A Unifying Framework of Off-Policy General Value Function Evaluation

Tengyu Xu
Meta Platforms, Inc
Menlo Park, CA 94025

Zhuoran Yang
Yale University
New Haven, CT 06520

Zhaoran Wang
Northwestern University
Evanston, IL 60208

Yingbin Liang
Ohio State University
Columbus, OH 43210

Abstract

General Value Function (GVF) is a powerful tool to represent both the *predictive* and *retrospective* knowledge in reinforcement learning (RL). In practice, often multiple interrelated GVFs need to be evaluated jointly with pre-collected off-policy samples. In the literature, the gradient temporal difference (GTD) learning method has been adopted to evaluate GVFs in the off-policy setting, but such an approach may suffer from a large estimation error even if the function approximation class is sufficiently expressive. Moreover, none of the previous work have formally established the convergence guarantee to the ground truth GVFs under the function approximation settings. In this paper, we address both issues through the lens of a class of GVFs with causal filtering, which cover a wide range of RL applications such as reward variance, value gradient, cost in anomaly detection, stationary distribution gradient, etc. We propose a new algorithm called GenTD for off-policy GVFs evaluation and show that GenTD learns multiple interrelated multi-dimensional GVFs as efficiently as a single canonical scalar value function. We further show that unlike GTD, the learned GVFs by GenTD are guaranteed to converge to the ground truth GVFs as long as the function approximation power is sufficiently large. To our best knowledge, GenTD is the first off-policy GVF evaluation algorithm that has global optimality guarantee.

1 Introduction

The value function, which represents the expected accumulation of reward [43], serves as a reliable performance metric of policy in the reinforcement learning (RL) tasks [42, 23]. In many RL applications, however, looking at only the value function is not enough. For example, in the risk-sensitive domains such as health care and financial assets, the variance of "reward-to-go" rather than the value function, i.e., the mean of "reward-to-go", is a more suitable performance metric. As another example, to obtain a variance-reduced or bias-reduced policy gradient estimator [14, 61, 18], in addition to the value function, the information of "gradient of value function" is also required. Moreover, in continuous control domain with differentiable and deterministic policy, the computation of policy gradient is only possible through "action/state-value gradient" [38, 8, 12], etc. All the aforementioned metrics can be viewed as *predicative* knowledge of certain multiple intercorrelated cumulative "signals" (possibly high-dimensional, e.g., the gradient of value function), and thus naturally fall into the framework of **forward GVFs** (refers to forward general value functions) [48, 57, 32, 33] (see Section 2.1 for the formal definition). One typical approach to evaluate GVFs, is to learn from samples that pre-collected from one or more behavior policies, which yields an

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

off-policy method. In practice, multiple forward GVFs are usually evaluated jointly at the same time due to their interrelationships [37, 33].

In contrast to forward GVFs defined based on predictive knowledge, the **backward GVFs** represents *retrospective* knowledge, which captures the accumulation of signals from the past to the present time [70] (see Section 2.2 for the formal definition). Although the concept of the backward GVFs has not been formally proposed until very recently [70], it is rooted in a number of important RL applications such as anomaly detection [70], emphatic weight learning [46, 70] and evaluation of gradient of logarithmic stationary distribution [26, 61, 18]. Different from the forward GVFs, for which the Bellman operator can be defined independently from the sampling distribution [37, 42, 44], the Bellman operator of the backward GVFs is only valid if the sampling exactly follows the on-policy stationary distribution [70]. Due to such a reason, *off-policy* evaluation of the backward GVFs is much more challenging than that of the forward GVFs.

In general, due to the high dimensionality and intercorrelation, it is very challenging to evaluate multiple GVFs simultaneously with standard policy evaluation approaches [55, 45, 23]. In previous studies, the gradient temporal difference (GTD) learning [45, 23], one of the most popular off-policy methods in value function evaluation, has been adopted to solve both the forward and backward GVFs evaluation problems [48, 37, 69]. GTD adopts the mean squared projected Bellman error (MSPBE) as its optimization objective and takes the expectation over the *behavior* policy, which does not exactly reflect the desirable evaluation under the *target* policy. As a result, GTD can encounter serious issues in GVFs evaluation problems. **First**, the optimal point to which GTD converges can be far away from the ground truth value of GVFs. It becomes worse when multiple GVFs are evaluated simultaneously, because the error of one GVFs evaluation can be further amplified across other GVFs' evaluation due to their inherent correlations. In the literature, no provable bound has been established on such an error, which can, in fact, be unbounded for some cases (see Example 1 in [19]). **Second**, for high-dimensional GVFs evaluations, the landscape geometry of the GTD objective function can be ill-conditioned [23], which could slow down the convergence of GTD significantly. As demonstrated by our empirical results in Section 5, GTD can suffer from both the large estimation error and the slow convergence rate, which further suggests that GTD may not be a good choice for GVFs evaluation tasks. This motivates our paper to address the following question:

- *Can we design a new off-policy approach for multiple interrelated and high-dimensional GVFs evaluation problems, which is guaranteed to converge fast and converge to the ground truth GVFs?*

Our Contributions. In this paper, we investigate the problem of evaluating multiple interrelated GVFs jointly. Rather than studying different GVFs on a case-by-case basis, we explore the class of "GVFs with causal filtering", which captures a common structural feature shared by GVFs in a wide range of RL applications (see Appendix C). **(a)** We prove that both forward and backward GVFs with causal filtering are the unique fixed point of their corresponding general Bellman operator (GBO) (defined for multiple high-dimensional GVFs), which is shown to have a contraction property with respect to a properly constructed norm metric. **(b)** Based on such a property of GVFs, we propose a new algorithm GenTD to solve off-policy GVFs evaluation problem. GenTD introduces a density ratio to adjust the behavior distribution and further incorporates a policy-agnostic approach GenDICE/GradientDICE [65, 70] for estimating the density ratio jointly with GVF evaluation. **(c)** In the linear function approximation setting, we show that GenTD converges to the globally optimal point at the rate of $\mathcal{O}(1/T)$, with conditional number independent from the dimension of GVFs. Such a result implies that GenTD learns multiple interrelated possibly high-dimensional GVFs as efficiently as TD learning for a single canonical scalar value function. **(d)** We further show that unlike GTD, GenTD is guaranteed to approximate the ground truth GVFs well as long as the function expressive power is sufficiently large. To our best knowledge, GenTD is the first off-policy GVF evaluation algorithm that has such a ground truth guarantee. **(e)** Our experiments further demonstrate that GenTD converges much faster than GTD, and more importantly, converges to ground truth closely, whereas GTD suffers from large approximation error.

Related Work. The forward GVF was first introduced in [48] to represent a set of accumulation of general signals with possibly time-varying discount factors. The forward GVF was later used to represent a set of interrelated predictions [37, 9, 41, 24, 33]. It has been observed that some RL metrics such as variance, gradient of value function, state/action value gradient can also be viewed as forward GVFs [51, 14, 61, 18, 38, 48, 57, 32, 8, 5]. In previous works, both TD learning and GTD have been used to evaluate forward GVFs in the on- and off-policy settings [48, 37, 33], respectively.

A more comprehensive review of studies of forward GVFs has been provided in [35]. The backward GVF was formally defined in [70]. Some previous works have also considered metrics that can be represented as accumulations of signals in the reverse time direction, such as emphatic weighting, page ranking cost, and derivative of logarithmic stationary distribution [66, 68, 26, 63, 10]. Another track of research has focused on evaluation of a general scalar function in the off-policy setting, a.k.a off-policy evaluation (OPE) [4, 64, 16]. However, since the focus of this paper is on the evaluation of multiple high-dimensional GVFs, results in OPE are not directly comparable with ours.

The theoretical studies of off-policy GVFs evaluation algorithms are rather limited. So far, only the asymptotic convergence guarantee (without the convergence rate characterization) of GTD has been established in both the forward and backward GVFs evaluation settings [37, 70]. The convergence rate of GTD has only been established in [62, 7, 17, 6, 58, 21] for the simple canonical value function evaluation setting, which is a special case of forward GVFs. However, as pointed out in [19, 10, 27], the optimal point of GTD may suffer from possibly unbounded approximation error, which is not desirable in practice. In contrast, we propose a new off-policy GVFs evaluation algorithm, which can solve a wide range of forward and backward GVFs evaluation problems, with convergence rate characterization and guaranteed optimality with respect to the ground truth GVFs value.

We note that the contraction property of GBO for forward GVFs has also been investigated in [33] with a structure called "acyclic graph", which is similar to "causal filtering" in our paper (see the footnote comment for Proposition 1). However, the results of backward GVFs are established in our work for the first time. The focus of this paper is on the finite time performance and optimality guarantee for a new off-policy GVFs evaluation algorithm GenTD, which was not studied in [33].

2 Markov Decision Process and General Value Function

We consider an infinite-horizon Markov Decision Process (MDP) with a state space $\mathcal{S}$, an action space $\mathcal{A}$, a reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, a transition kernel $\mathsf{P} : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$, a discounted factor $\gamma \in (0, 1)$, and an initial distribution $\mu_0 : \mathcal{S} \to [0, 1]$. An policy $\pi(a|s)$ is the probability of taking action a at state s. At time step t, an agent at a state s_t selects an action a_t according to $\pi(\cdot|s_t)$, receives a reward $r(s_t, a_t)$, and transits to state s_{t+1} according to $\mathsf{P}(\cdot|s_t, a_t)$. The state-action transition kernel is defined as $\mathsf{P}_\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$, in which $\mathsf{P}_\pi((s, a), (s', a')) = \mathsf{P}(s'|s, a)\pi(a'|s')$. When the MDP is ergodic, we define μ_π as the state-action stationary distribution which satisfies: $\mu_\pi^\top \mathsf{P}_\pi = \mu_\pi^\top$. For such an MDP, we define the discounted accumulation of reward as the "reward-to-go": $J_\pi = \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)$. The state-action value function (i.e., Q-function) is defined as $Q_\pi(s, a) = \mathbb{E}[J_\pi|(s_0, a_0) = (s, a)]$, and the state value function (i.e., V-function) is defined as $V_\pi(s) = \mathbb{E}[Q_\pi(s, a)|s]$. Note that $Q_\pi(s, a)$ satisfies the following Bellman equation

$$Q_\pi = \mathcal{T}_\pi Q_\pi = R + \gamma \mathsf{P}_\pi Q_\pi, \tag{1}$$

where $\mathcal{T}_\pi$ is the Bellman operator, and Q_π and $R \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ are vectors obtained via stacking $Q_\pi(s, a)$ and $r(s, a)$ over state-action space $\mathcal{S} \times \mathcal{A}$. We introduce a function of (s, a) (possibly in the vector form) as $v(s, a) \in \mathbb{R}^d$ ($d \geq 1$). Consider a distribution $\xi(\cdot)$ over $\mathcal{S} \times \mathcal{A}$. We define the ξ–norm of $v \in \mathbb{R}^{d|\mathcal{S}||\mathcal{A}|}$ as $\|v\|_\xi = \sqrt{\sum_{(s,a)} \xi(s, a) \|v(s, a)\|_2^2}$, where v is obtained by stacking the function $v(s, a)$ over $\mathcal{S} \times \mathcal{A}$. It has been proved that $\mathcal{T}_\pi$ is γ–contraction in μ_π–norm, i.e., $\|\mathcal{T}_\pi v - \mathcal{T}_\pi v'\|_{\mu_\pi} \leq \gamma \|v - v'\|_{\mu_\pi}$ and Q_π is the unique fixed point of $\mathcal{T}_\pi$ [44, 42, 55]. In the sequel, we denote I_d as the identity matrix with the dimension d and $\otimes$ as the Kronecker product. We further define $U_\pi = \text{diag}(U_{\pi,1}, \cdots, U_{\pi,k})$, in which $U_{\pi,i} = \text{diag}(\mu_\pi) \otimes \mathsf{I}_{d_i}$ for $i = \{1, \cdots, k\}$, and $P_\pi = \text{diag}(P_{\pi,1}, \cdots, P_{\pi,k})$, in which $P_{\pi,i} = \mathsf{P}_\pi \otimes \mathsf{I}_{d_i}$.

2.1 Forward General Value Function

Consider a set of the state-action general value functions (GVFs) $G_\pi = [G_{\pi,1}^\top, \cdots, G_{\pi,k}^\top]^\top$, where each GVF $G_{\pi,i}$ is defined as the accumulation of a corresponding signal $C_i(s, a) \in \mathbb{R}^{d_i}$ given by $G_{\pi,i}(s, a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma_i^t C_i(s_t, a_t)\middle|(s_0, a_0) = (s, a), \pi\right]$, where $\gamma_i \in (0, 1)$ is a discount factor associated with C_i. Since $C_i(s, a) \in \mathbb{R}^{d_i}$ can be high-dimensional, $G_{\pi,i}(s, a)$ can also be high-dimensional for each (s, a). Clearly, the Q-function is a special GVF associated with a scalar signal. Since G_π is defined as the accumulation of the signal C_i in a forward direction from the current time step t to the future ∞, we call G_π as "forward GVF".

In many RL applications, we are often interested in the case that G_π of GVFs have progressive dependence [48], i.e., each $C_i(s,a)$ (associated with $G_{\pi,i}$) depends on the lower-indexed value functions $G_{\pi,1},\cdots,G_{\pi,i-1}$ in the set. As a concrete example, suppose the policy is parametrized by a smooth function π_w, where the parameter $w\in\mathbb{R}^{d_w}$. In addition to the Q-function Q_π, the gradient $\nabla_w Q_\pi(s,a)$ of the Q-function w.r.t w arises as a GVF of interest in several applications. In such a case, $G_{\pi,1}=Q_\pi$ and $G_{\pi,2}=\nabla_w Q_\pi$. It has been shown in [5] that the reward $C_2(s,a)$ associated with $\nabla_w Q_\pi$ is given by $C_2(s,a)=\gamma\mathbb{E}[Q_\pi(s',a')\nabla_w\log(\pi_w(s',a'))|s,a]$, which depends on the lower-indexed $G_{\pi,1}=Q_\pi$. Hence, such defined GVFs vector has progressive dependence. Appendix C provides further details and more examples in RL. More formally, we refer to this structure of forward GVF with progressive dependence as casual filtering as defined below. Note that a similar structure was called acyclic graph in [33].

Definition 1 (Forward GVF with causal filtering)**.** *For a given policy π, a forward GVF $G_\pi=[G_{\pi,1}^\top,\cdots,G_{\pi,k}^\top]^\top$ with causal filtering are associated with signals satisfying*

$$C_i=B_i+\textstyle\sum_{j=1}^{i-1}A_{i,j}G_{\pi,j}\quad \textit{for}\ \ 2\le i\le k,$$

where C_i and $G_{\pi,j}$ are obtained by respectively stacking $C_i(s,a)\in\mathbb{R}^{d_i}$ and $G_{\pi,j}(s,a)\in\mathbb{R}^{d_i}$ over $\mathcal{S}\times\mathcal{A}$, $B_i\in\mathbb{R}^{d_i|\mathcal{S}||\mathcal{A}|}$ is an observable signal, and the coefficient matrix $A_{i,j}\in\mathbb{R}^{d_i|\mathcal{S}||\mathcal{A}|\times d_j|\mathcal{S}||\mathcal{A}|}$ captures how the j-th GVF $G_{\pi,j}$ affects the i-th accumulation signal C_i. Further, B_i and $A_{i,j}$ are bounded to ensure $G_{\pi,i}$ to be well defined.

Definition 1 indicates that all GVFs are interrelated with a causal filtering structure, i.e., each signal C_i is a linear function of all lower-indexed $G_{\pi,l}$ for $1\le l<i$. Definition 1 also implies that the forward GVF $G_\pi=[G_{\pi,1}^\top,\cdots,G_{\pi,k}^\top]^\top$ with causal filtering satisfies the following **lower-triangular Bellman equation** given by

$$G_\pi=\mathcal{T}_{G,\pi}G_\pi=B+M_\pi G_\pi,\tag{2}$$

where $\mathcal{T}_{G,\pi}$ denotes the forward general Bellman operator (GBO), $B=\left[B_1^\top,\cdots,B_k^\top\right]^\top$ and

$$M_\pi=\begin{bmatrix}\gamma_1\bar{P}_{\pi,1} & 0 & \cdots & 0\\ A_{2,1} & \gamma_2\bar{P}_{\pi,2} & \cdots & 0\\ \vdots & \vdots & & \vdots\\ A_{k,1} & A_{k,2} & \cdots & \gamma_k\bar{P}_{\pi,k}\end{bmatrix}.$$

where $\bar{P}_{\pi,i}=[\mathsf{P}_\pi\otimes\mathsf{I}_{d_i}]$. Clearly, the canonical value function Q_π and Bellman operator $\mathcal{T}_\pi$ defined in eq. (1) is a special case of G_π and $\mathcal{T}_{G,\pi}$ defined in eq. (2).

2.2 Backward General Value Function

In contrast to the forward GVF defined in the last section, which represents the *predictive knowledge*, in some RL scenarios, we also want to capture the *retrospective knowledge*, which represents the accumulation of signals that have been collected from the past. Consider a set of GVFs $\hat{G}_\pi=[\hat{G}_{\pi,1}^\top,\cdots,\hat{G}_{\pi,k}^\top]^\top$, where each GVF $\hat{G}_{\pi,i}$ is defined as the *backward* accumulation of a vector signal $\hat{C}_i(s,a)\in\mathbb{R}^{d_i}$ given by $\hat{G}_{\pi,i}(s,a)=\mathbb{E}\big[\sum_{t=-\infty}^{0}\gamma_i^{-t}\hat{C}_i(s_t,a_t)\big|(s_0,a_0)=(s,a),\pi\big]$. To distinguish from the forward GVF $G_{\pi,i}$ defined in Section 2.1, we denote $\hat{G}_{\pi,i}$ as the backward GVF. For general purpose, we also consider the causal filtering setting for $\hat{G}_\pi$, in which each $\hat{C}_i(s,a)$ depends on the lower-indexed value functions $\hat{G}_{\pi,1},\cdots,\hat{G}_{\pi,i-1}$ in the set. We define the backward GVF with causal filtering as follows.

Definition 2 (Backward GVF with causal filtering)**.** *For a given policy π, a backward GVF $\hat{G}_\pi=[\hat{G}_{\pi,1},\cdots,\hat{G}_{\pi,k}]$ with causal filtering are associated with signals satisfying*

$$\hat{C}_i=B_i+\textstyle\sum_{j=1}^{i-1}A_{i,j}\hat{G}_{\pi,j}\quad \textit{for}\ \ 2\le i\le k,$$

where $\hat{C}_i$ and $\hat{G}_{\pi,j}$ are obtained by respectively stacking $\hat{C}_i(s,a)\in\mathbb{R}^{d_i}$ and $\hat{G}_{\pi,j}(s,a)\in\mathbb{R}^{d_i}$ over $\mathcal{S}\times\mathcal{A}$, $B_i\in\mathbb{R}^{d_i|\mathcal{S}||\mathcal{A}|}$ is an observable signal, and the coefficient matrix $A_{i,j}\in\mathbb{R}^{d_i|\mathcal{S}||\mathcal{A}|\times d_j|\mathcal{S}||\mathcal{A}|}$ captures how the j-th GVF $\hat{G}_{\pi,j}$ affects the i-th accumulation signal $\hat{C}_i$. Further, B_i and $A_{i,j}$ are bounded to ensure $\hat{G}_{\pi,i}$ to be well defined.

For an ergodic MDP that starts from $-\infty$, we have $(s_{t-1}, a_{t-1}) \sim \mu_\pi(\cdot)$, $(s_t, a_t) \sim \mathsf{P}_\pi(\cdot|s_{t-1}, a_{t-1})$, and $(s_t, a_t) \sim \mu_\pi(\cdot)$ for all $-\infty < t < \infty$. The Bayes' theorem implies that

$$P((s_{t-1}, a_{t-1})|(s_t, a_t)) = \frac{\mu_\pi(s_{t-1}, a_{t-1})\mathsf{P}_\pi((s_t, a_t)|(s_{t-1}, a_{t-1}))}{\mu_\pi(s_t, a_t)}. \tag{3}$$

The reverse conditional probability in eq. (3) together with the definition of backward GVF in Definition 2 implies that the backward GVFs $\hat{G}_\pi = [\hat{G}_{\pi,1}^\top, \cdots, \hat{G}_{\pi,k}^\top]^\top$ with causal filtering satisfies

$$\hat{G}_\pi = \hat{\mathcal{T}}_{G,\pi}\hat{G}_\pi = B + \hat{M}_\pi \hat{G}_\pi, \tag{4}$$

where $\hat{\mathcal{T}}_{G,\pi}$ denotes the backward GBO, $B = \left[B_1^\top, \cdots, B_k^\top\right]^\top$, and

$$\hat{M}_\pi = \begin{bmatrix} \gamma_1 \hat{P}_{\pi,1} & 0 & \cdots & 0 \\ A_{2,1} & \gamma_2 \hat{P}_{\pi,2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ A_{k,1} & A_{k,2} & \cdots & \gamma_k \hat{P}_{\pi,k} \end{bmatrix},$$

where $\hat{P}_{\pi,i} = U_{\pi,i}^{-1}[\mathsf{P}_\pi \otimes \mathsf{I}_{d_i}]U_{\pi,i}$.

3 Off-Policy Evaluation of GVFs

3.1 Problem Formulation

In this paper, we study the GVFs evaluation problem for a target policy π. We focus on the *behavior-agnostic off-policy* setting, in which we have access only to samples generated from an off-policy (i.e., a behavior policy) with the distribution $\mathcal{D}$, i.e., $(s_j, a_j, B_j, s_j') \sim \mathcal{D}$ $(j > 0)$. Specifically, the state-action pair (s_j, a_j) is sampled from a possibly ***unknown*** distribution $D(\cdot) : \mathcal{S} \times \mathcal{A} \to [0, 1]$, $B_j = [B_1(s_j, a_j), \cdots, B_k(s_j, a_j)]$ is an observable signal vector, and the successor state s_i' is sampled from $\mathsf{P}(\cdot|s_i, a_i)$. Our goal is to design an efficient algorithm to estimate G_π (or $\hat{G}_\pi$) given the sample set $\{(s_j, a_j, B_j, s_j')\}_{j>0}$. We make the following dataset coverage assumption.

