1,m=1}^{k,M}$.

In the settings of Theorem H.1, we have $\mathbb{P}(\mathcal{E}_1) \geq 1-4\delta/c_0$ and some c_2 satisfies that $\mathbb{P}\left(\mathcal{E}_2{}^c|\mathcal{E}_1\right) \leq \delta/c_0$ due to Lemma H.7. Therefore, $\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \geq 1-5\delta/c_0$ holds due to Lemma C.1. Using the conditional Bernstein-type inequality (Lemma D.4) and taking the union bound over $k \in [K]$ and $(x,a) \in \mathcal{C}_f$, we have

$$\mathbb{P}\left(\left|\sum_{j=1}^{k} \alpha^{k-j}\varepsilon_j(x,a)\right| \geq \frac{8H\xi_{2,5}}{M} + 2\sqrt{2}\sqrt{\frac{\xi_{2,5}}{M}\left(1+\sum_{j=1}^{k}\alpha^{2(k-j)}\mathrm{Var}(v_{j-1})(x,a)\right)}\middle|\mathcal{E}_1\cap\mathcal{E}_2\right) \leq \frac{\delta}{c_0}\,, \tag{33}$$

for all $(x,a,k) \in \mathcal{C}_f \times [K]$. Here, $\xi_{2,5} = \iota_1 + \log(c_0/(c_0-5)) + \log\log_2(16KH^2)$ is due to the condition by $\mathcal{E}_1 \cap \mathcal{E}_2$. We used $\xi_{2,5}$ since $1/(1-5\delta/c_0) \leq c_0/(c_0-5)$.

Using the result, we have the following inequality with probability at least $1-\delta/c_0$. For all $(x,a,k) \in \mathcal{C}_f \times [K]$,

$$\begin{aligned}
&\max_{(y',b')\in\mathcal{C}_f} \frac{1}{f(y',b')}\left|\sum_{j=1}^{k}\alpha^{k-j}\varepsilon_j(y',b')\right| \\
&\overset{(a)}{\leq} \max_{(y',b')\in\mathcal{C}_f} \frac{1}{f(y',b')}\left(\frac{8H\xi_{2,5}}{M} + 2\sqrt{2}\sqrt{\frac{\xi_{2,5}}{M}\left(1+\sum_{j=1}^{k}\alpha^{2(k-j)}\mathrm{Var}(v_{j-1})(y',b')\right)}\right) \\
&\overset{(b)}{\leq} \max_{(y',b')\in\mathcal{C}_f} \frac{1}{f(y',b')}\left(\frac{8H\xi_{2,5}}{M} + 2\sqrt{2}\sqrt{\frac{\xi_{2,5}}{M}} + 2\sqrt{2}\sqrt{\frac{\xi_{2,5}}{M}\sum_{j=1}^{k}\alpha^{2(k-j)}\mathrm{Var}(v_{j-1})(y',b')}\right) \\
&\overset{(c)}{\leq} \frac{8H\xi_{2,5}}{Ml_f} + 2\sqrt{2}\sqrt{\frac{\xi_{2,5}}{Ml_f^2}} + 2\sqrt{2}\sqrt{\frac{\xi_{2,5}}{M}\sum_{j=1}^{k}\alpha^{2(k-j)}\max_{(y',b')\in\mathcal{C}_f}\frac{\mathrm{Var}(v_{j-1})(y',b')}{f(y',b')}}
\end{aligned}$$

where (a) is due to (33), (b) is due to Lemma C.2, and (c) uses $l_f = \min_{(x,a)\in\mathcal{X}\times\mathcal{A}} f(x,a)$. The claim holds by inserting the result into the inequality (32).

□

We are now ready to prove Theorem H.2.

H.5. Proof of Theorem H.2 (Step2)

Proof of Theorem H.2. We condition the proof by the event $\mathcal{E}_2\cap\mathcal{E}_3\cap\mathcal{E}_4$. Note that when `WLS-MDVI` is run with the settings defined in Theorem H.2, $\mathbb{P}(\mathcal{E}_2{}^c) \leq \delta/c_0$ due to Lemma H.6, $\mathbb{P}(\mathcal{E}_3{}^c|\mathcal{E}_2) \leq \delta/c_0$ due to Lemma H.8, and $\mathbb{P}(\mathcal{E}_4{}^c|\mathcal{E}_2) \leq \delta/c_0$ due to Lemma H.10 . Using Lemma C.1, these indicate that

$$\mathbb{P}(\mathcal{E}_2\cap\mathcal{E}_3\cap\mathcal{E}_4) \geq \mathbb{P}(\mathcal{E}_2) - \mathbb{P}\left((\mathcal{E}_3\cap\mathcal{E}_4)^c|\mathcal{E}_2\right) \geq \mathbb{P}(\mathcal{E}_2) - \mathbb{P}(\mathcal{E}_3{}^c|\mathcal{E}_2) - \mathbb{P}(\mathcal{E}_4{}^c|\mathcal{E}_2) \geq 1 - \frac{3\delta}{c_0}\,.$$

Therefore, $\mathcal{E}_2\cap\mathcal{E}_3\cap\mathcal{E}_4$ occurs with probability at least $1-2\delta/c_0$.

We now prove the claim by bounding $v_* - v_K$. Recall Lemma H.4 that, for any $k \in [K]$,

$$-2\gamma^k H\mathbf{1} - \sum_{j=0}^{k-1}\gamma^j\pi_{k-1}P_{k-j}^{k-1}\phi^\top W(f,\varepsilon_{k-j}) \leq v_* - v_k \leq \Gamma_{k-1} + 2H\gamma^k\mathbf{1} - \sum_{j=0}^{k-1}\gamma^j\pi_{k-1}P_{k-1-j}^{k-2}\phi^\top W(f,\varepsilon_{k-j})\,,$$

where

$$\Gamma_k \coloneqq \frac{1}{A_\infty}\sum_{j=0}^{k-1}\gamma^j\left(\pi_k P_{k-j}^{k-1} - \pi_* P_*^j\right)\phi^\top W(f,E_{k-j}) + 2H\left(\alpha^k + \frac{A_{\gamma,k}}{A_\infty}\right)\mathbf{1}\,.$$

When $\alpha = \gamma$, this bounds $\|v_* - v_K\|_\infty$ as

$$\|v_* - v_K\|_\infty \leq \underbrace{\frac{1}{H}\sum_{j=0}^{K-1}\gamma^j \left\|\left(\pi_K P_{K-j}^{K-1} - \pi_* P_*^j\right)\phi^\top W(f, E_{K-j})\right\|_\infty}_{\heartsuit} + \underbrace{\square(H+K)\gamma^K}_{\clubsuit} + \underbrace{H \max_{j\in[K]} \left\|\phi^\top W(\mathbf{1}, \varepsilon_j)\right\|_\infty}_{\diamondsuit} . \tag{34}$$

We bound for each of them. Note that $u_f/l_f = 1$, $K = \left\lceil \frac{3}{1-\alpha}\log c_3 H + 1\right\rceil$, and $M = \lceil c_4 dH^2 \iota_{2,5}/\varepsilon\rceil$ due to the settings of Theorem H.2.

