How to Improve Imitation Learning Performance with Sub-optimal Supplementary Data?

Ziniu Li [*] [1] [2] **Tian Xu** [*] [3] **Zeyu Qin** [4] **Yang Yu** [3] **Zhi-Quan Luo** [1] [2]

Abstract

Imitation learning (IL) is a machine learning technique that involves learning from examples provided by an expert. IL algorithms can solve the sequential decision-making tasks but their performance usually suffer when the amount of expert data is limited. To address this challenge, a new data-centric framework called (offline) IL with supplementary data has emerged, which *additionally* utilizes an imperfect dataset inexpensively collected from sub-optimal policies. However, the supplementary data may contain out-of-expert-distribution samples, making it tricky to utilize the supplementary data to improve performance. In this paper, we focus on a classic offline IL algorithm called behavioral cloning (BC) and its variants, studying the imitation gap bounds in the context of IL with supplementary data. Our theoretical results show that a naive method, which applies BC on the union of expert and supplementary data, has a non-vanishing imitation error. As a result, its performance may be worse than BC which relies solely on the expert data. To address this issue, we propose an importance-sampling-based approach for selecting in-expert-distribution samples from the supplementary dataset. The proposed method theoretically eliminates the gap of the naive method. Empirical studies demonstrate that our method can perform better than prior state-of-the-art methods on tasks including locomotion control, Atari games, and object recognition.

1. Introduction

Imitation learning (IL) is an essential technique within the field of artificial intelligence, allowing machines to learn and enhance their behavior by imitating expert demonstrations. IL algorithms have demonstrated significant success in training high-quality policies, as highlighted in (Argall et al., 2009; Osa et al., 2018). Among the IL approaches, behavioral cloning (BC) (Pomerleau, 1991) stands out as a popular method that achieves expert imitation through supervised learning. BC leverages state-action pairs extracted from trajectory data within the dataset, employing them as training samples to learn a mapping from states to actions. Consequently, IL expands upon the traditional supervised learning framework, enabling the acquisition of sequential decision-making capabilities.

The quantity of expert trajectories plays a crucial role in achieving satisfactory performance. Previous studies have shown that BC works well when the dataset contains a large number of expert-level trajectories (Spencer et al., 2021). However, the compounding errors issue (Ross & Bagnell, 2010) renders any offline IL algorithm, including BC, ineffective when the number of expert trajectories is small (Rajaraman et al., 2020; Xu et al., 2021). One naive solution to this problem is to collect more trajectories from the expert, but this approach is costly and impractical in certain domains, such as robotics and healthcare.

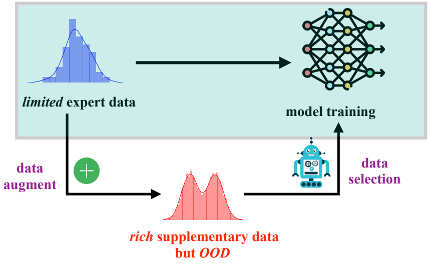

Figure 1. Compared with the standard IL framework (shown in cyan), the *supplementary data* helps address the expert data scarcity issue, and the *data selection* technique helps address the distribution shift issue in model training.

To overcome the challenge of scarce expert data, we focus

[*] Equal contribution [1] The Chinese University of Hong Kong, Shenzhen [2] Shenzhen Research Institute of Big Data [3] National Key Laboratory for Novel Software Technology, Nanjing University [4] Hong Kong University of Science and Technology. Correspondence to: Ziniu Li <ziniuli@Link.cuhk.edu.cn>, Tian Xu <xut@lamda.nju.edu.cn>.

Presented at Data-centric Machine Learning Research (DMLR) Workshop at ICML 2023. Copyright 2023 by the author(s).

on the offline setting (i.e., no online interaction) and adopt the IL with supplementary data framework, which has been recently proposed in (Kim et al., 2022b; Xu et al., 2022a). Under this framework, the learner can leverage an additional dataset, which can be cheaply obtained by executing sub-optimal policies, to supplement the expert dataset. Please refer to Figure 1 for illustration. However, the supplementary dataset introduces a distribution shift issue due to the presence of out-of-expert-distribution trajectories[1]. The distribution shift issue may hamper the model's performance in utilizing the supplementary data, as we will argue later.

We realize that a few empirical advances have been achieved in this direction (Kim et al., 2022b;a; Xu et al., 2022a; Ma et al., 2022). Most algorithms rely on a discriminator to distinguish between expert-style and sub-optimal samples, followed by optimization of a weighted BC objective to learn a good policy. For example, DemoDICE (Kim et al., 2022b) uses a regularized state-action distribution matching objective to train the discriminator, while DWBC (Xu et al., 2022a) employs a cooperative training framework for the policy and discriminator. Despite the empirical success of these methods in certain scenarios, there is a lack of systematic theoretical studies, particularly in terms of imitation gap (or equivalently sample complexity), which may hinder deep understanding and impede future algorithmic advances.

We aim to bridge the gap between theory and practice in the (offline) IL with supplementary data framework by designing effective algorithms and providing rigorous theoretical guarantees. To the best of our knowledge, only (Chang et al., 2021) provided imitation gap bounds for a model-based adversarial imitation learning approach in a similar problem. However, our focus is on the widely used and simpler BC and its variants, which are model-free in nature. Our contributions are summarized below.

- We establish theoretical bounds on the limitations of the IL with supplementary data framework, highlighting the impact of the distribution shift between expert data and supplementary data. A quick overview is provided in Table 1. Our analysis shows that the naively applying BC on the union of expert and supplementary data has a non-vanishing error term in the imitation gap bound. This means that naively using supplementary data may result in worse performance than BC which relies solely on the expert data.
- To address the distribution shift issue, in light of (Kim et al., 2022b; Xu et al., 2022a), we propose a new importance-sampling-based approach called ISW-BC. In contrast to prior methods (Kim et al., 2022b; Xu et al., 2022a) that use regularized weighting rules, ISW-BC corrects the loss function in an unbiased way. We provide the first imitation gap bound for this type of data selection method in imitation learning, which shows that ISW-BC eliminates the gap of the naive method and also has a better guarantee than BC.
- We validate our theoretical results through experiments on various tasks, including locomotion control, Atari games, and object recognition. Our results demonstrate that ISW-BC outperforms previous state-of-the-art methods, confirming the effectiveness of our proposed approach in addressing the distribution shift issue in IL with supplementary data.

Table 1. Theoretical guarantees of three methods: (1) BC, which relies solely on expert data, (2) NBCU, which naively utilizes supplementary data without selection, and (3) ISW-BC, a new method that employs importance sampling for data selection. NBCU suffers a non-vanishing error while ISW-BC does not. The capital notation N refers to the data size, and for the exact meaning of symbols, please refer to the main text.

	Imitation Gap
BC	$\mathcal{O}(\frac{\lvert\mathcal{S}\rvert H^2}{N_{\mathrm{E}}})$
NBCU	$\widetilde{\mathcal{O}}((1-\eta)(V(\pi^{\mathrm{E}}) - V(\pi^{\beta})) + \frac{\lvert\mathcal{S}\rvert H^2}{N_{\mathrm{tot}}})$
ISW-BC	$\mathcal{O}(\frac{\lvert\mathcal{S}\rvert H^2}{N_{\mathrm{E}}+N_{\mathrm{S}}/\mu})$

2. Related Work

We review broadly relevant studies in the main text and provide a detailed discussion in Appendix A.

Behavioral cloning (BC) is a popular algorithm in the offline setting, where the learner cannot interact with the environment. According to the learning theory in (Rajaraman et al., 2020), only using an expert dataset, BC has an imitation gap of $\mathcal{O}(|\mathcal{S}|H^2/N_{\mathrm{E}})$, where $|\mathcal{S}|$ is the state space size, H is the planning horizon, and N_{E} is the number of expert trajectories. Our work investigates the use of a supplementary dataset to enhance the dependence on the data size.

Our theoretical study is motivated by recent empirical advances in IL with supplementary data (Kim et al., 2022b; Xu et al., 2022a; Ma et al., 2022; Kim et al., 2022a). We have reviewed (Kim et al., 2022b; Xu et al., 2022a) and will not repeat their contributions again. Compared with (Kim et al., 2022b; Xu et al., 2022a), a related setting, learning from observation, where expert actions are missing, and only expert states are observed, is studied in (Ma et al., 2022; Kim et al., 2022a). The importance sampling technique used in our method for addressing distribution shift is also studied in (semi-)supervised learning (Sugiyama et al., 2007; Cortes et al., 2010; Liu & Tao, 2015; Fang et al., 2020). Our contri-

[1] This issue is separate from the *intrinsic* distribution shift problem that IL already faces, where the training and evaluation distributions differ (Ross & Bagnell, 2010).

bution is to validate this technique in the imitation learning set-up, where Markovian data needs to be considered.

3. Preliminary

Markov Decision Process. In this paper, we consider the episodic Markov decision process (MDP) framework (Puterman, 2014). An MDP is defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, H, \rho)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action space, respectively. H is the maximum length of a trajectory, and ρ is the initial state distribution. The non-stationary transition function is specified by $\mathcal{P} = \{P_1, \cdots, P_H\}$, where $P_h(s_{h+1}|s_h, a_h)$ determines the probability of transiting to state s_{h+1} given the current state s_h and action a_h in time step h, for $h \in [H]$. Here the symbol $[x]$ means the set of integers from 1 to x. Similarly, the reward function $r = \{r_1, \cdots, r_H\}$ specifies the reward received at each time step, where $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ for $h \in [H]$. A policy in an MDP is a function that maps each state to a probability distribution over actions. We consider time-dependent policies $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$, where $\Delta(\mathcal{A})$ is the probability simplex. The policy at each time step h is denoted as π_h, and we use π to denote the collection of time-dependent policies $\{\pi_h\}_{h=1}^H$ when the context is clear.

We measure the quality of a policy π by the policy value (i.e., environment-specific long-term return): $V(\pi) = \mathbb{E}\big[\sum_{h=1}^H r(s_h, a_h) \mid s_1 \sim \rho; a_h \sim \pi_h(\cdot|s_h), s_{h+1} \sim P_h(\cdot|s_h, a_h), \forall h \in [H]\big]$. To facilitate later analysis, we need to introduce the state-action distribution $d_h^\pi(s, a) = \mathbb{P}(s_h = s, a_h = a|\pi)$. We use the convention that d^π is the collection of all time-dependent state-action distributions.

Imitation Learning. Imitation learning (IL) aims to learn a policy that mimics an expert policy based on expert demonstrations. In this paper, we assume that there exists a good expert policy π^{E} that generates a dataset $\mathcal{D}^{\mathrm{E}}$ consisting of N_{E} trajectories of length H.

$$\mathcal{D}^{\mathrm{E}} = \big\{ \mathrm{tr} = (s_1, a_1, s_2, a_2, \cdots, s_H, a_H)\,; s_1 \sim \rho, \\ a_h \sim \pi_h^{\mathrm{E}}(\cdot|s_h), s_{h+1} \sim P_h(\cdot|s_h, a_h), \forall h \in [H] \big\}.$$

The learner aims to imitate the expert using the expert dataset $\mathcal{D}^{\mathrm{E}}$. The quality of the imitation is measured by the *imitation gap*, defined as $\mathbb{E}\left[V(\pi^{\mathrm{E}}) - V(\pi)\right]$, where the expectation is taken over the randomness of data collection. It is worth noting that in the training phase, IL algorithms do *not* have access to reward information. A good learner should closely mimic the expert, resulting in a small imitation gap. We assume that the expert policy is deterministic, a common assumption in the literature (Rajaraman et al., 2020; 2021a; Xu et al., 2021), and applicable to tasks such as MuJoCo locomotion control.

Behavioral Cloning. Behavioral cloning (BC) is a commonly used imitation learning algorithm that aims to learn a policy from an expert dataset $\mathcal{D}^{\mathrm{E}}$ via supervised learning. Specifically, BC seeks to find a policy π^{BC} that maximizes the log-likelihood of the expert actions in the dataset:

$$\pi^{\mathrm{BC}} \in \operatorname*{argmax}_{\pi} \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \widehat{d_h^{\mathrm{E}}}(s, a) \log \pi_h(a|s), \quad (1)$$

where $\widehat{d_h^{\mathrm{E}}}(s, a)$ is the empirical state-action distribution in the expert dataset. Through the maximum likelihood estimation (MLE), BC can make good decisions by duplicating expert actions on visited states. However, BC may take sub-optimal actions on non-visited states, resulting in compounding errors and a large imitation gap. This issue is significant when the expert data is limited.

4. IL with Supplementary Data

In this section, we consider the mentioned IL with supplementary data framework to address the challenge of limited availability of expert data. Following previous works (Kim et al., 2022b; Xu et al., 2022a), we assume that a supplementary dataset $\mathcal{D}^{\mathrm{S}} = \big\{ \mathrm{tr} = (s_1, a_1, s_2, a_2, \cdots, s_H, a_H) \big\}$ is collected by a (sub-optimal) behavior policy π^β. A naive approach is to perform MLE on the *union* of the expert and supplementary dataset $\mathcal{D}^{\mathrm{U}} = \mathcal{D}^{\mathrm{E}} \cup \mathcal{D}^{\mathrm{S}}$:

$$\pi^{\mathrm{NBCU}} \in \operatorname*{argmax}_{\pi} \sum_{h=1}^{H} \sum_{(s,a)} \widehat{d_h^{\mathrm{U}}}(s, a) \log \pi_h(a|s), \quad (2)$$

where $\widehat{d_h^{\mathrm{U}}}(s, a)$ is the empirical state-action distribution in $\mathcal{D}^{\mathrm{U}}$. We refer to this approach as NBCU (naive BC with the union dataset). NBCU treats these two datasets equally and is brittle to distribution shift, as we will demonstrate later. For theoretical analysis purposes, we give the dataset assumption below, in which we use $\eta \in [0, 1]$ to denote the fraction of expert data in the union dataset.

Assumption 1. *The expert dataset $\mathcal{D}^{\mathrm{E}}$ and supplementary dataset $\mathcal{D}^{\mathrm{S}}$ are collected in the following way: each time, we roll-out a behavior policy π^β with probability $1 - \eta$ and the expert policy with probability η. Such an experiment is independent and identically conducted by N_{tot} times.*

Under Assumption 1, we slightly overload our notations: we use N_{E} to denote the *expected* number of expert trajectories, which is given by $N_{\mathrm{E}} = \eta N_{\mathrm{tot}}$, and N_{S} to denote the *expected* number of supplementary trajectories, which is given by $N_{\mathrm{S}} = (1 - \eta) N_{\mathrm{tot}}$. Note that the conditional sampling procedure does not change the nature of our theoretical insights. In practice, one may collect a fixed number of expert and supplementary trajectories, respectively.

To establish a common ground, we begin by specifying the policy representations. Here, we adopt tabular representations, which assume that the parameterized functions can take any possible form. Specifically, we define

$\pi_h(a|s;\theta) = \langle \phi(s,a), \theta \rangle$, where $\phi(s,a) \in \mathbb{R}^d$ is the feature representation and $\theta \in \mathbb{R}^d$ is the parameter to optimize. In tabular representations, we use one-hot features for $\phi(s,a)$, which further implies that the expert policy is realizable and learnable within this function class. The tabular representations are widely considered in classical IL theory (Rajaraman et al., 2020; 2021b; Xu et al., 2022b; Shani et al., 2022). For a discussion on general function approximation schemes, please refer to Appendix D.

Imitation Gap of BC. In order to evaluate the usefulness of the supplementary dataset, we use BC with only the expert dataset as a baseline. The analysis of this approach has been done under the standard IL set-up in (Rajaraman et al., 2020), and we transfer their results to our setting.

Theorem 1. *Under Assumption 1, if we apply BC only on the expert dataset, we have that* $\mathbb{E}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{BC}})\right] = \mathcal{O}(\frac{|\mathcal{S}|H^2}{N_{\mathrm{E}}})$*, where the expectation is taken over the randomness in the dataset collection (same as other expectations).*

Proofs of Theorem 1 and other theoretical results are deferred to the Appendix. The proof of Theorem 1 builds on (Rajaraman et al., 2020), with the main difference being that the number of expert trajectories is a random variable in our set-up. We handle this difficulty by using Lemma 3 in the Appendix. The quadratic dependence on the planning horizon H indicates the compounding errors issue of BC. If the expert data is limited, BC may perform poorly.

Imitation Gap of NBCU. Guarantees of naively using the supplementary data are presented below.

Theorem 2. *Under Assumption 1, if we apply BC on the union dataset, we have* $\mathbb{E}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{NBCU}})\right] = \mathcal{O}((1-\eta)(V(\pi^{\mathrm{E}}) - V(\pi^{\beta})) + \frac{|\mathcal{S}|H^2\log(N_{\mathrm{tot}})}{N_{\mathrm{tot}}})$.

Remark 1. *In practice, the behavior policy may take non-expert actions, making it inferior to the expert policy. Therefore,* $V(\pi^{\mathrm{E}}) - V(\pi^{\beta}) > 0$ *is often the case. Even if* N_{tot} *is large enough to make the second term negligible, there is still a non-vanishing gap between* $V(\pi^{\mathrm{E}})$ *and* $V(\pi^{\beta})$ *due to the behavior policy's potential to collect non-expert actions. Consequently, the recovered policy may select a wrong action even on expert states, leading to sub-optimal performance of NBCU. Moreover, a previous work (Xu et al., 2021) has shown that*

$$V(\pi^{\mathrm{E}}) - V(\pi^{\beta}) = \mathcal{O}(H\varepsilon_d) = \mathcal{O}(H^2\varepsilon_\pi),$$

where $\varepsilon_d = \max_h \mathrm{TV}(d_h^{\pi^{\mathrm{E}}}, d_h^{\pi^{\beta}})$ *is the state-action distribution total variation (TV) distance and* $\varepsilon_\pi = \max_h \max_s \mathrm{TV}(\pi_h^{\mathrm{E}}(\cdot|s), \pi_h^{\beta}(\cdot|s))$ *is the policy distribution TV distance. Hence, we can also view Theorem 2 in the context of* *<u>state-action</u> or <u>policy</u> distribution shifts.*

The subsequent proposition establishes the inevitability of the gap $V(\pi^{\mathrm{E}}) - V(\pi^{\beta})$ in the worst case.

Proposition 1. *Under Assumption 1, there exists an MDP* $\mathcal{M}$*, an expert policy* π^{E} *and a behavior policy* π^{β}*, such that* $\mathbb{E}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{NBCU}})\right] = \Omega((1-\eta)(V(\pi^{\mathrm{E}}) - V(\pi^{\beta})))$.

The construction of the hard instance in Proposition 1 relies on the following insight: NBCU considers all action labels in the union dataset equally important and does not distinguish between them. Therefore, we can build an instance where the expert π selects a good action with a one-step reward of 1, while the behavior policy π^{β} chooses a bad action with a one-step reward of 0. The noise introduced by π^{β} results in incorrect learning goals, causing NBCU to make a mistake with probability $1-\eta$, which is the fraction of the noise in the union dataset. By putting extra effort into transition construction, we can obtain the expected bound in Proposition 1.

5. Addressing Distribution Shift with Importance Sampling

In this section, we propose a data selection approach to alleviate the distribution shift issue between expert data and supplementary data. Our approach is inspired by recent works (Kim et al., 2022b; Xu et al., 2022a), where a discriminator is trained to re-weight samples, and a weighted BC objective is used for policy optimization. Specifically, we define the weighted BC objective as follows:

$$\pi^{\text{ISW-BC}} \in \operatorname*{argmax}_{\pi} \sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \Big\{ \widehat{d_h^{\mathrm{U}}}(s,a) \times \left[w_h(s,a)\log\pi_h(a|s)\right] \times \mathbb{I}\left[w_h(s,a) \geq \delta\right] \Big\}, \quad (3)$$

where $\widehat{d_h^{\mathrm{U}}}(s,a)$ is the empirical state-action distribution of the union dataset, and $w_h(s,a) \in [0,\infty)$ is the weight decided by the discriminator. We introduce a hyper-parameter $\delta \in [0,\infty)$ to control the weight of samples (for theoretical analysis). However, in practice, we usually set $\delta = 0$.

We propose using the importance sampling technique (Shapiro, 2003, Chapter 9) to transfer samples in the union dataset to the expert policy distribution, which is the key idea behind our method. This technique helps address the failure mode of NBCU. In an ideal scenario where there are infinite samples (i.e., the population level), $\widehat{d_h^{\mathrm{U}}}$ would equal d_h^{U}. By setting $w_h(s,a) = d_h^{\mathrm{E}}(s,a)/d_h^{\mathrm{U}}(s,a)$, we obtain $\widehat{d_h^{\mathrm{U}}}(s,a)w_h(s,a) = d_h^{\mathrm{E}}(s,a)$, and the objective (3) enables the learning of a policy as if samples were collected by the expert policy. However, in practice, $d_h^{\mathrm{E}}(s,a)$ and $d_h^{\mathrm{U}}(s,a)$ are unknown, and we only have a finite number of samples from each of these distributions. Therefore, we must estimate the grounded importance sampling ratio $d_h^{\mathrm{E}}(s,a)/d_h^{\mathrm{U}}(s,a)$ from the expert data and union data.

We want to stress that estimating the probability densities

Algorithm 1 ISW-BC

```
Input: Expert dataset $\mathcal{D}^{\mathrm{E}}$ and supplementary dataset $\mathcal{D}^{\mathrm{S}}$.
1: $\mathcal{D}^{\mathrm{U}} \leftarrow \mathcal{D}^{\mathrm{E}} \cup \mathcal{D}^{\mathrm{S}}$.
2: Train a binary classifier $c$ with positive labels for $\mathcal{D}^{\mathrm{E}}$
   and negative labels for $\mathcal{D}^{\mathrm{U}}$.
3: Compute importance sampling ratio $w$ by Equation (5).
4: Apply BC to learn a policy $\pi$ by objective (3) with $\mathcal{D}^{\mathrm{U}}$.
```

of high-dimensional distributions separately for expert and union data and then calculating their quotient can be a challenging task. We take a different approach. Inspired by (Goodfellow et al., 2014), we directly train a discriminator to estimate the importance sampling ratio $d^{\mathrm{E}}_h(s,a)/d^{\mathrm{U}}_h(s,a)$. To this end, we introduce time-dependent parameterized discriminators $\{c_h : \mathcal{S} \times \mathcal{A} \to [0,1]\}_{h=1}^{H}$, each of which is optimized according to the objective function

$$\begin{aligned} \max_{c_h} & \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \widehat{d^{\mathrm{E}}_h}(s,a)\left[\log\left(c_h(s,a)\right)\right] \\ & + \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \widehat{d^{\mathrm{U}}_h}(s,a)\left[\log\left(1-c_h(s,a)\right)\right]. \end{aligned} \tag{4}$$

Solving the optimization problem in (4) is equivalent to training a binary classifier that assigns positive labels to expert data and negative labels to union data. We can obtain the optimal discriminator at the population level, from which we can derive the importance sampling ratio formula:

$$\begin{aligned} c^{\star}_h(s,a) &= \frac{d^{\mathrm{E}}_h(s,a)}{d^{\mathrm{E}}_h(s,a)+d^{\mathrm{U}}_h(s,a)}, \\ w_h(s,a) &= \frac{c^{\star}_h(s,a)}{1-c^{\star}_h(s,a)}. \end{aligned} \tag{5}$$

Based on the previous discussion, we present the implementation of our proposed method, named ISW-BC (importance-sampling-weighted BC), in Algorithm 1. It is worth noting that ISW-BC employs an unbiased weighting rule since it directly estimates the importance sampling ratio. In contrast, previous approaches such as (Kim et al., 2022b; Xu et al., 2022a) use regularized weighting rules that may fail to recover the expert policy even with infinite samples. For further details on the differences between our method and previous ones, please refer to Appendix A.