Assumption 1. *We assume that* $D(s, a) > 0$ *for all* $(s, a) \in \mathcal{S} \times \mathcal{A}$.

3.2 Linear Function Approximation

When $|\mathcal{S}|$ is large, a linear function can be used to approximate the GVF: $G_{\pi,i}(s, a) \approx G_{\pi,i}(\theta_i; s, a) = \theta_i^\top \phi_i(s, a) = [\phi_i(s, a)^\top \otimes \mathsf{I}_{d_i}]\mathrm{vec}(\theta_i^\top)$, where $\phi_i(s, a) \in \mathbb{R}^{K_i}$ is the feature vector, and $\theta_i \in \mathbb{R}^{K_i \times d_i}$ is a learnable weight matrix. In the sequel, we omit π in $G_{\pi,i}$ and use the notation G_i. We make the following assumption for linear feature ϕ, which is standard in linear function approximation setting [45, 23, 55].

Assumption 2. *We assume that* $\|\phi_i(s, a)\|_2 \leq 1$ *for all* $i = 1, \cdots, k$ *and* $(s, a) \in \mathcal{S} \times \mathcal{A}$.

The linear approximation can then be written as $G_i(\theta_i) = [\Phi_i \otimes \mathsf{I}_{d_i}]\mathrm{vec}(\theta_i^\top)$, where Φ_i is the base matrix obtained by stacking $\phi_i(s, a)^\top$ over $\mathcal{S} \times \mathcal{A}$. To ensure the uniqueness of the solution θ_i, we assume that Φ_i has linearly independent columns. The joint vector of GVFs can be denoted as $[G_1^\top(\theta_1), \cdots, G_k^\top(\theta_k)]^\top$, which is captured by the joint parameters $\theta = [\mathrm{vec}(\theta_1^\top)^\top, \cdots, \mathrm{vec}(\theta_k^\top)^\top]^\top \in \mathbb{R}^{\sum_{i=1}^k K_i d_i}$. Then the function approximation of GVFs can be written more compactly as $G(\theta) = \Phi\theta$, where $\Phi = \mathrm{diag}([\Phi_1 \otimes \mathsf{I}_{d_1}], \cdots, [\Phi_\mathrm{k} \otimes \mathsf{I}_{d_k}])$. For each (s, a), the linear function approximation associated with each (s, a) can be written as $G(\theta; s, a) = \phi(s, a)\theta$, where $\phi(s, a) = \mathrm{diag}([\phi_1(s, a)^\top \otimes \mathsf{I}_{d_1}], \cdots, [\phi_k(s, a)^\top \otimes \mathsf{I}_{d_k}])$. We define the linear function space spanned by the columns of the feature matrix Φ as $\mathcal{F}_\Phi = \{\Phi\theta | \theta \in R_\theta\}$, in which R_θ is a convex set. Given the function class $\mathcal{F}_\Phi$, the evaluation problem of GVFs amounts to searching for a parameter $\theta^* \in R_\theta$ such that $G(\theta^*)$ approximates G_π (or $\hat{G}_\pi$) well. In the sequel, we use $\bar{\mathcal{T}}_{G,\pi}$ to represent $\mathcal{T}_{G,\pi}$ or $\hat{\mathcal{T}}_{G,\pi}$, interchangeably, based on the context.

3.3 A New Off-policy GVF Evaluation Approach

Drawbacks of GTD. In previous works, the gradient TD (GTD) method [45, 23] has been used for policy evaluation (including GVF evaluation) in the off-policy setting [37, 69, 70, 61]. GTD adopts the Mean Squared Projected Bellman Error (MSPBE) for GVF evaluation with linear function approximation, which is given by

$$\hat{\theta}^* = \operatorname{argmin}_{\theta \in R_\theta} \mathrm{MSPBE}(\theta) \triangleq \mathbb{E}_D \left[\left\| G(\theta; s, a) - \Gamma_{\mathcal{F}_\Phi, D} \bar{\mathcal{T}}_{G,\pi} G(\theta; s, a) \right\|_2^2 \right]. \tag{5}$$

where $\Gamma_{\mathcal{F}_\Phi, D}$ denotes the projection operator onto the space $\mathcal{F}_\Phi$ w.r.t. the $\|\cdot\|_D$–norm, i.e., for any vector function $f(s, a)$ of (s, a), we have $\Gamma_{\mathcal{F}_\Phi, d} f = G(\theta_f)$, in which $\theta_f = \operatorname{argmin}_{\theta \in R_\theta} \|f - G(\theta)\|_D$. One drawback of GTD is that the expectation in the objective function is taken over the off-policy sampling distribution $D(\cdot)$, which does not exactly reflect the desirable evaluation under the **target** policy. As the result, the optimal point of GTD ($\hat{\theta}^*$) can still have a large approximation error with respect to the **ground truth** value of GVF, even if the approximation function class is arbitrarily expressive. More detailed discussion about GTD is provided in Appendix D.

Generalized Temporal Difference (GenTD) Learning. In this work, we propose a novel unified approach to evaluate both the forward and backward GVFs in the off-policy setting, which we refer to as generalized temporal difference (GenTD) learning. Specifically, we aim to learn θ^* for GVF evaluation by minimizing the mean-squared projected general Bellman error (MSPGBE) defined as

$$\theta^* = \operatorname{argmin}_{\theta \in R_\theta} \mathrm{MSPGBE}(\theta) \triangleq \mathbb{E}_{\mu_\pi} \left[\left\| G(\theta; s, a) - \Gamma_{\mathcal{F}_\Phi, \mu_\pi} \bar{\mathcal{T}}_{G,\pi} G(\theta; s, a) \right\|_2^2 \right], \tag{6}$$

where recall that $\bar{\mathcal{T}}_{G,\pi}$ represents the GBO of either forward or backward GVFs. In contrast to GTD, the objective function in eq. (6) takes the expectation over the stationary distribution μ_π of the target distribution, which precisely captures the desired goal of GVF evaluation under the target policy. On the other hand, such an objective does cause implementation challenge, because the data samples are generated by the behavior policy, so that estimators based on such data directly can incur a large bias error. To solve such an issue, we will apply the density ratio $\rho(s, a) = \mu_\pi(s, a)/D(s, a)$ to adjust the distribution and further adopt the GenDICE/GradientDICE method proposed in [65, 67] to estimate $\rho(s, a)$ during the execution of the algorithm.

To describe our algorithm GenTD (see Algorithm 1), we first note that eq. (6) implies the following optimality condition for θ^* and all $f \in \mathcal{F}_\Phi$,

$$\langle G(\theta^*; \cdot) - \bar{\mathcal{T}}_{G;\pi} G(\theta^*; \cdot), f(\cdot) - G(\theta^*;) \rangle_{\mu_\pi} \geq 0,$$

or equivalently

$$\langle g(\theta^*), \theta - \theta^* \rangle \geq 0, \quad \forall \theta \in R_\theta, \tag{7}$$

where $g(\theta) = \Phi^\top U_\pi (G(\theta) - \bar{\mathcal{T}}_{G,\pi} G(\theta))$. The variational inequality theory in (Chapter 3 [20]) suggests that under an appropriately chosen stepsize α_t, the update $\theta_{t+1} = \Gamma_{R_\theta}(\theta_t - \alpha_t g(\theta_t))$ converges to the optimal point θ^*, where Γ_{R_θ} denotes the projection operator onto the set R_θ in terms of the Euclidean norm. However, since it is intractable to explicitly compute $g(\theta)$ in practice, we usually estimate $g(\theta)$ using random samples. In the off-policy setting, consider a sample $x = (s, a, s', a')$, in which $(s, a) \sim D(\cdot)$, $s' \sim \mathsf{P}(\cdot|s, a)$, and $a' \sim \pi(\cdot|s')$, we can formulate the following update rule:

$$\theta_{t+1} = \theta_t - \alpha_t \hat{\rho}(s, a) g(x, \theta_t), \tag{8}$$

where $\hat{\rho}(s, a)$ is an approximation of the density ratio $\rho(s, a) = \mu_\pi(s, a)/D(s, a)$, $g(x, \theta) = -\phi(s, a)^\top \delta(x, \theta)$ for forward GVFs and $g(x, \theta) = -\phi(s', a')^\top \delta(x, \theta)$ for backward GVFs, where $\delta(x, \theta)$ is the temporal difference error defined as $\delta(x, \theta) = B(s, a) + m(x)\phi(s', a')\theta - \phi(s, a)\theta$ for forward GVFs, and $\delta(x, \theta) = B(s', a') + \hat{m}(x)\phi(s, a)\theta - \phi(s', a')\theta$ for backward GVFs. Here m and $\hat{m}$ are matrices that capture the correlations between difference estimations in forward and backward GVFs evaluation settings, respectively. Here we adopt the GenDICE/GradientDICE method that proposed in [65, 67] to learn $\rho(s, a)$. In previous works, GenDIC/GradientDICE has only been used for estimating the scalar value $J_\pi = \mathbb{E}_{\mu_\pi}[r(s, a)]$ in the off-policy setting [65, 70, 53]. Our work is the first to adapt this method to solve the more challenging off-policy GVFs evaluation problem.

Algorithm 1 Generalized TD Learning (GenTD)

Initialize: Approximator parameters $w_{f,0}$, $w_{\rho,0}$ and θ_0
for $t = 0, \cdots, T-1$ **do**
  Obtain sample $(s_t, a_t, C_t, s'_t) \sim \mathcal{D}_d$ and $a'_t \sim \pi(\cdot|s'_t)$
  $\bar{\delta}_t = \psi_t^\top \theta_{\rho,t}(\psi'_t - \psi_t)$
  $\eta_{t+1} = w_{\rho,t} + \beta_t(\psi_t^\top w_{\rho,t} - 1 - \eta_t)$
  $w_{f,t+1} = w_{f,t} + \beta_t(\bar{\delta}_t - \psi_t^\top w_{f,t}\psi_t)$
  $w_{\rho,t+1} = \Gamma_{R_\rho}\big(w_{\rho,t} - \beta_t(\psi_t'^\top w_{f,t}\psi_t - \psi_t^\top w_{f,t}\psi_t + \eta_t\psi_t)\big)$
  $\theta_{t+1} = \Gamma_{R_\theta}\big(\theta_t - \alpha_t[w_{\rho,t}^\top \psi(s_t, a_t)]g(x_t, \theta_t)\big)$
  Forward GVF: $g(x,\theta) = -\phi(s,a)^\top(B(s,a) + m(x)\phi(s',a')\theta - \phi(s,a)\theta)$
  Backward GVF: $g(x,\theta) = -\phi(s',a')^\top(B(s',a') + \hat{m}(x)\phi(s,a)\theta - \phi(s',a')\theta)$
end for

Learning Density Ratio. GenDICE/GradientDICE estimates the density ratio $\rho(s,a)$ via solving the following min-max problem [65, 67]:

$$\min_{\rho}\max_{f,\eta} L(\hat{\rho}, f, \eta) \coloneqq \mathbb{E}_{\mathcal{D}}[\hat{\rho}(f' - f)] - \frac{1}{2}\mathbb{E}_{\mathcal{D}}[f^2] + \mathbb{E}_{\mathcal{D}}[\eta\hat{\rho} - \eta] - \frac{1}{2}\eta^2. \tag{9}$$

We parameterize both ρ and f by linear function with linearly independent features $\psi \in \mathbb{R}^{d_\rho}$, i.e., $\hat{\rho}(s,a;w_\rho) = \psi(s,a)^\top w_\rho$ and $\hat{f}(s,a;w_f) = \psi(s,a)^\top w_f$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. To guarantees the stability of the density ratio learning, we assume that the matrix $A = \mathbb{E}_{\mathcal{D}\cdot\pi}[\psi(\psi - \psi')^\top]$ is non-singular. Note that this assumption can be removed by adding an l_2–regularizer in eq. (9). In GenTD (see Algorithm 1), we estimate the density ratio via updating the parameter $w_{\rho,t}$ iteratively. The density estimator $\hat{\rho}(s_t, a_t; w_{\rho_t}) = \psi(s_t, a_t)^\top w_{\rho_t}$ is then used to reweight the update $g(x_t, \theta_t)$.

Comparison between GenTD and GTD. Compared with GTD, our GenTD has the following two advantages. First, since GTD does not adjust the distribution mismatch of sampling, the optimal point of GTD can suffer from large approximation error with respect to the ground truth GVFs even with highly expressive function classes. In contrast, the optimum of GenTD is guaranteed to approximate the ground truth GVFs well with sufficiently expressive function classes. Second, GTD needs to update a high-dimensional auxiliary parameter w simultaneously with θ to stabilize the convergence, where $w \in \mathbb{R}^{\sum_{i=1}^k K_i d_i}$ has the same dimension as $\theta \in \mathbb{R}^{\sum_{i=1}^k K_i d_i}$ (note that $\sum_{i=1}^k K_i d_i$ can be very large in the high dimensional regime or when the number of GVFs k is very large). Such an update of w can be very costly. In contrast, GenTD introduces only low-dimensional auxiliary parameters $[w_\rho, w_f, \eta] \in \mathbb{R}^{2d_\rho+1}$ for density ratio estimation, which is more efficient than GTD since d_ρ could be much smaller than $\sum_{i=1}^k K_i d_i$.

4 Main Theorems

In this section, we develop the finite-time convergence rate for our Off-GenTD algorithm. To this end, we first want to establish a certain contraction property for the general Bellman operator of interest here. Although the contraction property has been proven in the canonical value function settings [55, 70], it is unclear whether such a property still holds for **multiple multi-dimensional and interrelated** GVFs. We will next establish that such a property still holds for both forward and backward GVFs with causal filtering, but needs to be under a properly chosen norm.

Consider the GVFs vector $G_\pi = [G_{\pi,1}^\top, \cdots, G_{\pi,k}^\top]^\top$. We define a norm $\|\cdot\|_{\mu_\pi,\alpha}$ associated with a weighting vector $\alpha = [\alpha_1, \cdots, \alpha_k] \in \Delta_k$, where Δ_k denotes the simplex in k-dimensional space, as $\|G_\pi\|_{\mu_\pi,\alpha} = \sum_{i=1}^k \alpha_i \|G_{\pi,i}\|_{\mu_\pi}$, where $0 < \alpha_i \leq 1$ for all i and $\sum_{i=1}^k \alpha_i = 1$. We also define $\gamma_{\max} := \max_{i=1\cdots,k}\gamma_i$, which is strictly less than 1.

Proposition 1 (Contraction of Forward/Backward GBO). [1] *For any $G_\pi, G'_\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\sum_{i=1}^k K_i d_i}$, there exists a weighting vector α such that*

$$\big\|\overline{\mathcal{T}}_{G,\pi}G_\pi - \overline{\mathcal{T}}_{G,\pi}G'_\pi\big\|_{\mu_\pi,\alpha} \leq \tfrac{1+\gamma_{\max}}{2}\left\|G_\pi - G'_\pi\right\|_{\mu_\pi,\alpha}, \tag{10}$$

[1] The contraction property of GBO for *forward* GVFs has been proved in [33] but under different assumptions and with respect to a different norm. The result for *backward* GVFs is first established in our work.

where $\overline{\mathcal{T}}_{G,\pi}$ can be either $\mathcal{T}_{G,\pi}$ (forward GBO, eq. (2)) or $\hat{\mathcal{T}}_{G,\pi}$ (backward GBO, eq. (4)).

Despite the correlations between GVFs, Proposition 1 shows that the contraction property is still preserved under a properly chosen norm for $\mathcal{T}_{G,\pi}$ and $\hat{\mathcal{T}}_{G,\pi}$ in forward and backward GVF settings, respectively. The norm can vary for different GVFs. Proposition 1 also implies that both forward and backward GVFs (G_π and $\hat{G}_\pi$) can be identified as unique fixed point of their corresponding GBOs.

Based on Proposition 1, we next establish the monotonicity property for our GenTD algorithm, if it takes the population update $g(\theta) = \Phi^\top U_\pi(G(\theta) - \bar{\mathcal{T}}_{G,\pi}G(\theta))$.

Proposition 2 (Monotonicity)**.** *Suppose Assumption 1 &2 hold. Consider the globally optimal point θ^* defined in eq. (7). There exists a constant λ_G such that for all $\theta \in R_\theta$, we have*

$$\langle g(\theta^*) - g(\theta), \theta^* - \theta\rangle \geq \lambda_G \left\|\theta - \theta^*\right\|_F^2, \tag{11}$$

where $\lambda_G := (1-\gamma_{\max})\min_{1\leq i\leq k}\zeta_i$ and $\zeta_i := \lambda_{\min}(\Phi_i^\top U_\pi \Phi_i)$.

Proposition 2 implies the contraction property of $g(\theta)$. It guarantees that θ moves towards a globally optimal point θ^* if it is updated along the direction $-g(\theta)$. Proposition 2 generalizes the monotonicity property to a much broader class of interrelated and multi-dimensional GVF evaluation, which is far more beyond TD learning for the value function evaluation studied in [55, 70]. The following theorem characterizes the convergence rate of GenTD.

Theorem 1. *Suppose Assumption 1 &2 hold. Consider the GenTD update in Algorithm 1. Let the stepsize $\alpha_t = \Theta(t^{-1})$ and $\beta_t = \Theta(t^{-1})$. We have*

$$\mathbb{E}[\|\theta_T - \theta^*\|_F^2] \leq \mathcal{O}\left(\tfrac{\|\theta_0-\theta^*\|_F^2}{T^2}\right) + \mathcal{O}\left(\tfrac{1}{\lambda_G^3 T}\right) + \mathcal{O}\left(\tfrac{\varepsilon_\rho}{\lambda_G^2}\right), \tag{12}$$

where $\varepsilon_\rho = \sqrt{\mathbb{E}_{\mathcal{D}\cdot\pi}[\hat{\rho}(s,a;w_\rho^*) - \rho(s,a)]^2}$ is the approximation error introduced by the density ratio learning, with w_ρ^* defined in eq. (9).

Theorem 1 shows that GenTD converges to the globally optimal point θ^* at a rate $\mathcal{O}(1/T)$. The convergence speed of θ also depends on the conditional number λ_G, where the converge becomes faster as λ_G increases. Specifically, the R.H.S. of eq. (12) consists of three terms. The first term corresponds to the initialization error, which delays as fast as $\mathcal{O}(1/T^2)$. The second term corresponds to the variance error, which dominates the convergence rate of GenTD to be $\mathcal{O}(1/T)$. The last term corresponds to a non-vanishing optimality gap, which is introduced by the function approximation error in the density ratio estimation, and decreases as the expressive power of the approximation function class $\{\hat{\rho}(w_\rho) : w_\rho \in R_\rho\}$ increases. For more discussion about this approximation error, please refer to [65, 67]. The convergence analysis of GenTD is more challenging than that of TD learning [2, 7, 40] and GTD [62, 17], as we need to handle an additional approximation error introduced by the *dynamically changing* density ratio estimator $\hat{\rho}(w_{\rho_t})$.

Theorem 1 establishes the convergence of GenTD to the globally optimal point θ^* of the objective function in eq. (6), which provides the value estimation $G(\theta^*)$ for the GVFs. We are then interested in characterizing how close such an estimation is to the ground truth GVF G_π, which is our ultimate goal of evaluation. We characterize this in the following theorem.

Theorem 2 (Convergence of GenTD to Ground Truth)**.** *Consider θ^* defined in eq. (6). Suppose the same conditions in Proposition 1 & 2 hold. We have*

$$\left\|G(\theta^*) - G_\pi\right\|_{\mu_\pi,\alpha} \leq \tfrac{1}{1-\gamma_G}\left\|\Gamma_{\mathcal{F}_\Phi,\mu_\pi}G_\pi - G_\pi\right\|_{\mu_\pi,\alpha}.$$

Theorem 2 indicates that the distance between the optimal estimation $G(\theta^*)$ and the true GVF G_π is upper bounded by the approximation error of the function class $\mathcal{F}_\Phi$ for the ground truth GVF G_π (note that $\Gamma_{\mathcal{F}_\Phi,\mu_\pi}G_\pi$ denotes the projection of G_π to the function approximation class $\mathcal{F}_\Phi$). Hence, Theorem 2 guarantees that $G(\theta^*)$ can be as close as possible to the true GVF G_π, as long as the function class $\mathcal{F}_\Phi$ is sufficiently expressive. In particular, if $\mathcal{F}_\Phi$ is complete, i.e., there exists $G_\theta \in \mathcal{F}_\Phi$ such that $G_\theta = G_\pi$, then GenTD is guaranteed to converge exactly to the ground truth G_π. Note that Theorem 2 is the first result of such a type developed for both forward and backward GVFs.

Comparison between GenTD and GTD. If $\mathcal{F}_\Phi$ is complete, GTD performs similarly to GenTD and is guaranteed to converge to the ground truth G_π (see Appendix D.2 for the proof). The major difference between GenTD and GTD occurs when $\mathcal{F}_\Phi$ is not complete. In such a case, our GenTD still maintains the desirable performance as guaranteed by Theorem 2, but the optimal point of GTD (i.e., $\hat{\theta}^*$ in eq. (5)) does not have guaranteed convergence to the ground truth. As shown in [19, 10, 27], even in the value function evaluation setting (a special case of forward GVF evaluation) the approximation error $||G(\hat{\theta}^*) - G_\pi||_D$ of GTD can be arbitrarily poor even if $\mathcal{F}_\Phi$ can represent the true value function arbitrarily well (but not exactly). Such a disadvantage of GTD is mainly due to the distribution mismatch in its objective function as we discuss in Section 3.3.