First, $\heartsuit$ can be bounded as

$$\heartsuit \leq \frac{2}{H}\sum_{j=0}^{K-1}\gamma^j \left\|\phi^\top W(f, E_{K-j})\right\|_\infty \overset{(a)}{\leq} \frac{2}{H}\sum_{j=0}^{K-1}\gamma^j \left(8H\sqrt{\frac{dH\iota_{2,5}}{M}}\right) \overset{(b)}{\leq} \square\sqrt{\frac{H}{c_4}}\varepsilon \overset{(c)}{\leq} \square\sqrt{\frac{H}{c_4}} ,$$

where (a) is due to $\mathcal{E}_3$, (b) is due to the value of M, and (c) is due to $\varepsilon \in (0, 1/H]$.

Second, Lemma C.7 with the value of K indicates that

$$\clubsuit \leq \frac{\square}{c_3} .$$

Finally, $\diamondsuit$ can be bounded as

$$\diamondsuit \overset{(a)}{\leq} 8\gamma H^2\sqrt{\frac{d\iota_{2,5}}{M}} \overset{(b)}{\leq} \square H\sqrt{\frac{\varepsilon}{c_4}} \overset{(c)}{\leq} \square\sqrt{\frac{H}{c_4}} ,$$

where (a) is due to $\mathcal{E}_4$, (b) is due to the value of M, and (c) is due to $\varepsilon \in (0, 1/H]$.

Inserting these results into the inequality (34), we have $\|v_* - v_K\|_\infty \leq \square\sqrt{H}(c_3^{-1} + c_4^{-0.5})$. Therefore, for some c_3 and c_4, the claim holds. □

H.6. Proof of Theorem H.1 (Step 3) and (Step 4)

As discribed in Appendix H.2, the proof requires tight bounds on $\heartsuit_k = H^{-1}\sum_{j=0}^{k-1}\gamma^j \pi_k P_{k-j}^{k-1}|\phi^\top W(f, E_{k-j})|$ and $\clubsuit_k = H^{-1}\sum_{j=0}^{k-1}\gamma^j \pi_* P_*^j|\phi^\top W(f, E_{k-j})|$. We first derive the bound of $\clubsuit_K$ and then derive the bound of $\heartsuit_K$.

H.6.1. $H^{-1}\sum_{j=0}^{K-1}\gamma^j \pi_* P_*^j|\phi^\top W(f, E_{K-j})|$ Bound ($\clubsuit_K$)

As discribed in **Step 3** of Appendix H.2, we need a bound of the discounted sum of $\sigma(v_* - v_k)$ for $\clubsuit_k$. Then, the following lemma is useful.

Lemma H.13. *Conditioned on $\mathcal{E}_2 \cap \mathcal{E}_3 \cap \mathcal{E}_4$,*

$$\|v_* - v_k\|_\infty < \min\{3H, \Psi_k\} \text{ and } \|\sigma(v_* - v_k)\|_\infty < \min\{3H, \Psi_k\} ,$$

where

$$\Psi_k = 3H\left(\max(\gamma, \alpha)^{k-1} + \frac{A_{\gamma,k-1}}{A_\infty}\right) + \frac{24H^2 u_f}{l_f}\sqrt{\frac{d\iota_2}{M}}\left(1 + \sqrt{\frac{1}{A_\infty}}\right)$$

for all $k \in [K]$.

Proof. Let $e_k \coloneqq \gamma^k H + H \max_{j\in[k]} \left\|\phi^\top W(f, \varepsilon_j)\right\|_\infty$. From Lemma H.4, for any $k \in [K]$,

$$v_* - v_k \geq -2\gamma^k H\mathbf{1} - \sum_{j=0}^{k-1} \gamma^j \pi_{k-1} P_{k-j}^{k-1} \phi^\top W(f, \varepsilon_{k-j}) \geq -2e_k \mathbf{1} ,$$

$$\begin{aligned} \text{and } v_* - v_k \leq& \Gamma_{k-1} + 2H\gamma^k \mathbf{1} - \sum_{j=0}^{k-1} \gamma^j \pi_{k-1} P_{k-1-j}^{k-2} \phi^\top W(f, \varepsilon_{k-j}) \\ \leq& 2H \left(\alpha^{k-1} + \frac{A_{\gamma,k-1}}{A_\infty} + \frac{1}{A_\infty} \max_{j\in[k-1]} \left\|\phi^\top W(f, E_j)\right\|_\infty \right) \mathbf{1} + 2e_k \mathbf{1} . \end{aligned}$$

Note that $\|v_* - v_k\|_\infty \leq 3H$ due to $\mathcal{E}_2$ for any $k \in [K]$. Also, due to $\mathcal{E}_4$ and $\mathcal{E}_3$, $\left\|\phi^\top W(f, \varepsilon_k)\right\|_\infty \leq (8Hu_f/l_f)\sqrt{d\iota_{2,5}/M}$ and $\left\|\phi^\top W(f, E_k)\right\|_\infty \leq (8Hu_f/l_f)\sqrt{dA_\infty \iota_{2,5}/M}$ for any $k \in [K]$. Therefore,

$$|v_* - v_k| \leq 3H \min\left\{1, \max(\gamma, \alpha)^{k-1} + \frac{A_{\gamma,k-1}}{A_\infty} + \frac{8Hu_f}{l_f}\sqrt{\frac{d\iota_{2,5}}{M}}\left(1 + \sqrt{\frac{1}{A_\infty}}\right)\right\}\mathbf{1}$$

for all $k \in [K]$. Also, due to Lemma D.5,

$$\sigma(v_* - v_k) \leq 3H \min\left\{1, 2\max(\gamma, \alpha)^{k-1} + \frac{A_{\gamma,k-1}}{A_\infty} + \frac{8Hu_f}{l_f}\sqrt{\frac{d\iota_{2,5}}{M}}\left(1 + \sqrt{\frac{1}{A_\infty}}\right)\right\}\mathbf{1} .$$

This concludes the proof. □

Now we have the following bound on $\clubsuit_K$.

Lemma H.14. *Assume that $\varepsilon \in (0, 1/H]$. Conditioned on $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \cap \mathcal{E}_4 \cap \mathcal{E}_5$, with the settings of Theorem H.1,*

$$\clubsuit_K = \frac{1}{H}\sum_{k=0}^{K-1} \gamma^k \pi_* P_*^k |\phi^\top W(f, E_{K-k})| \leq \square \left(c_1^{-1} + c_2^{-0.5}\right) \varepsilon \mathbf{1} .$$

Proof. Using the conditions $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \cap \mathcal{E}_4 \cap \mathcal{E}_5$, for all $k \in [K]$, we have

$$\max_{(x,a)\in\mathcal{C}_f} \frac{\sigma(v_k)(x,a)}{f(x,a)} \overset{(a)}{\leq} \max_{(x,a)\in\mathcal{C}_f} \underbrace{\frac{\sigma(v_*)(x,a)}{f(x,a)}}_{\leq 1 \text{ from } \mathcal{E}_1} + \frac{\sigma(v_* - v_k)(x,a)}{l_f} \overset{(b)}{\leq} 1 + \frac{\Psi_k}{\sqrt{H}} , \tag{35}$$

where (a) is due to Lemma E.1 and (b) is due to the conditions of $\mathcal{E}_2 \cap \mathcal{E}_3 \cap \mathcal{E}_4$ and Lemma H.13.