5.1. Negative Result of ISW-BC with Tabular Representations of Discriminator

We have not yet specified the representations of the discriminator. One natural choice is to use tabular representations, which correspond to linear function approximation with one-hot features. Tabular representations have a strong representation power since they can span all possible functions. However, surprisingly, we show that tabular representations can fail when considering generalization.

Proposition 2. *If the discriminator uses the one-hot feature with $\delta = 0$, we have $\pi^{\text{ISW-BC}} = \pi^{\text{BC}}$.*

Proposition 2 suggests that even if we have a large number of supplementary data and use importance sampling, ISW-BC is not guaranteed to outperform BC based on tabular representations. To illustrate, suppose we have a sample (s,a) that is an expert-style sample but only appears in the supplementary dataset, meaning that $d^{\mathrm{E}}_h(s,a) = 0, \widehat{d^{\mathrm{E}}_h}(s,a) = 0$ and $\widehat{d^{\mathrm{U}}_h}(s,a) > 0$. Using tabular representations, we can compute the closed-form solution $c^{\star}_h(s,a) = \widehat{d^{\mathrm{E}}_h}(s,a)/(\widehat{d^{\mathrm{E}}_h}(s,a) + \widehat{d^{\mathrm{U}}_h}(s,a)) = 0$. This implies that the importance sampling ratio $w_h(s,a) = c^{\star}_h(s,a)/(1-c^{\star}_h(s,a)) = 0$, so this good sample does not contribute to the learning objective (3). The failure of tabular representations is due to their discrete treatment of data, ignoring internal correlations. Consequently, although they work well in minimizing the empirical loss, they are not good at *generalization*. This kind of failure mode is also mentioned in the GAN literature (Arora et al., 2017).

5.2. Positive Result of ISW-BC with Function Approximation of Discriminator

In this section, we address the issue raised in the previous section by investigating ISW-BC with a specific function approximation. To avoid the limitations of tabular representations, we consider that the discriminator is parameterized by $c_h(s,a;\theta_h) = \frac{1}{1+\exp(-\langle\phi_h(s,a),\theta_h\rangle)}$, where $\theta_h \in \mathbb{R}^d$ is the parameter to be trained. Note that we require $d < |\mathcal{S}||\mathcal{A}|$ to avoid the tabular representations. Let $g(x) = \log(1+\exp(x))$. Then, the optimization problem of the discriminator becomes:

$$\begin{aligned} \min_{\theta_h} \mathcal{L}_h(\theta_h) \triangleq & \sum_{(s,a)} \widehat{d^{\mathrm{E}}_h}(s,a) g(-\langle\phi_h(s,a),\theta_h\rangle) \\ & + \sum_{(s,a)} \widehat{d^{\mathrm{U}}_h}(s,a) g(\langle\phi_h(s,a),\theta_h\rangle). \end{aligned} \tag{6}$$

Let $\theta^{\star} = \{\theta^{\star}_1, \cdots, \theta^{\star}_H\}$ be the optimal solution obtained from Equation (6). With the feature vector, samples are no longer treated independently, and the discriminator can perform *structured* estimation. To be consistent with the prior results, the policy is still based on tabular representations.

In the context of general linear function approximation, it is impossible to obtain a closed-form solution for $c^{\star}$. This raises the question: what can we infer about $c^{\star}$? Our intuition is as follows. We can envision the supplementary dataset containing two types of samples: some that were in-expert distribution, and others that were out-of-expert distribution. We expect that $w_h(s,a)$ is large in the former case and small in the latter case. Note that w_h is monotonic with respect to the inner product $\langle\phi_h(s,a),\theta\rangle$. Therefore, we conclude that a larger value of $\langle\phi_h(s,a),\theta\rangle$ implies a more significant contribution to the learning objective (3). In the following part, we demonstrate that the aforementioned intuition can be achieved under mild assumptions.

Assumption 2. *Let $\mathcal{D}^{\mathrm{S}}_h$ denote the set of state-action pairs in $\mathcal{D}^{\mathrm{S}}$ in h. Define $\mathcal{D}^{\mathrm{S},1}_h = \{(s,a) \in \mathcal{D}^{\mathrm{S}}_h : d^{\pi^{\mathrm{E}}}_h(s) > 0, a = \pi^{\mathrm{E}}_h(s)\}$ as the in-expert-distribution dataset in $\mathcal{D}^{\mathrm{S}}_h$ and $\mathcal{D}^{\mathrm{S},2}_h = \mathcal{D}^{\mathrm{S}}_h \setminus \mathcal{D}^{\mathrm{S},1}_h$ as the out-of-expert-distribution dataset. There exists a ground truth parameter $\bar{\theta}_h \in \mathbb{R}^d$, for any $(s,a) \in \mathcal{D}^{\mathrm{E}}_h \cup \mathcal{D}^{\mathrm{S},1}_h$ and $(s',a') \in \mathcal{D}^{\mathrm{S},2}_h$, it holds that $\langle \bar{\theta}_h, \phi_h(s,a) \rangle > 0$ and $\langle \bar{\theta}_h, \phi_h(s',a') \rangle < 0$.*

Readers may realize that Assumption 2 is closely related to the notion of "margin" in the classification problem. Define $\Delta_h(\theta) \triangleq \min_{(s,a)\in\mathcal{D}^{\mathrm{E}}_h\cup\mathcal{D}^{\mathrm{S},1}_h} \langle \theta, \phi_h(s,a)\rangle - \max_{(s',a')\in\mathcal{D}^{\mathrm{S},2}_h} \langle \theta, \phi_h(s',a')\rangle$. From Assumption 2, we have $\Delta_h(\bar{\theta}_h) > 0$. This means that there *exists* a classifier that recognizes samples from both $\mathcal{D}^{\mathrm{E}}_h$ and $\mathcal{D}^{\mathrm{S},1}_h$ as in-expert-distribution samples and samples from $\mathcal{D}^{\mathrm{S},2}_h$ as out-of-expert-distribution samples. Note that such a nice classifier is assumed to exist, which is not identical to what is learned via Equation (6). Before further discussion, we note that $\bar{\theta}_h$ is not unique if it exists. Without loss of generality, we define $\bar{\theta}_h$ as that can achieve the maximum margin (among all unit vectors).

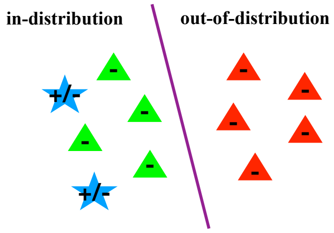

Figure 2. Illustration for ISW-BC.

Let us delve into the technical challenge. Although we assume two modes in the supplementary dataset, the learner is not aware of them beforehand. To gain a better understanding, refer to Figure 2, where the "star" corresponds to the expert data and the "triangle" corresponds to the supplementary data. The green and red parts of the triangle represent $\mathcal{D}^{\mathrm{S},1}$ and $\mathcal{D}^{\mathrm{S},2}$, respectively. While training the discriminator, we assign positive labels (shown in "+") to the expert data and negative labels (shown in "-") to the union data. Consequently, it becomes challenging to determine the learned decision boundary theoretically. To address this challenge, we develop the landscape properties, Lipschitz continuity and quadratic growth conditions, in Lemma 1 and Lemma 2, respectively. These terminologies are from the optimization literature (Karimi et al., 2016; Drusvyatskiy & Lewis, 2018). Incorporating these properties will aid in inferring the learned decision boundary.

Lemma 1. *For any $\theta \in \mathbb{R}^d$, the margin function is L_h-Lipschitz continuous in the sense that $\Delta_h(\bar{\theta}_h) - \Delta_h(\theta) \leq L_h \left\|\bar{\theta}_h - \theta\right\|$, where $L_h = \|\phi_h(s^1,a^1) - \phi_h(s^2,a^2)\|$ with $(s^1,a^1) \in \mathrm{argmin}_{(s,a)\in\mathcal{D}^{\mathrm{E}}_h\cup\mathcal{D}^{\mathrm{S},1}_h} \langle \theta, \phi_h(s,a)\rangle$ and $(s^2,a^2) \in \mathrm{argmax}_{(s,a)\in\mathcal{D}^{\mathrm{S},2}_h} \langle \theta, \phi_h(s,a)\rangle$.*

Lemma 2. *For any h, let $A_h \in \mathbb{R}^{N_{\mathrm{tot}}\times d}$ be the matrix that aggregates the feature vectors of samples in $\mathcal{D}^{\mathrm{U}}_h$. Assume that $\mathbf{rank}(A_h) = d$, then $\mathcal{L}_h$ has a (one-sided) quadratic growth condition. That is, there exists $\tau_h > 0$ such that $\mathcal{L}_h(\bar{\theta}_h) \geq \mathcal{L}_h(\theta^\star_h) + \frac{\tau_h}{2} \left\|\bar{\theta}_h - \theta^\star_h\right\|^2$.*

Using Lemma 1 and Lemma 2, we are ready to obtain the imitation gap bound of ISW-BC.

Theorem 3. *Under Assumptions 1 and 2, let $\mu = \max_{(s,h)\in\mathcal{S}\times[H]} d^{\pi^{\mathrm{E}}}_h(s,\pi^{\mathrm{E}}_h(s))/d^{\pi^{\beta}}_h(s,\pi^{\mathrm{E}}_h(s)) < \infty$, if the feature is designed such that $\sqrt{\frac{2\left(\mathcal{L}_h(\bar{\theta}_h)-\mathcal{L}_h(\theta^\star_h)\right)}{\tau_h}} < \frac{\Delta_h(\bar{\theta}_h)}{L_h}$ holds, then we have $\Delta_h(\theta^\star_h) > 0$. Furthermore, we have $\mathbb{E}[V(\pi^{\mathrm{E}}) - V(\pi^{\text{ISW-BC}})] = \mathcal{O}(\frac{H^2|\mathcal{S}|}{N_{\mathrm{E}}+N_{\mathrm{S}}/\mu})$.*

In order to interpret Theorem 3, it is important to note that $\Delta_h(\theta^\star_h) > 0$ means that there exists a $\delta > 0$ such that $w_h(s,a;\theta^\star_h) > \delta$ for $(s,a) \in \mathcal{D}^{\mathrm{E}}_h\cup\mathcal{D}^{\mathrm{S},1}_h$ and $w_h(s,a;\theta^\star_h) < \delta$ for $(s,a) \in \mathcal{D}^{\mathrm{S},2}_h$. As a result, all samples from $\mathcal{D}^{\mathrm{E}}_h$ and $\mathcal{D}^{\mathrm{S},1}_h$ are assigned with large weights, which allows ISW-BC to make use of additional samples and outperform BC.

We remark that the imitation gap bound of ISW-BC is dependent on the number of expert-style state-action pairs presented in the union of $\mathcal{D}^{\mathrm{E}}_h$ and $\mathcal{D}^{\mathrm{S},1}_h$. This number is represented as $N_{\mathrm{E}} + N_{\mathrm{S}}/\mu$, where μ is a state-action coverage parameter. It is important to mention that a similar notation is used in the literature of offline RL, as seen in (Munos & Szepesvári, 2008; Chen & Jiang, 2019). Additionally, ISW-BC has the ability to eliminate the gap of NBCU, meaning there is no non-vanishing error in Theorem 3. Moreover, ISW-BC can perform well even when mode-2 has noisy action labels, a scenario where NBCU may fail.

Although Theorem 3 produces desirable outcomes, it does have some limitations. First, the theoretical analysis necessitates knowledge of δ, which is typically challenging to determine beforehand. However, our empirical findings in Section 6 demonstrate that setting $\delta = 0$ is effective in practice. Second, Theorem 3 mandates the use of good smooth features to ensure the required inequality holds, thereby avoiding the undesirable case presented in Proposition 2. Our paper does not offer a solution for finding such feature representations. Nevertheless, our experiments indicate that neural networks can usually learn suitable features. We present a simple mathematical example corresponding to Theorem 3 in Appendix C.5. We defer more general results of ISW-BC to future work.

Table 2. Environment return of algorithms on 4 locomotion control tasks. Digits correspond to the mean performance over 5 random seeds and the subscript ± indicates the standard deviation. "Avg" computes the normalized score over environments. Same as the other tables.

		Ant	HalfCheetah	Hopper	Walker	Avg
	Random	-326	-280	-20	2	0%
	Expert	5229	11115	3589	5082	100%
	BC	1759$_{\pm 287}$	931$_{\pm 273}$	2468$_{\pm 164}$	1738$_{\pm 311}$	38%
Full Replay	NBCU	4932$_{\pm 148}$	10566$_{\pm 86}$	3241$_{\pm 276}$	4462$_{\pm 105}$	92%
	DemoDICE	**5000**$_{\pm 124}$	10781$_{\pm 67}$	3394$_{\pm 93}$	**4537**$_{\pm 125}$	**94%**
	DWBC	2951$_{\pm 155}$	1485$_{\pm 377}$	2567$_{\pm 88}$	1572$_{\pm 225}$	44%
	ISW-BC	4933$_{\pm 110}$	**10786**$_{\pm 56}$	**3434**$_{\pm 38}$	4475$_{\pm 164}$	**94%**
Noisy Expert	NBCU	3259$_{\pm 159}$	5561$_{\pm 539}$	558$_{\pm 23}$	518$_{\pm 56}$	35%
	DemoDICE	2523$_{\pm 244}$	6020$_{\pm 346}$	1990$_{\pm 90}$	1685$_{\pm 160}$	49%
	DWBC	**3270**$_{\pm 238}$	5688$_{\pm 557}$	**3317**$_{\pm 59}$	1985$_{\pm 175}$	62%
	ISW-BC	3075$_{\pm 268}$	**9284**$_{\pm 346}$	2624$_{\pm 249}$	**2859**$_{\pm 407}$	**69%**

6. Experiments

To validate the theoretical claims, we perform numerical experiments. We provide a brief overview of the experiment set-up below, and the details can be found in Appendix G due to space constraints.

6.1. Locomotion Control

In this section, we present our experiment on locomotion control, where we train a robot to run like a human in four environments from the Gym MuJoCo suite (Duan et al., 2016): `Ant`, `Hopper`, `Halfcheetah`, and `Walker`. We adopt online SAC (Haarnoja et al., 2018) to train an agent for each environment with 1M steps, and consider the resultant policy as the expert. For each environment, the expert data contains 1 trajectory collected by the expert policy. We consider two types of supplementary datasets:

- `Full Replay` (small distribution shift): the supplementary dataset (1 million samples) is directly sampled from the experience replay buffer of the online SAC agent, which is suggested by (Kim et al., 2022b). This setting has a small distribution shift as the online agent quickly converges to the expert policy (see Figure 5 in the Appendix), resulting in abundant expert trajectories in the replay buffer.

- `Noisy Expert` (large distribution shift): the supplementary dataset consists of 10 clean expert trajectories and 5 noisy expert trajectories where the action labels are corrupted (i.e., replaced by random actions). This introduces a large state-action distribution shift. For further discussion on dataset corruption and distribution shift, please refer to Appendix D.2.

Besides our proposed methods, we also evaluate two state-of-the-art methods in the locomotion control domain: DemoDICE (Kim et al., 2022b) and DWBC (Xu et al., 2022a). Please refer to Appendix G.1.1 for more experiment details.

We report the experiment results in Table 2. We observe that BC suffers since the amount of expert data is limited. In the full replay task, NBCU performs well due to the small distribution shift. However, in the noisy expert task, our results show that NBCU performs worse than BC, while ISW-BC outperforms NBCU significantly, demonstrating the robustness of ISW-BC to distribution shift. Note that among all evaluated methods, only our proposed method ISW-BC (with consistent parameters) performs well in both settings. Prior methods such as DemoDICE and DWBC only perform well in one of the two settings.

6.2. Atari Games

In this section, we evaluate algorithms on Atari games (Bellemare et al., 2013), which involve video frames as inputs and discrete controls as outputs. Furthermore, environment transitions are stochastic for these games. We consider 5 games, namely `Alien`, `MsPacman`, `Phoenix`, `Qbert`, and `SpaceInvaders`. We obtain the offline expert data and supplementary data from the replay buffer of an online DQN agent, as released by (Agarwal et al., 2020b). We use the expert data from the buffer with the last index, which only contains 50k frames, to create a challenging learning setting. To augment this data, we use earlier replay buffer data to obtain supplementary data with approximately 200k frames. We consider the same baselines as in Section 6.1. All methods build on the classical convolutional neural networks used in DQN.

Similar to Section 6.1, we consider two types of supplementary data. The `full replay` setting involves supplementary data that is close to the expert data, exhibiting a small distribution shift. The `noisy expert` setting has noisy action labels, leading to a large distribution shift. Experi-

Table 3. Environment return of algorithms on 5 Atari games.

		Alien	MsPacman	Phoenix	Qbert	SpaceInvaders	Avg
	Random	-228	307	761	164	148	0%
	Expert	2443	3601	4869	10955	1783	100%
	BC	$1051_{\pm 21}$	$1799_{\pm 27}$	$1520_{\pm 56}$	$4769_{\pm 111}$	$472_{\pm 10}$	32%
Full Replay	NBCU	$1405_{\pm 28}$	$2089_{\pm 48}$	$\mathbf{2431}_{\pm 104}$	$\mathbf{8065}_{\pm 109}$	$600_{\pm 13}$	**50%**
	DemoDICE	$1401_{\pm 16}$	$2146_{\pm 52}$	$2192_{\pm 72}$	$7820_{\pm 206}$	$558_{\pm 29}$	48%
	DWBC	$122_{\pm 4}$	$1251_{\pm 56}$	$583_{\pm 33}$	$1078_{\pm 50}$	$287_{\pm 6}$	7%
	ISW-BC	$\mathbf{1452}_{\pm 37}$	$\mathbf{2162}_{\pm 36}$	$2299_{\pm 76}$	$7848_{\pm 237}$	$\mathbf{613}_{\pm 16}$	**50%**
Noisy Expert	NBCU	$944_{\pm 22}$	$1378_{\pm 30}$	$1491_{\pm 55}$	$4366_{\pm 458}$	$418_{\pm 14}$	27%
	DemoDICE	$1054_{\pm 38}$	$1604_{\pm 59}$	$1448_{\pm 112}$	$\mathbf{5354}_{\pm 295}$	$395_{\pm 10}$	31%
	DWBC	$643_{\pm 18}$	$656_{\pm 16}$	$1165_{\pm 87}$	$3860_{\pm 104}$	$296_{\pm 5}$	16%
	ISW-BC	$\mathbf{1122}_{\pm 28}$	$\mathbf{1980}_{\pm 51}$	$\mathbf{1618}_{\pm 51}$	$5247_{\pm 328}$	$\mathbf{497}_{\pm 6}$	**36%**

ment details can be found in Appendix G.1.2. We report the game scores in Table 3. Our observations are consistent with those of the previous experiments. NBCU performs well when the distribution shift is small, while only ISW-BC is robust when the distribution shift is large.

6.3. Object Recognition

In our final experiment, we tackle an object recognition task that involves also image inputs. This task is a special type of imitation learning where the planning horizon is 1 and there are no environment transitions. The reward is classification accuracy. Please note that our main purpose here is to use the degraded one-step tasks to verify the theoretical results.

We use a famous dataset, DomainNet (Peng et al., 2019), which comprises 6 sub-datasets (clipart, infograph, painting, quickdraw, real, and sketch) that have different feature patterns and hence distribution shifts; see Figure 3 for an illustration. Following (Hong et al., 2022), our task is to perform 10-class object recognition (bird, feather, headphones, ice_cream, teapot, tiger, whale, windmill, wine_glass, and zebra) using 80% of the images for training and 20% for test. Each sub-dataset has roughly 2000-5000 images.

We build the classifier on the pretrained ResNet-18 (He et al., 2016), as directly training ResNet-18 on the DomainNet dataset failed. We then optimize a 2-hidden-layer neural network, where inputs are from the feature representations extracted by the pretrained and fixed ResNet-18. We create 6 sub-tasks, where one of the 6 sub-datasets is used as the expert data while the other 5 sub-datasets are used as the supplementary datasets. We evaluate the classification accuracy on the expert test data. Note that there is no natural extension of DemoDICE for this task. More details can be found in Appendix G.1.3.

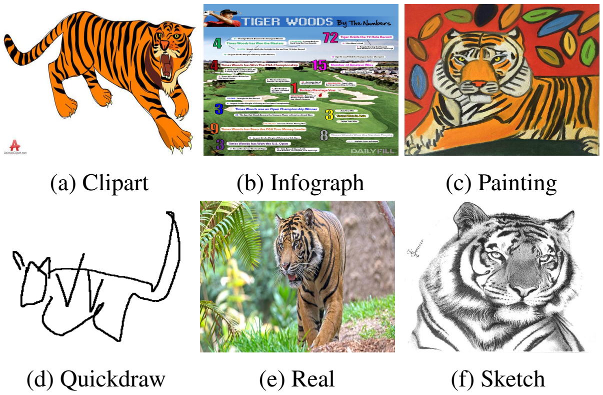
(a) Clipart (b) Infograph (c) Painting
(d) Quickdraw (e) Real (f) Sketch

Figure 3. Samples of tiger class from 6 sub-datasets of the DomainNet (Peng et al., 2019) dataset. Infograph and quickdraw have different patterns compared with the others.

The results of our experiment are presented in Table 4. We observe that due to the presence of distribution shift, NBCU performs even worse than BC. On the other hand, ISW-BC can improve the performance on 5 out of 6 tasks by re-weighting the supplementary data.