In the backward GVFs evaluation setting, GTD can perform even worse. As we show in the following example, GTD may fail to learn the ground truth G_π even if the function class $\mathcal{F}_\Phi$ is complete. Note that for such a case, GenTD converges to the ground truth as guaranteed by Theorem 2.

Example 1 (GTD Fails for Complete $\mathcal{F}_\Phi$)**.** *Consider a three-state Markov chain, with transition kernel* $\mathsf{P} = [[0.1, 0.9, 0], [0.1, 0, 0.9], [0, 0.1, 0.9]]^\top$*, discount factor* $\gamma = 0.99$*, and the reward function* $R = [1, 0, 1]^\top$*. The back value function in this MDP is given by* $\bar{V} = [8.1555, 9.0389, 9.0184]^\top$*. Suppose GTD is applied to solving the evaluation problem with the parameter space* $R_\theta = \mathbb{R}$*. Then, there exists an off-policy distribution* D *such that using the perfect bases* $\Phi = [8.1555, 9.0389, 9.0184]^\top$*, the optimal point* $\bar{\theta}^*$ *learned by GTD still has non-zero approximation error, i.e.,* $||\Phi\bar{\theta}^* - \bar{V}||_D \geq 3$*.*

5 Experiments

We conduct empirical experiments to answer the following two questions: (a) can GenTD evaluate both the forward and backward GVFs efficiently? (2) how does GenTD compare with GTD in terms of the convergence speed and the quality of the estimation results? In our experiments, we

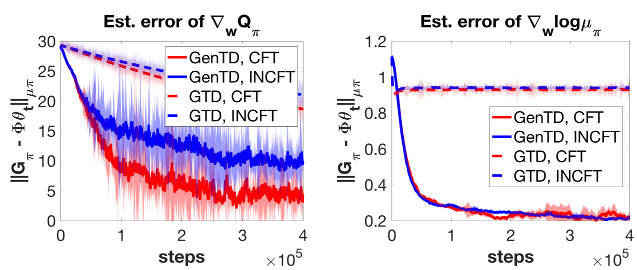

Figure 1: *Comparison between GenTD and GTD for the tasks of evaluating* $\nabla_w Q_\pi$ *and* $\nabla_w \log \mu_\pi$.

consider a variant of Baird's counterexample [1, 44] with 7 states and 2 actions (see Figure 2 in Appendix A). We study the problem of evaluating two high-dimensional GVFs, the gradient of Q-function: $\nabla_w Q_\pi \in \mathbb{R}^{14}$ (forward GVF), and the gradient of logarithmic stationary distribution: $\nabla_w \log(\mu_\pi) \in \mathbb{R}^{14}$ (backward GVF), associated with a soft-max policy parameterized by $w \in \mathbb{R}^{14}$. We consider two types of feature matrices Φ for estimating the GVFs: complete feature (CFT) and incomplete feature (INCFT), where CFT has large enough expressive power so that the ground true GVF can be fully expressed by the function class $\mathcal{F}_\Phi$, whereas INCFT does not have enough expressive power and cannot capture the ground true GVF exactly. The discount factor γ is set to be 0.99 in all tasks, and all curves in the plots are averaged over 20 independent runs. The detailed experimental setting is provided in Appendix A.

The learning curves for GenTD and GTD are provided in Figure 1. We evaluate their performances based on the estimation error with respect to the ground truth GVF: $\|\Phi\theta_t - G_\pi\|_{\mu_\pi}$. Note that both $\nabla_w Q_\pi$ and $\nabla_w \log \mu_\pi$ can be exactly computed in tabular setting, so that the estimator error of the ground truth can be computed. For the task of $\nabla_w Q$ evaluation, GenTD converges considerably faster and closer to the ground truth (i.e., smaller estimation error) than GTD. which can be attributed to the larger conditional number λ_G of GenTD. For the task of $\nabla_w \log \mu_\pi$ evaluation, GenTD moves fast towards the ground truth GVF, whereas GTD, although still converges, stays far away from the ground truth GVF even with CFT, which matches with our Example 1. As we discuss in Section 4, this is because GTD in the backward GVF evaluation setting has distribution mismatch in its objective function, which can significantly shift the optimal point from the ground truth GVF.

6 Conclusion

We studied the off-policy evaluation problem of both forward and backward GVFs. We focused on the class of GVFs with casual filtering, which covers a wide range of multiple interrelated and possibly high-dimensional GVFs. We first showed that GVFs in such a class is the fixed point of a general Bellman operator. Based on such a property, we proposed a new off-policy algorithm called GenTD. GenTD evaluates GVFs efficiently by jointly updating the GVF approximation parameter and a density ratio estimator, which adjusts the mismatch of the behavior policy and assists the convergence to the ground truth GVFs. We show that GenTD provably converges to the globally optimal point, and such an optimal point is guaranteed to converge to the ground truth GVFs as long as the function expressive power is sufficiently large. For future work, it is interesting to study nonlinear function approximation for GVFs evaluation.

7 Acknowledge

The work of T. Xu and Y. Liang was supported in part by the U.S. National Science Foundation under the grant 2148253 and 1761506.

References

[1] L. Baird. Residual algorithms: reinforcement learning with function approximation. In *Machine Learning Proceedings*, pages 30–37. 1995.

[2] J. Bhandari, D. Russo, and R. Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Proc. Conference on Learning Theory (COLT)*, pages 1691–1692, 2018.

[3] Q. Cai, Z. Yang, J. D. Lee, and Z. Wang. Neural temporal-difference and q-learning provably converge to global optima. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2019.

[4] Y. Chandak, S. Niekum, B. C. da Silva, E. Learned-Miller, E. Brunskill, and P. S. Thomas. Universal off-policy evaluation. *arXiv preprint arXiv:2104.12820*, 2021.

[5] G. Comanici, D. Precup, A. Barreto, D. K. Toyama, E. Aygün, P. Hamel, S. Vezhnevets, S. Hou, and S. Mourad. Knowledge representation for reinforcement learning using general value functions. *OpenReview*, 2018.

[6] G. Dalal, B. Szorenyi, and G. Thoppe. A tale of two-timescale reinforcement learning with the tightest finite-time bound. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 3701–3708.

[7] G. Dalal, B. Szörényi, G. Thoppe, and S. Mannor. Finite sample analyses for TD (0) with function approximation. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[8] P. D'Oro and W. Jaśkowski. How to learn a useful critic? model-based action-gradient-estimator policy optimization. *arXiv preprint arXiv:2004.14309*, 2020.

[9] C. Downey, A. Hefny, B. Li, B. Boots, and G. Gordon. Predictive state recurrent neural networks. In *Proc. Neural Information Processing Systems (NeurIPS)*, pages 6055–6066, 2017.

[10] A. Hallak and S. Mannor. Consistent on-line off-policy evaluation. In *Proc. International Conference on Machine Learning (ICML)*, pages 1372–1383, 2017.

[11] E. Hazan, S. Kakade, K. Singh, and A. Van Soest. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, pages 2681–2691, 2019.

[12] N. Heess, G. Wayne, D. Silver, T. Lillicrap, Y. Tassa, and T. Erez. Learning continuous control policies by stochastic value gradients. *arXiv preprint arXiv:1510.09142*, 2015.

[13] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge University Press, 2012.

[14] J. Huang and N. Jiang. From importance sampling to doubly robust policy gradient. In *Proc. International Conference on Machine Learning (ICML)*, pages 4434–4443, 2020.

[15] A. Jain, G. Patil, A. Jain, K. Khetarpal, and D. Precup. Variance penalized on-policy and off-policy actor-critic. *arXiv preprint arXiv:2102.01985*, 2021.

[16] N. Jiang and L. Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proc. International Conference on Machine Learning (ICML)*, pages 652–661. PMLR, 2016.

[17] M. Kaledin, E. Moulines, A. Naumov, V. Tadic, and H.-T. Wai. Finite time analysis of linear two-timescale stochastic approximation with markovian noise. In *Proc. Conference on Learning Theory (COLT)*, pages 2144–2203, 2020.

[18] N. Kallus and M. Uehara. Statistically efficient off-policy policy gradients. *arXiv preprint arXiv:2002.04014*, 2020.

[19] J. Kolter. The fixed points of off-policy TD. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, volume 24, pages 2169–2177, 2011.

[20] G. Lan. *First-order and Stochastic Optimization Methods for Machine Learning*. Springer, 2020.

[21] B. Liu, J. Liu, M. Ghavamzadeh, S. Mahadevan, and M. Petrik. Finite-sample analysis of proximal gradient TD algorithms. In *UAI*, pages 504–513, 2015.

[22] Q. Liu, L. Li, Z. Tang, and D. Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Proc. Advances in Neural Information Processing Systems (NIPS)*, 31, 2018.

[23] H. R. Maei. *Gradient temporal-difference learning algorithms*. PhD thesis, University of Alberta, 2011.

[24] A. R. Mahmood and R. S. Sutton. Representation search through generate and test. In *Proc. AAAI Workshop: Learning Rich Representations from Low-Level Sensors*, 2013.

[25] S. Mannor and J. N. Tsitsiklis. Algorithmic aspects of mean–variance optimization in markov decision processes. *European Journal of Operational Research*, 231(3):645–653, 2013.

[26] T. Morimura, E. Uchibe, J. Yoshimoto, J. Peters, and K. Doya. Derivatives of logarithmic stationary distributions for policy gradient reinforcement learning. *Neural computation*, 22(2):342–376, 2010.

[27] R. Munos. Error bounds for approximate policy iteration. In *Proc. International Conference on Machine Learning (ICML)*, volume 3, pages 560–567, 2003.

[28] O. Nachum, Y. Chow, B. Dai, and L. Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *Proc. Advances in Neural Information Processing Systems (NIPS)*, 32, 2019.

[29] D. Precup. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80, 2000.

[30] A. Ruszczyński. Risk-averse dynamic programming for markov decision processes. *Mathematical programming*, 125(2):235–261, 2010.

[31] M. Sato, H. Kimura, and S. Kobayashi. TD algorithm for the variance of return and mean-variance reinforcement learning. *Transactions of the Japanese Society for Artificial Intelligence*, 16(3):353–362, 2001.

[32] T. Schaul and M. Ring. Better generalization with forecasts. In *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1656–1662, 2013.

[33] M. Schlegel, A. Jacobsen, Z. Abbas, A. Patterson, A. White, and M. White. General value function networks. *Journal of Artificial Intelligence Research*, 70:497–543, 2021.

[34] W. F. Sharpe. Mutual fund performance. *The Journal of business*, 39(1):119–138, 1966.

[35] C. Sherstan. *Representation and general value functions*. PhD thesis, University of Alberta, 2020.

[36] S. M. Shortreed, E. Laber, D. J. Lizotte, T. S. Stroup, J. Pineau, and S. A. Murphy. Informing sequential clinical decision-making through reinforcement learning: an empirical study. *Machine Learning*, 84(1-2):109–136, 2011.

[37] D. Silver. Gradient temporal difference networks. In *European Workshop on Reinforcement Learning*, pages 117–130, 2013.

[38] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller. Deterministic policy gradient algorithms. In *Proc. International Conference on Machine Learning (ICML)*, pages 387–395, 2014.

[39] M. J. Sobel. The variance of discounted Markov decision processes. *Journal of Applied Probability*, pages 794–802, 1982.

[40] R. Srikant and L. Ying. Finite-time error bounds for linear stochastic approximation andtd learning. In *Proc. Conference on Learning Theory (COLT)*, pages 2803–2830, 2019.

[41] W. Sun, A. Venkatraman, B. Boots, and J. A. Bagnell. Learning to filter with predictive state inference machines. In *Proc. International Conference on Machine Learning (ICML)*, pages 1197–1205, 2016.

[42] R. S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44, 1988.

[43] R. S. Sutton. The grand challenge of predictive empirical abstract knowledge. In *Proc. IJCAI Workshop on Grand Challenges for Reasoning from Experiences*, 2009.

[44] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An introduction*. MIT press, 2018.

[45] R. S. Sutton, H. R. Maei, D. Precup, S. Bhatnagar, D. Silver, C. Szepesvári, and E. Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proc. International Conference on Machine Learning (ICML)*, pages 993–1000, 2009.

[46] R. S. Sutton, A. R. Mahmood, and M. White. An emphatic approach to the problem of off-policy temporal-difference learning. *The Journal of Machine Learning Research*, 17(1):2603–2631, 2016.

[47] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 1057–1063, 2000.

[48] R. S. Sutton, J. Modayil, M. Delp, T. Degris, P. M. Pilarski, A. White, and D. Precup. Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *Proc. International Conference on Autonomous Agents and Multiagent Systems*, pages 761–768, 2011.

[49] R. S. Sutton, D. Precup, and S. Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

[50] A. Tamar, D. Di Castro, and S. Mannor. Policy gradients with variance related risk criteria. In *Proc. International Conference on Machine Learning (ICML)*, pages 1651–1658, 2012.

[51] A. Tamar, D. Di Castro, and S. Mannor. Learning the variance of the reward-to-go. *The Journal of Machine Learning Research*, 17(1):361–396, 2016.

[52] A. Tamar and S. Mannor. Variance adjusted actor critic algorithms. *arXiv preprint arXiv:1310.3697*, 2013.

[53] Z. Tang, Y. Feng, L. Li, D. Zhou, and Q. Liu. Doubly robust bias reduction in infinite horizon off-policy estimation. *arXiv preprint arXiv:1910.07186*, 2019.

[54] P. Thomas and E. Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proc. International Conference on Machine Learning (ICML)*, pages 2139–2148. PMLR, 2016.

[55] J. N. Tsitsiklis and B. Van Roy. Analysis of temporal-diffference learning with function approximation. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 1075–1081, 1997.

[56] J. N. Tsitsiklis and B. Van Roy. Average cost temporal-difference learning. In *Proc. IEEE Conference on Decision and Control*, volume 1, pages 498–502, 1997.

[57] A. White et al. Developing a predictive approach to knowledge. 2015.

[58] T. Xu and Y. Liang. Sample complexity bounds for two timescale value-based reinforcement learning algorithms. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 811–819, 2021.

[59] T. Xu, Z. Wang, and Y. Liang. Improving sample complexity bounds for (natural) actor-critic algorithms. *In Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 33, 2020.

[60] T. Xu, Z. Wang, and Y. Liang. Non-asymptotic convergence analysis of two time-scale (natural) actor-critic algorithms. *arXiv preprint arXiv:2005.03557*, 2020.

[61] T. Xu, Z. Yang, Z. Wang, and Y. Liang. Doubly robust off-policy actor-critic: convergence and optimality, 2021.

[62] T. Xu, S. Zou, and Y. Liang. Two time-scale off-policy TD learning: Non-asymptotic analysis over markovian samples. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 10633–10643, 2019.

[63] H. Yao and D. Schuurmans. Reinforcement ranking. *arXiv preprint arXiv:1303.5988*, 2013.

[64] J. Zhang, A. Koppel, A. S. Bedi, C. Szepesvari, and M. Wang. Variational policy gradient method for reinforcement learning with general utilities. *arXiv preprint arXiv:2007.02151*, 2020.

[65] R. Zhang, B. Dai, L. Li, and D. Schuurmans. Gendice: generalized offline estimation of stationary values. *arXiv preprint arXiv:2002.09072*, 2020.

[66] S. Zhang, W. Boehmer, and S. Whiteson. Generalized off-policy actor-critic. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 2001–2011, 2019.

[67] S. Zhang, B. Liu, and S. Whiteson. GradientDICE: rethinking generalized offline estimation of stationary values. In *Proc. International Conference on Machine Learning (ICML)*, pages 11194–11203, 2020.

[68] S. Zhang, B. Liu, H. Yao, and S. Whiteson. Provably convergent off-policy actor-critic with function approximation. In *Proc. International Conference on Machine Learning (ICML)*, 2020.

[69] S. Zhang, B. Liu, H. Yao, and S. Whiteson. Provably convergent two-timescale off-policy actor-critic with function approximation. In *Proc. International Conference on Machine Learning (ICML)*, pages 11204–11213, 2020.

[70] S. Zhang, V. Veeriah, and S. Whiteson. Learning retrospective knowledge with reverse reinforcement learning. *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 33, 2020.

[71] S. Zhang and S. Whiteson. DAC: The double actor-critic architecture for learning options. *arXiv e-prints*, pages arXiv–1904, 2019.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [No]
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Supplementary Materials

A Specification of Experiments

The Baird's counterexample [1, 44] is shown in Figure 2. There are two actions represented by solid line and dash line, respectively. The the *dash* action leads to states 1-6 with equal probability and a reward $+1$, and *solid* action always leads to state 7 and a reward 0. The behavior distribution over the state-action space (s, a) is given as

$$D(\cdot) = \begin{cases} D(s_1, a_1) = 0.2, & D(s_1, a_2) = 0.1, \\ D(s_2, a_1) = 0.2, & D(s_2, a_2) = 0.1, \\ D(s_3, a_1) = 0.04, & D(s_3, a_2) = 0.04, \\ D(s_4, a_1) = 0.04, & D(s_4, a_2) = 0.04, \\ D(s_5, a_1) = 0.04, & D(s_5, a_2) = 0.04, \\ D(s_6, a_1) = 0.04, & D(s_6, a_2) = 0.04, \\ D(s_7, a_1) = 0.04, & D(s_7, a_2) = 0.04, \end{cases}$$

where s_i denotes state "i" ($i = 1, \cdots, 7$), a_1 denotes the *dash* action and a_2 denotes the *solid* action.

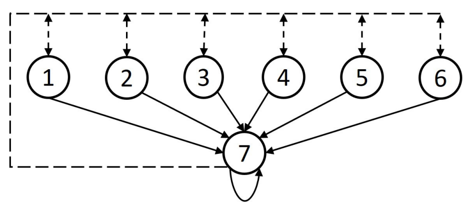

Figure 2: A variant of Baird's counterexample.

We consider the soft-max policy given as

$$\pi_w(s_i, a_j) = \frac{\exp(w_{2(i-1)+j})}{\exp(w_{2(i-1)+1}) + \exp(w_{2(i-1)+2})}, \quad i = \{1, \cdots, 7\}, \ j = \{1, 2\},$$

where $w \in \mathbb{R}^{14}$ is the parameter of the policy given as

$$w^\top = [0.0, 1.8, 0.0, 1.8, 0.0, 1.8, 0.0, 1.8, 0.0, 1.8, 0.0, 1.8, 0.0, 1.8].$$

The complete feature (CFT), incomplete feature (INCFT) and the learning rate for each task are given as follows:

- **Evaluation of $\nabla_w Q_\pi$ (forward GVF).** In this task, we need to evaluate both Q_π and $\nabla_w Q_\pi$ (see Appendix C.1 for detailed discussion about correlation between Q_π and $\nabla_w Q_\pi$). Let the complete feature matrix Φ be the identity matrix, i.e., $\Phi = I \in \mathbb{R}^{14\times 14}$. We let CFT for each (s, a) be one of the rows of Φ. We further remove one column of Φ to obtain the incomplete feature matrix $\Phi' \in \mathbb{R}^{14\times 13}$. We let INCFT for each (s, a) be one of the rows of Φ'. For GenTD, the learning rate for updating w_ρ and θ are 0.01 and 0.005, respectively. We use the same CFT and INCFT for the density ratio estimation as those for the $\nabla_w Q_\pi$ estimation. For GTD, the learning rate for both the main parameter θ and the auxiliary parameter w are 0.005.
- **Evaluation of $\nabla_w \log \mu_\pi$ (backward GVF).** Appendix C.2 has detailed discussion about such a GVF. Note that this task corresponds to the setting where $\gamma = 1$. As discussed in Appendix H, the ground true GVF in this setting is in the space perpendicular to the vector $e = [1, \cdots, 1]^\top \in \mathbb{R}^{14}$. Here we use singular value decomposition (SVD) to obtain the complete feature matrix $\Phi \in \mathbb{R}^{14\times 13}$ such that $\Phi\theta \neq Ce$ for all $\theta \neq 0$ and $C \neq 0$. We let CFT for each (s, a) be one of the rows of Φ. We further remove one column of Φ to obtain the incomplete feature matrix $\Phi' \in \mathbb{R}^{14\times 12}$. We let INCFT for each (s, a) be one of the rows of Φ'. For GenTD, the learning rate for updating w_ρ and θ are 0.05 and 0.005, respectively. We use the same CFT and INCFT for the density ratio estimation as those used in the above $\nabla_w Q_\pi$ estimation task. For GTD, the learning rate for both the main parameter θ and the auxiliary parameter w are 0.005.

A.1 Additional Experiments

In this subsection, we provide additional experiments in the task of evaluating $\nabla_w Q_\pi$ to explore the following two issues: (a) we demonstrate how GenTD performs with diminishing stepsize; and (b) we demonstrate how the bias error changes as the expressive power of the approximation function class changes. To this end, we remove one column of Φ with small weight and one column of Φ with large weight to obtain two incomplete feature matrices ${\Phi_1}'$ and ${\Phi_2}'$, respectively. We then let INCFT_1 and INCFT_2 for each (s,a) be one of the rows of ${\Phi_1}'$ and ${\Phi_2}'$, respectively. Here, the linear function class with base INCFT_1 has larger expressive power than that with INCFT_2. In the evaluation of $\nabla_w Q_\pi$, we let the learning rate for updating w_ρ and θ be $\alpha_t = \frac{1}{50+t}$ and $\beta_t = \frac{1}{100+t}$, respectively. The experiment result is given in Figure 3.