Note that with the conditions and the settings of $\varepsilon \in (0, 1/H]$, $\alpha = \gamma$, and $M = \left\lceil \frac{c_2 dH^2 \xi_{2,5}}{\varepsilon^2} \right\rceil$, we have $A_\infty = H$, $A_{\gamma,k} = k\gamma^k$, and $u_f/l_f \leq \sqrt{H}$. Therefore,

$$\begin{aligned} \frac{\Psi_k}{\sqrt{H}} =& 3\sqrt{H}\left(\max(\gamma, \alpha)^{k-1} + \frac{A_{\gamma,k-1}}{A_\infty}\right) + \frac{24Hu_f}{l_f}\sqrt{\frac{dH\iota_2}{M}}\left(1 + \sqrt{\frac{1}{A_\infty}}\right) \\ \leq& \square\left(\sqrt{H}\gamma^{k-1} + \frac{(k-1)\gamma^{k-1}}{\sqrt{H}} + \underbrace{H^2\sqrt{\frac{d\iota_{2,5}}{M}}}_{\leq \varepsilon H}\right) \leq \square\left(\sqrt{H}\gamma^{k-1} + \frac{(k-1)\gamma^{k-1}}{\sqrt{H}} + 1\right) , \end{aligned}$$

where the last inequality uses that $\varepsilon \in (0, 1/H]$. Using this result, for any $k \in [K]$,

$$\gamma^{2(K-k)}\left(1 + \frac{\Psi_k}{\sqrt{H}}\right)^2 \leq \square\left(\sqrt{H}\gamma^{K-1} + \frac{(k-1)\gamma^{K-1}}{\sqrt{H}} + \gamma^{K-k}\right)^2 \leq \square\left(H\gamma^{2K-2} + \frac{(k-1)^2\gamma^{2K-2}}{H} + \gamma^{2K-2k}\right)^2 \tag{36}$$

where the last inequality is due to the Cauchy-Schwarz inequality (Lemma C.4). This result implies that

$$
\begin{aligned}
V_K &= 2\sqrt{\frac{2\xi_{2,5}}{M}\sum_{j=1}^{K}\alpha^{2(K-j)}\max_{(y,b)\in\mathcal{C}_f}\frac{\sigma^2(v_{j-1})(y,b)}{f^2(y,b)}} \overset{(a)}{\leq} 2\sqrt{\frac{2\xi_{2,5}}{M}\sum_{j=1}^{K}\gamma^{2(K-j)}\left(1+\frac{\Psi_j}{\sqrt{H}}\right)^2} \\
&\overset{(b)}{\leq} \sqrt{\frac{\square\xi_{2,5}}{M}\left(HK\gamma^{2K-2}+\frac{\gamma^{2K-2}}{H}\underbrace{\sum_{i=1}^{K}(i-1)^2}_{\leq K^3 \text{ by Lemma C.8}}+\underbrace{\sum_{j=1}^{K}\gamma^{2(K-j)}}_{H}\right)} \qquad (37) \\
&\overset{(c)}{\leq} \sqrt{\frac{\square\xi_{2,5}}{M}}\left(\sqrt{HK}\gamma^{K-1}+\frac{K^{1.5}\gamma^{K-1}}{\sqrt{H}}+\sqrt{H}\right) \overset{(d)}{\leq} \sqrt{\frac{\square H\xi_{2,5}}{M}}\left(1+\frac{1}{c_1}\right) \overset{(e)}{\leq} \frac{\square\varepsilon}{\sqrt{c_2 H d}}\left(1+\frac{1}{c_1}\right)\,,
\end{aligned}
$$

where (a) is due to (35), (b) is due to (36), and (c) is due to Lemma C.2. (d) uses that $\sqrt{K}\gamma^{K-1} \leq K^{1.5}\gamma^{K-1} \leq \square/c_1$ due to the value of K and Lemma C.7, and (e) is due to the definition of M.

Finally,

$$
\begin{aligned}
\frac{1}{H}\sum_{k=0}^{K-1}\gamma^k\pi_* P_*^k\left|\phi^\top W(f,E_{K-k})\right| &\overset{(a)}{\leq} \frac{\sqrt{2d}}{H}\left(\frac{8H\xi_{2,5}}{l_f M}+2\sqrt{\frac{2\xi_{2,5}}{l_f^2 M}}+V_K\right)\sum_{k=0}^{K-1}\gamma^k\pi_* P_*^k f \\
&\overset{(b)}{\leq} \frac{\square\sqrt{d}}{H}\left(\frac{\xi_{2,5}\sqrt{H}}{M}+\frac{1}{H}\sqrt{\frac{\xi_{2,5}}{M}}+V_K\right)\sum_{k=0}^{K-1}\gamma^k\pi_* P_*^k f \\
&\overset{(c)}{\leq} \frac{\square\sqrt{d}}{H}\left(\frac{\xi_{2,5}\sqrt{H}}{M}+\frac{1}{H}\sqrt{\frac{\xi_{2,5}}{M}}+V_K\right)\left(H\sqrt{H}\mathbf{1}+\underbrace{\sum_{k=0}^{K-1}\gamma^k\pi_* P_*^k\sigma(v_*)}_{\leq\sqrt{2H^3}\mathbf{1} \text{ by Lemma E.2}}\right) \\
&\overset{(d)}{\leq} \frac{\square\sqrt{d}}{H}\left(\frac{\varepsilon^2}{c_2 dH\sqrt{H}}+\frac{\varepsilon}{H^2\sqrt{c_2 d}}+\frac{\varepsilon}{\sqrt{c_2 H d}}\left(1+\frac{1}{c_1}\right)\right)\mathbf{1} \\
&\leq \square\left(c_2^{-0.5}+c_1^{-1}c_2^{-0.5}\right)\varepsilon\mathbf{1} \\
&\leq \square\left(c_1^{-1}+c_2^{-0.5}\right)\varepsilon\mathbf{1}\,.
\end{aligned}
$$

where (a) is due to $\mathcal{E}_5$ and since V_k is increasing with respect to k, (b) is due to $l_f \geq \sqrt{H}$ by $\mathcal{E}_1$, (c) is due to $\sigma(v_*) \leq f \leq \sigma(v_*) + 2\sqrt{H}\mathbf{1}$ by $\mathcal{E}_1$, and (d) is due to $M = \left\lceil \frac{c_2 d H^2 \xi_{2,5}}{\varepsilon^2} \right\rceil$ with the inequality (37). This concludes the proof. □

H.6.2. $H^{-1}\sum_{k=0}^{K-1}\gamma^k\pi_K P_{K-k}^{K-1}|\phi^\top W(f,E_{K-k})|$ BOUND ($\heartsuit_K$)

As discribed in **Step 4** of Appendix H.2, we need a coarse bound of $\sigma(v_* - v_{\pi'_k})$ for $\heartsuit_k$. Then, the following lemma is useful.