7. Conclusion

In this paper, we investigate the imitation learning with supplementary data framework, which aims to address the expert data scarcity issue. We provide imitation gap bounds for three representative algorithms: BC, which relies solely on expert data, NBCU, which utilizes supplementary data without selection, and ISW-BC, a newly developed method that addresses distribution shift by employing importance sampling. Through theoretical analysis and empirical evaluations, we have shown that ISW-BC outperforms the other methods in robustness to distribution shift and effectiveness in utilizing supplementary data.

Future work could explore how our approach can be extended to other imitation learning algorithms and how it can be used in conjunction with other data-centric techniques

Table 4. Test classification accuracy (%) of algorithms on 6 types of expert and supplementary data.

	Clipart	Infograph	Painting	Quickdraw	Real	Sketch	Avg
BC	$89.31_{\pm 0.01}$	$55.80_{\pm 0.01}$	$90.14_{\pm 0.00}$	$\mathbf{85.61}_{\pm 0.01}$	$96.19_{\pm 0.00}$	$87.58_{\pm 0.01}$	84.10
NBCU	$89.16_{\pm 0.02}$	$56.32_{\pm 0.02}$	$88.29_{\pm 0.01}$	$84.78_{\pm 0.03}$	$95.31_{\pm 0.00}$	$87.57_{\pm 0.00}$	83.57
DWBC	$90.00_{\pm 0.09}$	$57.44_{\pm 0.06}$	$90.89_{\pm 0.04}$	$85.09_{\pm 0.09}$	$96.35_{\pm 0.01}$	$88.86_{\pm 0.09}$	84.77
ISW-BC	$\mathbf{90.86}_{\pm 0.00}$	$\mathbf{57.52}_{\pm 0.01}$	$\mathbf{91.78}_{\pm 0.01}$	$84.97_{\pm 0.01}$	$\mathbf{96.56}_{\pm 0.01}$	$\mathbf{89.63}_{\pm 0.06}$	**85.22**

(see e.g., (Polyzotis & Zaharia, 2021; Whang et al., 2023; Zha et al., 2023)) for improving imitation learning performance. Overall, our findings demonstrate the potential of using supplementary data to enhance imitation, and we hope this work can inspire further advances.

References

Agarwal, A., Kakade, S., Krishnamurthy, A., and Sun, W. Flambe: Structural complexity and representation learning of low rank mdps. *Advances in Neural Information Processing Systems 33*, pp. 20095–20107, 2020a.

Agarwal, R., Schuurmans, D., and Norouzi, M. An optimistic perspective on offline reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 104–114, 2020b.

Argall, B. D., Chernova, S., Veloso, M., and Browning, B. A survey of robot learning from demonstration. *Robotics and Autonomous Systems*, 57(5):469–483, 2009.

Arora, S., Ge, R., Liang, Y., Ma, T., and Zhang, Y. Generalization and equilibrium in generative adversarial nets (gans). In *Proceedings of the 34th International Conference on Machine Learning*, pp. 224–232, 2017.

Bellemare, M. G., Naddaf, Y., Veness, J., and Bowling, M. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

Brown, D., Goo, W., Nagarajan, P., and Niekum, S. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 783–792, 2019.

Chang, J., Uehara, M., Sreenivas, D., Kidambi, R., and Sun, W. Mitigating covariate shift in imitation learning via offline data with partial coverage. In *Advances in Neural Information Processing Systems 34*, pp. 965–979, 2021.

Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 1042–1051, 2019.

Cortes, C., Mansour, Y., and Mohri, M. Learning bounds for importance weighting. In *Advances in Neural Information Processing Systems 23*, pp. 442–450, 2010.

Diamond, S. and Boyd, S. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Drusvyatskiy, D. and Lewis, A. S. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018.

Duan, Y., Chen, X., Houthooft, R., Schulman, J., and Abbeel, P. Benchmarking deep reinforcement learning for continuous control. In *Proceedings of the 33nd International Conference on Machine Learning*, pp. 1329–1338, 2016.

Fang, T., Lu, N., Niu, G., and Sugiyama, M. Rethinking importance weighting for deep learning under distribution shift. In *Advances in Neural Information Processing Systems 33*, pp. 11996–12007, 2020.

Ghasemipour, S. K. S., Zemel, R. S., and Gu, S. A divergence minimization perspective on imitation learning methods. In *Proceedings of the 3rd Annual Conference on Robot Learning*, pp. 1259–1277, 2019.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pp. 2672–2680, 2014.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1856–1865, 2018.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Ho, J. and Ermon, S. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems 29*, pp. 4565–4573, 2016.

Hong, J., Lyu, L., Zhou, J., and Spranger, M. Outsourcing training without uploading data via efficient collaborative open-source sampling. In *Advances in Neural Information Processing Systems 35*, pp. 20133–20146, 2022.

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Proceeddings of The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pp. 795–811, 2016.

Kim, G.-H., Lee, J., Jang, Y., Yang, H., and Kim, K.-E. Lobsdice: Offline learning from observation via stationary distribution correction estimation. In *Advances in Neural Information Processing Systems 35*, pp. 8252–8264, 2022a.

Kim, G.-H., Seo, S., Lee, J., Jeon, W., Hwang, H., Yang, H., and Kim, K.-E. DemoDICE: Offline imitation learning with supplementary imperfect demonstrations. In *Proceedings of the 10th International Conference on Learning Representations*, 2022b.

Kostrikov, I., Agrawal, K. K., Dwibedi, D., Levine, S., and Tompson, J. Discriminator-actor-critic: Addressing sample inefficiency and reward bias in adversarial imitation learning. In *Proceedings of the 7th International Conference on Learning Representations*, 2019.

Li, Z., Xu, T., Yu, Y., and Luo, Z.-Q. Rethinking valuedice: Does it really improve performance? *arXiv preprint arXiv:2202.02468*, 2022.

Liu, L., Tang, Z., Li, L., and Luo, D. Robust imitation learning from corrupted demonstrations. *arXiv preprint arXiv:2201.12594*, 2022.

Liu, T. and Tao, D. Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):447–461, 2015.

Ma, Y. J., Shen, A., Jayaraman, D., and Bastani, O. Smodice: Versatile offline imitation learning via state occupancy matching. In *Prooceedings of the 39th International Conference on Machine Learning*, pp. 14639–14663, 2022.

Munos, R. and Szepesvári, C. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9 (5), 2008.

Osa, T., Pajarinen, J., Neumann, G., Bagnell, J. A., Abbeel, P., and Peters, J. An algorithmic perspective on imitation learning. *Foundations and Trends in Robotic*, 7(1-2): 1–179, 2018.

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., and Wang, B. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1406–1415, 2019.

Polyzotis, N. and Zaharia, M. What can data-centric ai learn from data and ml engineering? *arXiv*, 2112.06439, 2021.

Pomerleau, D. Efficient training of artificial neural networks for autonomous navigation. *Neural Computation*, 3(1): 88–97, 1991.

Puterman, M. L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

Rajaraman, N., Yang, L. F., Jiao, J., and Ramchandran, K. Toward the fundamental limits of imitation learning. In *Advances in Neural Information Processing Systems 33*, pp. 2914–2924, 2020.

Rajaraman, N., Han, Y., Yang, L., Liu, J., Jiao, J., and Ramchandran, K. On the value of interaction and function approximation in imitation learning. In *Advances in Neural Information Processing Systems 34*, pp. 1325–1336, 2021a.

Rajaraman, N., Han, Y., Yang, L. F., Ramchandran, K., and Jiao, J. Provably breaking the quadratic error compounding barrier in imitation learning, optimally. *arXiv preprint arXiv: 2102.12948*, 2021b.

Ross, S. and Bagnell, D. Efficient reductions for imitation learning. In *Proceedings of the 13rd International Conference on Artificial Intelligence and Statistics*, pp. 661–668, 2010.

Sasaki, F. and Yamashina, R. Behavioral cloning from noisy demonstrations. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.

Shani, L., Zahavy, T., and Mannor, S. Online apprenticeship learning. In *Proceedings of th 36th AAAI Conference on Artificial Intelligence*, pp. 8240–8248, 2022.

Shapiro, A. Monte carlo sampling methods. *Handbooks in operations research and management science*, 10:353–425, 2003.

Spencer, J., Choudhury, S., Venkatraman, A., Ziebart, B., and Bagnell, J. A. Feedback in imitation learning: The three regimes of covariate shift. *arXiv preprint arXiv:2102.02872*, 2021.

Sugiyama, M., Nakajima, S., Kashima, H., von Bünau, P., and Kawanabe, M. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems 20*, pp. 1433–1440, 2007.

Tangkaratt, V., Han, B., Khan, M. E., and Sugiyama, M. Variational imitation learning with diverse-quality demonstrations. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 9407–9417, 2020.

Wang, Y., Xu, C., Du, B., and Lee, H. Learning to weight imperfect demonstrations. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 10961–10970, 2021.

Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., and Weinberger, M. J. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Techical Report*, 2003.

Whang, S. E., Roh, Y., Song, H., and Lee, J.-G. Data collection and quality challenges in deep learning: A data-centric ai perspective. *The VLDB Journal*, pp. 1–23, 2023.

Wu, Y.-H., Charoenphakdee, N., Bao, H., Tangkaratt, V., and Sugiyama, M. Imitation learning from imperfect demonstration. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 6818–6827, 2019.

Xie, T., Jiang, N., Wang, H., Xiong, C., and Bai, Y. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in Neural Information Processing Systems 34*, pp. 27395–27407, 2021.

Xu, H., Zhan, X., Yin, H., and Qin, H. Discriminator-weighted offline imitation learning from suboptimal demonstrations. In *Prooceedings of the 39th International Conference on Machine Learning*, pp. 24725–24742, 2022a.

Xu, T., Li, Z., and Yu, Y. Error bounds of imitating policies and environments. In *Advances in Neural Information Processing Systems 33*, pp. 15737–15749, 2020.

Xu, T., Li, Z., and Yu, Y. Error bounds of imitating policies and environments for reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Xu, T., Li, Z., Yu, Y., and Luo, Z.-Q. Understanding adversarial imitation learning in small sample regime: A stage-coupled analysis. *arXiv preprint arXiv:2208.01899*, 2022b.

Zha, D., Bhat, Z. P., Lai, K.-H., Yang, F., Jiang, Z., Zhong, S., and Hu, X. Data-centric artificial intelligence: A survey. *arXiv*, 2303.10158, 2023.

A. Additional Related Work

In contrast to BC, adversarial imitation learning (AIL) methods, such as GAIL (Ho & Ermon, 2016), perform imitation through state-action distribution matching. It has been demonstrated both empirically and theoretically that AIL methods do not suffer from the compounding errors issue when the expert data is limited (Ho & Ermon, 2016; Ghasemipour et al., 2019; Kostrikov et al., 2019; Xu et al., 2020). Under mild conditions, (Xu et al., 2022b) provided a horizon-free bound of $\mathcal{O}(\min\{1, \sqrt{|\mathcal{S}|/N_{\mathrm{E}}}\})$, which is much better than BC in terms of H. However, AIL methods work naturally in the online setting (i.e., the interaction is allowed), which is not directly applicable in the offline setting that we study in this paper. We comment that the direct application of AIL in the offline setting is not effective, as pointed in (Li et al., 2022). Although the proposed method has a discriminator and a policy like AIL, our discriminator and policy are not designed to compete with each other adversarially, as we have explained in detail in the main text.

Our work builds upon previous research in IL with supplementary data, specifically the algorithms DemoDICE (Kim et al., 2022b) and DWBC (Xu et al., 2022a). These studies highlight the importance of careful data selection when using a supplementary dataset. In this vein, our method ISW-BC re-weights samples based on importance sampling, which we show to be theoretically sound. Notably, a significant distinction arises between ISW-BC and these two methods in terms of the weighting rule design. While DemoDICE and DWBC employ *regularized* weighting rules, our method directly estimates the importance sampling ratio. This fundamental difference can be critical as regularized weighting rules may struggle to recover the expert policy exactly even with infinite samples. We provide further elaboration on this point below.

First, DemoDICE also uses the weighted BC objective in Equation (3). But, DemoDICE uses the weighting rule of $\widetilde{w}(s,a) \propto d^{\star}(s,a)/d^{\mathrm{U}}(s,a)$ (refer to the formula between Equations (19)-(20) in (Kim et al., 2022b)), where $d^{\star}(s,a)$ is computed by the expert's state-action distribution matching objective regularized by a divergence to the union data distribution (refer to (Kim et al., 2022b, Equations (5)-(7)))[2]:

$$
\begin{aligned}
d^{\star} = & \operatorname*{argmin}_{d} D_{\mathrm{KL}}(d \| d^{\mathrm{E}}) + \alpha D_{\mathrm{KL}}(d \| d^{\mathrm{U}}) \\
\text{s.t.} \quad & d(s,a) \geq 0 \quad \forall s, a. \\
& \sum_{a} d(s,a) = (1-\gamma)\rho(s) + \gamma \sum_{s', a'} P(s|s', a') d(s', a') \quad \forall s.
\end{aligned}
$$

where $\gamma \in [0, 1)$ is the discount factor, $\alpha > 0$ is a hyper-parameter. Due to the regularization term in the objective, it holds that $d^{\star}(s,a) \neq d^{\pi^{\mathrm{E}}}(s,a)$, resulting in a biased weighting rule $\widetilde{w}(s,a)$.

Second, DWBC considers a different policy learning objective (refer to (Xu et al., 2022a, Equation (17))):

$$
\begin{aligned}
\min_{\pi} \quad & \alpha \sum_{(s,a) \in \mathcal{D}^{\mathrm{E}}} \left[-\log \pi(a|s) \right] - \sum_{(s,a) \in \mathcal{D}^{\mathrm{E}}} \left[-\log \pi(a|s) \cdot \frac{\lambda}{c(1-c)} \right] \\
& + \sum_{(s,a) \in \mathcal{D}^{\mathrm{S}}} \left[-\log \pi(a|s) \cdot \frac{1}{1-c} \right],
\end{aligned} \tag{7}
$$

where $\alpha > 0, \lambda > 0$ are hyper-parameters, and c is the output of the discriminator that is jointly trained with π (refer to (Xu et al., 2022a, Equation (8))):

$$
\begin{aligned}
\min_{c} \quad & \lambda \sum_{(s,a) \in \mathcal{D}^{\mathrm{E}}} \left[-\log c(s,a,\log \pi(a|s)) \right] + \sum_{(s,a) \in \mathcal{D}^{\mathrm{S}}} \left[-\log(1 - c(s,a,\log \pi(a|s))) \right] \\
& - \lambda \sum_{(s,a) \in \mathcal{D}^{\mathrm{E}}} \left[-\log(1 - c(s,a,\log \pi(a|s))) \right].
\end{aligned}
$$

Since its input additionally incorporates $\log \pi$, the discriminator is not guaranteed to estimate the state-action distribution. Thus, the weighting in Equation (7) loses a connection with the importance sampling ratio.

In addition to our work, (Chang et al., 2021) have also explored the use of supplementary data in the offline setting. However, their approach (called MILO) is based on adversarial imitation learning. Specifically, MILO learns a transition model from the supplementary dataset and performs adversarial imitation learning within the learned model. In contrast, our proposed

[2] For a moment, we use the notations in (Kim et al., 2022b) and present their results under the stationary and infinite-horizon MDPs. Same as the discussion of DWBC (Xu et al., 2022a).

method, ISW-BC, tackles the challenge of scarce expert data by identifying and utilizing expert-style samples that are hidden within the supplementary dataset. MILO has an imitation gap bound of $\mathcal{O}(H\sqrt{\frac{|\mathcal{S}|}{N_{\mathrm{E}}}} + H^2|\mathcal{S}|\sqrt{\frac{|\mathcal{A}|}{N_{\mathrm{S}}/\mu}})$ in theory. However, MILO makes different assumptions about the data collection procedure compared with ISW-BC. Consequently, the imitation gap bounds of MILO and ISW-BC are incomparable.

The problem considered in this paper is related to IL with a single imperfect dataset (Wu et al., 2019; Brown et al., 2019; Tangkaratt et al., 2020; Wang et al., 2021; Sasaki & Yamashina, 2021; Liu et al., 2022). In particular, the supplementary dataset in our set-up can also be viewed as imperfect demonstrations. However, our problem setting differs from IL with imperfect demonstrations in two key aspects. First, in IL with imperfect demonstrations, they either pose strong assumptions (Tangkaratt et al., 2020; Sasaki & Yamashina, 2021; Liu et al., 2022) or require auxiliary information (e.g., confidence scores on imperfect trajectories) on the imperfect dataset (Wu et al., 2019; Brown et al., 2019). In contrast, we assume access to a small number of expert trajectories to identify in-expert-distribution data. Second, most works (Wu et al., 2019; Brown et al., 2019; Tangkaratt et al., 2020; Wang et al., 2021) in IL with imperfect demonstrations require online environment interactions while we focus on the offline setting.

B. Proof of Results in Section 4

Recall the objective of BC in Equation (1):

$$\pi^{\mathrm{BC}} \in \max_{\pi} \sum_{h=1}^{H} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \widehat{d_h^{\mathrm{E}}}(s,a) \log \pi_h(a|s),$$

where $\widehat{d_h^{\mathrm{E}}}(s,a) = n_h^{\mathrm{E}}(s,a)/N_{\mathrm{tot}}$ is the empirical state-action distribution in the expert dataset, and $n_h^{\mathrm{E}}(s,a)$ is the number of expert trajectories such that their state-action pairs are equal to (s,a) in time step h. With the tabular representations, we can obtain a closed-formed solution to the above optimization problem.

$$\pi_h^{\mathrm{BC}}(a|s) = \begin{cases} \frac{n_h^{\mathrm{E}}(s,a)}{n_h^{\mathrm{E}}(s)} & \text{if } n_h^{\mathrm{E}}(s) > 0 \\ \frac{1}{|\mathcal{A}|} & \text{otherwise} \end{cases} \tag{8}$$

where $n_h^{\mathrm{E}}(s) \triangleq \sum_{a'} n_h^{\mathrm{E}}(s,a')$. Analogously, we also have a closed-form solution for NBCU in the tabular setting:

$$\pi_h^{\mathrm{NBCU}}(a|s) = \begin{cases} \frac{n_h^{\mathrm{U}}(s,a)}{n_h^{\mathrm{U}}(s)} & \text{if } n_h^{\mathrm{U}}(s) > 0 \\ \frac{1}{|\mathcal{A}|} & \text{otherwise} \end{cases} \tag{9}$$

We will discuss the generalization performance of NBCU later.

In the proof, we frequently use the notation $\lesssim$ and $\gtrsim$. In particular, $a(n) \lesssim b(n)$ means that there exist $C, n_0 > 0$ such that $a(n) \leq Cb(n)$ for all $n \geq n_0$. In our context, n usually refers to the number of trajectories. For any two distributions P and Q over a finite set $\mathcal{X}$, we define the total variation distance as

$$\mathrm{TV}(P,Q) = \frac{1}{2}\sum_{x\in\mathcal{X}} |P(x) - Q(x)| = \|P - Q\|_1.$$

B.1. Proof of Theorem 1

When $|\mathcal{D}^{\mathrm{E}}| \geq 1$, by (Rajaraman et al., 2020, Theorem 4.2), we have the following imitation gap bound for BC:

$$V(\pi^{\mathrm{E}}) - \mathbb{E}_{\mathcal{D}^{\mathrm{E}}}\left[V(\pi^{\mathrm{BC}})\right] \leq \frac{4|\mathcal{S}|H^2}{9|\mathcal{D}^{\mathrm{E}}|}.$$

When $|\mathcal{D}^{\mathrm{E}}| = 0$, we simply have that

$$V(\pi^{\mathrm{E}}) - \mathbb{E}_{\mathcal{D}^{\mathrm{E}}}\left[V(\pi^{\mathrm{BC}})\right] \leq H.$$

Therefore, we have the following unified bound.

$$V(\pi^{\mathrm{E}}) - \mathbb{E}_{\mathcal{D}^{\mathrm{E}}}\left[V(\pi^{\mathrm{BC}})\right] \leq \frac{|\mathcal{S}|H^2}{\max\{|\mathcal{D}^{\mathrm{E}}|, 1\}} \leq \frac{2|\mathcal{S}|H^2}{|\mathcal{D}^{\mathrm{E}}| + 1}.$$

The last inequality follows that $\max\{x, 1\} \geq (x+1)/2$ for any $x \geq 0$. Finally, notice that $|\mathcal{D}^{\mathrm{E}}|$ follows a binomial distribution by Assumption 1, i.e., $|\mathcal{D}^{\mathrm{E}}| \sim \mathrm{Bin}(N_{\mathrm{tot}}, \eta)$. By Lemma 3, we have that $\mathbb{E}[1/(|\mathcal{D}|^{E}+1)] \leq N_{\mathrm{tot}}\eta$, so

$$V(\pi^{\mathrm{E}}) - \mathbb{E}\left[V(\pi^{\mathrm{BC}})\right] \leq \mathbb{E}\left[\frac{2|\mathcal{S}|H^2}{|\mathcal{D}^{\mathrm{E}}|+1}\right] \leq \frac{2|\mathcal{S}|H^2}{N_{\mathrm{tot}}\eta} = \frac{2|\mathcal{S}|H^2}{N_{\mathrm{E}}},$$

which completes the proof.