In Figure 3, we first observe that GenTD is able to achieve near zero bias error with CFT under diminishing stepsize. Second, the bias error of GenTD increases only slightly from CFT to INCFT_1, where the expressive power of INCFT_1 is still large. The bias error increases further under INCFT_2, which has lower expressive power. Overall, the bias error increases as the expressive power of the function approximation class decreases. Both observations are consistent with Theorem 3.

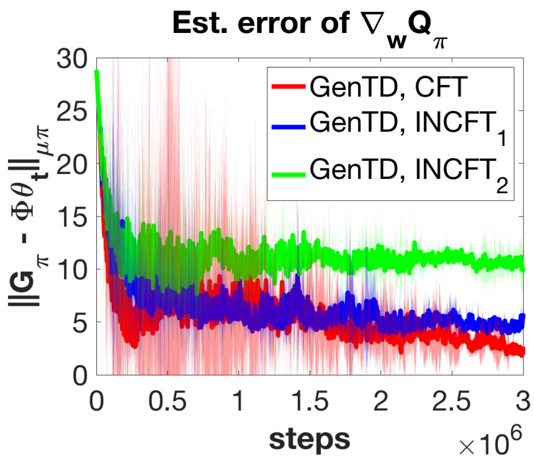

Figure 3: *Performance of GenTD in the task of evaluating $\nabla_w Q_\pi$ under diminishing stepsize.*

B Additional Related Works

The goal of OPE is to estimate the expected return of start states drawn randomly from a distribution. Importance sampling (IS) has been used for OPE in which sample rewards are reweighed to get unbiased value estimate of a new policy [29]. Later, doubly robust technique was proposed to reduce the variance of IS [16, 54, 22]. In the behavior policy agnostic setting, [28] proposed the GenDICE algorithm to estimate the IS with function approximation when performing the OPE, which also suffers less from the variance. Our approach GenTD is along the line of GenDICE in [28], which also adopts function approximation to estimate the density ratio. However, in our work we consider a more challenging setting in which we need to evaluate all the GVFs for each state-action pair instead of the mean of scalar value functions considered in [28].

C Examples of Forward and Backward GVFs

In this section, we present a number of example forward and backward GVF in RL applications.

C.1 Examples of Forward GVFs

The forward GVF in Definition 1 arises naturally in the following RL applications.

Case I: Variance of Reward-To-Go. In risk-sensitive domains such as finance, process control and clinical decision making [34, 25, 52, 50, 30, 31, 39, 15], in addition to the mean J_π

of the "reward-to-go", we are also interested in the variance of J_π [34, 36], which is given by $\mathrm{Var}[J_\pi|(s_0,a_0)=(s,a)] = H_\pi(s,a) - Q_\pi^2(s,a)$, where H_π is the second moment of J_π, i.e., $H_\pi(s,a) = \mathbb{E}[J_\pi^2|(s_0,a_0)=(s,a),\pi]$. [51] shows that H_π satisfies

$$H_\pi = R^2 + 2\gamma M_R \mathsf{P}_\pi Q_\pi + \gamma^2 \mathsf{P}_\pi H_\pi, \tag{13}$$

where R is defined in Section 2, $M_R = \mathrm{diag}(R) \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$. Equation (13) implies that H_π is the mean of the accumulation of signal $C(s,a) = r(s,a)^2 + 2\gamma\mathbb{E}[Q_\pi(s',a')|s,a]$ with discounted factor γ^2. Since $C(s,a)$ is a function of the reward $r(s,a)$ and value function $Q_\pi(s,a)$, we consider the joint vector of Q_π and H_π as G_π, i.e., $G_\pi = [Q_\pi^\top, H_\pi^\top]^\top$. We have that G_π satisfies the general Bellman equation in eq. (2) with B and M_π specified as

$$B = \begin{bmatrix} R \\ R^2 \end{bmatrix}, \quad M_\pi = \begin{bmatrix} \gamma\mathsf{P}_\pi & 0 \\ 2\gamma M_R \mathsf{P}_\pi & \gamma^2\mathsf{P}_\pi \end{bmatrix}. \tag{14}$$

We consider the setting in which reward is bounded, i.e., $r(s,a) \le C_R$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$.

Case II: Gradient of Q-function. Suppose that the policy is parametrized by a smooth function π_w, in which $w \in \mathbb{R}^{d_w}$ is the parameter. Then the gradient $\nabla_w Q_\pi(s,a)$ of the Q-function w.r.t w plays an important role in several RL applications such as variance reduced policy gradient [14] and on- and off-policy policy optimization [61, 38, 18, 5]. Specifically, [61, 18, 5] show that $\nabla_w Q_\pi$ satisfies:

$$\nabla_w Q_\pi = \gamma[\mathsf{P}_\pi \otimes \mathsf{I}_{d_w}][\nabla_w \Pi_{\pi_w} \cdot Q_\pi] + \gamma[\mathsf{P}_\pi \otimes \mathsf{I}_{d_w}]\nabla_w Q_\pi, \tag{15}$$

where $\nabla_w \Pi_{\pi_w} \in \mathbb{R}^{d_w|\mathcal{S}||\mathcal{A}|}$ is obtained by stacking $\nabla_w \log(\pi(s,a))$ over $\mathcal{S}\times\mathcal{A}$, i.e., $[\nabla_w \Pi_{\pi_w}](s,a) = \nabla_w \log(\pi_w(s,a))$, and $[\nabla_w \Pi_{\pi_w} \cdot Q_\pi] \in \mathbb{R}^{d_w|\mathcal{S}||\mathcal{A}|}$ is element-wise product between $\nabla_w \Pi_{\pi_w}$ and Q_π, i.e., $[\nabla_w \Pi_{\pi_w} \cdot Q_\pi](s,a) = \nabla_w \log(\pi_w(s,a)) Q_\pi(s,a)$. Equation (15) implies that $\nabla_w Q_\pi$ is the mean of the accumulation of signal $C(s,a) = \gamma\mathbb{E}[Q_\pi(s',a')\nabla_w \log(\pi_w(s',a'))|s,a]$ with the discounted factor γ. Let $G_\pi = [Q_\pi^\top, \nabla_w Q^\top]^\top$. We have that G_π satisfies the general Bellman equation in eq. (2) with B and M_π specified as

$$B = \begin{bmatrix} R \\ 0 \end{bmatrix}, \quad M_\pi = \begin{bmatrix} \gamma\mathsf{P}_\pi & 0 \\ \gamma[\mathsf{P}_\pi \otimes \mathsf{I}_{d_w}]\mathrm{diag}(\nabla_w \Pi_{\pi_w}) & \gamma\mathsf{P}_\pi \otimes \mathsf{I}_{d_w} \end{bmatrix}, \tag{16}$$

where $\mathrm{diag}(\nabla_w \Pi_{\pi_w}) \in \mathbb{R}^{d_w|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$ is obtained by arranging $\nabla_w \log(\pi(s,a)) \in \mathbb{R}^{d_w}$ diagonally. Without loss of generality, we assume that the score function is bounded [44, 47, 59], i.e., $\|\nabla_w \log(\pi_w(s,a))\|_2 \le C_\Pi$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$.

Case III: Stochastic Value Gradient. The stochastic value gradient (SVG) method combines advantages of model-based and model-free methods, in which both the estimated model and value function are updated to evaluate the policy gradient [12]. In the framework of SVG, the reward $r(s,a)$ is differentiable with respect to both $s \in \mathbb{R}^{d_s}$ an $a \in \mathbb{R}^{d_a}$, the stochastic policy takes the form $a = \pi(s,w) + \eta$, and the transition probability is modelled as $s' = f(s,a) + \xi$, where $\pi : \mathcal{S} \to \mathcal{A}$ and $f : \mathcal{S}\times\mathcal{A} \to \mathcal{S}$ are deterministic mappings, $w \in \mathbb{R}^{d_w}$ is the policy parameter, and $\eta \sim P(\eta)$ and $\xi \sim P(\xi)$ are noise variables. We abbreviate the partial differentiation using subscripts as $g_x \triangleq \partial g/\partial x$. The gradient of the Q-function w.r.t the policy parameter w is given by [12]

$$\begin{aligned} Q_s &= (R_s + \Pi_s R_a) + \gamma(\Pi_s F_a + F_s)([\mathsf{P}_\pi \otimes \mathsf{I}_{d_s}]Q_s), \\ Q_a &= R_a + \gamma F_a[\mathsf{P}_\pi \otimes \mathsf{I}_{d_s}]Q_s + \gamma[\mathsf{P}_\pi \otimes \mathsf{I}_{d_a}]Q_a, \\ \nabla_w Q_\pi &= \Pi_w Q_a + \gamma[\mathsf{P}_\pi \otimes \mathsf{I}_{d_w}]\nabla_w Q_\pi, \end{aligned} \tag{17}$$

where $R_s \in \mathbb{R}^{d_s|\mathcal{S}||\mathcal{A}|}$ and $R_a \in \mathbb{R}^{d_a|\mathcal{S}||\mathcal{A}|}$ are vectors obtained via stacking partial derivatives $r_s(s,a) \in \mathbb{R}^{d_s}$ and $r_a(s,a) \in \mathbb{R}^{d_a}$ over $(s,a) \in \mathcal{S}\times\mathcal{A}$, and $\Pi_s = \mathrm{diag}(\pi_s) \in \mathbb{R}^{d_s|\mathcal{S}||\mathcal{A}|\times d_a|\mathcal{S}||\mathcal{A}|}$, $\Pi_w = \mathrm{diag}(\pi_w) \in \mathbb{R}^{d_w|\mathcal{S}||\mathcal{A}|\times d_a|\mathcal{S}||\mathcal{A}|}$, $F_s = \mathrm{diag}(f_s) \in \mathbb{R}^{d_s|\mathcal{S}||\mathcal{A}|\times d_s|\mathcal{S}||\mathcal{A}|}$, and $F_a = \mathrm{diag}(f_a) \in \mathbb{R}^{d_a|\mathcal{S}||\mathcal{A}|\times d_s|\mathcal{S}||\mathcal{A}|}$ are Jacobian matrices. Consider GVF defined as $G_\pi = [Q_s^\top, Q_a^\top, \nabla_w Q_\pi^\top]^\top$. Consider the normalized setting in which $\Pi_s F_a + F_s = \mathbf{I}$. Then G_π satisfies the general Bellman equation in eq. (2) with B and M_π specified by

$$B = \begin{bmatrix} R_s + \Pi_s R_a \\ R_a \\ \Pi_w Q_a \end{bmatrix}, \quad M_\pi = \begin{bmatrix} \gamma[\mathsf{P}_\pi \otimes \mathsf{I}_s] & 0 & 0 \\ \gamma F_a[\mathsf{P}_\pi \otimes \mathsf{I}_s] & \gamma[\mathsf{P}_\pi \otimes \mathsf{I}_a] & 0 \\ 0 & \Pi_w & \gamma[\mathsf{P}_\pi \otimes \mathsf{I}_{d_w}] \end{bmatrix}. \tag{18}$$

We consider the setting in which $\|f_a(s,a)\|_F \le C_a$ and $\|\pi_w(s,a)\|_F \le C_w$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$. We make the following "non-expansive" assumption for the transition matrix $\gamma_1(s,a) = \gamma(\Pi_s F_a + F_s)$.

Assumption 3 (Normalization). *For any $v \in \mathbb{R}^{d_s \times 1}$, we have $\|(\Pi_s F_a + F_s)v\|_{U_\pi} \leq \|v\|_{U_\pi}$.*

Assumption 3 is the minimum requirement to guarantee the value of Q_s to be bounded. It can be satisfied by selecting appropriate policy class π_w and model approximation f together with the feature design of both state s and action a.

Case IV: Option Learning. In the option framework [49], an option is defined as $(\pi_o, \lambda_o, \mathcal{O})$, where $\pi_o : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is an intra-option policy, $\lambda_o : \mathcal{S} \to [0, 1]$ is a termination function, and $\mathcal{O}$ is the option set. In this framework, the policy π is defined over the option-state space, i.e., $\pi : \mathcal{O} \times \mathcal{S} \to [0, 1]$. At time step t, an agent at state s_t either terminates the previous option o_{t-1} with probability $\lambda_{o_{t-1}}(s_t)$ and initiates a new option o_t according to policy $\pi(\cdot|s_t)$, or proceeds with the previous option o_{t-1} with probability $1 - \lambda_{o_{t-1}}(s_t)$ and sets $o_t = o_{t-1}$. Then an action a_t is selected according to $\pi_{o_t}(\cdot|s_t)$. The agent receive a reward $r(s_t, a_t)$. Similar to regular MDP, here we define state-option value function as $Q_\pi(s, o) = \int_a \pi(a|s) Q_\pi(s, o, a) da$, where $Q_\pi(s, o, a) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)|s_0 = s, a_0 = a, o_0 = o]$. We consider evaluate the state-option value function $Q_\pi(s, o)$. [71] shows that

$$Q_\pi^{\mathcal{O}} = R + \gamma P_\pi^{\lambda, \mathcal{O}} Q_\pi^{\mathcal{O}}, \tag{19}$$

where $[P_\pi^{\lambda,\mathcal{O}}]((s, o), (s', o')) = [(1 - \lambda_o(s'))\mathbb{1}_{o'=o} + \lambda_o(s')\pi(o'|s')]\mathsf{P}(s'|s, o)$, where $\mathsf{P}(s'|s, o) = \int_a \pi_o(a|s)\mathsf{P}(s'|s, a)$. Let $Y_\pi(x) = Q_\pi(s, o)$, $C(x_t) = r(s_t, a_t)$ and $m(x, x_{t+1}) = \gamma$, it can be checked that Y_π satisfies the general Bellman equation defined in Section 2.1.

C.2 Examples of Backward GVFs

The backward GVF in Definition 2 also arises in the following important RL applications.

Case IV: Anomaly Detection. [70] has systemically discussed the application of retrospective knowledge in anomaly detection. Let $i(s, a)$ be the cost that an agent consumes when taking action a at state s, and $e_\pi(s, a)$ be the cost that an agent is expected to consume given the current status when following a predefined policy π. If the actual cost of the agent deviates too much from e_π, the agent may likely encounter anomalous events. For simplicity, we consider the setting when $\gamma(s, a) = \gamma$. It can be shown that e_π satisfies the following equation

$$e_\pi = i + \gamma U_\pi^{-1} \mathsf{P}_\pi^\top U_\pi e_\pi. \tag{20}$$

Clearly, eq. (20) satisfies the general backward Bellman equation in eq. (4) by letting $B = i$, $M_\pi = \gamma \mathsf{P}_\pi$, and $\hat{G}_\pi = e_\pi$.

Case V: Gradient of Logarithmic Stationary Distributions. In the policy parameterization setting, the gradient of logarithmic stationary distribution $\nabla_w \log \mu_\pi(s, a)$ has been used in policy gradient estimation [18, 61, 26] and maximum entropy exploration [11]. It has been shown in [26, 61] that $\nabla_w \log \mu_\pi(s, a)$ satisfies the following equation

$$\Psi_\pi = \nabla_w \Pi_{\pi_w} + U_\pi^{-1}[\mathsf{P}_\pi^\top \otimes \mathsf{I}_{d_w}]U_\pi \Psi_\pi, \tag{21}$$

where Ψ_π is obtained via stacking $\nabla_w \log(\mu_\pi(s, a))$ over $\mathcal{S} \times \mathcal{A}$, i.e., $[\Psi_\pi](s, a) = \nabla_w \log(\mu_{\pi_w}(s, a))$. Here, $\nabla_w \log \mu_\pi$ can be viewed as a backward accumulation of the signal $C(s, a) = \nabla_w \log(\pi(s, a))$ with the discounted factor $\gamma = 1$. Define the backward GVF as $\hat{G}_\pi = \Psi_\pi$. It is clear that $\hat{G}_\pi$ satisfies the general backward Bellman equation in eq. (4) with B and M_π specified by

$$B = \nabla_w \Pi_{\pi_w}, \qquad M_\pi = \mathsf{P}_\pi^\top \otimes \mathsf{I}_{d_w}. \tag{22}$$

Note that since $\gamma_{\max} = 1$ in the general Bellman equation in eq. (21), the result in Proposition 1 may not hold in such a setting, i.e., GBO may not be a contraction here. However, as we will show in appendix H, when the base matrix Φ satisfies the "non-constant parameterization" assumption, we can establish results similar to Proposition 2 and Theorem 3 for the evaluation of $\nabla_w \log \mu_\pi$.

D Gradient Temporal Difference Learning (GTD)

The GTD algorithm has been used for GVF evaluation in [48, 37]. So far, only the asymptomatic convergence (not the convergence rate) has been studied in [37]. In this section, we present the

Algorithm 2 GTD

Initialize: Approximator parameters w_0, and θ_0
for $t = 0, \cdots, T-1$ **do**
  Obtain sample $(s_t, a_t, B_t, s_t') \sim \mathcal{D}$ and $a_t' \sim \pi(\cdot|s_t')$
  $w_{t+1} = w_t - \beta_t(g(x_t, \theta_t) - l(x_t, w_t))$
    forward GVF: $l(x_t, w_t) = \phi(s_t, a_t)^\top \phi(s_t, a_t) w_t$
    backward GVF: $l(x_t, w_t) = \phi(s_t', a_t')^\top \phi(s_t', a_t') w_t$
  $\theta_{t+1} = \Gamma_{R_\theta}\left(\theta_t + \alpha_t(g(x_t, \theta_t) - h(x_t, \theta_t))\right)$
    forward GVF: $h(x_t, w_t) = \phi(s_t', a_t')^\top m(x_t) \phi(s_t, a_t) w_t$
    backward GVF: $h(x_t, w_t) = \phi(s_t, a_t)^\top \hat{m}(x_t) \phi(s_t', a_t') w_t$
end for

GTD algorithm for GVF evaluation and characterize the finite-time convergence rate for GTD. We define the dimension of parameter θ as $d_g = \sum_{i=1}^k K_i d_i$. In the sequel, we denote the MSPBE with parameter θ as $J(\theta)$. Note that the MSPBE in eq. (5) can be rewritten as

$$J(\theta) = \frac{1}{2} \left\|\mathbb{E}_D[g(x, \theta)]\right\|_{C^{-1}}^2, \ C \triangleq \begin{cases} \mathbb{E}_D[(\phi(s,a)^\top \phi(s,a))] & \text{(forward GVF)} \\ \mathbb{E}_D[(\phi(s',a')^\top \phi(s',a'))] & \text{(backward GVF)} \end{cases}, \tag{23}$$

where

$$g(x, \theta) \triangleq \begin{cases} -\phi(s,a)^\top (B(s,a) + m(x)\phi(s',a')\theta - \phi(s,a)\theta) & \text{(forward GVF)} \\ -\phi(s',a')^\top (B(s',a') + \hat{m}(x)\phi(s,a)\theta - \phi(s',a')\theta) & \text{(backward GVF)} \end{cases},$$

where the matrices $m(\cdot)$ and $\hat{m}(\cdot)$ are defined in eq. (8). The gradient of $J(\theta)$ is given as

$$-\nabla J(\theta) = \mathbb{E}_D[g(x,\theta) - h(x,\theta)], \ h(x,\theta) \triangleq \begin{cases} \phi(s',a')^\top m(x) \phi(s,a) w(\theta) & \text{(forward GVF)} \\ \phi(s,a)^\top \hat{m}(x) \phi(s',a') w(\theta) & \text{(backward GVF)} \end{cases}$$

in which $w(\theta) = C^{-1}\mathbb{E}_D[g(x,\theta)] \in \mathbb{R}^{d_g}$. Note that we can not estimate $\nabla J(\theta)$ directly due to the "double sampling" issue, i.e., $w(\theta)$ cannot be estimated via sampling. In GTD, an auxiliary parameter w_t is introduced, which is updated simultaneously with θ_t to approximate $w(\theta_t)$ [45, 23]. We present the update of GTD in Algorithm 2.

D.1 Convergence Rate of GTD

We make the following assumptions, which have also been adopted in the convergence analysis of GTD in the canonical value function evaluation setting [62, 60, 45, 23].

Assumption 4. *In both forward and backward GVF evaluation settings, the matrix C in eq.* (23) *is non-singular.*

Assumption 5. *We define the matrix $\hat{A}$ in the following way: (1) in the forward GVF evaluation setting: $\hat{A} = \mathbb{E}_{\mathcal{D}\cdot\pi}[\phi(s,a)^\top(m(x)\phi(s',a') - \phi(s,a))]$; (2) in the backward GVF evaluation setting, $\hat{A} = \mathbb{E}_{\mathcal{D}\cdot\pi}[\phi(s',a')(\hat{m}(x)\phi(s,a) - \phi(s',a'))]$. We require $\hat{A}$ to be non-singular in both the forward and backward GVF settings.*

We define the optimal point $\bar{\theta}^*$ for GTD as

$$\langle \nabla J(\bar{\theta}^*), \theta - \bar{\theta}^* \rangle \geq 0, \quad \forall \theta \in R_\theta,$$

which is the optimality condition for minimizing $J(\theta)$. The following theorem characterizes the convergence rate of GTD to $\bar{\theta}^*$.

Theorem 3. *Consider the GTD update in Algorithm 2. In both the forward and backward GVF evaluation settings, suppose Assumption 4-5 hold. Let the stepsize $\alpha_t = \Theta(t^{-1})$ and $\beta_t = \Theta(t^{-1})$. We have*

$$\mathbb{E}\left[\left\|\theta_T - \bar{\theta}^*\right\|_2^2\right] \leq \mathcal{O}\left(\frac{\left\|\theta_0 - \bar{\theta}^*\right\|_2^2}{T^2}\right) + \mathcal{O}\left(\frac{1}{\lambda_G'^2 T}\right),$$

where $\lambda_G' > 0$ is the conditional number of GTD defined in eq. (71) *of Appendix I.*

Theorem 1 shows that GTD converges to the globally optimal point $\bar{\theta}^*$ at a rate of $\mathcal{O}(1/T)$. The convergence speed of θ_t depends on the conditional number λ'_G, which decreases as λ'_G decreases. Differently from the conditional number λ_G of GenTD, which has a guaranteed lower bound from zero as given in Proposition 2, there exists no guaranteed lower bound for λ'_G even in the canonical value function evaluation setting. Thus, the converge speed of GTD could be very slow as λ'_G could be arbitrarily small.