Lemma H.15. *Conditioned on $\mathcal{E}_1 \cap \mathcal{E}_3$, with the settings of Theorem H.1, the output policies $(\pi_k)_{k=0}^K$ satisfy that $\|v_* - v_{\pi'_k}\|_\infty \leq \square/\sqrt{c_2}\mathbf{1} + 2(H+k)\gamma^k$ for all $k \in [K]$.*

Proof of Lemma H.15. Note that $A_\infty = H$ due to $\alpha = \gamma$ and $u_f/l_f \leq \sqrt{H}$ due to $\mathcal{E}_1$. Moreover, $\mathcal{E}_3$ and the setting of $M = \left\lceil \frac{c_2 d H^2 \xi_{2,5}}{\varepsilon^2} \right\rceil$ indicate that

$$
\left\|\phi^\top W(f,E_k)\right\|_\infty \leq \square\frac{u_f}{l_f}\sqrt{\frac{dA_\infty H^2\iota_{2,5}}{M}} \leq \frac{\square H\epsilon}{\sqrt{c_2}} \leq \frac{\square}{\sqrt{c_2}}\,,
$$

for any $k \in [K]$ where the last inequality is due to $\epsilon \in (0, 1/H]$.

Therefore,

$$
\frac{1}{A_\infty}\sum_{j=0}^{k-1}\gamma^j\left(\pi_k P_{k-j}^{k-1}-\pi_* P_*^j\right)\phi^T W(f,E_{k-j}) \leq \frac{2}{H}\sum_{j=0}^{k-1}\gamma^j\left\|\phi^T W(f,E_{k-j})\right\|_\infty\mathbf{1} \leq \frac{\square}{\sqrt{c_2}}\mathbf{1}\,,
$$

and thus $v_* - v_{\pi'_k} \leq \square/\sqrt{c_2}\mathbf{1} + 2(H+k)\gamma^k\mathbf{1}$ due to Lemma H.3. This concludes the proof. $\square$

Now we are ready to derive the bound of $\heartsuit_K$.

Lemma H.16. *Assume that $\varepsilon \in (0, 1/H]$. Conditioned on $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \cap \mathcal{E}_4 \cap \mathcal{E}_5$, with the settings of Theorem H.1,*

$$\heartsuit_K = \frac{1}{H}\sum_{k=0}^{K-1} \gamma^k \pi_K P_{K-k}^{K-1} \left|\phi^\top W(f, E_{K-k})\right| \leq \square\left(c_1^{-1} + c_2^{-0.5}\right)\varepsilon\mathbf{1}\,.$$

Proof. Following similar steps as in the proofs of Appendix H.6.1, we obtain the following bounds.

$$\frac{\Psi_k}{\sqrt{H}} \leq \square\left(\sqrt{H}\gamma^{k-1} + \frac{(k-1)\gamma^{k-1}}{\sqrt{H}} + 1\right) \quad \text{for any } k \in [K]\,,$$
$$V_K \leq \frac{\square\varepsilon}{\sqrt{c_2 H d}}\left(1 + \frac{1}{c_1}\right)\,,$$
$$\text{and } \frac{1}{H}\sum_{k=0}^{K-1}\gamma^k \pi_K P_{K-k}^{K-1}|\phi^\top W(f, E_{K-k})| \leq \frac{\square\sqrt{d}}{H}\left(\frac{\xi_{2,5}\sqrt{H}}{M} + \frac{1}{H}\sqrt{\frac{\xi_{2,5}}{M}} + V_K\right)\left(H\sqrt{H}\mathbf{1} + \sum_{k=0}^{K-1}\gamma^k\pi_K P_{K-k}^{K-1}\sigma(v_*)\right)\,. \tag{38}$$

We thus need to bound $\sum_{k=0}^{K-1}\gamma^k\pi_K P_{K-k}^{K-1}\sigma(v_*)$. Note that $\sigma(v_*)$ can be decomposed as

$$\sigma(v_*) \overset{(a)}{\leq} \sigma(v_* - v_{\pi'_k}) + \sigma(v_{\pi'_k}) \overset{(b)}{\leq} \|v_* - v_{\pi'_k}\|_\infty \mathbf{1} + \sigma(v_{\pi'_k}) \overset{(c)}{\leq} \frac{\square}{\sqrt{c_2}}\mathbf{1} + 2(H+k)\gamma^k\mathbf{1} + \sigma(v_{\pi'_k})\,,$$

where (a) is due to Lemma E.1, (b) is due Lemma D.5, and (c) is due to Lemma H.15. Accordingly,

$$\begin{aligned}
\sum_{k=0}^{K-1}\gamma^k\pi_K P_{K-k}^{K-1}\sigma(v_*) &\leq \sum_{k=0}^{K-1}\gamma^k\pi_K P_{K-k}^{K-1}\left(\frac{\square}{\sqrt{c_2}}\mathbf{1} + 2(H+K-k)\gamma^{K-k}\mathbf{1} + \sigma(v_{\pi'_{K-k}})\right)\\
&\leq \frac{\square H}{\sqrt{c_2}}\mathbf{1} + \square\left(HK\gamma^K + \gamma^K\underbrace{\sum_{k=0}^{K-1}(K-k)}_{K^2 \text{ by Lemma C.8}}\right)\mathbf{1} + \underbrace{\sum_{k=0}^{K-1}\gamma^k\pi_K P_{K-k}^{K-1}\sigma(v_{\pi'_{K-k}})}_{\sqrt{2H^3}\mathbf{1} \text{ by Lemma E.2}}\\
&\overset{(a)}{\leq} \square H\sqrt{H}\left(c_2^{-0.5} + c_1^{-1} + 1\right)\mathbf{1} \overset{(b)}{\leq} \square H\sqrt{H}\mathbf{1}
\end{aligned}$$

where (a) uses that $K\gamma^K \leq K^2\gamma^K \leq \square/c_1$ due to the value of K and Lemma C.7, and (b) uses $c_1, c_2 \geq 1$.

Inserting the result into the inequality (38), and following similar steps as in the proof of Appendix H.6.1, we have

$$\frac{1}{H}\sum_{k=0}^{K-1}\gamma^k\pi_K P_{K-k}^{K-1}|\phi^\top W(f, E_{K-k})| \leq \frac{\square\sqrt{d}}{H}\left(\frac{\varepsilon^2}{c_2 d H\sqrt{H}} + \frac{\varepsilon}{H^2\sqrt{c_2 d}} + \frac{\varepsilon}{\sqrt{c_2 H d}}\left(1 + \frac{1}{c_1}\right)\right)\mathbf{1} \leq \square\left(c_1^{-1} + c_2^{-0.5}\right)\varepsilon\mathbf{1}\,.$$

This concludes the proof. $\square$

H.6.3. Proof of Theorem H.1

The derived bounds of $\heartsuit_K$ and $\clubsuit_K$ yield the following proof of Theorem H.1.