B.2. Proof of Theorem 2

For analysis, we first define the mixture state-action distribution as follows.

$$d_h^{\mathrm{mix}}(s,a) \triangleq \eta d_h^{\pi^{\mathrm{E}}}(s,a) + (1-\eta) d_h^{\pi^{\beta}}(s,a),$$
$$d_h^{\mathrm{mix}}(s) \triangleq \sum_{a \in \mathcal{A}} d_h^{\mathrm{mix}}(s,a),\ \forall (s,a) \in \mathcal{S} \times \mathcal{A},\ \forall h \in [H].$$

By Assumption 1, in the population level, the marginal state-action distribution of union dataset $\mathcal{D}^{\mathrm{U}}$ in time step h is exactly d_h^{mix}. That is, $d_h^{\mathrm{U}}(s,a) = d_h^{\mathrm{mix}}(s,a),\ \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$. Then we define the mixture policy π^{mix} induced by d^{mix} as follows.

$$\pi_h^{\mathrm{mix}}(a|s) = \begin{cases} \frac{d_h^{\mathrm{mix}}(s,a)}{d_h^{\mathrm{mix}}(s)} & \text{if } d_h^{\mathrm{mix}}(s) > 0, \\ \frac{1}{|\mathcal{A}|} & \text{otherwise.} \end{cases} \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}\ , \forall h \in [H]. \tag{10}$$

From the theory of Markov Decision Processes, we know that (see, e.g., (Puterman, 2014))

$$\forall h \in [H], \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \quad d_h^{\pi^{\mathrm{mix}}}(s,a) = d_h^{\mathrm{mix}}(s,a).$$

Therefore, we can obtain that the marginal state-action distribution of union dataset $\mathcal{D}^{\mathrm{U}}$ in time step h is exactly $d_h^{\pi^{\mathrm{mix}}}$. Then we have the following decomposition.

$$\begin{aligned}\mathbb{E}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{NBCU}})\right] &= \mathbb{E}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{mix}}) + V(\pi^{\mathrm{mix}}) - V(\pi^{\mathrm{NBCU}})\right] \\ &= \mathbb{E}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{mix}})\right] + \mathbb{E}\left[V(\pi^{\mathrm{mix}}) - V(\pi^{\mathrm{NBCU}})\right] \\ &= V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{mix}}) + \mathbb{E}\left[V(\pi^{\mathrm{mix}}) - V(\pi^{\mathrm{NBCU}})\right].\end{aligned}$$

For $V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{mix}})$, we have that

$$\begin{aligned}V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{mix}}) &= \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left(d_h^{\pi^{\mathrm{E}}}(s,a) - d_h^{\pi^{\mathrm{mix}}}(s,a)\right) r_h(s,a) \\ &= \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left(d_h^{\pi^{\mathrm{E}}}(s,a) - d_h^{\mathrm{mix}}(s,a)\right) r_h(s,a) \\ &= (1-\eta) \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left(d_h^{\pi^{\mathrm{E}}}(s,a) - d_h^{\pi^{\beta}}(s,a)\right) r_h(s,a) \\ &= (1-\eta)\left(V(\pi^{\mathrm{E}}) - V(\pi^{\beta})\right).\end{aligned} \tag{11}$$

The last equation follows the dual formulation of policy value (see, e.g., (Puterman, 2014)), i.e., $V(\pi) = \sum_{h=1}^{H} \sum_{(s,a)} d_h^{\pi}(s,a) r_h(s,a)$ for any policy π. Besides, notice that $\mathbb{E}\left[V(\pi^{\mathrm{mix}}) - V(\pi^{\mathrm{NBCU}})\right]$ is exactly the imitation gap of BC when regarding π^{mix} and $\mathcal{D}^{\mathrm{U}}$ as the expert policy and expert dataset, respectively. Note that π^{mix} may be a stochastic policy. By (Rajaraman et al., 2020, Theorem 4.4), we have the following imtiation gap bound

$$\mathbb{E}\left[V(\pi^{\mathrm{mix}}) - V(\pi^{\mathrm{NBCU}})\right] \lesssim \frac{|\mathcal{S}|H^2 \log(N_{\mathrm{tot}})}{N_{\mathrm{tot}}}. \tag{12}$$

Combining Equation (11) and Equation (12) yields that

$$\mathbb{E}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{NBCU}})\right] \lesssim (1-\eta)\left(V(\pi^{\mathrm{E}}) - V(\pi^{\beta})\right) + \frac{|\mathcal{S}|H^2 \log(N_{\mathrm{tot}})}{N_{\mathrm{tot}}}.$$

B.3. Proof of Proposition 1

The hard instance in Proposition 1 builds on the Standard Imitation MDP proposed in (Xu et al., 2021); see Figure 4 for illustration. For this MDP, each state is an absorbing state, i.e., $P_h(s|s,a)=1$ for any s and a. This property is mainly used to facilitate probability calculation and does not change the nature of our analysis. Furthermore, by only taking the action a^1 (shown in green), the agent can obtain a reward of $+1$. Otherwise, the agent obtains a reward of 0 for the other action $a \neq a^1$. The initial state distribution is a uniform distribution, i.e., $\rho(s)=1/|\mathcal{S}|$ for any $s\in\mathcal{S}$.

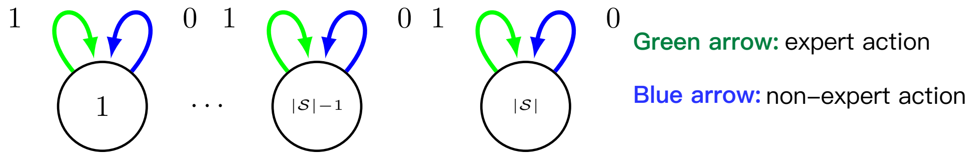

Figure 4. The Standard Imitation MDP in (Xu et al., 2021) corresponding to prove Proposition 1.

We consider that the expert policy π^{E} always takes the action a^1 (shown in green) while the behavioral policy π^{β} always takes another action a^2 (shown in blue). Formally, $\pi_h^{\mathrm{E}}(a^1|s)=1$ and $\pi_h^{\beta}(a^2|s)=1$ for any $s\in\mathcal{S}$ and $h\in[H]$. It is direct to calculate that $V(\pi^{\mathrm{E}})=H$ and $V(\pi^{\beta})=0$. The supplementary dataset $\mathcal{D}^{\mathrm{S}}$ and the expert dataset $\mathcal{D}^{\mathrm{E}}$ are collected according to Assumption 1. The mixture state-action distribution (introduced in Appendix B.2) can be calculated as for any $s\in\mathcal{S}$ and $h\in[H]$:

$$
\begin{aligned}
d_h^{\mathrm{mix}}(s,a^1) &= \eta d_h^{\pi^{\mathrm{E}}}(s,a^1)+(1-\eta)d_h^{\pi^{\beta}}(s,a^1)=\eta d_h^{\pi^{\mathrm{E}}}(s,a^1)=\eta\rho(s),\\
d_h^{\mathrm{mix}}(s,a^2) &= \eta d_h^{\pi^{\mathrm{E}}}(s,a^2)+(1-\eta)d_h^{\pi^{\beta}}(s,a^2)=(1-\eta)d_h^{\pi^{\beta}}(s,a^2)=(1-\eta)\rho(s).
\end{aligned}
$$

Note that in the population level, the marginal distribution of the union dataset $\mathcal{D}^{\mathrm{U}}$ in time step h is exactly d_h^{mix}. The mixture policy induced by d^{mix} (introduced in Appendix B.2) can be formulated as

$$
\pi_h^{\mathrm{mix}}(a^1|s)=\eta, \pi_h^{\mathrm{mix}}(a^2|s)=1-\eta, \forall s\in\mathcal{S}, h\in[H].
$$

Just like before, we have $d_h^{\pi^{\mathrm{mix}}}(s,a)=d_h^{\mathrm{mix}}(s,a)$. The policy value of π^{mix} can be calculated as

$$
V(\pi^{\mathrm{mix}})=\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h^{\mathrm{mix}}(s,a)r_h(s,a)=\sum_{h=1}^{H}\sum_{s\in\mathcal{S}} d_h^{\mathrm{mix}}(s,a^1)=\eta H.
$$

Recall from Equation (9) that π^{NBCU} can be formulated as

$$
\forall h\in[H],\quad \pi_h^{\mathrm{NBCU}}(a|s)=\begin{cases}\frac{n_h^{\mathrm{U}}(s,a)}{\sum_{a'} n_h^{\mathrm{U}}(s,a')} & \text{if} \sum_{a'} n_h^{\mathrm{U}}(s,a')>0\\ \frac{1}{|\mathcal{A}|} & \text{otherwise}\end{cases} \tag{13}
$$

We can view that the BC's policy learned on the union dataset mimics the mixture policy π^{mix}. In the following part, we analyze the lower bound on the imitation gap of π^{NBCU}.

$$
\begin{aligned}
\mathbb{E}\left[V(\pi^{\mathrm{E}})-V(\pi^{\mathrm{NBCU}})\right] &= V(\pi^{\mathrm{E}})-V(\pi^{\mathrm{mix}})+\mathbb{E}\left[V(\pi^{\mathrm{mix}})-V(\pi^{\mathrm{NBCU}})\right]\\
&= H-\eta H+\mathbb{E}\left[V(\pi^{\mathrm{mix}})-V(\pi^{\mathrm{NBCU}})\right]\\
&= (1-\eta)(V(\pi^{\mathrm{E}})-V(\pi^{\beta}))+\mathbb{E}\left[V(\pi^{\mathrm{mix}})-V(\pi^{\mathrm{NBCU}})\right].
\end{aligned}
$$

Then we consider the term $\mathbb{E}\left[V(\pi^{\mathrm{mix}})-V(\pi^{\mathrm{NBCU}})\right]$.

$$
\begin{aligned}
&V(\pi^{\mathrm{mix}})-V(\pi^{\mathrm{NBCU}})\\
&=\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left(d_h^{\pi^{\mathrm{mix}}}(s,a)-d_h^{\pi^{\mathrm{NBCU}}}(s,a)\right)r_h(s,a)\\
&=\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\rho(s)\left(\pi_h^{\mathrm{mix}}(a|s)-\pi_h^{\mathrm{NBCU}}(a|s)\right)r_h(s,a)
\end{aligned}
$$

$$= \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \rho(s) \left(\pi_h^{\mathrm{mix}}(a|s) - \pi_h^{\mathrm{NBCU}}(a|s) \right) r_h(s,a) \mathbb{I}\{n_h^{\mathrm{U}}(s) > 0\}$$
$$+ \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \rho(s) \left(\pi_h^{\mathrm{mix}}(a|s) - \pi_h^{\mathrm{NBCU}}(a|s) \right) r_h(s,a) \mathbb{I}\{n_h^{\mathrm{U}}(s) = 0\}.$$

We take expectation over the randomness in $\mathcal{D}^{\mathrm{U}}$ on both sides and obtain that

$$\mathbb{E}\left[V(\pi^{\mathrm{mix}}) - V(\pi^{\mathrm{NBCU}}) \right] \tag{14}$$
$$= \mathbb{E}\left[\sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \rho(s) \left(\pi_h^{\mathrm{mix}}(a|s) - \pi_h^{\mathrm{NBCU}}(a|s) \right) r_h(s,a) \mathbb{I}\{n_h^{\mathrm{U}}(s) > 0\} \right]$$
$$+ \mathbb{E}\left[\sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \rho(s) \left(\pi_h^{\mathrm{mix}}(a|s) - \pi_h^{\mathrm{NBCU}}(a|s) \right) r_h(s,a) \mathbb{I}\{n_h^{\mathrm{U}}(s) = 0\} \right]. \tag{15}$$

For the first term in RHS, we have that

$$\mathbb{E}\left[\sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \rho(s) \left(\pi_h^{\mathrm{mix}}(a|s) - \pi_h^{\mathrm{NBCU}}(a|s) \right) r_h(s,a) \mathbb{I}\{n_h^{\mathrm{U}}(s) > 0\} \right]$$
$$= \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \rho(s) r_h(s,a) \mathbb{E}\left[\left(\pi_h^{\mathrm{mix}}(a|s) - \pi_h^{\mathrm{NBCU}}(a|s) \right) \mathbb{I}\{n_h^{\mathrm{U}}(s) > 0\} \right]$$
$$= \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \rho(s) r_h(s,a) \mathbb{P}\left(n_h^{\mathrm{U}}(s) > 0 \right) \mathbb{E}\left[\pi_h^{\mathrm{mix}}(a|s) - \pi_h^{\mathrm{NBCU}}(a|s) \mid n_h^{\mathrm{U}}(s) > 0 \right]$$
$$= 0.$$

The last equation follows the fact that $\pi_h^{\mathrm{NBCU}}(a|s)$ is an unbiased estimation of $\pi_h^{\mathrm{mix}}(a|s)$, so $\mathbb{E}[\pi_h^{\mathrm{mix}}(a|s) - \pi_h^{\mathrm{NBCU}}(a|s) \mid n_h^{\mathrm{U}}(s) > 0]$. For the second term in Equation (15), we have that

$$\mathbb{E}\left[\sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \rho(s) \left(\pi_h^{\mathrm{mix}}(a|s) - \pi_h^{\mathrm{NBCU}}(a|s) \right) r_h(s,a) \mathbb{I}\{n_h^{\mathrm{U}}(s) = 0\} \right]$$
$$= \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \rho(s) r_h(s,a) \mathbb{E}\left[\left(\pi_h^{\mathrm{mix}}(a|s) - \pi_h^{\mathrm{NBCU}}(a|s) \right) \mathbb{I}\{n_h^{\mathrm{U}}(s) = 0\} \right]$$
$$= \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \rho(s) r_h(s,a) \mathbb{P}\left(n_h^{\mathrm{U}}(s) = 0 \right) \mathbb{E}\left[\pi_h^{\mathrm{mix}}(a|s) - \pi_h^{\mathrm{NBCU}}(a|s) \mid n_h^{\mathrm{U}}(s) = 0 \right]$$
$$= \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \rho(s) r_h(s,a) \mathbb{P}\left(n_h^{\mathrm{U}}(s) = 0 \right) \left(\pi_h^{\mathrm{mix}}(a|s) - \frac{1}{|\mathcal{A}|} \right)$$
$$\overset{(a)}{=} \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} \rho(s) \mathbb{P}\left(n_h^{\mathrm{U}}(s) = 0 \right) \left(\eta - \frac{1}{|\mathcal{A}|} \right)$$
$$\overset{(b)}{=} H \left(\eta - \frac{1}{|\mathcal{A}|} \right) \sum_{s \in \mathcal{S}} \rho(s) \mathbb{P}\left(n_1^{\mathrm{U}}(s) = 0 \right).$$

In the equation (a), we use the fact that $r_h(s, a^1) = 1$ but $r_h(s,a) = 0$ for any $a \neq a^1$. In the equation (b), since each state is an absorbing state, we have that $\mathbb{P}(n_h^{\mathrm{U}}(s) = 0) = \mathbb{P}(n_1^{\mathrm{U}}(s) = 0)$ for any $h \in [H]$. We consider two cases to address RHS of equation (b). In the first case of $\eta \geq 1/|\mathcal{A}|$, we directly have that

$$\mathbb{E}\left[\sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\rho(s)\left(\pi_h^{\mathrm{mix}}(a|s)-\pi_h^{\mathrm{NBCU}}(a|s)\right)r_h(s,a)\mathbb{I}\{n_h^{\mathrm{U}}(s)=0\}\right]\geq 0.$$

By Equation (15), we have that

$$\mathbb{E}\left[V(\pi^{\mathrm{mix}})-V(\pi^{\mathrm{NBCU}})\right]\geq 0,$$

which implies that

$$\mathbb{E}\left[V(\pi^{\mathrm{E}})-V(\pi^{\mathrm{NBCU}})\right]\geq (1-\eta)(V(\pi^{\mathrm{E}})-V(\pi^{\beta})).$$

In the second case of $\eta < 1/|\mathcal{A}|$, we have that

$$\begin{aligned}H\left(\eta-\frac{1}{|\mathcal{A}|}\right)\sum_{s\in\mathcal{S}}\rho(s)\mathbb{P}\left(n_1^{\mathrm{U}}(s)=0\right)&\overset{(a)}{\geq}-\left(\frac{1}{|\mathcal{A}|}-\eta\right)H\exp\left(-\frac{N_{\mathrm{tot}}}{|\mathcal{S}|}\right)\\&\geq-(1-\eta)H\exp\left(-\frac{N_{\mathrm{tot}}}{|\mathcal{S}|}\right)\\&\overset{(b)}{\geq}-\frac{(1-\eta)H}{2}.\end{aligned}$$

In the inequality (a), we use that

$$\sum_{s\in\mathcal{S}}\rho(s)\mathbb{P}\left(n_1^{\mathrm{U}}(s)=0\right)=\sum_{s\in\mathcal{S}}\rho(s)(1-\rho(s))^{N_{\mathrm{tot}}}=\left(1-\frac{1}{|\mathcal{S}|}\right)^{N_{\mathrm{tot}}}\leq\exp\left(-\frac{N_{\mathrm{tot}}}{|\mathcal{S}|}\right).$$

The inequality (b) holds since we consider the range where $N_{\mathrm{tot}}\geq|\mathcal{S}|\log(2)$. By Equation (15), we have that

$$\mathbb{E}\left[V(\pi^{\mathrm{mix}})-V(\pi^{\mathrm{NBCU}})\right]\geq-\frac{(1-\eta)H}{2}.$$

This implies that

$$\begin{aligned}\mathbb{E}\left[V(\pi^{\mathrm{E}})-V(\pi^{\mathrm{NBCU}})\right]&\geq(1-\eta)(V(\pi^{\mathrm{E}})-V(\pi^{\beta}))-\frac{(1-\eta)H}{2}\\&=\frac{(1-\eta)}{2}(V(\pi^{\mathrm{E}})-V(\pi^{\beta})).\end{aligned}$$

In both cases, we prove that $\mathbb{E}\left[V(\pi^{\mathrm{E}})-V(\pi^{\mathrm{NBCU}})\right]\gtrsim(1-\eta)(V(\pi^{\mathrm{E}})-V(\pi^{\beta}))$ and thus complete the proof.

C. Proof of Results in Section 5

C.1. Proof of Proposition 2

In the tabular case, with the first-order optimality condition, we have $c_h^{\star}(s,a)=\widehat{d_h^{\mathrm{E}}}(s,a)/(\widehat{d_h^{\mathrm{E}}}(s,a)+\widehat{d_h^{\mathrm{U}}}(s,a))$. By Equation (5), we have

$$\widehat{d_h^{\mathrm{U}}}(s,a)w_h(s,a)=\widehat{d_h^{\mathrm{U}}}(s,a)\times\frac{\widehat{d_h^{\mathrm{E}}}(s,a)}{\widehat{d_h^{\mathrm{U}}}(s,a)}=\widehat{d_h^{\mathrm{E}}}(s,a).$$

Hence, the learning objective (3) reduces to (1).

C.2. Proof of Lemma 1

Recall that

$$\Delta_h(\theta)=\min_{(s,a)\in\mathcal{D}_h^{\mathrm{E}}\cup\mathcal{D}_h^{\mathrm{S},1}}\langle\theta,\phi_h(s,a)\rangle-\max_{(s',a')\in\mathcal{D}_h^{\mathrm{S},2}}\langle\theta,\phi_h(s',a')\rangle.$$

Then we have that

$$\Delta_h(\bar{\theta}_h)-\Delta_h(\theta)=\min_{(s,a)\in\mathcal{D}_h^{\mathrm{E}}\cup\mathcal{D}_h^{\mathrm{S},1}}\langle\bar{\theta}_h,\phi_h(s,a)\rangle-\max_{(s',a')\in\mathcal{D}_h^{\mathrm{S},2}}\langle\bar{\theta}_h,\phi_h(s',a')\rangle$$

$$
\begin{aligned}
&\quad - \min_{(s,a)\in\mathcal{D}_h^{\mathrm{E}}\cup\mathcal{D}_h^{\mathrm{S},1}} \langle \theta, \phi_h(s,a)\rangle + \max_{(s',a')\in\mathcal{D}_h^{\mathrm{S},2}} \langle \theta, \phi_h(s',a')\rangle \\
&\overset{(a)}{\leq} \langle \bar{\theta}_h, \phi_h(s^1,a^1)\rangle - \langle \bar{\theta}_h, \phi_h(s^2,a^2)\rangle - \langle \theta, \phi_h(s^1,a^1)\rangle + \langle \theta, \phi_h(s^2,a^2)\rangle \\
&= \langle \bar{\theta}_h - \theta, \phi_h(s^1,a^1) - \phi_h(s^2,a^2)\rangle \\
&\overset{(b)}{\leq} \left\| \bar{\theta}_h - \theta \right\| \left\| \phi_h(s^1,a^1) - \phi_h(s^2,a^2) \right\|.
\end{aligned}
$$

In inequality (a), we utilize the facts that $(s^1,a^1) \in \mathrm{argmin}_{(s,a)\in\mathcal{D}_h^{\mathrm{E}}\cup\mathcal{D}_h^{\mathrm{S},1}} \langle \theta_h, \phi_h(s,a)\rangle$ and $(s^2,a^2) \in \mathrm{argmax}_{(s,a)\in\mathcal{D}_h^{\mathrm{S},2}} \langle \theta_h, \phi_h(s,a)\rangle$. Inequality (b) follows the Cauchy–Schwarz inequality. Let $L_h = \left\| \phi_h(s^1,a^1) - \phi_h(s^2,a^2) \right\|$ and we finish the proof.