D.2 Global Optimum of GTD and Proof of Example 1

For simplicity, we consider scenarios when the function approximation class $\mathcal{F}_\Phi$ is complete. We show that the global optimum of GTD exhibits very different properties in the forward and backward GVF evaluation settings.

We first show that in the forward GVF evaluation setting, the global optimum $\Phi\bar{\theta}^*$ of GTD equals the ground truth GVF. Since the function space $\mathcal{F}_\Phi$ is complete, there exists a parameter $\theta_{\text{true}} \in R_\theta$ such that $\Phi\theta_{\text{true}} = G_\pi$, which implies $J(\theta_{\text{true}}) = 0$. Since $J(\theta) \geq 0$ for all $\theta \in R_\theta$ and $J(\theta)$ is strongly-convex, $J(\theta) = 0$ if and only if $\theta = \bar{\theta}^*$, which implies $\theta_{\text{true}} = \bar{\theta}^*$.

In the backward GVFs evaluation setting, we provide an example (see Example 1 in Section 4) to show that GTD can fail to learn the ground truth G_π even if the function class $\mathcal{F}_\Phi$ is complete. We next present the proof for such an example.

Proof of Example 1. The backward value function can be obtained as follows

$$\bar{V} = U^{-1}(I - \gamma \mathsf{P}^\top)^{-1}\mathsf{P}^\top U R = [8.1555, 9.0389, 9.0184]^\top.$$

The fixed point of GTD is given by

$$\bar{\theta}^* = \bar{A}^{-1}\bar{b},$$

where

$$\bar{A} = \gamma \Phi^\top \mathsf{P}^\top \bar{D}'\Phi - \Phi^\top \bar{D}'\Phi, \qquad \bar{b} = \Phi^\top \mathsf{P}^\top \bar{D} R, \tag{24}$$

where $\bar{D} = \text{diag}(D)$, $\bar{D}' = \text{diag}(D')$ and $D'^\top = D^\top P$. Also note that the base matrix $\Phi = [8.1555, 9.0389, 9.0184]^\top$ and the off-policy sampling distribution $D = [1/3, 1/3, 1/3]^\top$. We can obtain

$$\bar{A} = -9.9422, \qquad \bar{b} = 5.9904,$$

which implies

$$\bar{\theta}^* = \frac{-b}{A} = 0.6025. \tag{25}$$

Note that the perfect base matrix $[8.1555, 9.0389, 9.0184]^\top$ can fully represent $\bar{V}_\pi$, with parameter $\theta_{\text{true}} = 1$. However, eq. (25) shows that the global optimum of GTD $\bar{\theta}^* \neq \theta_{\text{true}}$, which introduces a non-zero approximation error:

$$\left\|\Phi\bar{\theta}^* - \bar{V}\right\|_D = 3.7848.$$

$\square$

The above example demonstrates a drawback of GTD, which can fail to learn the ground truth G_π even if the function class $\mathcal{F}_\Phi$ is complete. Such an issue does not occur for the GenTD algorithm proposed in this paper, which converges to the ground truth as guaranteed by Theorem 2 in Section 4.

E Proofs of Propositions 1 and 2

E.1 Supporting Lemmas

We provide the following lemmas, which are useful for the proofs of Propositions 1 and 2.

Lemma 1. *For any $v \in \mathbb{R}^{d|\mathcal{S}||\mathcal{A}|}$,we have $\|[\mathsf{P}_\pi \otimes \mathsf{I}_d]v\|_{U_\pi} \leq \|v\|_{U_\pi}$.*

Proof. Consider the square of $\|[\mathsf{P}_\pi \otimes \mathsf{I}_d]v\|_{U_\pi}$. We have

$$\begin{aligned}
\|[\mathsf{P}_\pi \otimes \mathsf{I}_d]v\|_{U_\pi}^2 &= v^\top[\mathsf{P}_\pi^\top \otimes \mathsf{I}_d]U_\pi[\mathsf{P}_\pi \otimes \mathsf{I}_d]v \\
&= \sum_{i=1}^{|\mathcal{S}||\mathcal{A}|} \mu_\pi(i) \left\| \sum_{j=1}^{|\mathcal{S}||\mathcal{A}|} \mathsf{P}_\pi(j|i)v_j \right\|_2^2 \\
&\le \sum_{i=1}^{|\mathcal{S}||\mathcal{A}|} \mu_\pi(i) \sum_{j=1}^{|\mathcal{S}||\mathcal{A}|} \mathsf{P}_\pi(j|i)\left\|v_j\right\|_2^2 \\
&= \sum_{i=1}^{|\mathcal{S}||\mathcal{A}|} \sum_{j=1}^{|\mathcal{S}||\mathcal{A}|} \mu_\pi(i)\mathsf{P}_\pi(j|i)\left\|v_j\right\|_2^2 \\
&= \sum_{j=1}^{|\mathcal{S}||\mathcal{A}|} \mu_\pi(i)\left\|v_j\right\|_2^2 \\
&= \|v\|_{U_\pi}^2,
\end{aligned}$$

where the first inequality follows from Jensen's inequality and the fourth equality follows from the property of the stationary distribution $\mu_\pi^\top \mathsf{P}_\pi = \mu_\pi^\top$. ☐

We provide the follow lemma to characterize a similar property in backward GVF evaluation setting.

Lemma 2. *For any $v \in \mathbb{R}^{d|\mathcal{S}||\mathcal{A}|}$,we have $\left\|U_\pi^{-1}[\mathsf{P}_\pi^\top \otimes \mathsf{I}_d]U_\pi v\right\|_{U_\pi} \le \|v\|_{U_\pi}$.*

Proof. Consider the square of $\left\|U_\pi^{-1}[\mathsf{P}_\pi^\top \otimes \mathsf{I}_d]U_\pi v\right\|_{U_\pi}$. We have

$$\begin{aligned}
\left\|U_\pi^{-1}[\mathsf{P}_\pi \otimes \mathsf{I}_d]U_\pi v\right\|_{U_\pi}^2 &= v^\top[U_\pi^{-1}[\mathsf{P}_\pi^\top \otimes \mathsf{I}_d]U_\pi]^\top U_\pi[U_\pi^{-1}[\mathsf{P}_\pi^\top \otimes \mathsf{I}_d]U_\pi]v \\
&= \sum_{j=1}^{|\mathcal{S}||\mathcal{A}|} \mu_\pi(j) \left\| \sum_{i=1}^{|\mathcal{S}||\mathcal{A}|} \frac{\mu_\pi(i)\mathsf{P}_\pi(j|i)}{\mu_\pi(j)} v(i) \right\|_2^2 \\
&\le \sum_{j=1}^{|\mathcal{S}||\mathcal{A}|} \mu_\pi(j) \sum_{i=1}^{|\mathcal{S}||\mathcal{A}|} \frac{\mu_\pi(i)\mathsf{P}_\pi(j|i)}{\mu_\pi(j)} \left\|v(i)\right\|_2^2 \\
&= \sum_{j=1}^{|\mathcal{S}||\mathcal{A}|} \mu_\pi(i)\left\|v(i)\right\|_2^2 \sum_{i=1}^{|\mathcal{S}||\mathcal{A}|} \mathsf{P}_\pi(j|i) \\
&= \sum_{j=1}^{|\mathcal{S}||\mathcal{A}|} \mu_\pi(i)\left\|v(i)\right\|_2^2 \\
&= \|v\|_{U_\pi}^2,
\end{aligned}$$

where the first inequality follows from the Jensen's inequality. ☐

E.2 Proof of Proposition 1

We first consider the **forward GBO setting**. Recall the following definition of GBO $\mathcal{T}_{G,\pi}$ in eq. (2)

$$G_\pi = \mathcal{T}_{G,\pi}G_\pi = B + M_\pi G_\pi,$$

where

$$B = \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_k \end{bmatrix}, \quad M_\pi = \begin{bmatrix} \gamma_1[\mathsf{P}_\pi \otimes \mathsf{I}_{d_1}] & 0 & \cdots & 0 \\ A_{2,1} & \gamma_2[\mathsf{P}_\pi \otimes \mathsf{I}_{d_2}] & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ A_{k,1} & A_{k,2} & \cdots & \gamma_k[\mathsf{P}_\pi \otimes \mathsf{I}_{d_k}] \end{bmatrix}.$$

Let $G'_\pi, G''_\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\sum_{i=1}^k K_i d_i}$ be two vectors, and let $\Delta_G = [\Delta_1, \cdots, \Delta_k]$, where $\Delta_i = G'_{\pi,i} - G''_{\pi,i}$. We have

$$\mathcal{T}_\pi G'_\pi - \mathcal{T}_\pi G''_\pi = M_\pi \Delta_G = \begin{bmatrix} \gamma_1[\mathsf{P}_\pi \otimes \mathsf{I}_{d_1}]\Delta_1 \\ A_{2,1}\Delta_1 + \gamma_2[\mathsf{P}_\pi \otimes \mathsf{I}_{d_2}]\Delta_2 \\ \vdots \\ \sum_{j=1}^{k-1} A_{k,j}\Delta_j + \gamma_k[\mathsf{P}_\pi \otimes \mathsf{I}_{d_k}]\Delta_k \end{bmatrix}. \tag{26}$$

Recall that $A_{i,j}$ is bounded for all i, j. Thus, there exists a constant $0 < C_A < \infty$ such that $\|A_{i,j}\|_{U_\pi} \leq C_A$ for all i, j. Without loss of generality, we assume $C_A > 1$. Let α be the solution of the following matrix function

$$Fx = f, \tag{27}$$

where $F \in \mathbb{R}^{k\times k}$ and $f \in \mathbb{R}^k$ are specified as

$$F = \begin{bmatrix} -\frac{1-\gamma}{2} & C_A & C_A & \cdots & C_A & C_A \\ 0 & -\frac{1-\gamma}{2} & C_A & \cdots & C_A & C_A \\ \vdots & \vdots & & \vdots & & \\ 0 & 0 & 0 & \cdots & -\frac{1-\gamma}{2} & C_A \\ 1 & 1 & 1 & \cdots & 1 & 1 \end{bmatrix}, \qquad f = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}.$$

It can be checked that the solution of eq. (27) is strictly positive, i.e., if $F\alpha = f$, then we have $\alpha_l > 0$ for $1 \leq l \leq k$. Recalling the definition of $\|\cdot\|_{\mu_\pi,\alpha}$– norm, we have

$$\begin{aligned}
&\|M_\pi \Delta_G\|_{\mu_\pi,\alpha} \\
&= \gamma_1\alpha_1 \|[\mathsf{P}_\pi \otimes \mathsf{I}_{d_1}]\Delta_1\|_{\mu_\pi} + \alpha_2 \|A_{2,1}\Delta_1 + \gamma_2[\mathsf{P}_\pi \otimes \mathsf{I}_{d_2}]\Delta_2\|_{\mu_\pi} \\
&\quad + \cdots + \alpha_k \left\| \sum_{j=1}^{k-1} A_{k,j}\Delta_j + \gamma_k[\mathsf{P}_\pi \otimes \mathsf{I}_{d_k}]\Delta_k \right\|_{\mu_\pi} \\
&\leq \gamma_1\alpha_1 \|[\mathsf{P}_\pi \otimes \mathsf{I}_{d_1}]\Delta_1\|_{\mu_\pi} + \alpha_2 \|A_{2,1}\Delta_1\|_{\mu_\pi} + \cdots + \alpha_k \|A_{k,1}\Delta_1\|_{\mu_\pi} \\
&\quad + \gamma_2\alpha_2 \|[\mathsf{P}_\pi \otimes \mathsf{I}_{d_2}]\Delta_2\|_{\mu_\pi} + \alpha_3 \|A_{3,2}\Delta_2\|_{\mu_\pi} + \cdots + \alpha_k \|A_{k,2}\Delta_2\|_{\mu_\pi} \\
&\quad + \cdots \\
&\quad + \gamma_k\alpha_k \|[\mathsf{P}_\pi \otimes \mathsf{I}_{d_k}]\Delta_k\|_{\mu_\pi} \\
&\leq \left(\gamma\alpha_1 + C_A \sum_{i=2}^k \alpha_i\right) \|\Delta_1\|_{\mu_\pi} + \left(\gamma\alpha_2 + C_A \sum_{i=3}^k \alpha_i\right) \|\Delta_2\|_{\mu_\pi} + \cdots + \gamma\alpha_k \|\Delta_k\|_{\mu_\pi} \\
&\leq \frac{1+\gamma}{2}\alpha_1 \|\Delta_1\|_{\mu_\pi} + \frac{1+\gamma}{2}\alpha_2 \|\Delta_2\|_{\mu_\pi} + \cdots + \gamma\alpha_k \|\Delta_k\|_{\mu_\pi} \\
&\leq \frac{1+\gamma}{2} \left(\sum_{i=1}^k \alpha_i \|\Delta_i\|_{\mu_\pi} \right) \\
&= \frac{1+\gamma}{2} \|\Delta_G\|_{\mu_\pi,\alpha},
\end{aligned} \tag{28}$$

where the first inequality follows from the triangle inequality, the second inequality follows from the fact that $A_{i,j}$ is bounded and Lemma 1, and the third inequality follows from the definition of γ and the fact that α is the solution of eq. (27). Obviously, eq. (28) implies the following property,

$$\|\mathcal{T}_\pi G'_\pi - \mathcal{T}_\pi G''_\pi\|_{\mu_\pi,\alpha} \leq \frac{1+\gamma}{2} \|G'_\pi - G''_\pi\|_{\mu_\pi,\alpha},$$

which completes the proof in the forward GBO evaluation setting.

We next consider the **backward GBO setting**, where $\hat{\mathcal{T}}_{G,\pi}$ is defined in eq. (3). Following steps similar to those from eq. (26) – eq. (28), we can obtain

$$\left\| \hat{\mathcal{T}}_\pi G_\pi - \hat{\mathcal{T}}_\pi G'_\pi \right\|_{\mu_\pi,\alpha} = \|M_\pi \Delta_G\|_{\mu_\pi,\alpha}$$

$$
\begin{aligned}
&= \gamma_1\alpha_1 \left\| U_{\pi,1}^{-1}[\mathsf{P}_\pi \otimes \mathsf{I}_{d_1}]U_{\pi,1}\Delta_1 \right\|_{\mu_\pi} + \alpha_2 \left\| A_{2,1}\Delta_1 + \gamma_2 U_{\pi,2}^{-1}[\mathsf{P}_\pi \otimes \mathsf{I}_{d_2}]U_{\pi,2}\Delta_2 \right\|_{\mu_\pi} + \cdots \\
&\quad + \alpha_k \left\| \sum_{j=1}^{k-1} A_{k,j}\Delta_j + \gamma_k U_{\pi,k}^{-1}[\mathsf{P}_\pi \otimes \mathsf{I}_{d_k}]U_{\pi,k}\Delta_k \right\|_{\mu_\pi} \\
&\leq \gamma_1\alpha_1 \left\| U_{\pi,1}^{-1}[\mathsf{P}_\pi \otimes \mathsf{I}_{d_1}]U_{\pi,1}\Delta_1 \right\|_{\mu_\pi} + \alpha_2 \left\| A_{2,1}\Delta_1 \right\|_{\mu_\pi} + \cdots + \alpha_k \left\| A_{k,1}\Delta_1 \right\|_{\mu_\pi} \\
&\quad + \gamma_2\alpha_2 \left\| U_{\pi,2}^{-1}[\mathsf{P}_\pi \otimes \mathsf{I}_{d_2}]U_{\pi,2}\Delta_2 \right\|_{\mu_\pi} + \alpha_3 \left\| A_{3,2}\Delta_2 \right\|_{\mu_\pi} + \cdots + \alpha_k \left\| A_{k,2}\Delta_2 \right\|_{\mu_\pi} \\
&\quad + \cdots \\
&\quad + \gamma_k\alpha_k \left\| U_{\pi,k}^{-1}[\mathsf{P}_\pi \otimes \mathsf{I}_{d_k}]U_{\pi,k}\Delta_k \right\|_{\mu_\pi} \\
&\leq \left(\gamma\alpha_1 + C\sum_{i=2}^{k}\alpha_i \right) \left\|\Delta_1\right\|_{\mu_\pi} + \left(\gamma\alpha_2 + C\sum_{i=3}^{k}\alpha_i \right) \left\|\Delta_2\right\|_{\mu_\pi} + \cdots + \gamma\alpha_k \left\|\Delta_k\right\|_{\mu_\pi} \\
&\leq \frac{1+\gamma}{2}\alpha_1 \left\|\Delta_1\right\|_{\mu_\pi} + \frac{1+\gamma}{2}\alpha_2 \left\|\Delta_2\right\|_{\mu_\pi} + \cdots + \gamma\alpha_k \left\|\Delta_k\right\|_{\mu_\pi} \\
&\leq \frac{1+\gamma}{2} \left(\sum_{i=1}^{k} \alpha_i \left\|\Delta_i\right\|_{\mu_\pi} \right) \\
&= \frac{1+\gamma}{2} \left\|\Delta_G\right\|_{\mu_\pi,\alpha}, \qquad (29)
\end{aligned}
$$

where the first inequality follows from the triangle inequality, the second inequality follows from the fact that $A_{i,j}$ is bounded and Lemma 2, and the third inequality follows from the definition of γ and the fact that α is the solution of eq. (27). Equation (29) implies the following

$$
\left\| \hat{\mathcal{T}}_\pi G'_\pi - \hat{\mathcal{T}}_\pi G''_\pi \right\|_{\mu_\pi,\alpha} \leq \frac{1+\gamma}{2} \left\| G'_\pi - G''_\pi \right\|_{\mu_\pi,\alpha},
$$

which completes the proof in the backward GBO evaluation setting.

E.3 Proof of Proposition 2

We first consider the **forward GFV setting**. Recall the linear function approximation of G_π is given by

$$
G(\theta) = \Phi\theta,
$$

where

$$
\Phi = \begin{bmatrix} \Phi_1 \otimes \mathsf{I}_{d_1} & 0 & \cdots & 0 \\ 0 & \Phi_2 \otimes \mathsf{I}_{d_2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \Phi_k \otimes \mathsf{I}_{d_k} \end{bmatrix}, \quad \theta = \begin{bmatrix} \mathrm{vec}(\theta_1^\top) \\ \mathrm{vec}(\theta_2^\top) \\ \vdots \\ \mathrm{vec}(\theta_k^\top) \end{bmatrix}.
$$

Folloing the definition of $g(\theta)$ in eq. (7), we have

$$
\begin{aligned}
-g(\theta) &= \Phi^\top U_\pi(\mathcal{T}_{G,\pi}G(\theta) - G(\theta)) \\
&= \Phi^\top U_\pi((M_\pi - I)\Phi\theta + B) \\
&= G\theta + g,
\end{aligned}
$$

where $G = \Phi^\top U_\pi(M_\pi - I)\Phi$ and $g = \Phi^\top U_\pi B$. Since the monotonicity depends only on the matrix G, we next proceed to show that G is Hurwitz. For the matrix G, we have

$$
G = \Phi^\top U_\pi(M_\pi - I)\Phi = \begin{bmatrix} A_1 & 0 & \cdots & 0 \\ N_{2,1} & A_2 & \cdots & 0 \\ \vdots & \vdots & & \\ N_{k,1} & N_{k,2} & \cdots & A_k \end{bmatrix}, \qquad (30)
$$

where $A_i = [\Phi_i^\top U_{\pi,i}(\gamma_i \mathsf{P}_\pi - I)\Phi_i] \otimes \mathsf{I}_{d_i}$ and $N_{i,j}$ is a matrix that depends on Φ_i, Φ_j, P_π and μ_π. We have the following equations hold:

$$\mathrm{eig}(G) = \{\mathrm{eig}(A_1), \cdots, \mathrm{eig}(A_k)\}, \tag{31}$$

$$\mathrm{eig}(A_i) = \mathrm{eig}(\Phi_i^\top U_{\pi,i}(\gamma_i \mathsf{P}_\pi - I)\Phi_i), \tag{32}$$

$$\max\{\mathrm{eig}(\Phi_i^\top U_{\pi,i}(\gamma_i \mathsf{P}_\pi - I)\Phi_i)\} = -(1-\gamma)\zeta_i, \tag{33}$$

where ζ_i is defined in Proposition 2, the first equation follows because the eigenvalue of a matrix is determined by the eigenvalues of its diagonal block matrices [13], the second equation follows from the fact that $\mathrm{eig}(M \otimes \mathsf{I}_d) = \mathrm{eig}(M)$ for any matrix M and positive integer d, and the last follows from Lemma 1 and Lemma 3 in [2]. Combining eq. (31)–(33), we can obtain equation

$$\max\{\mathrm{eig}(G)\} \le -(1-\gamma)\min_i \zeta_i = -\lambda_G < 0, \tag{34}$$

which completes the proof in the forward GVF setting.