Proof of Theorem H.1. We condition the proof by the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \cap \mathcal{E}_4 \cap \mathcal{E}_5$. Note that when `WLS-MDVI` is run with the settings defined in Theorem H.1,

$$\mathbb{P}(\mathcal{E}_1) \geq 1 - 4\delta/c_0\,, \quad \underbrace{\mathbb{P}\left({\mathcal{E}_2}^c|\mathcal{E}_1\right) \leq \delta/c_0}_{\text{from Lemma H.7}}\,, \quad \underbrace{\mathbb{P}\left({\mathcal{E}_3}^c|\mathcal{E}_1 \cap \mathcal{E}_2\right) \leq \delta/c_0}_{\text{from Lemma H.9}}\,,$$
$$\underbrace{\mathbb{P}\left({\mathcal{E}_4}^c|\mathcal{E}_1 \cap \mathcal{E}_2\right) \leq \delta/c_0}_{\text{from Lemma H.11}} \quad \underbrace{\mathbb{P}\left({\mathcal{E}_5}^c|\mathcal{E}_1 \cap \mathcal{E}_2\right) \leq \delta/c_0}_{\text{from Lemma H.12}}\,.$$

With Lemma C.1, these indicates that

$$\begin{aligned}\mathbb{P}(\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3\cap\mathcal{E}_4\cap\mathcal{E}_5) &\geq \mathbb{P}(\mathcal{E}_1\cap\mathcal{E}_2) - \mathbb{P}((\mathcal{E}_3\cap\mathcal{E}_4\cap\mathcal{E}_5)^c|\mathcal{E}_1\cap\mathcal{E}_2)\\ &\geq \mathbb{P}(\mathcal{E}_1) - \mathbb{P}({\mathcal{E}_2}^c|\mathcal{E}_1)\\ &\quad - \mathbb{P}({\mathcal{E}_3}^c|\mathcal{E}_1\cap\mathcal{E}_2) - \mathbb{P}({\mathcal{E}_4}^c|\mathcal{E}_1\cap\mathcal{E}_2) - \mathbb{P}({\mathcal{E}_5}^c|\mathcal{E}_1\cap\mathcal{E}_2)\\ &\geq 1-8\delta/c_0\end{aligned}$$

Therefore, these events occur with probability at least $1-8\delta/c_0$.

Note that under the current settings of Theorem H.1, $A_\infty = H$ and $2(H+K)\gamma^K \leq \square/c_1$ due to Lemma C.7. Combined with Lemma H.3, Lemma H.14, and Lemma H.16, we have

$$\begin{aligned}\left|v_* - v_{\pi'_K}\right| &\leq \underbrace{\frac{1}{A_\infty}\sum_{i=0}^{K-1}\gamma^i \pi_K P_{K-i}^{K-1}\left|\phi^\top W(f, E_{K-i})\right|}_{\leq \square\left(c_1^{-1}+c_2^{-0.5}\right)\epsilon\mathbf{1}\ \text{ due to Lemma H.16}} + \underbrace{\frac{1}{A_\infty}\sum_{j=0}^{K-1}\gamma^j \pi_* P_*^j\left|\phi^\top W(f, E_{K-j})\right|}_{\leq \square\left(c_1^{-1}+c_2^{-0.5}\right)\epsilon\mathbf{1}\ \text{ due to Lemma H.14}} + \underbrace{2\left(H+K\right)\gamma^K\mathbf{1}}_{\leq \square c_1^{-1}\mathbf{1}}\\ &\leq \square\left(\frac{1}{c_1}+\frac{1}{\sqrt{c_2}}\right)\epsilon\mathbf{1}.\end{aligned}$$

Therefore, for some c_1 and c_2, the claim holds. $\square$

I. Formal Theorem and Proof of Theorem 5.2

Instead of the informal theorem Theorem 5.2, we are going to prove the following formal theorem.

Theorem I.1 (Accuracy of `VarianceEstimation`)**.** *Let c_0 be a positive constant such that $8 \geq c_0 \geq 6$ and $v \in \mathcal{F}_v$ be a random variable. Assume that an event*

$$\left\|v_* - v\right\|_\infty \leq \frac{1}{2}\sqrt{H}$$

occurs with probability at least $1-3\delta/c_0$. With a positive constant $c_5 \geq 1$, define

$$M^{var} := \left\lceil c_5 dH^2 \log\frac{2c_0^2 u_{\mathcal{C}} K}{(c_0-3)\delta}\right\rceil .$$

When `VarianceEstimation` is run with the settings $v_\sigma = v$ and $M_\sigma = M^{var}$, there exists $c_5 \geq 1$ independent of d, H, X, A, and δ such that the output ω satisfies $\sigma(v_*) \leq \sqrt{\max(\phi^\top\omega, 0)} + \sqrt{H} \leq \sigma(v_*) + 2\sqrt{H}$ with probability at least $1-4\delta/c_0$, using $\widetilde{\mathcal{O}}\left(2u_{\mathcal{C}}M_\sigma\right) = \widetilde{\mathcal{O}}\left(d^2H^2\right)$ samples from the generative model.

We let $\mathbf{F}_m$ be the σ-algebra generated by random variables $\{v\}\cup\{y_{n,x,a}|(n,x,a)\in[m-1]\times\mathcal{X}\times\mathcal{A}\}\cup\{z_{n,x,a}|(n,x,a)\in[m-1]\times\mathcal{X}\times\mathcal{A}\}$. Recall that $v\in\mathcal{F}_v$ is the random variable that is inputted to `VarianceEstimation` as $v_\sigma = v$ and $\omega\in\mathbb{R}^d$ is the parameter to approximate the variance as $\mathrm{Var}_\omega := \phi^\top\omega$.

We first introduce the necessary events.

Event 6. *$\mathcal{E}_6$ denotes the event $|v_* - v_\sigma| \leq \frac{1}{2}\sqrt{H}\mathbf{1}$.*

Event 7. *$\mathcal{E}_7$ denotes the event $\left|\sigma_* - \sqrt{\max(\phi^\top\omega, 0)}\right| \leq \sqrt{H}\mathbf{1}$.*

Due to the setting of Theorem I.1, $\mathbb{P}(\mathcal{E}_6) \geq 1-3\delta/c_0$. We need the following pivotal lemma to show Theorem I.1.

Lemma I.2. *When `VarianceEstimation` is run with the settings of Theorem I.1, there exists c_5 independent of d, H, X, A, and δ such that $\mathbb{P}({\mathcal{E}_7}^c|\mathcal{E}_6) \leq \delta/c_0$.*

Proof. For the input v_σ, we write $\mathrm{Var}(v_\sigma)$ as Var_v by abuse of notation. Let ω^* be the unknown underlying parameter that satisfies $\phi^\top\omega^* = \mathrm{Var}_v$. This is ensured to exist by Assumption 3.2.