C.3. Proof of Lemma 2

First, by Taylor's Theorem, there exists $\theta_h^\prime \in \{\theta \in \mathbb{R}^d : \theta^t = \theta_h^\star + t(\bar{\theta}_h - \theta_h^\star), \ \forall t \in [0,1]\}$ such that

$$
\begin{aligned}
\mathcal{L}_h(\bar{\theta}_h) &= \mathcal{L}_h(\theta_h^\star) + \langle \nabla \mathcal{L}_h(\theta_h^\star), \bar{\theta}_h - \theta_h^\star \rangle + \frac{1}{2}\left(\bar{\theta}_h - \theta_h^\star\right)^\top \nabla^2 \mathcal{L}_h(\theta_h^\prime)\left(\bar{\theta}_h - \theta_h^\star\right) \\
&= \mathcal{L}_h(\theta_h^\star) + \frac{1}{2}\left(\bar{\theta}_h - \theta_h^\star\right)^\top \nabla^2 \mathcal{L}_h(\theta_h^\prime)\left(\bar{\theta}_h - \theta_h^\star\right). \qquad (16)
\end{aligned}
$$

The last equality follows the optimality condition that $\nabla \mathcal{L}_h(\theta_h^\star) = 0$. Then, our strategy is to prove that the smallest eigenvalue of the Hessian matrix $\nabla^2 \mathcal{L}_h(\theta_h^\prime)$ is positive, i.e., $\lambda_{\min}(\nabla^2 \mathcal{L}_h(\theta_h^\prime)) > 0$. We first calculate the Hessian matrix $\nabla^2 \mathcal{L}_h(\theta_h^\prime)$. Given $\mathcal{D}^{\mathrm{E}}$ and $\mathcal{D}^{\mathrm{U}}$, we define the function $G : \mathbb{R}^{(|\mathcal{D}^{\mathrm{E}}|+|\mathcal{D}^{\mathrm{U}}|)} \to \mathbb{R}$ as

$$
G(v) \triangleq \frac{1}{|\mathcal{D}^{\mathrm{E}}|}\sum_{i=1}^{|\mathcal{D}^{\mathrm{E}}|} g(v_i) + \frac{1}{|\mathcal{D}^{\mathrm{U}}|}\sum_{j=1}^{|\mathcal{D}^{\mathrm{U}}|} g(v_j),
$$

where v_i is the i-th element in the vector $v \in \mathbb{R}^{(|\mathcal{D}^{\mathrm{E}}|+|\mathcal{D}^{\mathrm{U}}|)}$ and $g(x) = \log(1+\exp(x))$ is a real-valued function. Besides, we use $B_h \in \mathbb{R}^{(|\mathcal{D}^{\mathrm{E}}|+|\mathcal{D}^{\mathrm{U}}|)\times d}$ to denote the matrix whose i-th row $B_{h,i} = -y_i \phi_h(s^i,a^i)^\top$, and $y_i = 1$ if $(s^i,a^i) \in \mathcal{D}_h^{\mathrm{E}}$, $y_i = -1$ if $(s^i,a^i) \notin \mathcal{D}_h^{\mathrm{E}}$. Then the objective function can be reformulated as

$$
\begin{aligned}
&\mathcal{L}_h(\theta_h) \\
&= \sum_{(s,a)} \widehat{d_h^{\mathrm{E}}}(s,a)\left[\log\left(1+\exp\left(-\langle \phi_h(s,a), \theta_h\rangle\right)\right)\right] + \sum_{(s,a)} \widehat{d_h^{\mathrm{U}}}(s,a)\left[\log\left(1+\exp\left(\langle \phi_h(s,a), \theta_h\rangle\right)\right)\right] \\
&= \frac{1}{|\mathcal{D}^{\mathrm{E}}|}\sum_{(s,a)\in\mathcal{D}^{\mathrm{E}}} \log\left(1+\exp\left(-\langle \phi_h(s,a), \theta_h\rangle\right)\right) + \frac{1}{|\mathcal{D}^{\mathrm{U}}|}\sum_{(s,a)\in\mathcal{D}^{\mathrm{U}}} \log\left(1+\exp\left(\langle \phi_h(s,a), \theta_h\rangle\right)\right) \\
&= G(B_h\theta_h).
\end{aligned}
$$

Then we have that $\nabla^2 \mathcal{L}_h(\theta_h) = B_h^\top \nabla^2 G(B_h\theta_h) B_h$, where

$$
\begin{aligned}
&\nabla^2 G(B_h\theta_h) \\
&= \mathbf{diag}\left(\frac{g''((B_h\theta_h)_1)}{|\mathcal{D}^{\mathrm{E}}|}, \ldots, \frac{g''((B_h\theta_h)_{|\mathcal{D}^{\mathrm{E}}|})}{|\mathcal{D}^{\mathrm{E}}|}, \frac{g''((B_h\theta_h)_{|\mathcal{D}^{\mathrm{E}}|+1})}{|\mathcal{D}^{\mathrm{E}}|+|\mathcal{D}^{\mathrm{U}}|}, \ldots, \frac{g''((B_h\theta_h)_{|\mathcal{D}^{\mathrm{E}}|+|\mathcal{D}^{\mathrm{U}}|})}{|\mathcal{D}^{\mathrm{E}}|+|\mathcal{D}^{\mathrm{U}}|}\right).
\end{aligned}
$$

Here $g''(x) = \sigma(x)(1-\sigma(x))$, where $\sigma(x) = 1/(1+\exp(-x))$ is the sigmoid function. The eigenvalues of $\nabla^2 G(B_h\theta_h)$ are

$$
\left\{\frac{g''((B_h\theta_h)_1)}{|\mathcal{D}^{\mathrm{E}}|}, \ldots, \frac{g''((B_h\theta_h)_{|\mathcal{D}^{\mathrm{E}}|})}{|\mathcal{D}^{\mathrm{E}}|}, \frac{g''((B_h\theta_h)_{|\mathcal{D}^{\mathrm{E}}|+1})}{|\mathcal{D}^{\mathrm{E}}|+|\mathcal{D}^{\mathrm{U}}|}, \ldots, \frac{g''((B_h\theta_h)_{|\mathcal{D}^{\mathrm{E}}|+|\mathcal{D}^{\mathrm{U}}|})}{|\mathcal{D}^{\mathrm{E}}|+|\mathcal{D}^{\mathrm{U}}|}\right\}.
$$

Notice that $\theta_h^\prime \in \{\theta \in \mathbb{R}^d : \theta^t = \theta_h^\star + t(\bar{\theta}_h - \theta_h^\star), \ \forall t \in [0,1]\}$. For a matrix A, we use $\lambda_{\min}(A)$ to denote the minimal eigenvalue of A. Here we claim that the minimum of the minimal eigenvalues of $\nabla^2 G(B_h\theta^t)$ over $t \in [0,1]$ is achieved at $t = 0$ or $t = 1$. That is,

$$
\min\{\lambda_{\min}(\nabla^2 G(B_h\theta^t)) : \forall t \in [0,1]\} = \min\{\lambda_{\min}(\nabla^2 G(B_h\theta^0)), \lambda_{\min}(\nabla^2 G(B_h\theta^1))\}.
$$

We prove this claim as follows. For any $t \in [0,1]$, we use $\{\lambda_1(t), \ldots, \lambda_{|\mathcal{D}^{\mathrm{E}}|+|\mathcal{D}^{\mathrm{U}}|}(t)\}$ to denote the eigenvalues of $\nabla^2 G(B_h \theta^t)$. For each $i \in [|\mathcal{D}^{\mathrm{E}}| + |\mathcal{D}^{\mathrm{U}}|]$, we consider $\lambda_i(t) : [0,1] \to \mathbb{R}$ as a function of t. Specifically,

$$\lambda_i(t) = \begin{cases} \frac{g''((B_h\theta_h^\star)_i + t(B_h(\bar{\theta}_h - \theta_h^\star))_i)}{|\mathcal{D}^{\mathrm{E}}|}, & \text{if } i \in [|\mathcal{D}^{\mathrm{E}}|] \\ \frac{g''((B_h\theta_h^\star)_i + t(B_h(\bar{\theta}_h - \theta_h^\star))_i)}{|\mathcal{D}^{\mathrm{E}}|+|\mathcal{D}^{\mathrm{U}}|}, & \text{otherwise.} \end{cases}$$

We observe that $g'''(x) = \sigma(x)(1-\sigma(x))(1-2\sigma(x))$ which satisfies that $\forall x \leq 0,\ g'''(x) \geq 0$, and $\forall x \geq 0,\ g'''(x) \leq 0$. Therefore, we have that the minimum of $\lambda_i(t)$ over $t \in [0,1]$ must be achieved at $t=0$ or $t=1$. That is,

$$\min_{t \in [0,1]} \lambda_i(t) = \min\{\lambda_i(0), \lambda_i(1)\}. \tag{17}$$

For any $t \in [0,1]$, we define $i^t \in [|\mathcal{D}^{\mathrm{E}}| + |\mathcal{D}^{\mathrm{U}}|]$ as the index of the minimal eigenvalue of $\nabla^2 G(B_h\theta^t)$, i.e., $\lambda_{i^t}(t) = \lambda_{\min}(\nabla^2 G(B_h\theta^t))$. Then we have that

$$\begin{aligned} \min\{\lambda_{\min}(\nabla^2 G(B_h\theta^t)) : \forall t \in [0,1]\} &= \min\{\lambda_{i^t}(t) : \forall t \in [0,1]\} \\ &\overset{(a)}{=} \min\{\min\{\lambda_{i^t}(0), \lambda_{i^t}(1)\} : \forall t \in [0,1]\} \\ &= \min\{\lambda_{i^0}(0), \lambda_{i^1}(1)\} \\ &\overset{(b)}{=} \min\{\lambda_{\min}(\nabla^2 G(B_h\theta^0)), \lambda_{\min}(\nabla^2 G(B_h\theta^1))\} \end{aligned}$$

Equality (a) follows (17) and equality (b) follows that $\lambda_{i^0}(0)$ and $\lambda_{i^1}(1)$ are the minimal eigenvalues of $\nabla^2 G(B_h\theta^0)$ and $\nabla^2 G(B_h\theta^1)$, respectively.

In summary, we derive that

$$\min\{\lambda_{\min}(\nabla^2 G(B_h\theta^t)) : \forall t \in [0,1]\} = \min\{\lambda_{\min}(\nabla^2 G(B_h\theta^0)), \lambda_{\min}(\nabla^2 G(B_h\theta^1))\}, \tag{18}$$

which proves the previous claim.

Further, we consider $\lambda_{\min}\left(\nabla^2 \mathcal{L}_h(\theta_h)\right)$.

$$\begin{aligned} \lambda_{\min}\left(\nabla^2 \mathcal{L}_h(\theta_h)\right) &= \inf_{x \in \mathbb{R}^d : \|x\|=1} x^\top \nabla^2 \mathcal{L}_h(\theta_h) x \\ &= \inf_{x \in \mathbb{R}^d : \|x\|=1} (B_h x)^\top \nabla^2 G(B_h\theta_h)(B_h x) \\ &= \inf_{z \in \mathrm{Im}(B_h)} z^\top \nabla^2 G(B_h\theta_h) z \\ &= \left(\inf_{z \in \mathrm{Im}(B_h)} \|z\|\right)^2 \lambda_{\min}(\nabla^2 G(B_h\theta_h)) \\ &\geq \left(\inf_{z \in \mathrm{Im}(B_h)} \|z\|\right)^2 \min\{\lambda_{\min}(\nabla^2 G(B_h\theta^0)), \lambda_{\min}(\nabla^2 G(B_h\theta^1))\}. \end{aligned}$$

Here $\mathrm{Im}(B_h) = \{z \in \mathbb{R}^d : z = B_h x, \|x\| = 1\}$. The last inequality follows Equation (18).

Recall we assume that $\mathbf{rank}(A_h) = d$, so we have that $\mathbf{rank}(B_h) = d$. Thus, $\mathrm{Im}(B_h)$ is a set of vectors with positive norms, i.e., $\inf_{z \in \mathrm{Im}(B_h)} \|z\| > 0$. Besides, since $g''(x) = \sigma(x)(1-\sigma(x)) > 0$, we also have that

$$\min\{\lambda_{\min}(\nabla^2 G(B_h\theta^0)), \lambda_{\min}(\nabla^2 G(B_h\theta^1))\} > 0.$$

In summary, we obtain that

$$\lambda_{\min}\left(\nabla^2 \mathcal{L}_h(\theta_h)\right) \geq \left(\inf_{z \in \mathrm{Im}(B_h)} \|z\|\right)^2 \min\{\lambda_{\min}(\nabla^2 G(B_h\theta^0)), \lambda_{\min}(\nabla^2 G(B_h\theta^1))\} > 0.$$

Then, with Equation (16), there exists

$$\tau_h = \left(\inf_{z \in \mathrm{Im}(B_h)} \|z\|\right)^2 \min\{\lambda_{\min}(\nabla^2 G(B_h\theta^0)), \lambda_{\min}(\nabla^2 G(B_h\theta^1))\} > 0$$

such that

$$\mathcal{L}_h(\bar{\theta}_h) \geq \mathcal{L}_h(\theta_h^\star) + \frac{\tau_h}{2} \left\|\bar{\theta}_h - \theta_h^\star\right\|^2,$$

which completes the proof.

C.4. Proof of Theorem 3

First, invoking Lemma 1 with $\theta = \theta_h^\star$ yields that

$$\Delta_h(\theta_h^\star) \geq \Delta_h(\bar{\theta}_h) - L_h \left\| \bar{\theta}_h - \theta_h^\star \right\| .$$

Here $L_h = \| \phi_h(s, a) - \phi_h(s', a') \|$ with $(s, a) \in \operatorname{argmin}_{(s,a) \in \mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S},1}} \langle \theta_h^\star, \phi_h(s, a) \rangle$ and $(s', a') \in \operatorname{argmax}_{(s,a) \in \mathcal{D}_h^{\mathrm{S},2}} \langle \theta_h^\star, \phi_h(s, a) \rangle$. Then, by Lemma 2, there exists $\tau_h > 0$ such that

$$\mathcal{L}_h(\bar{\theta}_h) \geq \mathcal{L}_h(\theta_h^\star) + \frac{\tau_h}{2} \left\| \bar{\theta}_h - \theta_h^\star \right\|^2 .$$

This directly implies an upper bound of the distance between $\bar{\theta}_h$ and $\theta_h^\star$.

$$\left\| \bar{\theta}_h - \theta_h^\star \right\| \leq \sqrt{\frac{2 \left(\mathcal{L}_h(\bar{\theta}_h) - \mathcal{L}_h(\theta_h^\star) \right)}{\tau_h}} .$$

If the feature is designed such that $\sqrt{\frac{2 \left(\mathcal{L}_h(\bar{\theta}_h) - \mathcal{L}_h(\theta_h^\star) \right)}{\tau_h}} < \frac{\Delta_h(\bar{\theta}_h)}{L_h}$ holds, we further have that $\left\| \bar{\theta}_h - \theta_h^\star \right\| < \Delta_h(\bar{\theta}_h) / L_h$. Then we get that

$$\Delta_h(\theta_h^\star) \geq \Delta_h(\bar{\theta}_h) - L_h \left\| \bar{\theta}_h - \theta_h^\star \right\| > 0,$$

which completes the proof of the first statement.

Then we proceed to prove the imitation gap bound. We first identify the property of $\pi^{\text{ISW-BC}}$. Recall the objective of WBCU.

$$\pi^{\text{ISW-BC}} \in \operatorname*{argmax}_{\pi} \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left\{ \widehat{d_h^{\mathrm{U}}}(s, a) \times \left[w_h(s, a) \log \pi_h(a|s) \right] \times \mathbb{I} \left[w_h(s, a) \geq \delta \right] \right\}.$$

For any state s with $\sum_{a \in \mathcal{A}} \widehat{d_h^{\mathrm{U}}}(s, a) w_h(s, a) \mathbb{I} \left[w_h(s, a) \geq \delta \right] > 0$, with the first-order optimality condition, we have

$$\pi_h^{\text{ISW-BC}}(a|s) = \frac{\widehat{d_h^{\mathrm{U}}}(s, a) w_h(s, a) \mathbb{I} \left[w_h(s, a) \geq \delta \right]}{\sum_{a \in \mathcal{A}} \widehat{d_h^{\mathrm{U}}}(s, a) w_h(s, a) \mathbb{I} \left[w_h(s, a) \geq \delta \right]} .$$

For an expert state s with $d_h^{\pi^{\mathrm{E}}}(s) > 0$, if $(s, \pi_h^{\mathrm{E}}(s)) \in \mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S},1}$, we have that

$$\langle \theta_h^\star, \phi_h(s, \pi_h^{\mathrm{E}}(s)) \rangle > \langle \theta_h^\star, \phi_h(s, a) \rangle, \quad \forall (s, a) \in \mathcal{D}_h^{\mathrm{S},2}.$$

This is due to the first statement that $\Delta_h(\theta_h^\star) > 0$ in this theorem. Recall that

$$c_h(s, a; \theta_h^\star) = \frac{1}{1 + \exp(-\langle \phi_h(s, a), \theta_h^\star \rangle)} \quad \text{and} \quad w_h(s, a) = \frac{c_h(s, a; \theta_h^\star)}{1 - c_h(s, a; \theta_h^\star)} .$$

We can further obtain that $w_h(s, \pi_h^{\mathrm{E}}(s)) > w_h(s, a)$ for any $(s, a) \in \mathcal{D}_h^{\mathrm{S},2}$. This implies that we can find a δ such that $\mathbb{I} \left[w_h(s, \pi_h^{\mathrm{E}}(s)) \geq \delta \right] = 1$ for any $(s, \pi_h^{\mathrm{E}}(s)) \in \mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S},1}$ and $\mathbb{I} \left[w_h(s, a) \geq \delta \right] = 0$ for any $(s, a) \in \mathcal{D}_h^{\mathrm{S},2}$. Based on the above analytical form of $\pi^{\text{ISW-BC}}$, we have that $\pi_h^{\text{ISW-BC}}(\pi_h^{\mathrm{E}}(s)|s) = 1$ for any $(s, \pi_h^{\mathrm{E}}(s)) \in \mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S},1}$. In summary, for any state s with $(s, \pi_h^{\mathrm{E}}(s)) \in \mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S},1}$, we have that $\pi_h^{\text{ISW-BC}}(\pi_h^{\mathrm{E}}(s)|s) = 1$.

With the above property of $\pi^{\text{ISW-BC}}$, we proceed to analyze the policy value gap. According to (Rajaraman et al., 2020, Lemma 4.3), we have

$$V(\pi^{\mathrm{E}}) - V(\pi^{\text{ISW-BC}}) \leq H \sum_{h=1}^{H} \mathbb{E}_{s \sim d_h^{\pi^{\mathrm{E}}}(\cdot)} \left[\mathrm{TV} \left(\pi_h^{\mathrm{E}}(\cdot|s), \pi_h^{\text{ISW-BC}}(\cdot|s) \right) \right] .$$

Since π^{E} is assumed to be deterministic, we have

$$V(\pi^{\mathrm{E}}) - V(\pi^{\text{ISW-BC}}) \leq H \sum_{h=1}^{H} \mathbb{E}_{s \sim d_h^{\pi^{\mathrm{E}}}(\cdot)} \left[\mathbb{E}_{a \sim \pi_h^{\text{ISW-BC}}(\cdot|s)} \left[\mathbb{I} \left\{ a \neq \pi_h^{\mathrm{E}}(s) \right\} \right] \right]$$

$$
\begin{aligned}
&\overset{(a)}{\leq} H \sum_{h=1}^{H} \mathbb{E}_{s \sim d_h^{\pi^{\mathrm{E}}}(\cdot)} \left[\mathbb{I}\left\{ (s, \pi_h^{\mathrm{E}}(s)) \notin \mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S},1} \right\} \right] \\
&\overset{(b)}{=} H \sum_{h=1}^{H} \mathbb{E}_{s \sim d_h^{\pi^{\mathrm{E}}}(\cdot)} \left[\mathbb{I}\left\{ (s, \pi_h^{\mathrm{E}}(s)) \notin \mathcal{D}_h^{\mathrm{U}} \right\} \right].
\end{aligned}
$$

Inequality (a) follows the property of $\pi^{\text{ISW-BC}}$ derived above. In particular, for any state s with $(s, \pi_h^{\mathrm{E}}(s)) \in \mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S},1}$, we have that $\pi_h^{\text{ISW-BC}}(\pi_h^{\mathrm{E}}(s)|s) = 1$. Equation (b) holds due to the Assumption 2. In particular, for an expert state s that $d_h^{\pi^{\mathrm{E}}}(s) > 0$, the events of $(s, \pi_h^{\mathrm{E}}(s)) \notin \mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S},1}$ and $(s, \pi_h^{\mathrm{E}}(s)) \notin \mathcal{D}_h^{\mathrm{U}}$ are equivalent.

Moreover, we take the expectation over $\mathcal{D}^{\mathrm{U}}$ on both sides and obtain that

$$
\begin{aligned}
\mathbb{E}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\text{ISW-BC}}) \right] &\leq H \sum_{h=1}^{H} \mathbb{E}_{s \sim d_h^{\pi^{\mathrm{E}}}(\cdot)} \left[\mathbb{P}\left((s, \pi_h^{\mathrm{E}}(s)) \notin \mathcal{D}_h^{\mathrm{U}} \right) \right] \\
&= H \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} d_h^{\pi^{\mathrm{E}}}(s) \mathbb{P}\left((s, \pi_h^{\mathrm{E}}(s)) \notin \mathcal{D}_h^{\mathrm{U}} \right).
\end{aligned}
$$

According to Assumption 1, we have that

$$
\begin{aligned}
d_h^{\mathrm{U}}(s, \pi_h^{\mathrm{E}}(s)) &= \eta d_h^{\pi^{\mathrm{E}}}(s, \pi_h^{\mathrm{E}}(s)) + (1-\eta) d_h^{\pi^{\beta}}(s, \pi_h^{\mathrm{E}}(s)) \\
&\overset{(a)}{\geq} \eta d_h^{\pi^{\mathrm{E}}}(s, \pi_h^{\mathrm{E}}(s)) + \frac{(1-\eta)}{\mu} d_h^{\pi^{\mathrm{E}}}(s, \pi_h^{\mathrm{E}}(s)) \\
&= \left(\eta + \frac{(1-\eta)}{\mu} \right) d_h^{\pi^{\mathrm{E}}}(s, \pi_h^{\mathrm{E}}(s)).
\end{aligned}
$$

Inequality (a) follows the definition of μ in Theorem 3: for any $(s, h) \in \mathcal{S} \times [H]$, we have $d_h^{\pi^{\mathrm{E}}}(s, \pi_h^{\mathrm{E}}(s)) / d_h^{\pi^{\beta}}(s, \pi_h^{\mathrm{E}}(s)) \leq \mu$. Then we obtain that

$$
\begin{aligned}
\mathbb{E}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\text{ISW-BC}}) \right] &\leq H \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} d_h^{\pi^{\mathrm{E}}}(s) (1 - d_h^{\mathrm{U}}(s, \pi_h^{\mathrm{E}}(s)))^{N_{\mathrm{tot}}} \\
&\leq \left(\frac{1}{\eta + (1-\eta)/\mu} \right) H \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} d_h^{\mathrm{U}}(s, \pi_h^{\mathrm{E}}(s)) \mathbb{P}\left((s, \pi_h^{\mathrm{E}}(s)) \notin \mathcal{D}_h^{\mathrm{U}} \right).
\end{aligned}
$$

For each $(s, h) \in \mathcal{S} \times [H]$, we observe that

$$
d_h^{\mathrm{U}}(s, \pi_h^{\mathrm{E}}(s)) \mathbb{P}\left((s, \pi_h^{\mathrm{E}}(s)) \notin \mathcal{D}_h^{\mathrm{U}} \right) = d_h^{\mathrm{U}}(s, \pi_h^{\mathrm{E}}(s)) \left(1 - d_h^{\mathrm{U}}(s, \pi_h^{\mathrm{E}}(s)) \right)^{N_{\mathrm{tot}}} \leq \frac{4}{9 N_{\mathrm{tot}}}.
$$

Here the last inequality follows Lemma 5. Consequently, we can derive that

$$
\sum_{h=1}^{H} \sum_{s \in \mathcal{S}} d_h^{\mathrm{U}}(s, \pi_h^{\mathrm{E}}(s)) \mathbb{P}\left((s, \pi_h^{\mathrm{E}}(s)) \notin \mathcal{D}_h^{\mathrm{U}} \right) \leq \frac{4 H |\mathcal{S}|}{9 N_{\mathrm{tot}}},
$$

which further implies that

$$
\mathbb{E}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\text{ISW-BC}}) \right] \leq \left(\frac{1}{\eta + (1-\eta)/\mu} \right) \frac{4 H^2 |\mathcal{S}|}{9 N_{\mathrm{tot}}} = \frac{4 H^2 |\mathcal{S}|}{9 \left(N_{\mathrm{E}} + N_{\mathrm{S}}/\mu \right)}.
$$

We complete the proof.

C.5. An Example Corresponding to Theorem 3

In this section, we provide an example that illustrates the required feature design in Theorem 3 can hold.