We next consider the **backward GVF setting**. Following the steps similar to those for deriving eq. (30), we can obtain $-g(\theta) = \hat{G}\theta + \hat{g}$, where $\hat{g} = \Phi^\top P_\pi^\top U_\pi B$. For the matrix $\hat{G}$, we have

$$\hat{G} = \Phi^\top P_\pi^\top (\hat{M}_\pi - I) U_\pi \Phi = \begin{bmatrix} \hat{A}_1 & 0 & \cdots & 0 \\ \hat{N}_{2,1} & \hat{A}_2 & \cdots & 0 \\ \vdots & \vdots & & \\ \hat{N}_{k,1} & \hat{N}_{k,2} & \cdots & \hat{A}_k \end{bmatrix}, \tag{35}$$

where $\hat{A}_i = [\Phi_i^\top (\gamma_i \mathsf{P}_\pi^\top - I) U_{\pi,i}\Phi_i] \otimes \mathsf{I}_{d_i}$ and $\hat{N}_{i,j}$ is a matrix that depends on Φ_i, Φ_j, P_π and μ_π. Following the steps similar to those in eq. (31)–(34) and using the result in the verification of item (c) in Assumption 2 in [70], we have

$$\max\{\mathrm{eig}(G)\} \le -(1-\gamma)\min_i \zeta_i = -\lambda_G < 0,$$

which completes the proof in the backward GVF setting.

F Proof of Theorem 1

F.1 Supporting Lemmas

We first develop the property for the update of w_ρ in Algorithm 1. Given a sample $(s_t, a_t, B_t, s_t') \sim \mathcal{D}$ and $a_t' \sim \pi(\cdot|s_t')$, we introduce the following definitions.

$$P_t = \begin{bmatrix} \psi_t^\top \psi_t & (\psi_t - \psi_t')\psi_t^\top & 0 \\ -\psi_t(\psi_t^\top - \psi_t'^\top) & 0 & \psi_t \\ 0 & -\psi_t^\top & 1 \end{bmatrix}, \qquad p_t = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Consider the matrix P_t and vector p_t, we have the following holds

$$\begin{aligned} \|P_t\|_F^2 &= \left\|\psi_t^\top \psi_t\right\|_F^2 + 2\left\|(\psi_t - \psi_t')\psi_t^\top\right\|_F^2 + 2\left\|\psi_t\right\|_F^2 + 1 \\ &\le 9C_\psi^4 + 2C_\psi^2 + 1, \end{aligned} \tag{36}$$

where C_ψ is the upper bound on the feature fector $\psi(\cdot)$, i.e., $\|\psi(s,a)\|_2 \le C_\psi$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, which implies $\|P_t\|_F \le C_P$, where

$$C_P = \sqrt{9C_\psi^4 + 2C_\psi^2 + 1}. \tag{37}$$

For the vector p_t, it can be checked easily that $\|p_t\|_2 \le 1$.

We also define $P = \mathbb{E}_{\mathcal{D}\cdot\pi}[P_t]$ and $p = \mathbb{E}_{\mathcal{D}\cdot\pi}[p_t]$, i.e.,

$$P = \begin{bmatrix} \mathbb{E}_{\mathcal{D}\cdot\pi}[\psi^\top \psi] & \mathbb{E}_{\mathcal{D}\cdot\pi}[(\psi - \psi')\psi^\top] & 0 \\ -\mathbb{E}_{\mathcal{D}\cdot\pi}[\psi(\psi^\top - \psi'^\top)] & 0 & \mathbb{E}_{\mathcal{D}\cdot\pi}[\psi] \\ 0 & -\mathbb{E}_{\mathcal{D}\cdot\pi}[\psi^\top] & 1 \end{bmatrix}, \qquad p = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Note that

$$\left[\begin{array}{c} \nabla_{w_f} L(\hat{\rho}, \hat{f}, \eta) \\ \nabla_{w_\rho} L(\hat{\rho}, \hat{f}, \eta) \\ \nabla_{\eta} L(\hat{\rho}, \hat{f}, \eta) \end{array}\right] = -(P\kappa + p).$$

It has been shown in Theorem 2 in [67] that the real parts of all eigenvalues of P are strictly positive, which guarantees that there exists a positive constant λ_P such that

$$\langle Px, x\rangle \geq \lambda_P \|x\|_2^2 \quad \text{for all} \quad x \in \mathbb{R}^{2d_\rho}. \tag{38}$$

We also define $\kappa_t = [w_{f,t}^\top, w_{\rho,t}^\top, \eta_t]^\top$. The update of density ratio learning can be rewritten as

$$\kappa_{t+1} = \kappa_t - \beta_t \zeta(x_t, \kappa_t),$$

where $\zeta(x_t, \kappa_t) = P_t\kappa_t + p_t$. We also define the population update as $\zeta(\kappa_t) = \mathbb{E}_{D\cdot\pi}[\zeta(x, \kappa_t)] = P\kappa_t + p$. Without loss of generality, we assume that there exists a positive constant C_κ such that $\|\kappa^*\|_2 \leq C_\kappa$, where κ^* is the global optimum of the density ratio learning defined as

$$\langle \zeta(\kappa^*), \kappa - \kappa^*\rangle \leq 0, \quad \forall \kappa \in \mathbb{R}^{d_\rho} \times R_\rho \times \mathbb{R}. \tag{39}$$

The following lemma, often referred to as the "three-points" lemma, characterizes the incremental updating progress of κ_t with projection, a proof of which can be found in Lemma 3.1 in [20].

Lemma 3. *Consider the update of $w_{f,t}$, $w_{\rho,t}$ and η_t in Algorithm 1. For all $\kappa \in \mathbb{R}^M \times R_\rho \times \mathbb{R}$, we have the following holds*

$$\beta_t \langle \zeta(x_t, \kappa_t), \kappa_{t+1} - \kappa\rangle + \frac{1}{2}\|\kappa_{t+1} - \kappa_t\|_2^2 \leq \frac{1}{2}\|\kappa_t - \kappa\|_2^2 - \frac{1}{2}\|\kappa_{t+1} - \kappa\|_2^2. \tag{40}$$

Similarly to Lemma 3, we also have the following "three-points lemma" for the iteration of θ_t.

Lemma 4. *Consider the update of θ_t in Algorithm 1. For all $\theta \in R_\theta$, we have the following holds*

$$-\alpha_t \langle \hat{\rho}(x_t, w_{\rho,t}) g(x_t, \theta_t), \theta_{t+1} - \theta\rangle + \frac{1}{2}\|\theta_{t+1} - \theta_t\|_2^2 \leq \frac{1}{2}\|\theta_t - \theta\|_2^2 - \frac{1}{2}\|\theta_{t+1} - \theta\|_2^2, \tag{41}$$

where $\hat{\rho}(x_t, w_{\rho,t})$ is defined in eq. (8)*.*

The following lemma characterizes the smoothness of $\zeta(\cdot)$.

Lemma 5. *For any $\kappa, \kappa' \in \mathbb{R}^{d_\rho} \times R_\rho \times \mathbb{R}$, we have*

$$\|\zeta(\kappa) - \zeta(\kappa')\|_2 \leq C_P \|\kappa - \kappa'\|_2,$$

where C_P is defined in eq. (37)*.*

Proof. Recalling the definition of $\zeta(\kappa) = P\kappa + p$, we can obtain the following

$$\|\zeta(\kappa) - \zeta(\kappa')\|_2 = \|P(\kappa - \kappa')\|_2 \leq \|P\|_2 \|\kappa - \kappa'\|_2 \leq C_P \|\kappa - \kappa'\|_2,$$

which completes the proof. ☐

Similarly, the following lemma characterizes the smoothness of $g(\theta) = \mathbb{E}_{\mu_\pi}[g(x, \theta)]$.

Lemma 6. *In both the forward and backward GVF evaluation settings, for any $\theta, \theta' \in \mathbb{R}^{d_\rho}$, we have*

$$\|g(\theta) - g(\theta')\|_2 \leq C_g \|\theta - \theta'\|_F,$$

where $C_g = (d_g C_\phi C_m + 1)C_\phi$.

Proof. First consider the forward GVF evaluation setting. Recall the definition of $g(\theta)$ and $x = (s, a, s', a')$, we have

$$g(\theta) = \mathbb{E}_{\mu_\pi}[\phi(s,a)^\top (B(x) + m(x)\phi(s', a')\theta - \phi(s,a)\theta)],$$

which implies

$$\begin{aligned} \|g(\theta) - g(\theta')\|_2 &= \|\mathbb{E}_{\mu_\pi}[\phi(s,a)(m(x)\phi(s',a')(\theta - \theta') + \phi(s,a)(\theta' - \theta))]\|_2 \\ &\leq \mathbb{E}_{\mu_\pi}[(\|\phi(s,a)\|_F \|m(x)\|_F + 1)\|\theta' - \theta\|_2 \|\phi(s,a)\|_F] \\ &\leq C_g \|\theta - \theta'\|_2. \end{aligned} \tag{42}$$

Following the steps similar to those in eq. (42), we can also prove that $\|g(\theta) - g(\theta')\|_2 \leq C_g \|\theta - \theta'\|_F$ holds in the backward GVF evaluation setting. ☐

The following lemma characterizes the monotonicity of $\zeta(\cdot)$.

Lemma 7. *We have the following holds*

$$\langle \zeta(\kappa), \kappa - \kappa^* \rangle \geq \lambda_P \left\| \kappa - \kappa^* \right\|_2^2, \quad \forall \kappa \in \mathbb{R}^{d_\rho} \times R_\rho \times \mathbb{R}.$$

Proof. Recall that P is strictly positive defined (eq. (38)). We have

$$\begin{aligned} \langle \zeta(\kappa), \kappa - \kappa^* \rangle &= \langle \zeta(\kappa^*), \kappa - \kappa^* \rangle + \langle \zeta(\kappa) - \zeta(\kappa^*), \kappa - \kappa^* \rangle \\ &\geq \langle \zeta(\kappa) - \zeta(\kappa^*), \kappa - \kappa^* \rangle \\ &= \langle P(\kappa - \kappa^*), \kappa - \kappa^* \rangle \\ &\geq \lambda_P \left\| \kappa - \kappa^* \right\|_2^2, \end{aligned} \tag{43}$$

which completes the proof. □

The next lemma bounds the per-iteration variance of the update of κ_t.

Lemma 8. *Given a sample $(s_t, a_t, B_t, s_t') \sim \mathcal{D}_d$ and $a_t' \sim \pi(\cdot|s_t')$ and any $\kappa \in \mathbb{R}^{d_\rho} \times R_\rho \times \mathbb{R}$, we have the following holds*

$$\left\| \zeta(x_t, \kappa) - \zeta(\kappa) \right\|_2^2 \leq 8C_P^2 \left\| \kappa - \kappa^* \right\|_2^2 + 8C_P^2 C_\kappa^2.$$

Proof. Recalling the definitions of $\zeta(x_t, \kappa) = P_t \kappa + p_t$ and $\zeta(\kappa) = P\kappa + p$, we can obtain the following

$$\begin{aligned} \left\| \zeta(x_t, \kappa) - \zeta(\kappa) \right\|_2^2 &= \left\| (P_t - P)\kappa \right\|_2^2 = 2 \left\| (P_t - P)(\kappa - \kappa^*) \right\|_2^2 + 2 \left\| (P_t - P)\kappa^* \right\|_2^2 \\ &\leq 8C_P^2 \left\| \kappa - \kappa^* \right\|_2^2 + 8C_P^2 C_\kappa^2. \end{aligned}$$

□

The following lemma bounds the norm of the stochastic update $g(x, \theta)$ and the per-iteration variance of GenTD update with density ratio $\rho(s, a)$.

Lemma 9. *Given a sample $(s_t, a_t, B_t, s_t') \sim \mathcal{D}$ and $a_t' \sim \pi(\cdot|s_t')$ and any $\theta \in R_\theta$, we have the following holds*

$$\left\| g(x_t, \theta) \right\|_2 \leq D_g, \quad \textit{and} \quad \mathbb{E}[\left\| \rho(s_t, a_t) g(x_t, \theta) - g(\theta) \right\|_2^2] \leq V_g,$$

where $D_g = d_g C_\phi [C_{\max} + (C_m + 1) D_\theta C_\phi]$ and $V_g = 2\rho_{\max} D_g$.

Proof. We prove the first result as follows,

$$\begin{aligned} \left\| g(x_t, \theta) \right\|_2 &= \left\| \phi(s,a)^\top (B(x) + m(s', a')\phi(s', a')\theta - \phi(s,a)\theta) \right\|_2 \\ &\leq \left\| \phi(s,a)^\top B(x) \right\|_F + \left\| \phi(s,a) \right\|_F \left\| m(s', a')\phi(s', a')\theta - \phi(s,a)\theta \right\|_F \\ &\leq d_g C_\phi [C_{\max} + (C_m + 1) D_\theta C_\phi], \end{aligned}$$

where the last inequality follows from the boundness of the set R_θ. Here we consider $\left\| \theta \right\|_2 \leq D_\theta$ for all $\theta \in R_\theta$. The second result can be obtained as follows

$$\left\| \rho(s_t, a_t) g(x_t, \theta) - g(\theta) \right\|_2^2 \leq |\rho(s_t, a_t)| \left(\left\| g(x_t, \theta) \right\|_2 + \left\| g(\theta) \right\|_2 \right) \leq 2\rho_{\max} D_g. \tag{44}$$

□

We next bound the convergence rate of $w_{\rho,t}$.

Lemma 10. *Consider $w_{f,t}$, $w_{\rho,t}$ and η_t in Algorithm 1. Let stepsize $\beta_t = \frac{2}{\lambda_P (t + t_0 + 1)}$ where $t_0 = \frac{36 C_P^2}{\lambda_P^2}$. For any $\kappa \in \mathbb{R}^{d_\rho} \times R_\rho \times \mathbb{R}$, we have*

$$\mathbb{E}[\left\| \kappa_T - \kappa^* \right\|_2^2] \leq \frac{(1 + 16\beta_0^2 C_P^2)(t_0 + 1)^2 \left\| \kappa_0 - \kappa^* \right\|_2^2}{(T + t_0 - 1)(T + t_0)} + \frac{64 C_P^2 C_\kappa^2}{(T + t_0)\lambda_P^2}.$$

Proof. The inner product in eq. (40) can be equivalently written as

$$
\begin{aligned}
&\langle \zeta(x_t,\kappa_t), \kappa_{t+1}-\kappa\rangle \\
&\quad= \langle \zeta(\kappa_{t+1}), \kappa_{t+1}-\kappa\rangle + \langle \zeta(\kappa_t)-\zeta(\kappa_{t+1}), \kappa_{t+1}-\kappa\rangle + \langle \zeta(x_t,\kappa_t)-\zeta(\kappa_t), \kappa_t-\kappa\rangle \\
&\qquad+ \langle \zeta(x_t,\kappa_t)-\zeta(\kappa_t), \kappa_{t+1}-\kappa_t\rangle \\
&\quad\geq \langle \zeta(\kappa_{t+1}), \kappa_{t+1}-\kappa\rangle - \left\|\zeta(\kappa_t)-\zeta(\kappa_{t+1})\right\|_2 \left\|\kappa_{t+1}-\kappa\right\|_2 + \langle \zeta(x_t,\kappa_t)-\zeta(\kappa_t), \kappa_t-\kappa\rangle \\
&\qquad- \left\|\zeta(x_t,\kappa_t)-\zeta(\kappa_t)\right\|_2 \left\|\kappa_{t+1}-\kappa_t\right\|_2 \\
&\quad\geq \langle \zeta(\kappa_{t+1}), \kappa_{t+1}-\kappa\rangle - C_P \left\|\kappa_t-\kappa_{t+1}\right\|_2 \left\|\kappa_{t+1}-\kappa\right\|_2 + \langle \zeta(x_t,\kappa_t)-\zeta(\kappa_t), \kappa_t-\kappa\rangle \\
&\qquad- \left\|\zeta(x_t,\kappa_t)-\zeta(\kappa_t)\right\|_2 \left\|\kappa_{t+1}-\kappa_t\right\|_2,
\end{aligned}\tag{45}
$$

where the last inequality follows from Lemma 5. Substituting eq. (45) into eq. (40), we obtain

$$
\begin{aligned}
&\frac{1}{2}\left\|\kappa_t-\kappa\right\|_2^2 - \frac{1}{2}\left\|\kappa_{t+1}-\kappa\right\|_2^2 \\
&\quad\geq \beta_t\langle \zeta(\kappa_{t+1}), \kappa_{t+1}-\kappa\rangle - \beta_t C_P \left\|\kappa_t-\kappa_{t+1}\right\|_2 \left\|\kappa_{t+1}-\kappa\right\|_2 + \beta_t\langle \zeta(x_t,\kappa_t)-\zeta(\kappa_t), \kappa_t-\kappa\rangle \\
&\qquad- \beta_t \left\|\zeta(x_t,\kappa_t)-\zeta(\kappa_t)\right\|_2 \left\|\kappa_{t+1}-\kappa_t\right\|_2 + \frac{1}{2}\left\|\kappa_{t+1}-\kappa_t\right\|_2^2.
\end{aligned}\tag{46}
$$

Note that we have the following holds

$$
\begin{aligned}
&\frac{1}{2}\left\|\kappa_{t+1}-\kappa_t\right\|_2^2 - \beta_t C_P \left\|\kappa_t-\kappa_{t+1}\right\|_2 \left\|\kappa_{t+1}-\kappa\right\|_2 - \beta_t \left\|\zeta(x_t,\kappa_t)-\zeta(\kappa_t)\right\|_2 \left\|\kappa_{t+1}-\kappa_t\right\|_2 \\
&= \frac{1}{4}\left\|\kappa_{t+1}-\kappa_t\right\|_2^2 - \beta_t C_P \left\|\kappa_t-\kappa_{t+1}\right\|_2 \left\|\kappa_{t+1}-\kappa\right\|_2 + \frac{1}{4}\left\|\kappa_{t+1}-\kappa_t\right\|_2^2 \\
&\quad- \beta_t \left\|\zeta(x_t,\kappa_t)-\zeta(\kappa_t)\right\|_2 \left\|\kappa_{t+1}-\kappa_t\right\|_2 \\
&\geq -\beta_t^2 C_P^2 \left\|\kappa_{t+1}-\kappa\right\|_2^2 - \beta_t^2 \left\|\zeta(x_t,\kappa_t)-\zeta(\kappa_t)\right\|_2^2.
\end{aligned}\tag{47}
$$

Substituting eq. (47) in eq. (46) yields

$$
\begin{aligned}
&\frac{1}{2}\left\|\kappa_t-\kappa\right\|_2^2 - \frac{1}{2}\left\|\kappa_{t+1}-\kappa\right\|_2^2 \\
&\quad\geq \beta_t\langle \zeta(\kappa_{t+1}), \kappa_{t+1}-\kappa\rangle + \beta_t\langle \zeta(x_t,\kappa_t)-\zeta(\kappa_t), \kappa_t-\kappa\rangle - \beta_t^2 C_P^2 \left\|\kappa_{t+1}-\kappa\right\|_2^2 \\
&\qquad- \beta_t^2 \left\|\zeta(x_t,\kappa_t)-\zeta(\kappa_t)\right\|_2^2.
\end{aligned}\tag{48}
$$

Rearranging eq. (48) and letting $\kappa=\kappa^*$ yield

$$
\begin{aligned}
&\left\|\kappa_t-\kappa^*\right\|_2^2 + 2\beta_t^2 \left\|\zeta(x_t,\kappa_t)-\zeta(\kappa_t)\right\|_2^2 \\
&\geq (1-2\beta_t^2 C_P^2)\left\|\kappa_{t+1}-\kappa^*\right\|_2^2 + 2\beta_t\langle \zeta(\kappa_{t+1}), \kappa_{t+1}-\kappa^*\rangle + 2\beta_t\langle \zeta(x_t,\kappa_t)-\zeta(\kappa_t), \kappa_t-\kappa^*\rangle \\
&\geq (1+2\beta_t\lambda_P-2\beta_t^2 C_P^2)\left\|\kappa_{t+1}-\kappa^*\right\|_2^2 + 2\beta_t\langle \zeta(x_t,\kappa_t)-\zeta(\kappa_t), \kappa_t-\kappa^*\rangle,
\end{aligned}\tag{49}
$$

where the last inequality follows from $\langle \zeta(\kappa_{t+1}), \kappa_{t+1}-\kappa^*\rangle \leq \lambda_P \left\|\kappa_{t+1}-\kappa^*\right\|_2^2$. Taking expectation on both sides of eq. (49), and noting that $\mathbb{E}[\langle \zeta(x_t,\kappa_t)-\zeta(\kappa_t), \kappa_t-\kappa^*\rangle|\mathcal{F}_t]=0$, we obtain

$$
\begin{aligned}
&(1+2\beta_t\lambda_P-2\beta_t^2 C_P^2)\mathbb{E}[\left\|\kappa_{t+1}-\kappa^*\right\|_2^2] \\
&\quad\leq \mathbb{E}[\left\|\kappa_t-\kappa^*\right\|_2^2] + 2\beta_t^2\mathbb{E}[\left\|\zeta(x_t,\kappa_t)-\zeta(\kappa_t)\right\|_2^2] \\
&\quad\leq (1+16\beta_t^2 C_P^2)\mathbb{E}[\left\|\kappa_t-\kappa^*\right\|_2^2] + 16\beta_t^2 C_P^2 C_\kappa^2.
\end{aligned}\tag{50}
$$

Multiplying both sides of eq. (50) with I_t and summing over $t=0,\cdots,T-1$ yield

$$
\sum_{t=0}^{T-1} a_t\mathbb{E}[\left\|\kappa_{t+1}-\kappa^*\right\|_2^2] \leq \sum_{t=0}^{T-1} b_t\mathbb{E}[\left\|\kappa_t-\kappa^*\right\|_2^2] + c,\tag{51}
$$

where

$$
\begin{aligned}
a_t &= (1+2\beta_t\lambda_P-2\beta_t^2 C_P^2)I_t, \\
b_t &= (1+16\beta_t^2 C_P^2)I_t, \\
c &= 16 C_P^2 C_\kappa^2 \sum_{t=0}^{T-1}\beta_t^2 I_t.
\end{aligned}
$$

We further let

$$
\begin{aligned}
I_t &= (t+t_0)(t+t_0+1),\\
\beta_t &= \frac{2}{\lambda_P(t+t_0-1)},\\
t_0 &= \frac{36C_P^2}{\lambda_P^2}+1.
\end{aligned}
$$