The weighted KW bound (Lemma 4.3) indicates that

$$\left|\mathrm{Var}_\omega - \mathrm{Var}_v\right| = \left|\phi^\top\omega - \phi^\top\omega^*\right| = \left|\phi^\top W\left(\mathbf{1}, \widehat{\mathrm{Var}} - \mathrm{Var}_v\right)\right| \leq \sqrt{2d}\mathbf{1}\max_{(y',b')\in\mathcal{C}}\underbrace{\left|\widehat{\mathrm{Var}}(y',b') - \mathrm{Var}_v(y',b')\right|}_{\heartsuit}. \tag{39}$$

We are going to bound $\left|\widehat{\mathrm{Var}}(x,a) - \mathrm{Var}_v(x,a)\right|$ for $(x,a) \in \mathcal{C}$. Note that $\Big(v_\sigma\left(y_{m,x,a}\right) - v_\sigma(z_{m,x,a})\Big)^2/2$ is the unbiased estimator of Var_v since

$$\begin{aligned}\mathbb{E}\left[\Big(v_\sigma(y_{m,x,a}) - v_\sigma(z_{m,x,a})\Big)^2\right] &= \mathbb{E}\left[\Big(v_\sigma\left(y_{m,x,a}\right) - Pv_\sigma(x,a) + Pv_\sigma(x,a) - v_\sigma(z_{m,x,a})\Big)^2\right] \\ &= \mathbb{E}\left[\Big(v_\sigma\left(y_{m,x,a}\right) - Pv_\sigma(x,a)\Big)^2\right] + \mathbb{E}\left[\Big(v_\sigma(z_{m,x,a}) - Pv_\sigma(x,a)\Big)^2\right] \\ &\quad - 2\mathbb{E}\left[\Big(v_\sigma\left(y_{m,x,a}\right) - Pv_\sigma(x,a)\Big)\Big(v_\sigma(z_{m,x,a}) - Pv_\sigma(x,a)\Big)\right] \\ &= 2\mathrm{Var}_v(x,a)\,.\end{aligned}$$

Moreover, $\mathcal{E}_6$ implies $|v_\sigma| \leq |v_\sigma - v_*| + |v_*| \leq \frac{3}{2}H$. Also, due to Lemma D.5, $\mathrm{Var}_v \leq \frac{9}{4}H^2 \leq 3H^2$. For a fixed $(x,a) \in \mathcal{C}$,

$$\widehat{\mathrm{Var}}(x,a) - \mathrm{Var}_v(x,a) = \frac{1}{M_\sigma}\sum_{m=1}^{M_\sigma} \underbrace{\Big(\left(v_\sigma\left(y_{m,x,a}\right) - v_\sigma(z_{m,x,a})\right)^2/2 - \mathrm{Var}_v(x,a)\Big)}_{\text{bounded by } 8H^2}$$

is a sum of bounded martingale differences with respect to $(\mathbf{F}_m)_{m=1}^{M_\sigma}$. Therefore, using the conditional Azuma-Hoeffding inequality (Lemma D.2) and taking the union bound over $(x,a) \in \mathcal{C}$, we have

$$\mathbb{P}\left(\exists (x,a,k) \in \mathcal{C} \text{ s.t. } \left|\widehat{\mathrm{Var}}(x,a) - \mathrm{Var}_v(x,a)\right| \geq H^2\sqrt{\frac{128\iota_{2,3}}{M_\sigma}} \middle| \mathcal{E}_6\right) \leq \frac{\delta}{c_0}\,,$$

where $\iota_{2,3} = \iota_1 + \log(c_0/(c_0-3))$ is due to the condition by $\mathcal{E}_6$ with $\mathbb{P}(\mathcal{E}_6) \geq 1 - 3\delta/c_0$. We used $\iota_{2,3}$ since $1/(1-3\delta/c_0) \leq c_0/(c_0-3)$. Inserting the result into (39), with probability $1 - \delta/c_0$,

$$|\mathrm{Var}_\omega - \mathrm{Var}_v| \leq 16H^2\sqrt{\frac{d\iota_{2,3}}{M_\sigma}}\mathbf{1}\,.$$

Due to the setting of $M_\sigma = \lceil c_5 dH^2\iota_{2,3}\rceil$, some c_5 exists such that $|\phi^\top\omega - \mathrm{Var}_v| \leq \frac{1}{4}H\mathbf{1}$. This implies that $\left|\max(\phi^\top\omega, 0) - \mathrm{Var}_v\right| \leq \frac{1}{4}H\mathbf{1}$ since $\mathrm{Var}_v \geq 0$, and furthermore, $\left|\sqrt{\max(\mathrm{Var}_\theta, 0)} - \sqrt{\mathrm{Var}_v}\right| \leq \frac{1}{2}\sqrt{H}\mathbf{1}$ due to Lemma C.3. Finally,

$$\begin{aligned}\left|\sqrt{\max(\phi^\top\omega, 0)} - \sigma(v_*)\right| &\overset{(a)}{\leq} \left|\sqrt{\max(\phi^\top\omega, 0)} - \sqrt{\mathrm{Var}_v}\right| + \left|\sigma(v_*) - \sqrt{\mathrm{Var}_v}\right| \\ &\overset{(b)}{\leq} \left|\sqrt{\max(\phi^\top\omega, 0)} - \sqrt{\mathrm{Var}_v}\right| + \underbrace{|v_* - v_\sigma|}_{\leq \frac{1}{2}\sqrt{H}\mathbf{1} \text{ due to } \mathcal{E}_6} \leq \sqrt{H}\mathbf{1}\,,\end{aligned}$$

where (a) is due to Lemma E.1, (b) is due to Lemma D.5. This concludes the proof.

□

We are now ready to prove Theorem I.1.

Proof of Theorem I.1. The claim holds by showing that the event $\mathcal{E}_6 \cap \mathcal{E}_7$ occurs with high probability. Note that when `VarianceEstimation` is run with the settings defined in Theorem I.1, $\mathbb{P}({\mathcal{E}_6}^c) \leq 3\delta/c_0$ and $\mathbb{P}({\mathcal{E}_7}^c|\mathcal{E}_6) \leq \delta/c_0$. According to Lemma C.1, we have

$$\mathbb{P}(\mathcal{E}_6 \cap \mathcal{E}_7) \geq \mathbb{P}(\mathcal{E}_6) - \mathbb{P}({\mathcal{E}_7}^c|\mathcal{E}_6) \geq 1 - 4\delta/c_0\,.$$

Therefore, these events occur with probability at least $1 - 4\delta/c_0$ and thus the claim holds. □

J. Formal Theorem and Proof of Theorem 5.3

Instead of the informal Theorem J.1, we prove the following formal theorem. In the theorem, we denote K^{ls} and M^{ls} as the values defined in Theorem H.2, K^{wls} and M^{wls} as the values defined in Theorem H.1, and M^{var} as the value defined in Theorem I.1.

Theorem J.1 (Sample complexity of `VWLS-MDVI`)**.** *Assume that $\varepsilon \in (0, 1/H]$ and $c_0 = 8$. There exist positive constants $c_1, c_2, c_3, c_4, c_5 \geq 1$ independent of d, H, X, A, ε and δ such that when `VWLS-MDVI` is run with the settings $\alpha = \gamma$, $K = K^{ls}$, $M = M^{ls}$, $\widetilde{K} = K^{wls}$, $\widetilde{M} = M^{wls}$, and $M_\sigma = M^{var}$, the output sequence of policies π' satisfy $\|v_* - v_{\pi'}\|_\infty \leq \varepsilon$ with probability at least $1 - \delta$, using*

$$\widetilde{\mathcal{O}}\left(u_{\mathcal{C}} KM + u_{\mathcal{C}} \widetilde{K}\widetilde{M} + u_{\mathcal{C}}(M_\sigma + 1)\right) = \widetilde{\mathcal{O}}\left(d^2 H^3 / \varepsilon^2\right)$$

samples from the generative model.