Example 1. *To illustrate Theorem 3, we consider an example in the feature space $\mathbb{R}^2$. In particular, for time step $h \in [H]$, we have the expert dataset and supplementary dataset as follows.*

$$
\mathcal{D}_h^{\mathrm{E}} = \left\{ \left(s^{(1)}, a^{(1)} \right), \left(s^{(4)}, a^{(4)} \right) \right\}, \; \mathcal{D}_h^{\mathrm{S}} = \left\{ \left(s^{(2)}, a^{(2)} \right), \left(s^{(3)}, a^{(3)} \right) \right\},
$$

$$\mathcal{D}_h^{\mathrm{S},1} = \left\{ \left(s^{(2)}, a^{(2)} \right) \right\}, \; \mathcal{D}_h^{\mathrm{S},2} = \left\{ \left(s^{(3)}, a^{(3)} \right) \right\}.$$

The corresponding features are

$$\phi_h\left(s^{(1)}, a^{(1)}\right) = (0,1)^\top, \; \phi_h\left(s^{(2)}, a^{(2)}\right) = \left(-\frac{1}{2}, 0\right)^\top,$$
$$\phi_h\left(s^{(3)}, a^{(3)}\right) = \left(0, -\frac{1}{2}\right)^\top, \; \phi_h\left(s^{(4)}, a^{(4)}\right) = (-1,0)^\top.$$

Notice that the set of expert-style samples is $\mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S},1} = \{(s^{(1)}, a^{(1)}), (s^{(2)}, a^{(2)}), (s^{(4)}, a^{(4)})\}$ *and the set of non-expert-style samples is* $\mathcal{D}_h^{\mathrm{S},2} = \{(s^{(3)}, a^{(3)})\}$*. It is direct to calculate that the ground-truth parameter that achieves the maximum margin among unit vectors is* $\bar{\theta}_h = (-\sqrt{2}/2, \sqrt{2}/2)^\top$ *and the maximum margin is* $\Delta_h(\bar{\theta}_h) = \sqrt{2}/2$*. According to Equation (6), for* $\theta_h = (\theta_{h,1}, \theta_{h,2})^\top$*, the optimization objective is*

$$\begin{aligned}
&\mathcal{L}_h(\theta_h) \\
&= \sum_{(s,a)} \widehat{d_h^{\mathrm{E}}}(s,a) \left[\log\left(1 + \exp\left(-\langle \phi_h(s,a), \theta_h \rangle\right)\right)\right] + \sum_{(s,a)} \widehat{d_h^{\mathrm{U}}}(s,a) \left[\log\left(1 + \exp\left(\langle \phi_h(s,a), \theta_h \rangle\right)\right)\right] \\
&= \frac{1}{2}\left(\log\left(1 + \exp\left(-\theta_{h,2}\right)\right) + \log\left(1 + \exp\left(\theta_{h,1}\right)\right)\right) \\
&\quad + \frac{1}{4}\left(\log\left(1 + \exp\left(\theta_{h,2}\right)\right) + \log\left(1 + \exp\left(-\frac{1}{2}\theta_{h,1}\right)\right)\right) \\
&\quad + \frac{1}{4}\left(\log\left(1 + \exp\left(-\frac{1}{2}\theta_{h,2}\right)\right) + \log\left(1 + \exp\left(-\theta_{h,1}\right)\right)\right).
\end{aligned}$$

We apply CVXPY (Diamond & Boyd, 2016) to calculate the optimal solution $\theta_h^\star \approx (-0.310, 0.993)^\top$ *and the objective values* $\mathcal{L}_h(\theta_h^\star) \approx 1.287, \; \mathcal{L}_h(\bar{\theta}_h) \approx 1.309$*. Furthermore, we calculate the Lipschitz coefficient* L_h *appears in Lemma 1.*

$$(s^{(2)}, a^{(2)}) = \mathop{\mathrm{argmin}}_{(s,a) \in \mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S},1}} \langle \theta_h^\star, \phi_h(s,a) \rangle, \; (s^{(3)}, a^{(3)}) \in \mathop{\mathrm{argmax}}_{(s,a) \in \mathcal{D}_h^{\mathrm{S},2}} \langle \theta_h^\star, \phi_h(s,a) \rangle,$$
$$L_h = \left\| \phi_h(s^{(2)}, a^{(2)}) - \phi_h(s^{(3)}, a^{(3)}) \right\| = \frac{\sqrt{2}}{2}.$$

Then we calculate the parameter of strong convexity τ_h *appears in Lemma 2. Based on the proof of Lemma 2, our strategy is to calculate the minimal eigenvalue of the Hessian matrix.*

First, for $\theta_h = (\theta_{h,1}, \theta_{h,2})^\top$*, the gradient of* $\mathcal{L}_h(\theta_h)$ *is*

$$\begin{aligned}
&\nabla \mathcal{L}_h(\theta_h) \\
&= -\sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \widehat{d_h^{\mathrm{E}}}(s,a) \sigma(-\langle \phi_h(s,a), \theta_h \rangle) + \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \widehat{d_h^{\mathrm{U}}}(s,a) \sigma\left(\langle \phi_h(s,a), \theta_h \rangle\right) \\
&= \left(\frac{1}{2}\sigma(\theta_{h,1}) - \frac{1}{4}\sigma(-\theta_{h,1}) - \frac{1}{8}\sigma(-\frac{1}{2}\theta_{h,1}), \frac{1}{4}\sigma\left(\theta_{h,2}\right) - \frac{1}{2}\sigma\left(-\theta_{h,2}\right) - \frac{1}{8}\sigma(-\frac{1}{2}\theta_{h,2})\right)^\top.
\end{aligned}$$

Here $\sigma(x) = 1/(1 + \exp(-x))$ *for* $x \in \mathbb{R}$ *is the sigmoid function. Then the Hessian matrix at* θ_h *is*

$$\nabla^2 \mathcal{L}_h(\theta_h) = \begin{pmatrix} \frac{3}{4}f(\theta_{h,1}) + \frac{1}{16}f\left(\frac{1}{2}\theta_{h,1}\right) & 0 \\ 0 & \frac{3}{4}f(\theta_{h,2}) + \frac{1}{16}f\left(\frac{1}{2}\theta_{h,2}\right) \end{pmatrix},$$

where $f(x) = \sigma(x)(1 - \sigma(x))$ *and* $f(x) = f(-x)$*. For any* $t \in [0,1]$*, the eigenvalues of the Hessian matrix at* $\theta_h^t = \bar{\theta}_h + t(\theta_h^\star - \bar{\theta}_h)$ *are*

$$\frac{3}{4}f(\theta_{h,1}^t) + \frac{1}{16}f\left(\frac{1}{2}\theta_{h,1}^t\right), \; \frac{3}{4}f(\theta_{h,2}^t) + \frac{1}{16}f\left(\frac{1}{2}\theta_{h,2}^t\right).$$

Now, we calculate the minimal eigenvalues of $\nabla^2 \mathcal{L}_h(\theta_h^t)$*. We consider the function*

$$g(x) = \frac{3}{4}f(x) + \frac{1}{16}f\left(\frac{1}{2}x\right), \; \forall x \in [a,b].$$

The gradient is

$$g'(x) = \frac{3}{4}\sigma(x)(1-\sigma(x))(1-2\sigma(x)) + \frac{1}{32}\sigma\left(\frac{1}{2}x\right)\left(1-\sigma\left(\frac{1}{2}x\right)\right)\left(1-2\sigma\left(\frac{1}{2}x\right)\right).$$

We observe that $\forall x \leq 0,\ g'(x) \geq 0$, *and* $\forall x \geq 0,\ g'(x) \leq 0$. *Thus, we have that the minimum of* $g(x)$ *must be achieved at* $x = a$ *or* $x = b$. *Besides, we have that* $g(x) = g(-x)$. *With the above arguments, we know that the minimal eigenvalue is* $g(0.993) \approx 0.163$ *and* $\tau_h \approx 0.163$. *Then we can calculate that*

$$\sqrt{\frac{2\left(\mathcal{L}_h(\bar{\theta}_h) - \mathcal{L}_h(\theta_h^\star)\right)}{\tau_h}} \approx 0.520, \quad \frac{\Delta_h(\bar{\theta}_h)}{L_h} = 1.$$

The inequality in Theorem 3 holds.

D. Discussion

In the main text, we focus on the tabular representations for policies. Furthermore, we consider a trajectory sampling procedure for behavior policy in collecting the supplementary dataset. We present two possible extensions in this section.

D.1. Function Approximation of Policies

Assume that the learner is access to a finite function class $\Pi = \{\pi = (\pi_1, \pi_2, \ldots, \pi_h)\}$, where $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$ could be any function (e.g., neural networks). For simplicity of analysis, we assume that Π is a finite class. Notice that the algorithms considered in this paper are BC and its variants, which all take the principle of maximum likelihood estimation (MLE). The theoretical analysis of these algorithms is based on the following inequality:

$$V(\pi^{\mathrm{E}}) - V(\pi) \leq H \sum_{h=1}^{H} \mathbb{E}_{s \sim d_h^{\pi^{\mathrm{E}}}(\cdot)}\left[\mathrm{TV}\left(\pi_h^{\mathrm{E}}(\cdot|s), \pi_h(\cdot|s)\right)\right].$$

Therefore, the key is to upper bound the TV distance. Take BC as an example (i.e., $\pi = \pi^{\mathrm{BC}}$). By using the concentration inequality in (Agarwal et al., 2020a, Theorem 21), we obtain that for any $\delta \in (0,1)$, when $|\mathcal{D}^{\mathrm{E}}| \geq 1$, with probability at least $1-\delta$ over the randomness within $\mathcal{D}^{\mathrm{E}}$,

$$\mathbb{E}_{s \sim d_h^{\pi^{\mathrm{E}}}(\cdot)}\left[\mathrm{TV}^2\left(\pi_h^{\mathrm{E}}(\cdot|s), \pi_h^{\mathrm{BC}}(\cdot|s)\right)\right] \leq 2\frac{\log(|\Pi|/\delta)}{|\mathcal{D}^{\mathrm{E}}|}. \tag{19}$$

With additional efforts (by using union bound and Jensen's inequality), we have the following result.

Theorem 4 (BC with Function Approximation)**.** *Under Assumption 1. In the general function approximation setting, additionally assume that* $\pi^{\mathrm{E}} \in \Pi$. *If we apply BC on the expert data, we have*

$$\mathbb{E}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{BC}})\right] = \mathcal{O}\left(H^2\sqrt{\frac{\log(|\Pi| H N_{\mathrm{E}})}{N_{\mathrm{E}}}}\right),$$

where the expectation is taken over the randomness in the dataset collection.

The detailed proof is deferred to Appendix E. Compared with Theorem 1, we notice that the change in theoretical bound is that $\mathcal{O}(|\mathcal{S}|/N_{\mathrm{E}})$ is replaced by $\mathcal{O}(\sqrt{\log(|\Pi| H N_{\mathrm{E}})/N_{\mathrm{E}}})$.

NBCU can be analyzed in a similar way in the function approximation setting.

Theorem 5 (NBCU with Function Approximation)**.** *Under Assumption 1. In the general function approximation setting, additionally assume that the realizable policy class* Π *is realizable, i.e.,* $\pi^{\mathrm{mix}} \in \Pi$, *where* π^{mix} *is defined in Equation (10). If we apply BC on the union dataset, we have*

$$\mathbb{E}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{NBCU}})\right] = \mathcal{O}\left((1-\eta)(V(\pi^{\mathrm{E}}) - V(\pi^{\beta})) + H^2\sqrt{\frac{\log(|\Pi| H N_{\mathrm{tot}})}{N_{\mathrm{tot}}}}\right).$$

The proof of Theorem 5 is deferred to Appendix E. We use Theorem 4 to help prove Theorem 5.

Unfornatunetly, the analysis of ISW-BC with function approximation is much more complicated since the maximum likelihood estimation is performed in a weighted manner.In the following part, we make a conjecture on the theoretical guarantee of the weighted maximum likelihood estimation. With such a conjecture, we can derive the imitation gap of ISW-BC with general function approximation. We leave the proof of the conjecture and other proof possibilities for future works.

Recall the objective of ISW-BC.

$$\pi^{\text{ISW-BC}} \in \operatorname*{argmax}_{\pi \in \Pi} \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left\{ \widehat{d_h^{\mathrm{U}}}(s,a) \times \left[w_h(s,a) \log \pi_h(a|s) \right] \times \mathbb{I}\left[w_h(s,a) \geq \delta \right] \right\},$$

Notice that the analysis of the discriminators is independent of the function approximation of policies. Therefore, we can follow the analysis of the discriminators in the proof of Theorem 3. Importantly, we can derive that there exists δ such that $\mathbb{I}\left[w_h(s, \pi_h^{\mathrm{E}}(s)) \geq \delta\right] = 1$ for any $(s, \pi_h^{\mathrm{E}}(s)) \in \mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S},1}$ and $\mathbb{I}\left[w_h(s,a) \geq \delta\right] = 0$ for any $\forall (s,a) \in \mathcal{D}_h^{\mathrm{S},2}$. Then we can obtain that

$$\begin{aligned} &\sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left\{ \widehat{d_h^{\mathrm{U}}}(s,a) \times \left[w_h(s,a) \log \pi_h(a|s) \right] \times \mathbb{I}\left[w_h(s,a) \geq \delta \right] \right\} \\ &= \sum_{h=1}^{H} \sum_{s \in \mathcal{S}_h^{\mathrm{E}}, a = \pi_h^{\mathrm{E}}(s)} \widehat{d_h^{\mathrm{U}}}(s,a) \times \left[w_h(s,a) \log \pi_h(a|s) \right]. \end{aligned}$$

Here $\mathcal{S}_h^{\mathrm{E}} = \{ s \in \mathcal{S} : d_h^{\pi^{\mathrm{E}}}(s) > 0 \}$. Then we have that

$$\pi^{\text{ISW-BC}} \in \operatorname*{argmax}_{\pi \in \Pi} \sum_{h=1}^{H} \sum_{s \in \mathcal{S}_h^{\mathrm{E}}, a = \pi_h^{\mathrm{E}}(s)} \widehat{d_h^{\mathrm{U}}}(s,a) \times \left[w_h(s,a) \log \pi_h(a|s) \right].$$

We conjecture that $\pi^{\text{ISW-BC}}$ learned by the above weighted maximum likelihood holds the following theoretical guarantee. For any $\delta \in (0,1)$, with probability at least $1-\delta$, we have that

$$\sum_{s \in \mathcal{S}_h^{\mathrm{E}}} d_h^{\mathrm{U}}(s, \pi_h^{\mathrm{E}}(s)) \mathrm{TV}^2 \left(\pi_h^{\mathrm{E}}(\cdot|s), \pi_h^{\text{ISW-BC}}(\cdot|s) \right) = \mathcal{O}\left(\frac{\log\left(|\Pi| N_{\text{tot}} \right)}{N_{\text{tot}}} \right). \tag{20}$$

This conjecture corresponds to (19) in the unweighted maximum likelihood estimation. With this conjecture, we can derive the imitation gap of ISW-BC with function approximation.

Conjecture 1 (Imitation Gap of ISW-BC with Function Approximation)**.** *Under Assumptions 1 and 2, let* $\mu = \max_{(s,h) \in \mathcal{S} \times [H]} d_h^{\pi^{\mathrm{E}}}(s, \pi_h^{\mathrm{E}}(s)) / d_h^{\pi^{\beta}}(s, \pi_h^{\mathrm{E}}(s))$. *In the general function approximation setting with the realizable policy class* Π, *i.e.,* $\pi^{\mathrm{E}} \in \Pi$. *Furthermore, assume that the feature is designed such that* $\sqrt{\frac{2\left(\mathcal{L}_h(\bar{\theta}_h) - \mathcal{L}_h(\theta_h^{\star})\right)}{\tau_h}} < \frac{\Delta_h(\bar{\theta}_h)}{L_h}$ *holds and the conjecture in (20) holds. Then, we have the imitation gap bound*

$$\mathbb{E}[V(\pi^{\mathrm{E}}) - V(\pi^{\text{ISW-BC}})] = \mathcal{O}\left(H^2 \sqrt{\frac{\log(|\Pi| H N_{\text{tot}})}{N_{\mathrm{E}} + N_{\mathrm{S}}/\mu}} \right).$$

D.2. Supplementary Data with Corruption

In the main text, we consider the trajectory sampling procedure in Assumption 1. However, in some cases, the supplementary data can be poisoned and corrupted by an adversary. For example, although the human expert demonstrates an optimal trajectory, the recorder or the recording system possibly corrupts the data by accident or on purpose. Data corruption is one of the main security threats to imitation learning methods (Liu et al., 2022). Therefore, it is valuable to investigate the robustness of the presented algorithms in this poison setting. Supplementary data with corruption is partially investigated in our experiments under the noisy expert setting, which we argue have a large state-action distribution shift.

Assumption 3 (Poison Setting)**.** *The supplementary dataset* $\mathcal{D}^{\mathrm{S}}$ *and expert dataset* $\mathcal{D}^{\mathrm{E}}$ *are collected in the following way: each time, with probability* η, *we rollout the expert policy to collect a trajectory. With probability* $1-\eta$, *we still rollout the expert policy to collect a trajectory but with probability* $1-\eta'$, *the actions along the sampled trajectory are replaced with actions uniformly sampled from the action space. Such an experiment is independent and identically conducted by* N_{tot} *times.*

Theorem 6 (NBCU in the Poison Setting)**.** *Under Assumption 3. In the tabular case, for any $\eta \in (0, 1]$, we have*

$$\mathbb{E}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{NBCU}})\right] = \mathcal{O}\left((1-\eta)(1-\eta')H^2\left(1 - \frac{1}{|\mathcal{A}|}\right) + H^2\sqrt{\frac{|\mathcal{S}||\mathcal{A}|}{N_{\mathrm{tot}}}}\right),$$

where the expectation is taken over the randomness in the dataset collection.

Theorem 7 (ISW-BC in the Poison Setting)**.** *Under Assumptions 2 and 3, if the feature is designed such that $\sqrt{\frac{2\left(\mathcal{L}_h(\bar{\theta}_h) - \mathcal{L}_h(\theta_h^\star)\right)}{\tau_h}} < \frac{\Delta_h(\bar{\theta}_h)}{L_h}$ holds, we have the imitation gap bound*

$$\mathbb{E}[V(\pi^{\mathrm{E}}) - V(\pi^{\text{ISW-BC}})] = \mathcal{O}\left(\frac{H^2|\mathcal{S}|}{N_{\mathrm{E}} + N_{\mathrm{S}}\eta'}\right).$$

Proofs of Theorem 6 and Theorem 7 can be found in Appendix E. Compared with the imitation gap of NBCU, there is no non-vanishing gap due to the corrupted actions in the imitation gap of ISW-BC. This means that ISW-BC is still robust in this setting.

E. Proof of Results in Section D

E.1. Proof of Theorem 4

According to (Rajaraman et al., 2020, Lemma 4.3), we have

$$V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{BC}}) \leq H\sum_{h=1}^{H}\mathbb{E}_{s\sim d_h^{\pi^{\mathrm{E}}}(\cdot)}\left[\mathrm{TV}\left(\pi_h^{\mathrm{E}}(\cdot|s), \pi_h^{\mathrm{BC}}(\cdot|s)\right)\right].$$

With (Agarwal et al., 2020a, Theorem 21), when $|\mathcal{D}^{\mathrm{E}}| \geq 1$, for any $\delta \in (0, 1)$, with probability at least $1-\delta$ over the randomness within $\mathcal{D}^{\mathrm{E}}$, we have that

$$\mathbb{E}_{s\sim d_h^{\pi^{\mathrm{E}}}(\cdot)}\left[\mathrm{TV}^2\left(\pi_h^{\mathrm{E}}(\cdot|s), \pi_h^{\mathrm{BC}}(\cdot|s)\right)\right] \leq 2\frac{\log(|\Pi|/\delta)}{|\mathcal{D}^{\mathrm{E}}|}.$$

With union bound, with probability at least $1-\delta$, for all $h \in [H]$, it holds that

$$\mathbb{E}_{s\sim d_h^{\pi^{\mathrm{E}}}(\cdot)}\left[\mathrm{TV}^2\left(\pi_h^{\mathrm{E}}(\cdot|s), \pi_h^{\mathrm{BC}}(\cdot|s)\right)\right] \leq 2\frac{\log(|\Pi|H/\delta)}{|\mathcal{D}^{\mathrm{E}}|},$$

which implies that

$$\begin{aligned}
V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{BC}}) &\leq H\sum_{h=1}^{H}\mathbb{E}_{s\sim d_h^{\pi^{\mathrm{E}}}(\cdot)}\left[\mathrm{TV}\left(\pi_h^{\mathrm{E}}(\cdot|s), \pi_h^{\mathrm{BC}}(\cdot|s)\right)\right] \\
&\overset{(a)}{\leq} H\sum_{h=1}^{H}\sqrt{\mathbb{E}_{s\sim d_h^{\pi^{\mathrm{E}}}(\cdot)}\left[\mathrm{TV}^2\left(\pi_h^{\mathrm{E}}(\cdot|s), \pi_h^{\mathrm{BC}}(\cdot|s)\right)\right]} \\
&\leq \sqrt{2}H^2\sqrt{\frac{\log(|\Pi|H/\delta)}{|\mathcal{D}^{\mathrm{E}}|}}.
\end{aligned}$$

Inequality (a) follows Jensen's inequality. Taking expectation over the randomness within $\mathcal{D}^{\mathrm{E}}$ yields that

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}^{\mathrm{E}}}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{BC}})\right] &\leq \delta H + (1-\delta)\sqrt{2}H^2\sqrt{\frac{\log(|\Pi|H/\delta)}{|\mathcal{D}^{\mathrm{E}}|}} \\
&\overset{(a)}{=} \frac{H}{2|\mathcal{D}^{\mathrm{E}}|} + \left(1 - \frac{1}{2|\mathcal{D}^{\mathrm{E}}|}\right)\sqrt{2}H^2\sqrt{\frac{\log(2|\Pi|H|\mathcal{D}^{\mathrm{E}}|)}{|\mathcal{D}^{\mathrm{E}}|}} \\
&\leq \left(\sqrt{2}+1\right)H^2\sqrt{\frac{\log(2|\Pi|H|\mathcal{D}^{\mathrm{E}}|)}{|\mathcal{D}^{\mathrm{E}}|}}
\end{aligned}$$

$$\leq 4H^2\sqrt{\frac{\log(4|\Pi|H|\mathcal{D}^{\mathrm{E}}|)}{|\mathcal{D}^{\mathrm{E}}|}}.$$

Equation (a) holds due to the choice that $\delta = 1/(2|\mathcal{D}^{\mathrm{E}}|)$. For $|\mathcal{D}^{\mathrm{E}}| = 0$, we directly have that

$$\mathbb{E}_{\mathcal{D}^{\mathrm{E}}}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{BC}})\right] \leq H.$$