We can obtain the following

$$
\begin{aligned}
a_t - b_{t+1} &= (1+2\beta_t\lambda_P - 2\beta_t^2 C_P^2)I_t - (1+16\beta_{t+1}^2 C_P^2)s_{t+1}\\
&\geq (1+2\beta_t\lambda_P)I_t - (1+2\beta_t^2+16\beta_{t+1}^2 C_P^2)s_{t+1}\\
&\geq (1+2\beta_t\lambda_P)I_t - (1+18\beta_t^2 C_P^2)s_{t+1}\\
&\geq (1+2\beta_t\lambda_P)I_t - (1+\beta_t\lambda_P)s_{t+1}\\
&\geq (t+t_0+1)\frac{(t+t_0+3)(t+t_0)-(t+t_0+2)^2}{t+t_0-1}\\
&\geq 0.
\end{aligned}
$$

Substituting the above results into eq. (51) yields

$$
a_{T-1}t\mathbb{E}[\|\kappa_T-\kappa^*\|_2^2] \leq b_0\|\kappa_0-\kappa^*\|_2^2 + c,
$$

which implies

$$
\begin{aligned}
\mathbb{E}[\|\kappa_T-\kappa^*\|_2^2] &\leq \frac{b_0\|\kappa_0-\kappa^*\|_2^2}{A_{T-1}} + \frac{c}{a_{T-1}}\\
&= \frac{(1+16\beta_0^2C_P^2)s_0\|\kappa_0-\kappa^*\|_2^2}{(1+2\beta_{T-1}\lambda_P-2\beta_{T-1}^2C_P^2)s_{T-1}} + \frac{16C_P^2C_\kappa^2\sum_{t=0}^{T-1}\beta_t^2 I_t}{(1+2\beta_{T-1}\lambda_P-2\beta_{T-1}^2C_P^2)s_{T-1}}\\
&\leq \frac{(1+16\beta_0^2C_P^2)(t_0+1)^2\|\kappa_0-\kappa^*\|_2^2}{(T+t_0-1)(T+t_0)} + \frac{64C_P^2C_\kappa^2}{(T+t_0)\lambda_P^2},
\end{aligned}
$$

which completes the proof. □

Note that Lemma 10 implies that there exists a positive number D_ρ such that

$$
\mathbb{E}[\|w_{\rho,t}-w_\rho^*\|_2^2] \leq \frac{D_\rho}{t+t_0}. \tag{52}
$$

F.2 Proof of Theorem 1

Consider the inner product term in eq. (41). We have

$$
\begin{aligned}
&-\langle\hat{\rho}(x_t,w_{\rho,t})g(x_t,\theta_t),\theta_{t+1}-\theta)\rangle\\
&\quad= -\langle g(\theta_{t+1}),\theta_{t+1}-\theta)\rangle - \langle g(\theta_t - g(\theta_{t+1}),\theta_{t+1}-\theta)\rangle\\
&\quad\quad - \langle\rho(x_t)g(x_t,\theta_t)-g(\theta_t),\theta_{t+1}-\theta_t\rangle - \langle\rho(x_t)g(x_t,\theta_t)-g(\theta_t),\theta_t-\theta\rangle\\
&\quad\quad - \langle(\hat{\rho}(x_t,w_\rho^*)-\rho(x_t))g(x_t,\theta_t),\theta_{t+1}-\theta\rangle\\
&\quad\quad - \langle(\hat{\rho}(x_t,w_{\rho,t})-\hat{\rho}(x_t,w_\rho^*))g(x_t,\theta_t),\theta_{t+1}^\top-\theta\rangle\\
&\quad\geq -\langle g(\theta_{t+1}),\theta_{t+1}^\top-\theta\rangle - C_g\|\theta_t-\theta_{t+1}\|_2\|\theta_{t+1}-\theta\|_2\\
&\quad\quad - \|\rho(x_t)g(x_t,\theta_t)-g(\theta_t)\|_2\|\theta_{t+1}-\theta_t\|_2 - \langle\rho(x_t)g(x_t,\theta_t)-g(\theta_t),\theta_t^\top-\theta\rangle\\
&\quad\quad - \left|\hat{\rho}(x_t,w_\rho^*)-\rho(x_t)\right|\|g(x_t,\theta_t)\|_2\|\theta_{t+1}-\theta\|_2\\
&\quad\quad - \left|\hat{\rho}(x_t,w_{\rho,t})-\hat{\rho}(x_t,w_\rho^*)\right|\|g(x_t,\theta_t)\|_2\|\theta_{t+1}-\theta\|_2,
\end{aligned} \tag{53}
$$

where the last inequality follows from Lemma 6. Substituting eq. (53) into eq. (41) yields

$$
\frac{1}{2}\|\theta_t-\theta\|_2^2 - \frac{1}{2}\|\theta_{t+1}-\theta\|_2^2
$$

$$
\begin{aligned}
&\geq -\alpha_t\langle g(\theta_{t+1}), \theta_{t+1}^\top - \theta\rangle - \alpha_t C_g \left\|\theta_t - \theta_{t+1}\right\|_2 \left\|\theta_{t+1} - \theta\right\|_2 \\
&\quad - \alpha_t \left\|\rho(x_t)g(x_t,\theta_t) - g(\theta_t)\right\|_2 \left\|\theta_{t+1} - \theta_t\right\|_2 - \alpha_t\langle \rho(x_t)g(x_t,\theta_t) - g(\theta_t), \theta_t^\top - \theta^\top\rangle \\
&\quad - \alpha_t \left|\hat{\rho}(x_t, w_\rho^*) - \rho(x_t)\right| \left\|g(x_t,\theta_t)\right\|_2 \left\|\theta_{t+1} - \theta\right\|_2 \\
&\quad - \alpha_t \left|\hat{\rho}(x_t, w_{\rho,t}) - \hat{\rho}(x_t, w_\rho^*)\right| \left\|g(x_t,\theta_t)\right\|_2 \left\|\theta_{t+1} - \theta\right\|_2 + \frac{1}{2}\left\|\theta_{t+1} - \theta_t\right\|_2^2. \qquad (54)
\end{aligned}
$$

We have the following holds

$$
\begin{aligned}
&\frac{1}{2}\left\|\theta_{t+1} - \theta_t\right\|_2^2 - \alpha_t C_g \left\|\theta_t - \theta_{t+1}\right\|_2 \left\|\theta_{t+1} - \theta\right\|_2 - \alpha_t \left\|\rho(x_t)g(x_t,\theta_t) - g(\theta_t)\right\|_2 \left\|\theta_{t+1} - \theta_t\right\|_2 \\
&\quad = \frac{1}{4}\left\|\theta_{t+1} - \theta_t\right\|_2^2 - \alpha_t C_g \left\|\theta_t - \theta_{t+1}\right\|_2 \left\|\theta_{t+1} - \theta\right\|_2 + \frac{1}{4}\left\|\theta_{t+1} - \theta_t\right\|_2^2 \\
&\quad\quad - \alpha_t \left\|\rho(x_t)g(x_t,\theta_t) - g(\theta_t)\right\|_2 \left\|\theta_{t+1} - \theta_t\right\|_2 \\
&\quad \geq -\alpha_t^2 C_g^2 \left\|\theta_{t+1} - \theta\right\|_2^2 - \alpha_t^2 \left\|\rho(x_t)g(x_t,\theta_t) - g(\theta_t)\right\|_2^2, \qquad (55)
\end{aligned}
$$

which implies

$$
\begin{aligned}
&\frac{1}{2}\left\|\theta_t - \theta\right\|_2^2 - \frac{1}{2}\left\|\theta_{t+1} - \theta\right\|_2^2 \\
&\quad \geq -\alpha_t\langle g(\theta_{t+1}), \theta_{t+1} - \theta\rangle - \alpha_t^2 C_g^2 \left\|\theta_{t+1} - \theta\right\|_2^2 - \alpha_t^2 \left\|\rho(x_t)g(x_t,\theta_t) - g(\theta_t)\right\|_2^2 \\
&\quad\quad - \alpha_t\langle \rho(x_t)g(x_t,\theta_t) - g(\theta_t), \theta_t - \theta\rangle - \alpha_t D_g \left|\hat{\rho}(x_t, w_\rho^*) - \rho(x_t)\right| \left\|\theta_{t+1} - \theta\right\|_2 \\
&\quad\quad - \alpha_t D_g \left|\hat{\rho}(x_t, w_{\rho,t}) - \hat{\rho}(x_t, w_\rho^*)\right| \left\|\theta_{t+1} - \theta\right\|_2, \qquad (56)
\end{aligned}
$$

where we use the fact that $\left\|g(x_t,\theta_t)\right\|_2 \leq D_g$ in Lemma 9. Rearranging eq. (56) and letting $\theta = \theta^*$ yield

$$
\begin{aligned}
&\left\|\theta_t - \theta^*\right\|_2^2 + 2\alpha_t^2 \left\|\rho(x_t)g(x_t,\theta_t) - g(\theta_t)\right\|_2^2 \\
&\quad \geq \left\|\theta_{t+1} - \theta^*\right\|_2^2 - 2\alpha_t\langle g(\theta_{t+1}), \theta_{t+1} - \theta^*\rangle - 2\alpha_t^2 C_g^2 \left\|\theta_{t+1} - \theta^*\right\|_2^2 \\
&\quad\quad - 2\alpha_t\langle \rho(x_t)g(x_t,\theta_t) - g(\theta_t), \theta_t - \theta^*\rangle - 2\alpha_t D_g \left|\hat{\rho}(x_t, w_\rho^*) - \rho(x_t)\right| \left\|\theta_{t+1} - \theta^*\right\|_2 \\
&\quad\quad - 2\alpha_t D_g \left|\hat{\rho}(x_t, w_{\rho,t}) - \hat{\rho}(x_t, w_\rho^*)\right| \left\|\theta_{t+1} - \theta\right\|_2 \\
&\quad \geq (1 + 2\alpha_t\lambda_g - 2\alpha_t^2 C_g^2) \left\|\theta_{t+1} - \theta^*\right\|_2^2 - 2\alpha_t\langle \rho(x_t)g(x_t,\theta_t) - g(\theta_t), \theta_t - \theta^*\rangle \\
&\quad\quad - 2\alpha_t D_g \left|\hat{\rho}(x_t, w_\rho^*) - \rho(x_t)\right| \left\|\theta_{t+1} - \theta^*\right\|_2 \\
&\quad\quad - 2\alpha_t D_g \left|\hat{\rho}(x_t, w_{\rho,t}) - \hat{\rho}(x_t, w_\rho^*)\right| \left\|\theta_{t+1} - \theta\right\|_2 \\
&\quad \geq (1 + 2\alpha_t\lambda_g - 2\alpha_t^2 C_g^2) \left\|\theta_{t+1} - \theta^*\right\|_2^2 - 2\alpha_t\langle \rho(x_t)g(x_t,\theta_t) - g(\theta_t), \theta_t - \theta^*\rangle \\
&\quad\quad - \frac{1}{2}\alpha_t\lambda_g \left\|\theta_{t+1} - \theta^*\right\|_2^2 - \frac{2\alpha_t D_g^2}{\lambda_g}\left|\hat{\rho}(x_t, w_\rho^*) - \rho(x_t)\right|^2 \\
&\quad\quad - \frac{1}{2}\alpha_t\lambda_g \left\|\theta_{t+1} - \theta^*\right\|_2^2 - \frac{2\alpha_t D_g^2}{\lambda_g}\left|\hat{\rho}(x_t, w_{\rho,t}) - \hat{\rho}(x_t, w_\rho^*)\right|^2 \\
&\quad = (1 + \alpha_t\lambda_g - 2\alpha_t^2 C_g^2) \left\|\theta_{t+1} - \theta^*\right\|_2^2 - 2\alpha_t\langle \rho(x_t)g(x_t,\theta_t) - g(\theta_t), \theta_t - \theta^*\rangle \\
&\quad\quad - \frac{2\alpha_t D_g^2}{\lambda_g}\left|\hat{\rho}(x_t, w_\rho^*) - \rho(x_t)\right|^2 - \frac{2\alpha_t D_g^2}{\lambda_g}\left|\hat{\rho}(x_t, w_{\rho,t}) - \hat{\rho}(x_t, w_\rho^*)\right|^2, \qquad (57)
\end{aligned}
$$

where the first inequality follows from Lemma 4, and the third inequality follows from Young's inequality. Taking expectation on both sides of eq. (57) yields

$$
\begin{aligned}
&(1 + \alpha_t\lambda_g - 2\alpha_t^2 C_g^2)\mathbb{E}[\left\|\theta_{t+1} - \theta^*\right\|_2^2] \\
&\quad \leq \mathbb{E}[\left\|\theta_t - \theta^*\right\|_2^2] + 2\alpha_t^2\mathbb{E}[\left\|\rho(x_t)g(x_t,\theta_t) - g(\theta_t)\right\|_2^2] + \frac{2\alpha_t D_g^2}{\lambda_g}\mathbb{E}\left[\left|\hat{\rho}(x_t, w_\rho^*) - \rho(x_t)\right|^2\right] \\
&\quad\quad + \frac{2\alpha_t D_g^2 C_\psi^2}{\lambda_g}\mathbb{E}\left[\left|\hat{\rho}(x_t, w_{\rho,t}) - \hat{\rho}(x_t, w_\rho^*)\right|^2\right]
\end{aligned}
$$

$$\leq \mathbb{E}[\|\theta_t - \theta^*\|_2^2] + 2V_g\alpha_t^2 + \frac{2\alpha_t D_g^2 C_\psi^2}{\lambda_g}\mathbb{E}\left[\left\|w_{\rho,t} - w_\rho^*\right\|_2^2\right] + \frac{2D_g^2\alpha_t\varepsilon_\rho}{\lambda_g}, \tag{58}$$

where the last inequality follows from Lemma 9.

Substituting eq. (52) into eq. (58) yields

$$\begin{aligned}(1 + \alpha_t\lambda_g - &2\alpha_t^2 C_g^2)\mathbb{E}[\|\theta_{t+1} - \theta^*\|_2^2] \\ &\leq \mathbb{E}[\|\theta_t - \theta^*\|_2^2] + 2V_g\alpha_t^2 + \frac{2\alpha_t D_g^2 D_\rho C_\psi^2}{\lambda_g(t+t_0)} + \frac{2D_g^2\alpha_t\varepsilon_\rho}{\lambda_g}. \end{aligned} \tag{59}$$

Multiplying both sides of eq. (59) with r_t and summing over $t = 0, \cdots, T-1$ yield

$$\begin{aligned}&\sum_{t=0}^{T-1} a_t' \mathbb{E}[\|\theta_{t+1} - \theta^*\|_2^2] \\ &\leq \sum_{t=0}^{T-1} r_t \mathbb{E}[\|\theta_t - \theta^*\|_2^2] + 2V_g\sum_{t=0}^{T-1} r_t\alpha_t^2 + \frac{2D_g^2 D_\rho C_\psi^2}{\lambda_g(t+t_0)}\sum_{t=0}^{T-1} r_t\alpha_t + \frac{2D_g^2\varepsilon_\rho}{\lambda_g}\sum_{t=0}^{T-1} r_t\alpha_t, \end{aligned} \tag{60}$$

where

$$a_t' = (1 + \alpha_t\lambda_g - 2\alpha_t^2 C_g^2) r_t.$$

Now we let

$$\begin{aligned} r_t &= (t+t_1)(t+t_1+1), \\ \alpha_t &= \frac{4}{\lambda_g(t+t_1-1)}, \\ t_1 &= \frac{16C_g^2}{\lambda_g^2} + 1. \end{aligned}$$

We can obtain the following

$$\begin{aligned} a_t' - r_{t+1} &= (1 + \alpha_t\lambda_g - 2\alpha_t^2 C_g^2) r_t - r_{t+1} \\ &\geq (1 + \alpha_t\lambda_g) r_t - (1 + 2\alpha_t^2 C_g^2) r_{t+1} \\ &\geq (1 + \alpha_t\lambda_g) r_t - \left(1 + \frac{1}{2}\alpha_t\lambda_g\right) r_{t+1} \\ &\geq (t+t_1+1)\frac{(t+t_1)(t+t_1+3) - (t+t_1+2)^2}{t+t_1+1} \\ &\geq 0, \end{aligned}$$

where the second inequality follows from the fact that $\alpha_t \leq \frac{\lambda_g}{4C_g^2}$.

Substituting the above result to eq. (60) yields

$$a_{T-1}'\mathbb{E}[\|\theta_T - \theta^*\|_2^2] \leq r_0\|\theta_0 - \theta^*\|_2^2 + 2V_g\sum_{t=0}^{T-1} r_t\alpha_t^2 + \frac{2D_g^2 D_\rho C_\psi^2}{\lambda_g(t+t_0)}\sum_{t=0}^{T-1} r_t\alpha_t + \frac{2D_g^2\varepsilon_\rho}{\lambda_g}\sum_{t=0}^{T-1} r_t\alpha_t.$$

The above inequality implies the following convergence rate

$$\mathbb{E}[\|\theta_T - \theta^*\|_2^2] \leq \frac{r_0\|\theta_0 - \theta^*\|_2^2}{(T+t_1-1))(T+t_1)} + \frac{128V_g}{\lambda_g^2(T+t_1)} + \frac{64D_g^2 C_g^2 C_\psi^2\lambda_P^2}{9C_P^2\lambda_g^3(T+t_1)} + \frac{16D_g^2}{\lambda_g^2}\varepsilon_\rho,$$

which completes the proof.

G Proof of Theorem 2

Following the similar argument similar to that in Lemma 4.2 in [3] and Theorem 1 in [55], we can prove that $\Phi\theta^*$ is the fixed point of the composite operator $\Gamma_{\Phi,\mu_\pi}\bar{\mathcal{T}}_\pi$. We then proceed as follows

$$\|\Phi\theta^* - G_\pi\|_{\mu_\pi,\alpha} = \left\|\Gamma_{\Phi,\mu_\pi}\bar{\mathcal{T}}_\pi\Phi\theta^* - \Gamma_{\Phi,\mu_\pi}G_\pi + \Gamma_{\Phi,\mu_\pi}G_\pi - G_\pi\right\|_{\mu_\pi,\alpha}$$

$$
\begin{aligned}
&\leq \left\|\Gamma_{\Phi,\mu_\pi}\bar{\mathcal{T}}_\pi\Phi\theta^* - \Gamma_{\Phi,\mu_\pi}G_\pi\right\|_{\mu_\pi,\alpha} + \left\|\Gamma_{\Phi,\mu_\pi}G_\pi - G_\pi\right\|_{\mu_\pi,\alpha} \\
&= \left\|\Gamma_{\Phi,\mu_\pi}\bar{\mathcal{T}}_\pi\Phi\theta^* - \Gamma_{\Phi,\mu_\pi}\bar{\mathcal{T}}_\pi G_\pi\right\|_{\mu_\pi,\alpha} + \left\|\Gamma_{\Phi,\mu_\pi}G_\pi - G_\pi\right\|_{\mu_\pi,\alpha} \\
&\leq \left\|\Gamma_{\Phi,\mu_\pi}[\bar{\mathcal{T}}_\pi\Phi\theta^* - \bar{\mathcal{T}}_\pi G_\pi]\right\|_{\mu_\pi,\alpha} + \left\|\Gamma_{\Phi,\mu_\pi}G_\pi - G_\pi\right\|_{\mu_\pi,\alpha} \\
&\leq \left\|\bar{\mathcal{T}}_\pi\Phi\theta^* - \bar{\mathcal{T}}_\pi G_\pi\right\|_{\mu_\pi,\alpha} + \left\|\Gamma_{\Phi,\mu_\pi}G_\pi - G_\pi\right\|_{\mu_\pi,\alpha} \\
&\leq \gamma_G \left\|\Phi\theta^* - G_\pi\right\|_{\mu_\pi,\alpha} + \left\|\Gamma_{\Phi,\mu_\pi}G_\pi - G_\pi\right\|_{\mu_\pi,\alpha}, \qquad (61)
\end{aligned}
$$

where the first equality follows from the fact that $\Gamma_{\Phi,\mu_\pi}\bar{\mathcal{T}}_\pi\Phi\theta^* = \Phi\theta^*$, the second equality follows from the fact that $\bar{\mathcal{T}}_\pi G_\pi = G_\pi$, the third inequality follows from the non-expansive property of the projection operator Γ_{Φ,μ_π}, and the last inequality follows from Proposition 1. Equation (61) implies the following result

$$
\left\|\Phi\theta^* - G_\pi\right\|_{\mu_\pi,\alpha} \leq \frac{1}{1-\gamma_G}\left\|\Gamma_{\Phi,\mu_\pi}G_\pi - G_\pi\right\|_{\mu_\pi,\alpha}.
$$

H Extension to Case $\gamma_{\max} = 1$

As shown in Proposition 1, the operator $\bar{\mathcal{T}}_{G,\pi}$ is not necessarily a contraction when $\gamma_{\max} = 1$. The uniqueness of G_π and $\hat{G}_\pi$ is not guaranteed in this case. We next consider the following assumption for the base matrix Φ_i, which can yield a desired property as we show below. Such an assumption has also been considered in the average reward MDP setting [56].

Assumption 6 (Non-constant Parameterization)**.** *For all $i = 1, \cdots, k$, we have $\Phi_i\theta_i \neq c\mathbf{1}$ for any $\theta_i \in \mathbb{R}^{d_i}$ and $c \in \mathbb{R}/0$.*

Despite the non-contraction nature of $\bar{\mathcal{T}}_{G,\pi}$, if the base function Φ_i satisfies Assumption 6, we can show that the monotonicity condition of $g(\theta)$ in Proposition 2 still holds with a positive constant λ_G. As a result, the convergence bound in eq. (12) of theorem 1 is directly applicable to this setting with the corresponding value of λ_G. We can then further establish a result similar to Theorem 2 for the case with $\gamma_{\max} = 1$ under Assumption 6.