Proof. The claim is easily seen from Theorem H.2, Theorem I.1, and Theorem H.1.

From Theorem H.2, the first `WLS-MDVI` in `VWLS-MDVI` outputs v_K such that $\|v_* - v_K\|_\infty \leq 1/2\sqrt{H}$ with probability $1 - 3\delta/c_0$. This v_K satisfies the requirement of Theorem I.1.

According to Theorem I.1, `VarianceEstimation` in `VWLS-MDVI` outputs ω such that $\sigma(v_*) \leq \sqrt{\max(\phi^T\omega, \mathbf{0})} + \sqrt{H}\mathbf{1} \leq \sigma(v_*) + 2\sqrt{H}\mathbf{1}$ with probability $1 - 4\delta/c_0$. Therefore, $\widetilde{\sigma} = \min\left(\sqrt{\max(\phi^T\omega, \mathbf{0})} + \sqrt{H}\mathbf{1}, H\mathbf{1}\right)$ defined in the algorithm can be used as the weighting function of Theorem H.1.

Finally, Theorem H.1 indicates that the second `WLS-MDVI` in `VWLS-MDVI` outputs the ε-optimal policy with probability $1 - 8\delta/c_0$. When $c_0 = 8$, `VWLS-MDVI` outputs the ε-optimal policy with probability at least $1 - \delta$. ☐

K. Pseudocode of Missing Algorithms

Algorithm 5 Tabular MDVI (α, K, M)

Input: $\alpha \in [0, 1)$, K, and M.
Initialize $s_0 = \mathbf{0} \in \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$ and $w_0 = w_{-1} = \mathbf{0} \in \mathbb{R}^{\mathcal{X}}$.
for $k = 0$ **to** $K - 1$ **do**
  $v_k = w_k - \alpha w_{k-1}$.
  for each state-action pair $(x, a) \in \mathcal{X} \times \mathcal{A}$ **do**
    $q_{k+1}(x, a) = r(x, a) + \gamma \widehat{P}_k(M) v_k(x, a)$.
    $s_{k+1} = q_{k+1} + \alpha s_k$ and $w_{k+1}(x) = \max_{a \in \mathcal{A}} s_{k+1}(x, a)$ for each $x \in \mathcal{X}$.
  end for
end for
Return: $(\pi_k)_{k=0}^K$, where π_k is greedy policy with respect to s_k.

Algorithm 6 InitializeDesign

Choose an arbitrary nonzero $c_0 \in \mathbb{R}^d$
for $j = 0$ **to** $d - 1$ **do**
  $(\overline{x}_j, \overline{a}_j) = \arg\max_{(x,a)\in\mathcal{X}\times\mathcal{A}} c_j^\top \phi(x, a)$.
  $(\underline{x}_j, \underline{a}_j) = \arg\min_{(x,a)\in\mathcal{X}\times\mathcal{A}} c_j^\top \phi(x, a)$.
  $y_j = \phi(\overline{x}_j, \overline{a}_j) - \phi(\underline{x}_j, \underline{a}_j)$.
  Choose an arbitrary nonzero c_{j+1} orthogonal to $y_0, \ldots, y_j$.
end for
Let $Z := \{(\overline{x}_j, \overline{a}_j), (\underline{x}_j, \underline{a}_j) \mid j = 0, \ldots, d-1\}$.
Choose ρ to put equal weight on each of the distinct points of Z.
Return: ρ.

Algorithm 7 Frank-Wolfe for finite $\mathcal{X}$ $(f, \varepsilon^{\mathrm{FW}})$

 # We write X be the size of $\mathcal{X}$. Without loss of generality, we assume $\mathcal{X} = [X]$ and $\mathcal{A} = [A]$.
Input: $f : \mathcal{X} \times \mathcal{A} \to (0, \infty)$, $\varepsilon^{\mathrm{FW}} \in \mathbb{R}$.
$\rho = \texttt{InitializeDesign}()$ by Algorithm 6.
Let $U : \rho \mapsto \mathrm{diag}(\rho) \in \mathbb{R}^{XA \times XA}$ where diag constructs a diagonal matrix with elements of ρ.
For $(x, a) \in \mathcal{X} \times \mathcal{A}$, let $\Phi_f \in \mathbb{R}^{XA \times d}$ be a matrix where its $(xA + a)$ th row is $\phi(x, a)/f(x, a)$.
Let $H : \rho \mapsto (\Phi_f U(\rho) \Phi_f)^{-1}$.
Let $\omega : (x, a, \rho) \mapsto \phi(x, a)^\top H(\rho) \phi(x, a)$
Let $\delta : \rho \mapsto \max_{(x,a) \in \mathcal{X} \times \mathcal{A}} (\omega(x, a, \rho) - d)/d$
while $\delta(\rho) > \varepsilon^{\mathrm{FW}}$ **do**
  Let $(y, b) := \arg\max_{(x,a) \in \mathcal{X} \times \mathcal{A}} \omega(x, a)$
  Let $\lambda^* := (\omega(y, b) - d) \,/\, ((d - 1)\omega(y, b))$
  $\rho(y, b) \leftarrow \rho(y, b) + \lambda^*$
  $\rho \leftarrow \rho/(1 + \lambda^*)$
end while
$C = \left\{ (x, a) \mid \; \omega(x, a, \rho) \geq d \left(1 + \frac{\delta(\rho) d}{2} - \sqrt{\delta(\rho)(d - 1) + \frac{\delta(\rho)^2 d^2}{4}} \right) \right\}$
$G = \sum_{(y,b) \in \mathcal{C}} \frac{\rho(y,b)}{f^2(y,b)} \phi(y, b) \phi(y, b)^\top$
Return: $\rho, \mathcal{C}, G$.

Algorithm 8 DVW for online (Munchausen-)DQN

Input: $K \in \mathbb{N}$ the number of update iteration, $F \in \mathbb{N}$ the target update interval, $T \in \mathbb{N}$ the number of environment steps in one iteration.
Initialize θ and ω at random. $\overline{\theta} = \widehat{\theta} = \theta$ and $\overline{\omega} = \omega$.
Initialize $\eta = 1$.
Initialize $\mathcal{B} = \{\}$.
for $k = 0$ **to** $K - 1$ **do**
  for $t = 0$ **to** $T - 1$ **do**
    Collect a transition $b = (x, a, r, x')$ from the environment.
    $\mathcal{B} \leftarrow \mathcal{B} \cup \{b\}$.
  end for
  Sample a random batch of transition $B_k \in \mathcal{B}$.
  On B_k, update ω with one step of SGD on $\mathcal{L}(\omega)$, see Equation (9).
  On B_k, update η with one step of SGD on $\mathcal{L}(\eta)$, see Equation (11).
  On B_k and with f of Equation (10), update θ with one step of SGD on $\mathcal{L}(\theta)$, see Equation (8).
  if $k \mod F = 0$ **then**
    $\widehat{\theta} \leftarrow \overline{\theta}$.
    $\overline{\theta} \leftarrow \theta$.
    $\overline{\omega} \leftarrow \omega$.
  end if
end for
Return: A greedy policy with respect to q_θ

L. Experiment Details

The source code for all the experiments is available at `https://github.com/matsuolab/Variance-Weighted-MDVI`.