Therefore, for any $|\mathcal{D}^{\mathrm{E}}| \geq 0$, we have that

$$\mathbb{E}_{\mathcal{D}^{\mathrm{E}}}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{BC}})\right] \leq 4H^2\sqrt{\frac{\log(4|\Pi|H\max\{|\mathcal{D}^{\mathrm{E}}|, 1\})}{\max\{|\mathcal{D}^{\mathrm{E}}|, 1\}}}.$$

We consider a real-valued function $f(x) = \log(cx)/x$ for $x \geq 1$, where $c = 4|\Pi|H > 4$. Its gradient function is $f'(x) = (1 - \log(cx))/x^2 \leq 0$ when $x \geq 1$. Then we know that $f(x)$ is decreasing as x increases. Furthermore, we have that $\max\{|\mathcal{D}^{\mathrm{E}}|, 1\} \geq (|\mathcal{D}^{\mathrm{E}}| + 1)/2$ when $|\mathcal{D}^{\mathrm{E}}| \geq 0$. Then we obtain

$$\begin{aligned}\mathbb{E}_{\mathcal{D}^{\mathrm{E}}}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{BC}})\right] &\leq 4H^2\sqrt{\frac{\log(4|\Pi|H\max\{|\mathcal{D}^{\mathrm{E}}|, 1\})}{\max\{|\mathcal{D}^{\mathrm{E}}|, 1\}}} \\ &\leq 4H^2\sqrt{\frac{2\log(4|\Pi|H(|\mathcal{D}^{\mathrm{E}}|+1))}{|\mathcal{D}^{\mathrm{E}}|+1}}.\end{aligned}$$

Taking expectation over the random variable $|\mathcal{D}^{\mathrm{E}}| \sim \mathrm{Bin}(N_{\mathrm{tot}}, \eta)$ yields that

$$\begin{aligned}\mathbb{E}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{BC}})\right] &\leq 4H^2\mathbb{E}\left[\sqrt{\frac{2\log(4|\Pi|H(|\mathcal{D}^{\mathrm{E}}|+1))}{|\mathcal{D}^{\mathrm{E}}|+1}}\right] \\ &\overset{(a)}{\leq} 4H^2\sqrt{\mathbb{E}\left[\frac{2\log(4|\Pi|H(|\mathcal{D}^{\mathrm{E}}|+1))}{|\mathcal{D}^{\mathrm{E}}|+1}\right]}.\end{aligned}$$

Inequality (a) follows Jensen's inequality. We consider the function $g(x) = -x\log(x/c)$ for $x \in (0, 1]$, where $c = 4|\Pi|H$.

$$g'(x) = -(\log(x/c) + 1) \geq 0,\ g''(x) = -\frac{1}{x} \leq 0, \quad \forall x \in (0, 1].$$

Thus, $g(x)$ is a concave function. By Jensen's inequality, we have that $\mathbb{E}[g(x)] \leq g(\mathbb{E}[x])$. Then we can derive that

$$\begin{aligned}\mathbb{E}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{BC}})\right] &\leq 4H^2\sqrt{\mathbb{E}\left[\frac{2\log(4|\Pi|H(|\mathcal{D}^{\mathrm{E}}|+1))}{|\mathcal{D}^{\mathrm{E}}|+1}\right]} \\ &= 4\sqrt{2}H^2\sqrt{\mathbb{E}\left[g\left(\frac{1}{|\mathcal{D}^{\mathrm{E}}|+1}\right)\right]} \\ &\leq 4\sqrt{2}H^2\sqrt{g\left(\mathbb{E}\left[\frac{1}{|\mathcal{D}^{\mathrm{E}}|+1}\right]\right)} \\ &\overset{(a)}{\leq} 4\sqrt{2}H^2\sqrt{g\left(\frac{1}{N_{\mathrm{E}}}\right)} \\ &\leq 4\sqrt{2}H^2\sqrt{\frac{\log(4|\Pi|HN_{\mathrm{E}})}{N_{\mathrm{E}}}}.\end{aligned}$$

In inequality (a), we use the facts that $g'(x) \geq 0$ and $\mathbb{E}\left[1/(|\mathcal{D}^{\mathrm{E}}|+1)\right] \leq 1/N_{\mathrm{E}}$ from Lemma 3. We complete the proof.

E.2. Proof of Theorem 5

Despite the function approximation scheme, we can perform the same decomposition analysis as in the proof of Theorem 2. Therefore, we can obtain that

$$\mathbb{E}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{NBCU}})\right] = (1-\eta)\left(V(\pi^{\mathrm{E}}) - V(\pi^{\beta})\right) + \mathbb{E}\left[V(\pi^{\mathrm{mix}}) - V(\pi^{\mathrm{NBCU}})\right].$$

Recall that

$$\pi^{\text{NBCU}} \in \max_{\pi \in \Pi} \sum_{h=1}^{H} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \widehat{d_h^{\text{U}}}(s,a) \log \pi_h(a|s).$$

In the proof of Theorem 2, we have shown that $d_h^{\text{U}}(s,a) = d_h^{\pi^{\text{mix}}}(s,a)$, meaning that the state-action distribution of the union dataset equals the state-action distribution of the policy π^{mix}. Therefore, we can regard that π^{NBCU} is obtained by performing BC on the dataset generated by π^{mix}. Consequently, we can apply Theorem 4 to obtain that[3]

$$\mathbb{E}\left[V(\pi^{\text{mix}}) - V(\pi^{\text{NBCU}})\right] \leq 4\sqrt{2}H^2 \sqrt{\frac{\log(4|\Pi| H N_{\text{tot}})}{N_{\text{tot}}}}.$$

Finally, we arrive at

$$\mathbb{E}\left[V(\pi^{\text{E}}) - V(\pi^{\text{NBCU}})\right] = (1-\eta)\left(V(\pi^{\text{E}}) - V(\pi^{\beta})\right) + 4\sqrt{2}H^2 \sqrt{\frac{\log(4|\Pi| H N_{\text{tot}})}{N_{\text{tot}}}},$$

which completes the proof.

E.3. Proof of Theorem 6

We first analyze the data distribution in $\mathcal{D}^{\text{U}}$. According to Assumption 3, we summarize the sampling procedure of trajectories in $\mathcal{D}^{\text{U}}$ as follows. Each time, we rollout the expert policy to collect a trajectory. Furthermore, with the probability of $(1-\eta)(1-\eta')$, the actions along the sampled expert trajectory are replaced with actions uniformly sampled from the action space. Then we put this poisoned expert trajectory into $\mathcal{D}^{\text{U}}$. Otherwise, with the probability of $1-(1-\eta)(1-\eta')$, we directly put the original expert trajectory into $\mathcal{D}^{\text{U}}$. Therefore, we can formulate the marginal distribution of the state-action pairs in time step h in $\mathcal{D}^{\text{U}}$. For each $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$,

$$\begin{aligned} d_h^{\text{U}}(s,a) &= \left(1-(1-\eta)(1-\eta')\right) d_h^{\pi^{\text{E}}}(s,a) + (1-\eta)(1-\eta') d_h^{\pi^{\text{E}}}(s) \frac{1}{|\mathcal{A}|}, \\ d_h^{\text{U}}(s) &= \sum_{a \in \mathcal{A}} d_h^{\text{U}}(s,a) = d_h^{\pi^{\text{E}}}(s). \end{aligned}$$

Then we proceed to analyze the imitation gap. Similar to the proof of Theorem 2, according to (Rajaraman et al., 2020, Lemma 4.3), we have

$$V(\pi^{\text{E}}) - V(\pi^{\text{NBCU}}) \leq H \sum_{h=1}^{H} \mathbb{E}_{s \sim d_h^{\pi^{\text{E}}}(\cdot)}\left[\text{TV}\left(\pi_h^{\text{E}}(\cdot|s), \pi_h^{\text{NBCU}}(\cdot|s)\right)\right].$$

Again, we introduce the definition of the policy π^{mix}.

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}, \forall h \in [H], \ \pi_h^{\text{mix}}(a|s) = \begin{cases} \frac{d_h^{\text{U}}(s,a)}{d_h^{\text{U}}(s)} & \text{if } d_h^{\text{U}}(s) = d_h^{\pi^{\text{E}}}(s) > 0, \\ \frac{1}{|\mathcal{A}|} & \text{otherwise.} \end{cases}$$

In particular, if $d_h^{\text{U}}(s) > 0$, we have that

$$\pi_h^{\text{mix}}(a|s) = \frac{d_h^{\text{U}}(s,a)}{d_h^{\text{U}}(s)} = \left(1-(1-\eta)(1-\eta')\right) \pi_h^{\text{E}}(a|s) + (1-\eta)(1-\eta') \frac{1}{|\mathcal{A}|}.$$

Then we decompose the imitation gap into two parts.

$$\begin{aligned} & V(\pi^{\text{E}}) - V(\pi^{\text{NBCU}}) \\ & \leq H \sum_{h=1}^{H} \mathbb{E}_{s \sim d_h^{\pi^{\text{E}}}(\cdot)}\left[\text{TV}\left(\pi_h^{\text{E}}(\cdot|s), \pi_h^{\text{NBCU}}(\cdot|s)\right)\right] \\ & \leq H \sum_{h=1}^{H} \mathbb{E}_{s \sim d_h^{\pi^{\text{E}}}(\cdot)}\left[\text{TV}\left(\pi_h^{\text{E}}(\cdot|s), \pi_h^{\text{mix}}(\cdot|s)\right)\right] + H \sum_{h=1}^{H} \mathbb{E}_{s \sim d_h^{\pi^{\text{E}}}(\cdot)}\left[\text{TV}\left(\pi_h^{\text{mix}}(\cdot|s), \pi_h^{\text{NBCU}}(\cdot|s)\right)\right]. \end{aligned}$$

[3] Note that Theorem 4 holds for the case where the expert policy is stochastic.

We first analyze the first term in RHS. For certain (s, h) such $d^{\mathrm{U}}_h(s) = d^{\pi^{\mathrm{E}}}_h(s) > 0$, we have that

$$\begin{aligned} \mathrm{TV}\left(\pi^{\mathrm{E}}_h(\cdot|s), \pi^{\mathrm{mix}}_h(\cdot|s)\right) &= \sum_{a \neq \pi^{\mathrm{E}}_h(s)} \pi^{\mathrm{mix}}_h(a|s) \\ &= \sum_{a \neq \pi^{\mathrm{E}}_h(s)} \left(1 - (1-\eta)(1-\eta')\right) \pi^{\mathrm{E}}_h(a|s) + (1-\eta)(1-\eta') \frac{1}{|\mathcal{A}|} \\ &= (1-\eta)(1-\eta')\left(1 - \frac{1}{|\mathcal{A}|}\right). \end{aligned}$$

Therefore, we can derive that

$$H \sum_{h=1}^{H} \mathbb{E}_{s \sim d^{\pi^{\mathrm{E}}}_h(\cdot)} \left[\mathrm{TV}\left(\pi^{\mathrm{E}}_h(\cdot|s), \pi^{\mathrm{mix}}_h(\cdot|s)\right)\right] \leq (1-\eta)(1-\eta') H^2 \left(1 - \frac{1}{|\mathcal{A}|}\right).$$

Now we analyze the second term of

$$H \sum_{h=1}^{H} \mathbb{E}_{s \sim d^{\pi^{\mathrm{E}}}_h(\cdot)} \left[\mathrm{TV}\left(\pi^{\mathrm{mix}}_h(\cdot|s), \pi^{\mathrm{NBCU}}_h(\cdot|s)\right)\right].$$

Recall the formula of π^{NBCU}.

$$\pi^{\mathrm{NBCU}}_h(a|s) = \begin{cases} \frac{n^{\mathrm{U}}_h(s,a)}{n^{\mathrm{U}}_h(s)} & \text{if } n^{\mathrm{U}}_h(s) > 0 \\ \frac{1}{|\mathcal{A}|} & \text{otherwise} \end{cases}$$

Notice that π^{NBCU} is the maximum likelihood estimation of π^{mix}. According to the concentration inequality of total variation (Weissman et al., 2003), for each $(s, h) \in \mathcal{S} \times [H]$, for any fixed $\delta \in (0, 1)$, when $n^{\mathrm{U}}_h(s) > 0$, with probability at least $1 - \delta$, we have

$$\mathrm{TV}\left(\pi^{\mathrm{mix}}_h(\cdot|s), \pi^{\mathrm{NBCU}}_h(\cdot|s)\right) \leq \sqrt{\frac{|\mathcal{A}| \log(3/\delta)}{n^{\mathrm{U}}_h(s)}}.$$

When $n^{\mathrm{U}}_h(s) = 0$, we have that

$$\mathrm{TV}\left(\pi^{\mathrm{mix}}_h(\cdot|s), \pi^{\mathrm{NBCU}}_h(\cdot|s)\right) \leq 1 \leq \sqrt{|\mathcal{A}| \log(3/\delta)}.$$

By combining the above two inequalities, for each $(s, h) \in \mathcal{S} \times [H]$, with probability at least $1 - \delta$, we have

$$\mathrm{TV}\left(\pi^{\mathrm{mix}}_h(\cdot|s), \pi^{\mathrm{NBCU}}_h(\cdot|s)\right) \leq \sqrt{\frac{|\mathcal{A}| \log(3/\delta)}{\max\{n^{\mathrm{U}}_h(s), 1\}}}.$$

Applying union bound yields that with probability at least $1 - \delta/2$, for all $(s, h) \in \mathcal{S} \times [H]$,

$$\mathrm{TV}\left(\pi^{\mathrm{mix}}_h(\cdot|s), \pi^{\mathrm{NBCU}}_h(\cdot|s)\right) \leq \sqrt{\frac{|\mathcal{A}| \log(6|\mathcal{S}|H/\delta)}{\max\{n^{\mathrm{U}}_h(s), 1\}}}.$$

Then we have that

$$\begin{aligned} &H \sum_{h=1}^{H} \mathbb{E}_{s \sim d^{\pi^{\mathrm{E}}}_h(\cdot)} \left[\mathrm{TV}\left(\pi^{\mathrm{mix}}_h(\cdot|s), \pi^{\mathrm{NBCU}}_h(\cdot|s)\right)\right] \\ &\leq H \sum_{h=1}^{H} \mathbb{E}_{s \sim d^{\pi^{\mathrm{E}}}_h(\cdot)} \left[\sqrt{\frac{|\mathcal{A}| \log(6|\mathcal{S}|H/\delta)}{\max\{n^{\mathrm{U}}_h(s), 1\}}}\right] \\ &= H \sqrt{|\mathcal{A}| \log(6|\mathcal{S}|H/\delta)} \sum_{h=1}^{H} \mathbb{E}_{s \sim d^{\pi^{\mathrm{E}}}_h(\cdot)} \left[\sqrt{\frac{1}{\max\{n^{\mathrm{U}}_h(s), 1\}}}\right] \\ &= H \sqrt{|\mathcal{A}| \log(6|\mathcal{S}|H/\delta)} \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} \sqrt{d^{\pi^{\mathrm{E}}}_h(s)} \sqrt{\frac{d^{\pi^{\mathrm{E}}}_h(s)}{\max\{n^{\mathrm{U}}_h(s), 1\}}} \end{aligned}$$

$$\leq H\sqrt{|\mathcal{A}| \log(6|\mathcal{S}|H/\delta)} \sum_{h=1}^{H} \sqrt{\sum_{s \in \mathcal{S}} \frac{d_h^{\pi^{\mathrm{E}}}(s)}{\max\{n_h^{\mathrm{U}}(s), 1\}}}.$$

Here the last inequality follows Cauchy-Swartz inequality. Notice that $n_h^{\mathrm{U}}(s)$ is the number of times that the state s appears in $\mathcal{D}^{\mathrm{U}}$ in time step h and thus follows the Binomial distribution of $\mathrm{Bin}(N_{\mathrm{tot}}, d_h^{\pi^{\mathrm{E}}}(s))$. By applying Lemma 4, for each (s, h), with probability at least $1-\delta$, we have

$$\frac{d_h^{\pi^{\mathrm{E}}}(s)}{\max\{n_h^{\mathrm{U}}(s), 1\}} \leq \frac{8\log(1/\delta)}{N_{\mathrm{tot}}}.$$

By union bound, with probability at least $1-\delta/2$, for all $(s, h) \in \mathcal{S} \times [H]$,

$$\frac{d_h^{\pi^{\mathrm{E}}}(s)}{\max\{n_h^{\mathrm{U}}(s), 1\}} \leq \frac{8\log(2|\mathcal{S}|H/\delta)}{N_{\mathrm{tot}}}.$$

Then, with probability at least $1-\delta$, we have

$$H \sum_{h=1}^{H} \mathbb{E}_{s \sim d_h^{\pi^{\mathrm{E}}}(\cdot)} \left[\mathrm{TV}\left(\pi_h^{\mathrm{mix}}(\cdot|s), \pi_h^{\mathrm{NBCU}}(\cdot|s) \right) \right] \leq H^2 \sqrt{\frac{8|\mathcal{S}||\mathcal{A}| \log^2(6|\mathcal{S}|H/\delta)}{N_{\mathrm{tot}}}}.$$

Finally, we upper bound the imitation gap. With probability at least $1-\delta$, we have

$$V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{NBCU}}) \leq (1-\eta)(1-\eta')\left(1 - \frac{1}{|\mathcal{A}|}\right) + H^2 \sqrt{\frac{8|\mathcal{S}||\mathcal{A}| \log^2(6|\mathcal{S}|H/\delta)}{N_{\mathrm{tot}}}}.$$

We set $\delta = H/N_{\mathrm{tot}}$ and obtain that

$$\begin{aligned}
&\mathbb{E}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{NBCU}}) \right] \\
&\leq \delta H + (1-\delta)\left((1-\eta)(1-\eta')\left(1 - \frac{1}{|\mathcal{A}|}\right) + H^2 \sqrt{\frac{8|\mathcal{S}||\mathcal{A}| \log^2(6|\mathcal{S}|H/\delta)}{N_{\mathrm{tot}}}} \right) \\
&\leq \frac{H^2}{N_{\mathrm{tot}}} + (1-\eta)(1-\eta')\left(1 - \frac{1}{|\mathcal{A}|}\right) + H^2 \sqrt{\frac{8|\mathcal{S}||\mathcal{A}| \log^2(6|\mathcal{A}|N_{\mathrm{tot}})}{N_{\mathrm{tot}}}} \\
&\leq (1-\eta)(1-\eta')\left(1 - \frac{1}{|\mathcal{A}|}\right) + 4H^2 \sqrt{\frac{2|\mathcal{S}||\mathcal{A}| \log^2(6|\mathcal{A}|N_{\mathrm{tot}})}{N_{\mathrm{tot}}}}.
\end{aligned}$$

On the other hand, we directly have $\mathbb{E}[V(\pi^{\mathrm{E}}) - V(\pi^{\mathrm{NBCU}})] \leq H$. We complete the proof.

E.4. Proof of Theorem 7

In the poison setting, we can conduct the same analysis as in the proof of Theorem 3 and demonstrate that $\pi^{\text{ISW-BC}}(\pi_h^{\mathrm{E}}(s)|s) = 1, \ \forall (s, \pi_h^{\mathrm{E}}(s)) \in \mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S},1}$, where $\mathcal{D}_h^{\mathrm{E}}$ is the set of state-action pairs in $\mathcal{D}^{\mathrm{E}}$ in time step h and $\mathcal{D}_h^{\mathrm{S},1} = \{(s,a) \in \mathcal{D}_h^{\mathrm{S}} : d_h^{\pi^{\mathrm{E}}}(s) > 0, a = \pi_h^{\mathrm{E}}(s)\}$. According to (Rajaraman et al., 2020, Lemma 4.3), we have

$$V(\pi^{\mathrm{E}}) - V(\pi^{\text{ISW-BC}}) \leq H \sum_{h=1}^{H} \mathbb{E}_{s \sim d_h^{\pi^{\mathrm{E}}}(\cdot)} \left[\mathrm{TV}\left(\pi_h^{\mathrm{E}}(\cdot|s), \pi_h^{\text{ISW-BC}}(\cdot|s) \right) \right].$$

Since the expert policy is assumed to be deterministic, we can obtain

$$\begin{aligned}
V(\pi^{\mathrm{E}}) - V(\pi^{\text{ISW-BC}}) &\leq H \sum_{h=1}^{H} \mathbb{E}_{s \sim d_h^{\pi^{\mathrm{E}}}(\cdot)} \left[\mathbb{E}_{a \sim \pi_h^{\text{ISW-BC}}(\cdot|s)} \left[\mathbb{I}\left\{ a \neq \pi_h^{\mathrm{E}}(s) \right\} \right] \right] \\
&\leq H \sum_{h=1}^{H} \mathbb{E}_{s \sim d_h^{\pi^{\mathrm{E}}}(\cdot)} \left[\mathbb{I}\left\{ (s, \pi_h^{\mathrm{E}}(s)) \notin \mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S},1} \right\} \right].
\end{aligned}$$

Let $\mathcal{D}^{\mathrm{S,clean}}$ denote the non-corrupted dataset in $\mathcal{D}^{\mathrm{S}}$. Then we can obtain that

$$\begin{aligned} V(\pi^{\mathrm{E}}) - V(\pi^{\text{ISW-BC}}) &\overset{(a)}{\leq} H \sum_{h=1}^{H} \mathbb{E}_{s \sim d_h^{\pi^{\mathrm{E}}}(\cdot)} \left[\mathbb{I}\left\{ (s, \pi_h^{\mathrm{E}}(s)) \notin \mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S,clean}} \right\} \right] \\ &= H \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} d_h^{\pi^{\mathrm{E}}}(s) \mathbb{I}\left\{ (s, \pi_h^{\mathrm{E}}(s)) \notin \mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S,clean}} \right\}, \end{aligned}$$

where $\mathcal{D}_h^{\mathrm{S,clean}}$ denotes the set of state-action pairs in $\mathcal{D}^{\mathrm{S,clean}}$ in time step h. Inequality (a) follows that $\mathcal{D}_h^{\mathrm{S,clean}} \subseteq \mathcal{D}_h^{\mathrm{S,1}}$ since $\mathcal{D}^{\mathrm{S,clean}}$ is collected by the expert policy. Taking expectation over the randomness in $\mathcal{D}^{\mathrm{E}}$ and $\mathcal{D}^{\mathrm{S,clean}}$ on both sides yields that

$$\mathbb{E}_{\mathcal{D}^{\mathrm{E}}, \mathcal{D}^{\mathrm{S,clean}}} \left[V(\pi^{\mathrm{E}}) - V(\pi^{\text{ISW-BC}}) \right] \leq H \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} d_h^{\pi^{\mathrm{E}}}(s) \mathbb{P}\left((s, \pi_h^{\mathrm{E}}(s)) \notin \mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S,clean}} \right).$$