We first extend Proposition 2 and Theorem 2 to the case in which $\gamma_{\max} = 1$. Without loss of generality, we consider $\gamma_i = 1$ for all $i = 1, \cdots, k$.

Forward GVF. We first verify Proposition 2. In this setting, we can still obtain the same result for G as in eq. (30), but with $A_i = [\Phi_i^\top \bar{U}_\pi(\mathsf{P}_\pi - I)\Phi_i] \otimes \mathsf{I}_{d_i}$, where $\bar{U}_\pi = \mathrm{diag}(\mu_\pi)$. As shown in Lemma 7 in [56], the matrix $[\Phi_i^\top U_{\pi,i}(\mathsf{P}_\pi - I)\Phi_i]$ is Hurwitz when the base matrix Φ_i satisfies Assumption 6. Following the steps similar to those in eq. (31) - (34), we can conclude that the matrix G is also Hurwitz, which completes the proof.

We then verify Theorem 2. We proceed as follows,

$$
\begin{aligned}
&\left\|\Phi\theta^* - G_\pi\right\|_{\mu_\pi,\alpha} \\
&\quad= \left\|\Gamma_{\Phi,\mu_\pi}\mathcal{T}_\pi\Phi\theta^* - \mathcal{T}_\pi G_\pi\right\|_{\mu_\pi,\alpha} \\
&\quad\leq \left\|\Gamma_{\Phi,\mu_\pi}\mathcal{T}_\pi\Phi\theta^* - \Gamma_{\Phi,\mu_\pi}\mathcal{T}_\pi G_\pi\right\|_{\mu_\pi,\alpha} + \left\|\Gamma_{\Phi,\mu_\pi}\mathcal{T}_\pi G_\pi - \mathcal{T}_\pi G_\pi\right\|_{U_\pi} \\
&\quad\leq \left\|\Gamma_{\Phi,\mu_\pi}M_\pi(\Phi\theta^* - G_\pi)\right\|_{U_\pi} + \left\|\Gamma_{\Phi,\mu_\pi}\mathcal{T}_\pi G_\pi - \mathcal{T}_\pi G_\pi\right\|_{\mu_\pi,\alpha} \\
&\quad\leq \left\|M_\pi(\Phi\theta^* - G_\pi)\right\|_{U_\pi} + \left\|\Gamma_{\Phi,\mu_\pi}\mathcal{T}_\pi G_\pi - \mathcal{T}_\pi G_\pi\right\|_{\mu_\pi,\alpha} \\
&\quad\leq C_\zeta \left\|\Phi\theta^* - G_\pi\right\|_{U_\pi} + \left\|\Gamma_{\Phi,\mu_\pi}\mathcal{T}_\pi G_\pi - \mathcal{T}_\pi G_\pi\right\|_{\mu_\pi,\alpha}, \qquad (62)
\end{aligned}
$$

where the last inequality in eq. (62) can be obtained as follows. Following the steps similar to those in eq. (31)-(34), we can conclude that $\mathrm{eig}(M_\pi) = \mathrm{eig}(\mathsf{P}_\pi)$. For an ergodic MDP, we have $\max[\mathrm{eig}(\mathsf{P}_\pi)] = \max[\mathrm{eig}(\mathsf{P}_\pi^\top)] = 1$. Let $i = \mathrm{argmax}_j[\mathrm{eig}(\mathsf{P}_\pi)_j]$. We then have $\max_{j\neq i}\mathrm{eig}(\mathsf{P}_\pi)_j < 1$. Let G_π be the fixed point of $\mathcal{T}_\pi$ that is perpendicular to $[c_1\mathbf{1}_{d_1}, \cdots, c_k\mathbf{1}_{d_k}]$, where $c_1, \cdots, c_k$ could be any constant. The vector $\Phi\theta^* - G_\pi$ is perpendicular to the space spanned by the eigenvectors of M_π associated with the eigenvalue 1. Thus, there exists a positive constant $C_\zeta < 1$ such that $\left\|M_\pi(\Phi\theta^* - G_\pi)\right\|_{U_\pi} \leq C_\zeta \left\|\Phi\theta^* - G_\pi\right\|_{U_\pi}$, which yields the following results

$$
\left\|\Phi\theta^* - G_\pi\right\|_{\mu_\pi,\alpha} \leq \frac{1}{1-C_\zeta}\left\|\Gamma_{\Phi,\mu_\pi}\mathcal{T}_\pi G_\pi - \mathcal{T}_\pi G_\pi\right\|_{\mu_\pi,\alpha}. \qquad (63)
$$

Bakcward GVF. To verify Proposition 2, we can obtain the same result for G as in eq. (30) with $A_i = [\Phi_i^\top (\mathsf{P}_\pi^\top - I)\bar{U}_\pi \Phi_i] \otimes \mathsf{I}_{d_i}$. Define $\bar{A}_i = \Phi_i^\top (\mathsf{P}_\pi^\top - I) U_\pi \Phi_i$. We next show that $\bar{A}_i$ is Hurwitz. Note that $\bar{A}_i = \mathbb{E}_{\mu_\pi}[\phi'(\phi - \phi')]$. Let z be a non-constant function on the state-action space. Then we have

$$
\begin{aligned}
0 &< \frac{1}{2}\mathbb{E}_{\mu_\pi}[(z(s,a) - z(s',a'))^2] \\
&= \mathbb{E}_{\mu_\pi}[z(s,a)^2] - \mathbb{E}[z(s,a)z(s',a')] \\
&= z^\top \bar{U}_\pi z - z^\top \mathsf{P}_\pi \bar{U}_\pi z \\
&= z^\top (I - \mathsf{P}_\pi)\bar{U}_\pi z.
\end{aligned} \tag{64}
$$

For a vector $v \in \mathbb{R}^{K_i}$, we have

$$
v^\top \bar{A}_i v = v \Phi_i^\top (\mathsf{P}_\pi^\top - I) U_\pi \Phi_i v. \tag{65}
$$

Since Φv is a non-constant function, eq. (64) and eq. (65) together imply that

$$
v^\top \bar{A}_i v < 0 \quad \text{for all} \quad v \in \mathbb{R}^{K_i}.
$$

Thus, the matrix $\bar{A}_i$ is Hurwitz, which further implies that A_i is also Hurwitz. Following the steps similar to those in eq. (31) - (34), we can conclude that the matrix G is also Hurwitz, which completes the proof.

We then verify Theorem 2. We proceed as follows,

$$
\begin{aligned}
&\|\Phi\theta^* - G_\pi\|_{\mu_\pi,\alpha} \\
&\quad = \|\Gamma_{\Phi,\mu_\pi}\mathcal{T}_\pi \Phi\theta^* - \mathcal{T}_\pi G_\pi\|_{\mu_\pi,\alpha} \\
&\quad \le \|\Gamma_{\Phi,\mu_\pi}\mathcal{T}_\pi \Phi\theta^* - \Gamma_{\Phi,\mu_\pi}\mathcal{T}_\pi G_\pi\|_{\mu_\pi,\alpha} + \|\Gamma_{\Phi,\mu_\pi}\mathcal{T}_\pi G_\pi - \mathcal{T}_\pi G_\pi\|_{U_\pi} \\
&\quad \le \left\|\Gamma_{\Phi,\mu_\pi}\hat{M}_\pi(\Phi\theta^* - G_\pi)\right\|_{U_\pi} + \|\Gamma_{\Phi,\mu_\pi}\mathcal{T}_\pi G_\pi - \mathcal{T}_\pi G_\pi\|_{\mu_\pi,\alpha} \\
&\quad \le \left\|\hat{M}_\pi(\Phi\theta^* - G_\pi)\right\|_{U_\pi} + \|\Gamma_{\Phi,\mu_\pi}\mathcal{T}_\pi G_\pi - \mathcal{T}_\pi G_\pi\|_{\mu_\pi,\alpha} \\
&\quad \le C_\zeta \|\Phi\theta^* - G_\pi\|_{U_\pi} + \|\Gamma_{\Phi,\mu_\pi}\mathcal{T}_\pi G_\pi - \mathcal{T}_\pi G_\pi\|_{\mu_\pi,\alpha},
\end{aligned} \tag{66}
$$

where the last inequality in eq. (66) can be obtained as follows. Using Theorem 1.3.22 in [13], we have

$$
\mathrm{eig}(U_{\pi,i}^{-1}[\mathsf{P}_{\pi,i}^\top \otimes \mathsf{I}_{d_i}]U_{\pi,i}) = \mathrm{eig}([\mathsf{P}_{\pi,i}^\top \otimes \mathsf{I}_{d_i}]U_{\pi,i}U_{\pi,i}^{-1}) = \mathrm{eig}([\mathsf{P}_\pi^\top \otimes \mathsf{I}_{d_i}]) = \mathrm{eig}(\mathsf{P}_\pi^\top) = \mathrm{eig}(\mathsf{P}_\pi).
$$

Following the steps similar to those in eq. (31)-(34), we can conclude that $\mathrm{eig}(\hat{M}_\pi) = \mathrm{eig}(\mathsf{P}_\pi)$. Following the steps similar to those for obtaining eq. (63). We have

$$
\|\Phi\theta^* - G_\pi\|_{\mu_\pi,\alpha} \le \frac{1}{1 - C_\zeta}\|\Gamma_{\Phi,\mu_\pi}\mathcal{T}_\pi G_\pi - \mathcal{T}_\pi G_\pi\|_{\mu_\pi,\alpha},
$$

where $0 < C_\zeta < 1$, which completes the proof.

I Proof of Theorem 3

We first define the matrix B in the following way:

- Forward GVF: $B = \mathbb{E}_{\mathcal{D}\cdot\pi}[[\phi(s',a') \otimes \mathsf{I}_d]m(x)[\phi(s,a) \otimes \mathsf{I}_d]]$
- Backward GVF: $B = \mathbb{E}_{\mathcal{D}\cdot\pi}[[\phi(s,a) \otimes \mathsf{I}_d]m(x)[\phi(s',a') \otimes \mathsf{I}_d]]$.

We further define the following stochastic matrices in both the forward and backward GVF evaluation settings. Recall that $(s_t, a_t) \sim D(\cdot)$, $s_t' \sim \mathsf{P}(\cdot|s_t, a_t)$ and $a_t' \sim \pi(\cdot|s_t'$.

- Forward GVF:

$$
\begin{aligned}
A_t &= [\phi(s_t,a_t) \otimes \mathsf{I}_d](m(x_t)[\phi(s_t',a_t') \otimes \mathsf{I}_d]^\top - [\phi(s_t,a_t) \otimes \mathsf{I}_d]^\top), \\
B_t &= [\phi(s_t',a_t') \otimes \mathsf{I}_d]m(x_t)[\phi(s_t,a_t) \otimes \mathsf{I}_d], \\
C_t &= (\phi(s_t,a_t)\phi(s_t,a_t)^\top) \otimes \mathsf{I}_d, \\
b_t &= [\phi(s_t,a_t) \otimes \mathsf{I}_d]C(x_t).
\end{aligned} \tag{67}
$$

- Backward GVF:

$$
\begin{aligned}
A_t &= [\phi(s'_t, a'_t) \otimes \mathsf{I}_d](m(x_t)[\phi(s_t, a_t) \otimes \mathsf{I}_d]^\top - [\phi(s'_t, a'_t) \otimes \mathsf{I}_d]^\top), \\
B_t &= [\phi(s_t, a_t) \otimes \mathsf{I}_d] m(x_t)[\phi(s'_t, a'_t) \otimes \mathsf{I}_d], \\
C_t &= (\phi(s'_t, a'_t)\phi(s_t, a_t)^\top) \otimes \mathsf{I}_d, \\
b_t &= [\phi(s'_t, a'_t) \otimes \mathsf{I}_d] C(x_t).
\end{aligned} \tag{68}
$$

Recall the matrices A and C defined in Appendix D. For a constant $\xi > 0$, we define

$$
H_t = \begin{bmatrix} A_t & B_t \\ \xi A_t & \xi C_t \end{bmatrix}, \qquad h_t = \begin{bmatrix} b_t \\ 0 \end{bmatrix}.
$$

and

$$
H = \begin{bmatrix} A & B \\ \xi A & \xi C \end{bmatrix}, \qquad h = \begin{bmatrix} b \\ 0 \end{bmatrix},
$$

where $A = \mathbb{E}[A_t]$, $B = \mathbb{E}[B_t]$, $C = \mathbb{E}[C_t]$ and $b = \mathbb{E}[b_t]$.

For the matrix H_t, we have the following holds

$$
\begin{aligned}
\|H_t\|_F^2 &= (1 + \xi^2) \|A_t\|_F^2 + \|B_t\|_F^2 + \xi^2 \|C_t\|_F^2 \\
&\leq (1 + \xi^2)[d^2 C_\phi^2 (C_m + 1)]^2 + d^2 C_\phi^2 C_m^2 + \xi^2 C_\phi^4 d^2.
\end{aligned} \tag{69}
$$

which implies that $\|H_t\|_F \leq C_H$, where

$$
C_H = \sqrt{(1 + \xi^2)[d^2 C_\phi^2 (C_m + 1)]^2 + d^2 C_\phi^2 C_m^2 + \xi^2 C_\phi^4 d^2}.
$$

For the vector h_t, we can obtain $\|h_t\|_2 \leq C_h = d C_\phi R_C$ by following the steps similar to those for obtaining eq. (69).

The update in Algorithm 2 can be rewritten as

$$
v_{t+1} = \Gamma_{R_v} \left(v_t + \alpha_t (H_t v_t + h_t) \right), \tag{70}
$$

where $v_t = [\theta_t^\top, w_t^\top]^\top$, and $R_v = R_\theta \times \mathbb{R}^{K d_g \times 1}$. Following the proof similar to those in Theorem 3 of Section 5.3.3 in [23], we can show that the matrix H is Hurwitz under Assumption 4 and Assumption 5 with an appropriately chosen $\xi > \max\{0, -\mathrm{eig}_{\min}(C^{-1}[(A + A^\top)/2])\}$.

We define the following optimal point $v^* = [\bar{\theta}^{*\top}, w^{*\top}]^\top$ for the linear SA defined in eq. (70)

$$
\langle \varphi(v^*), v - v^* \rangle \leq 0, \quad \forall v \in R_v,
$$

where $\varphi(v) = Hv + b$. We also define $C_v = \|v^*\|_2$. It can be checked that there exist a positive constant λ'_G such that

$$
\langle \varphi(v^*) - \varphi(v), v^* - v \rangle \leq -\lambda'_G \|v - v^*\|_2^2. \tag{71}
$$

We further define $\varphi(x_t, v) = H_t v + h_t$.

Following the steps similar to those for proving Lemma 7 and Lemma 8, we can obtain the following two lemmas.

Lemma 11. *Given a sample $(s_t, a_t, B_t, s'_t) \sim \mathcal{D}_d$ and $a'_t \sim \pi(\cdot|s'_t)$ and any $v \in R_v$, we have the following holds*

$$
\|\varphi(x_t, v) - \varphi(v)\|_2^2 \leq 16 C_H^2 \|\kappa - \kappa^*\|_2^2 + 16 C^2 h P C_v^2 + 8 C_h^2.
$$

Proof. Based on the definition of $\varphi(x_t, v)$ and $\varphi(v)$, we can obtain the following

$$
\begin{aligned}
\|\varphi(x_t, v) - \varphi(v)\|_2^2 &\leq 2 \|(H_t - H)v\|_2^2 + 2 \|h_t - h\|_2^2 \\
&\leq 4 \|(H_t - H)(v - v^*)\|_2^2 + 4 \|(H_t - H)v^*\|_2^2 + 2 \|h_t - h\|_2^2 \\
&\leq 16 C_H^2 \|\kappa - \kappa^*\|_2^2 + 16 C_h^2 C_v^2 + 8 C_h^2.
\end{aligned}
$$

□

Lemma 12. *Consider the population GTD update $\varphi(v) = Hv + b$. We have*

$$\langle -\varphi(v), v - v^* \rangle \geq \lambda'_G \left\| v - v^* \right\|_2^2, \quad \forall v \in R_v.$$

We also have the following "three-point lemma" holds for the GTD update.

Lemma 13. *Consider the update of w_t and θ_t in Algorithm 2. For all $v \in R_v$, we have the following holds*

$$-\alpha_t \langle \varphi(x_t, v_t), v_{t+1} - v \rangle + \frac{1}{2} \left\| v_{t+1} - v_t \right\|_2^2 \leq \frac{1}{2} \left\| v_t - v \right\|_2^2 - \frac{1}{2} \left\| v_{t+1} - v \right\|_2^2. \tag{72}$$

Using Lemma 13 and following the steps similar to those from eq. (45) to eq. (48), we can obtain

$$\begin{aligned}
&\frac{1}{2} \left\| v_t - v \right\|_2^2 - \frac{1}{2} \left\| v_{t+1} - v \right\|_2^2 \\
&\quad \geq -\alpha_t \langle \varphi(v_{t+1}), v_{t+1} - v \rangle - \alpha_t \langle \varphi(x_t, v_t) - \varphi(v_t), v_t - v \rangle - \alpha_t^2 C_H^2 \left\| v_{t+1} - v \right\|_2^2 \\
&\qquad - \alpha_t^2 \left\| \varphi(x_t, \kappa_t) - \varphi(\kappa_t) \right\|_2^2.
\end{aligned} \tag{73}$$

Taking expectation on both sides of eq. (73), letting $v = v^*$, and using the fact that $-\langle \varphi(v_{t+1}), v_{t+1} - v^* \rangle \leq \lambda'_G \left\| v_{t+1} - v^* \right\|_2$ yield

$$\begin{aligned}
(1 + 2\alpha_t \lambda'_G - 2\alpha_t^2 C_H^2) \mathbb{E}[\| v_{t+1} - v^* \|_2^2] &\leq \mathbb{E}[\| v_t - v^* \|_2^2] + 2\alpha_t^2 \mathbb{E}[\| \varphi(x_t, \kappa_t) - \varphi(\kappa_t) \|_2^2] \\
&\leq (1 + 32\alpha_t^2 C_H^2) \mathbb{E}[\| v_t - v^* \|_2^2] + +32 C_h^2 C_v^2 + 16 C_h^2,
\end{aligned} \tag{74}$$

where the second inequality follows from Lemma 12. Multiplying both sides of eq. (74) by o_t and summing over iterations $t = 0, \cdots, T-1$ yield

$$\sum_{t=0}^{T-1} a''_t \mathbb{E}[\| v_{t+1} - v^* \|_2^2] \leq \sum_{t=0}^{T-1} b''_t \mathbb{E}[\| v_t - v^* \|_2^2] + c'', \tag{75}$$

where

$$\begin{aligned}
a''_t &= (1 + 2\alpha_t \lambda'_G - 2\alpha_t^2 C_H^2) o_t, \\
b''_t &= (1 + 32\alpha_t^2 C_H^2) o_t, \\
c'' &= (32 C^2 h P C_v^2 + 16 C_h^2) \sum_{t=0}^{T-1} \alpha_t^2 o_t.
\end{aligned}$$

Now we let

$$\begin{aligned}
o_t &= (t + t_2)(t + t_2 + 1), \\
\alpha_t &= \frac{4}{\lambda'_G (t + t_2 - 1)}, \\
t_2 &= \frac{34 C_H^2}{\lambda_J^2} + 1.
\end{aligned}$$

Then, we can obtain the following

$$\begin{aligned}
a''_t - b''_{t+1} &= (1 + 2\alpha_t \lambda'_G - 2\alpha_t^2 C_H^2) s_t - (1 + 32\alpha_{t+1}^2 C_H^2) o_{t+1} \\
&\geq (1 + 2\alpha_t \lambda'_G) o_t - (1 + 2\alpha_t^2 C_H^2 + 32\alpha_{t+1}^2 C_H^2) o_{t+1} \\
&\geq (1 + 2\alpha_t \lambda'_G) o_t - (1 + 34\alpha_t^2 C_H^2) o_{t+1} \\
&\geq (1 + 2\alpha_t \lambda'_G) o_t - (1 + \alpha_t \lambda'_G) o_{t+1} \\
&\geq (t + t_2 + 1) \frac{(t + t_2 + 3)(t + t_2) - (t + t_2 + 2)^2}{t + t_2 - 1} \\
&\geq 0,
\end{aligned}$$

where the second inequality follows from the fact that $\alpha_t \leq \frac{\lambda'_G}{34 C_H^2}$.

Applying the above property to eq. (75) yields

$$a''_{T-1}t\mathbb{E}[\|v_T - v^*\|_2^2] \le b''_0 \|v_0 - v^*\|_2^2 + c'',$$

which implies

$$\begin{aligned}\mathbb{E}[\|v_T - v^*\|_2^2] &\le \frac{b''_0 \|v_0 - v^*\|_2^2}{a''_{T-1}} + \frac{c''}{a''_{T-1}} \\ &\le \frac{(1+16\alpha_0^2 C_H^2)(t_2+1)^2 \|v_0 - v^*\|_2^2}{(T+t_2-1)(T+t_2)} + \frac{128C_h^2C_v^2 + 64C_h^2}{(T+t_2)\lambda_J^2}.\end{aligned}$$

Using the fact $\left\|\theta_T - \bar{\theta}^*\right\|_F^2 \le \|v_T - v^*\|_2^2$, we have

$$\mathbb{E}[\left\|\theta_T - \bar{\theta}^*\right\|_F^2] \le \frac{(1+16\alpha_0^2 C_H^2)(t_2+1)^2 \|v_0 - v^*\|_2^2}{(T+t_2-1)(T+t_2)} + \frac{128C_h^2C_v^2 + 64C_h^2}{(T+t_2)\lambda_J^2},$$

which completes the proof.