L.1. Details of Section 7.1

Hard linear MDP. The hard linear MDP we used is based on the **Theorem H.3** in Weisz et al. (2022). Specifically, the MDP has two states: $\mathcal{X} = \{x_0, x_1\}$ with x_0 being the initial state and x_1 being the absorbing state. To add randomness to the MDP, the action space is constructed as $\mathcal{A} = \{a_0, a_1, \ldots, a_A\}$ where $a_i \in \mathbb{R}^{d-2}$ for $i = [A]$ is randomly sampled from a multivariate uniform distribution of $\mathcal{U}(\mathbf{0}, \mathbf{1})$ with $d-2$ dimension. Same as Weisz et al. (2022), for all a, the feature map is defined as

$$\phi(x_0, a) = \left(1, 0, a^\top\right)^\top \quad \text{and} \quad \phi(x_1, a) = (0, 1, 0, \ldots, 0)^\top .$$

Using $\psi = (1, 0, \ldots, 0)^\top$, we make the state x_0 be the rewarding state as

$$r(x_0, a) = \phi(x_0, a)^\top \psi = 1 \quad \text{and} \quad r(x_1, a) = \phi(x_0, a)^\top \psi = 0 .$$

Let $\mu(x_0) = \left(\gamma, 0, 0.01 a_0^\top\right)^\top$ and $\mu(x_1) = \left(1-\gamma, 1, -0.01 a_0^\top\right)^\top$ be the design parameters for the transition probability kernel. This implies that

$$\begin{aligned} P(x_0 \mid x_0, a) &= \gamma + 0.01 \cdot a_0^\top a, \qquad & P(x_1 \mid x_0, a) &= 1 - \gamma - 0.01 \cdot a_0^\top a, \\ P(x_0 \mid x_1, a) &= 0, & P(x_1 \mid x_1, a) &= 1 . \end{aligned}$$

Intuitively, choosing an action similar to a_0 increases the probability of transitioning to x_0 and yields a higher return. We provide the hyperparameters of the MDP in Table 3.

Table 3. Hyperparameters of hard linear MDP in Section 7.1

Parameter	Value
MDP parameter	
action space size	$A = 30$
dimension of the feature map	$d = 4$
discount factor	$\gamma = 0.9$
Algorithm parameter	
weight for MDVI update	$\alpha = 0.9$
accuracy of the Frank-Wolfe algorithm	$\varepsilon^{\mathrm{FW}} = 0.01$

Algorithm implementations. The algorithms `WLS-MDVI` and `VWLS-MDVI` are implemented according to Algorithm 1 and Algorithm 3. The optimal designs are computed using the Frank-Wolfe (FW) algorithm (Algorithm 7). We provide the hyperparameters of the algorithms in Table 3.

L.2. Details of Section 7.2.1

Gridworld environment. The gridworld environment we used is a 25×25 grid with randomly placed 8 pitfalls. This is similar to the gridworld environment of Fu et al. (2019), but there are some differences.

The agent starts from the top left grid and can move to any of its neighboring grids with success probability 0.6, or to a different random direction with probability 0.4. The agent receives $+1$ reward when it reaches the goal grid located at the bottom right grid. Other rewards are set to 0. When the agent enters a pitfall, the agent can no longer move and receives 0 reward until the environment terminates. We use $\gamma = 0.995$ and the environment terminates after 200 steps.

Algorithm implementations. We implement the environment and algorithms using ShinRL (Kitamura & Yonetani, 2021). For the implementation of M-DQN, same as Vieillard et al. (2020b), we clip the value of log-policy term by $\max(\log \pi_{\overline{\theta}}, l_0)$ with $l_0 > 0$ to avoid numerical issues in Equation (8). We provide the hyperparameters used in the experiment in Table 4. In the table, we denote FC(n) be a fully convolutional layer with n neurons.

Table 4. Hyperparameters of algorithms in Section 7.2.1

Parameter	Value
Shared	
optimizer	Adam
iteration (K)	2000000
target update interval (F)	100
learning rate	10^{-3}
discount factor (γ)	0.995
horizon (H)	200
q-network structure	FC(128) $-$ FC(128) $-$ FC$(\|\mathcal{A}\|)$
activations	Relu
Munchausen-DQN parameters	
entropy regularization coefficient (κ)	10^{-5}
KL regularization coefficient (τ)	$\kappa\gamma/(1-\gamma)$
clipping value (l_0)	-1
DVW parameters	
ω and η-optimizer	Adam
activations	Relu
Var-network structure	FC(128) $-$ FC(128) $-$ FC$(\|\mathcal{A}\|)$
lower threshold parameter ($\underline{c}_f$)	0.1
upper threshold parameter ($\overline{c}_f$)	0.1
learning rate of Var$_\omega$	10^{-3}
learning rate of η	5.0×10^{-3}

L.3. Details of Section 7.2.2

Algorithm implementation. We implement algorithms as variations of DQN from CleanRL (Huang et al., 2022). For a fair comparison, all the algorithms use the same epsilon-greedy exploration strategy. It randomly chooses an action with probability e_t otherwise chooses a greedy action w.r.t. q_θ. For the implementation of M-DQN, same as Vieillard et al. (2020b), we clip the value of log-policy term by $\max(\log \pi_{\overline{\theta}}, l_0)$ with $l_0 > 0$ to avoid numerical issues in Equation (8).

We provide the hyperparameters used in the experiment in Table 5. In the table, we denote FC(n) be a fully convolutional layer with n neurons, and Conv$_{a,b}^d(c)$ be a 2D convolutional layer with c filters of size $a \times b$ and of stride d.

Table 5. Hyperparameters of algorithms in Section 7.2.2

Parameter	Value
Shared	
e_k (random actions rate)	start from 1.0 and linearly decay to 0.1 until the period of 10^6 steps
θ-optimizer	Adam
iteration (K)	10^7
target update interval (F)	1000
learning rate of q_θ	2.5×10^{-4}
replay buffer size ($\lvert\mathcal{B}\rvert$)	10^5
batch size ($\lvert\mathcal{B}_k\rvert$)	32
train frequency (T)	4
discount factor (γ)	0.99
q-network structure	Conv$_{3,3}^1(16)$ $-$ FC(128) $-$ FC$(\lvert\mathcal{A}\rvert)$
activations	Relu
Munchausen-DQN parameters	
entropy regularization coefficient (κ)	0.003
KL regularization coefficient (τ)	0.027
clipping value (l_0)	-1
DVW parameters	
ω and η-optimizer	Adam
Var-network structure	Conv$_{3,3}^1(16)$ $-$ FC(128) $-$ FC$(\lvert\mathcal{A}\rvert)$
lower threshold parameter ($\underline{c}_f$)	0.1
upper threshold parameter ($\overline{c}_f$)	0.1
learning rate of Var$_\omega$	2.5×10^{-4}
learning rate of η	10^{-3}