Notice that both $\mathcal{D}^{\mathrm{E}}$ and $\mathcal{D}^{\mathrm{S,clean}}$ are collected by the expert policy. Then if $|\mathcal{D}^{\mathrm{E}}| + |\mathcal{D}^{\mathrm{S,clean}}| \geq 1$, we can calculate that for each $(s, h) \in \mathcal{S} \times [H]$,

$$\begin{aligned} d_h^{\pi^{\mathrm{E}}}(s) \mathbb{P}\left((s, \pi_h^{\mathrm{E}}(s)) \notin \mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S,clean}} \right) &= d_h^{\pi^{\mathrm{E}}}(s) \left(1 - d_h^{\pi^{\mathrm{E}}}(s) \right)^{|\mathcal{D}^{\mathrm{E}}| + |\mathcal{D}^{\mathrm{S,clean}}|} \\ &\leq \frac{4}{9(|\mathcal{D}^{\mathrm{E}}| + |\mathcal{D}^{\mathrm{S,clean}}|)}, \end{aligned}$$

where the last inequality follows Lemma 5. If $|\mathcal{D}^{\mathrm{E}}| + |\mathcal{D}^{\mathrm{S,clean}}| = 0$, we directly have that

$$d_h^{\pi^{\mathrm{E}}}(s) \mathbb{P}\left((s, \pi_h^{\mathrm{E}}(s)) \notin \mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S,clean}} \right) \leq 1 = \frac{1}{\max\{ |\mathcal{D}^{\mathrm{E}}| + |\mathcal{D}^{\mathrm{S,clean}}|, 1 \}}.$$

We unify the above two inequalities and get that

$$d_h^{\pi^{\mathrm{E}}}(s) \mathbb{P}\left((s, \pi_h^{\mathrm{E}}(s)) \notin \mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S,clean}} \right) \leq \frac{1}{\max\{ |\mathcal{D}^{\mathrm{E}}| + |\mathcal{D}^{\mathrm{S,clean}}|, 1 \}}.$$

Now we proceed to upper bound the imitation gap.

$$\begin{aligned} \mathbb{E}_{\mathcal{D}^{\mathrm{E}}, \mathcal{D}^{\mathrm{S,clean}}} \left[V(\pi^{\mathrm{E}}) - V(\pi^{\text{ISW-BC}}) \right] &\leq H \sum_{h=1}^{H} \sum_{s \in \mathcal{S}} d_h^{\pi^{\mathrm{E}}}(s) \mathbb{P}\left((s, \pi_h^{\mathrm{E}}(s)) \notin \mathcal{D}_h^{\mathrm{E}} \cup \mathcal{D}_h^{\mathrm{S,clean}} \right) \\ &\leq \frac{|\mathcal{S}| H^2}{\max\{ |\mathcal{D}^{\mathrm{E}}| + |\mathcal{D}^{\mathrm{S,clean}}|, 1 \}}. \end{aligned}$$

Note that $|\mathcal{D}^{\mathrm{E}}| + |\mathcal{D}^{\mathrm{S,clean}}| \sim \mathrm{Bin}(N_{\mathrm{tot}}, \eta + (1-\eta)\eta')$. Taking expectation with respect to $|\mathcal{D}^{\mathrm{E}}| + |\mathcal{D}^{\mathrm{S,clean}}|$ yields that

$$\begin{aligned} \mathbb{E}\left[V(\pi^{\mathrm{E}}) - V(\pi^{\text{ISW-BC}}) \right] &\leq \mathbb{E}\left[\frac{|\mathcal{S}| H^2}{\max\{ |\mathcal{D}^{\mathrm{E}}| + |\mathcal{D}^{\mathrm{S,clean}}|, 1 \}} \right] \\ &\leq \mathbb{E}\left[\frac{2|\mathcal{S}| H^2}{|\mathcal{D}^{\mathrm{E}}| + |\mathcal{D}^{\mathrm{S,clean}}| + 1} \right] \\ &\overset{(a)}{\leq} \frac{2|\mathcal{S}| H^2}{N_{\mathrm{tot}}(\eta + (1-\eta)\eta')} \\ &= \frac{2|\mathcal{S}| H^2}{N_{\mathrm{E}} + \eta' N_{\mathrm{S}}}. \end{aligned}$$

Inequality (a) follows Lemma 3. We finish the proof.

F. Technical Lemmas

Lemma 3. *For any $N \in \mathbb{N}_+$ and $p \in (0,1)$, if the random variable X follows the binomial distribution, i.e., $X \sim \mathrm{Bin}(N, p)$, then we have that*

$$\mathbb{E}\left[\frac{1}{X+1} \right] \leq \frac{1}{Np}.$$

Proof.

$$
\begin{aligned}
\mathbb{E}\left[\frac{1}{X+1}\right] &= \sum_{x=0}^{N} \left(\frac{1}{x+1}\right) \frac{N!}{x!(N-x)!} p^x (1-p)^{N-x} \\
&= \frac{1}{(N+1)p} \sum_{x=1}^{N+1} \left(\frac{(N+1)!}{x!(N+1-x)!}\right) p^x (1-p)^{N+1-x} \\
&= \frac{1}{(N+1)p} \left(1 - (1-p)^{N+1}\right) \leq \frac{1}{Np}.
\end{aligned}
$$

□

Lemma 4 (Binomial concentration (Lemma A.1 in (Xie et al., 2021)))**.** *For any* $N \in \mathbb{N}_+$ *and* $p \in (0, 1)$*, suppose* $X \sim \mathrm{Bin}(N, p)$*. Then with probability at least* $1 - \delta$*, we have*

$$
\frac{p}{\max\{X, 1\}} \leq \frac{8 \log(1/\delta)}{N}.
$$

Lemma 5. *For any* $N \in \mathbb{N}_+$ *and* $x \in [0, 1]$*, consider the function* $f(x) := x(1-x)^N$*, then we have*

$$
\forall x \in [0, 1], \;\; f(x) \leq \frac{4}{9N}.
$$

Proof. We calculate that $f'(x) = (1-x)^{N-1}(1-(N+1)x)$. It is direct to have that $f(x)$ achieves its maximum at $x^\star = 1/(N+1)$. Furthermore, we have

$$
f\left(\frac{1}{N+1}\right) = \frac{1}{N}\left(1 - \frac{1}{N+1}\right)^{N+1} \overset{(a)}{\leq} \frac{1}{eN} \leq \frac{4}{9N}.
$$

Inequality (a) follows that $(1 + x/N)^N \leq \exp(x), \; \forall N \geq 1, |x| \leq N$. We complete the proof. □

G. Experiments Details and Additional Results

G.1. Experiment Details

In this section, we present the experiment details to facilitate the replication of our results. Our codebase will be made available for public access at a later stage. The experiments are conducted on a machine comprising 48 CPU cores and 4 V100 GPU cores. We repeat each experiment 5 times using different random seeds (2021, 2022, 2023, 2024, and 2025).

G.1.1. Locomotion Control

In this study, we evaluate the performance of various imitation learning algorithms on four locomotion control tasks from the MuJoCo suite: `Ant-v2`, `HalfCheetah-v2`, `Hopper-v2`, and `Walker2d-v2`. These tasks are widely used in the literature and are considered challenging benchmarks.

To train the expert policy, we use the online Soft Actor-Critic (SAC) algorithm (Haarnoja et al., 2018) with 1 million training steps. We implement the algorithm using the rlkit codebase, which is available at `https://github.com/rail-berkeley/rlkit`. The training curves of the online SAC agent are shown in Figure 5. We treat the resulting policy as the expert policy and use it to generate expert trajectories.

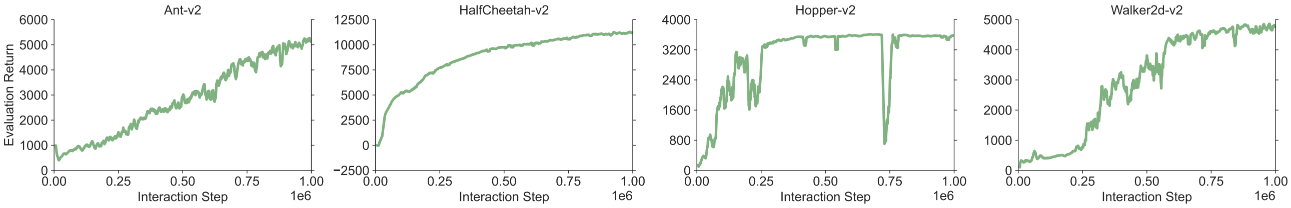

Figure 5. Training curves of online SAC on 4 locomotion control environments.

In our experimental setup, we utilize an expert dataset comprising of 1 expert trajectory collected by the trained SAC agent. Additionally, all algorithms are provided with a supplementary dataset. There are two setting of the supplementary data.

- `Full Replay`. The supplementary dataset is obtained from the replay buffer of the online SAC agent, which has over one million samples, equivalent to 1000+ trajectories. The rapid convergence of online SAC, as illustrated in Figure 5, implies that the replay buffer is enriched with a substantial number of expert-level trajectories. As a result, we expect that utilizing the supplementary data without any modification may lead to desirable results.
- `Noisy Expert`. The supplementary dataset comprises of 10 clean expert trajectories and 5 noisy expert trajectories. In this case, we replace the action labels in the noisy trajectories with random actions drawn from $[-1, 1]$. This replacement creates noisy action labels for the expert states, leading to a significant distribution shift at the state-action level, as noted in Remark 1. The high degree of distribution shift makes it challenging for using the supplementary data.

We use a 2-hidden-layer multi-layer perceptron (MLP) with hidden size 256 and ReLU activation for all algorithms, as the state information in locomotion control tasks is informative by design. The codebase of DemoDICE is based on the original authors' work, which can be accessed at `https://github.com/KAIST-AILab/imitation-dice`. For DWBC, we also use the authors' codebase, which is available at `https://github.com/ryanxhr/DWBC`. We experimented with different hyper-parameters for both algorithms but found that the default parameters provided by the authors work well. We normalize state observations in the dataset before training all algorithms, following (Kim et al., 2022b). This is crucial for achieving satisfactory performance.

In training the discriminator of ISW-BC, we use the gradient penalty (GP) regularization, as recommended by (Kim et al., 2022b). We add the following loss to the original loss (4) to enforce 1-Lipschitz continuity:

$$\min_{\theta} \sum_{(s,a) \in \mathcal{B}} \left(\| g(s, a; \theta) \| - 1 \right)^2 ,$$

where g is the gradient of the discriminator $c(s, a; \theta)$, and $\mathcal{B}$ is a mini-batch. This promotes the learning of smooth features and can improve generalization performance.

In our implementation of ISW-BC, we employ 2-hidden-layer MLPs with 256 hidden units and ReLU activation for both the discriminator and policy networks. We use a batch size of 256 and Adam optimizer with a learning rate of 0.0003 for training both networks. The training objective is to maximize the log-likelihood. We set δ to 0 and use a gradient penalty coefficient of 8 by default, unless otherwise stated. The training process is carried out for 1 million iterations. We evaluate the performance every 10k iterations with 10 episodes. The normalized score in the last column of Table 2 is computed in the following way:

$$\text{Normalized score} = \frac{\text{Expert performance} - \text{Agent performance}}{\text{Expert performance} - \text{Random policy performance}}. \tag{21}$$

G.1.2. Atari Games

We evaluate algorithms on a set of 5 Atari games from the standard benchmark: `Alien`, `MsPacman`, `Phoenix`, `Qbert`, and `SpaceInvaders`. We preprocess the game environments using a standard set of procedures, including sticky actions with a probability of 0.25, grayscaling, downsampling to an image size of [84, 84], and stacking frames of 4. These procedures follow the instructions provided by the dopamine codebase, which is available at `https://github.com/google/dopamine/blob/master/dopamine/discrete_domains/atari_lib.py`. The final image inputs are of shape (84, 84, 4).

We use the replay buffer data from an online DQN agent, which is publicly available at `https://console.cloud.google.com/storage/browser/atari-replay-datasets`, thanks to the work of (Agarwal et al., 2020b). The dataset consists of 200 million frames, divided into 50 indexed buckets (ranging from 0 to 49). However, using the entire dataset is computationally infeasible[4] and unnecessary for our task. Therefore, we select specific buffer buckets for imitation learning.

We choose the expert data from bucket index 49, using only the first 400K frames for training. This makes the task challenging (we find that BC performs well with 1M frames of expert data). For the full replay setting, we select supplementary data

[4] Loading 200M frames requires over 500GB memory.

from buffer indices 45 to 48, using the first 400K frames from each bucket. This yields a supplementary dataset that is 4 times larger than the expert data. In the noisy task setting, we follow the same procedure for selecting supplementary data, but replace the action labels with random labels on buffer index 45.

All agents employ the same convolutional neural network (CNN) architecture as the DQN agent, consisting of three convolutional blocks. The first block applies a filter size of 8, a stride of 4, and has a channel size of 32. The second block uses a filter size of 4, a stride of 4, and a channel size of 64, while the third block applies a filter size of 3, a stride of 4, and has a channel size of 64. All blocks use the ReLU activation function. The feature representations are flattened to a vector, on which a 1-hidden-layer MLP with a hidden size of 512 and ReLU activation function is applied. Finally, the outputs are passed through a softmax function to obtain a probability distribution.

Atari games are not considered in (Kim et al., 2022b; Xu et al., 2022a) and public implementations of DemoDICE and DWBC for Atari games are not available. To use these methods in the Atari environment, we extend their original implementation by replacing the MLP used in locomotion control with the CNN described earlier. Implementing ISW-BC is a little more complicated. We use the same CNN policy network as in the other methods, but find that directly training the discriminator from scratch is less effective. This is because the discriminator tends to focus on irrelevant background information instead of the decision-centric part. To overcome this issue, we build the discriminator upon the feature extractor of the policy network, leveraging its ability to extract useful information. The discriminator is an MLP with ReLU activation and a hidden size of 1024: the image feature representation has a dimension 512 and the action feature representation also has a dimension 512 (we randomly project one-hot discrete actions to a 512-dimension space). We find that the depth of the MLP is crucial for performance, using a depth of 1 for the full replay setting and 3 for the noisy expert setting. We clip the importance sampling ratio for numerical stability, using a minimum value of 0 and a maximum value of 5 for the full replay setting, and a minimum value of 0.2 and a maximum value of 5 for the noisy expert setting. We provide ablation studies of these hyperparameters in Appendix G.3.1.

All methods were optimized using the Adam optimizer with a learning rate of 0.00025 and a batch size of 256. The training objective is to maximize the log-likelihood. The training process consisted of 200K gradient steps. Every 2K gradient steps, the algorithms were evaluated by running 10 episodes and computing the raw game scores. The normalized score in the last column of Table 3 is computed by Equation (21).

G.1.3. Object Recognition

We utilize the publicly available DomainNet dataset (Peng et al., 2019) for our experiments, which can be accessed at `http://csr.bu.edu/ftp/visda/2019/multi-source`. This dataset comprises six sub-datasets: `clipart`, `infograph`, `painting`, `quickdraw`, `real`, and `sketch`, with 2103, 2626, 2472, 4000, 4864, and 2213 images, respectively. Our task involves recognizing objects from 10 different classes: `bird`, `feather`, `headphones`, `ice_cream`, `teapot`, `tiger`, `whale`, `windmill`, `wine_glass`, and `zebra`. We divided the images into training and test sets, with 80% for training and 20% for testing.

We employ a 2-hidden-layer neural network with a hidden size of 512 and ReLU activation as the classifier. To extract features from images, we utilize the pretrained ResNet-18 model (trained on ImageNet), which has a feature dimension of 512. The ResNet-18 model can be accessed at `https://pytorch.org/vision/main/models/generated/torchvision.models.resnet18.html`. We opted for this approach as training such a large convolutional neural network directly on the DomainNet dataset proved to be ineffective. The training objective is to minimize the cross-entropy loss. To optimize the network parameters, we use the stochastic gradient descent (SGD) optimizer with a learning rate of 0.01 and momentum of 0.9. Additionally, we apply weight decay with a coefficient of 0.0005. The models are trained for 100 epochs with a batch size of 100, following the standard practice.

The discriminators used in ISW-BC and DWBC are implemented as 2-hidden-layer neural networks with ReLU activation. It's important to note that these discriminators take both the image and label as inputs. The image input is processed by the pre-trained and fixed ResNet-18, while the label input is projected to the same dimension (512) by a random projection matrix. The hidden size for the discriminator is set to 1024 for ISW-BC and 1025 for DWBC, as the discriminator in DWBC also takes the log-likelihood as an input. For ISW-BC, the discriminator is trained independently for 100 epochs with the same optimization configuration as the classifier. Afterward, the discriminator is fixed, and its output is used to compute the importance sampling ratio, which is then used to train the classifier.

G.2. Additional Results

G.3. Training Curves

Training curves. The training curves on the MuJoco locomotion control tasks are displayed in Figure 6 and Figure 7. The training curves on Atari games are displayed in Figure 8 and Figure 9. The training curves on the object recognition task are displayed in Figure 10.

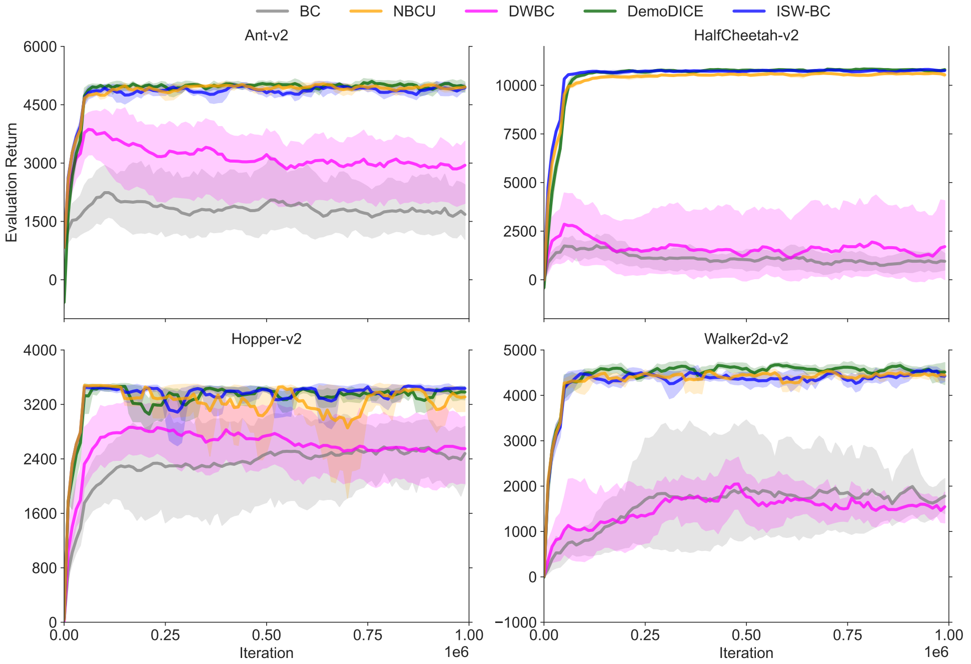

Figure 6. Training curves of algorithms on the locomotion control task with the full replay setting. Solid lines correspond to the mean performance and shaded regions correspond to the 95% confidence interval. Same as other figures.

G.3.1. Ablation Study

In this section, we present ablation studies conducted on Atari games, aiming to provide insights into the underlying working scheme of our method. We specifically emphasize Atari games due to their high-dimensional image inputs, making these tasks particularly challenging. In contrast, the other two tasks, locomotion control and object recognition, involve informative vector inputs, setting them apart from the unique characteristics of Atari games.

Ablation Study on Feature Representations of Discriminator Network. Our study reveals that employing a separate CNN for the discriminator yields inferior results compared to utilizing the feature extractor of the policy network. Please refer to Figure 11. Our conjecture is that training the discriminator independently may cause it to fit noise information (e.g., background). In contrast, the policy CNN network is capable of learning decision-centric information, enabling an effective approach to building the discriminator network through the feature extractor of the policy network.

Ablation Study on Depth of Discriminator Network. We have discovered that the number of discriminator layers plays a crucial role in the performance of Atari games. The training curves, depicted in both Figure 12 and Figure 13, illustrate the performance variation based on the number of layers in the discriminator network. Notably, a 1-hidden-layer neural network yields the best results for the full replay setting, while a 3-hidden-layer neural network performs optimally in the noisy expert setting. It is important to note that this phenomenon is specific to Atari games. We do not have a good explanation yet. We believe this deserves further investigation in the future work.

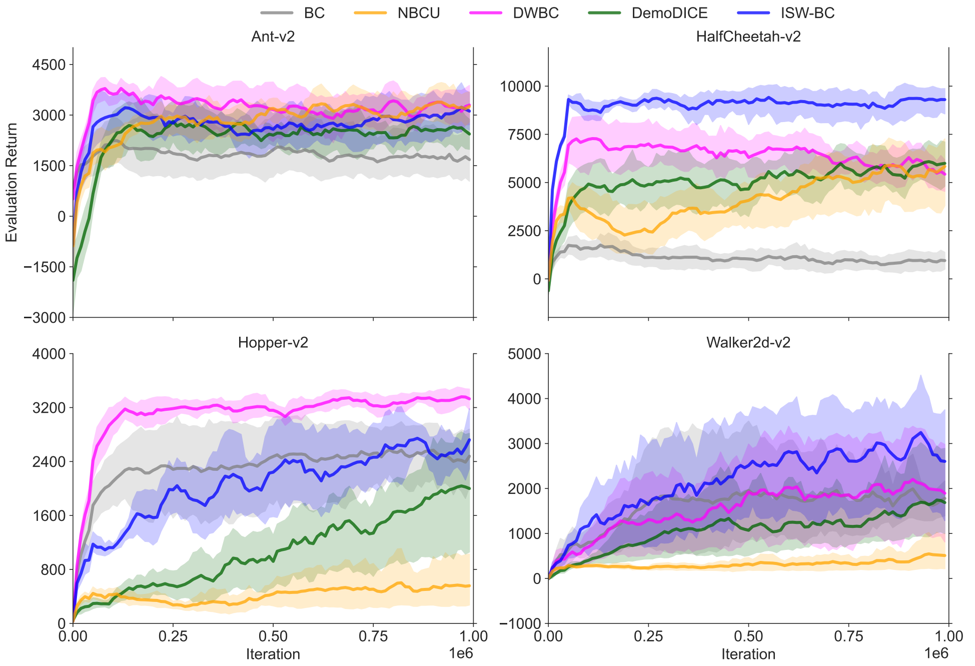

Figure 7. Training curves of algorithms on the locomotion control task with the noisy expert setting.

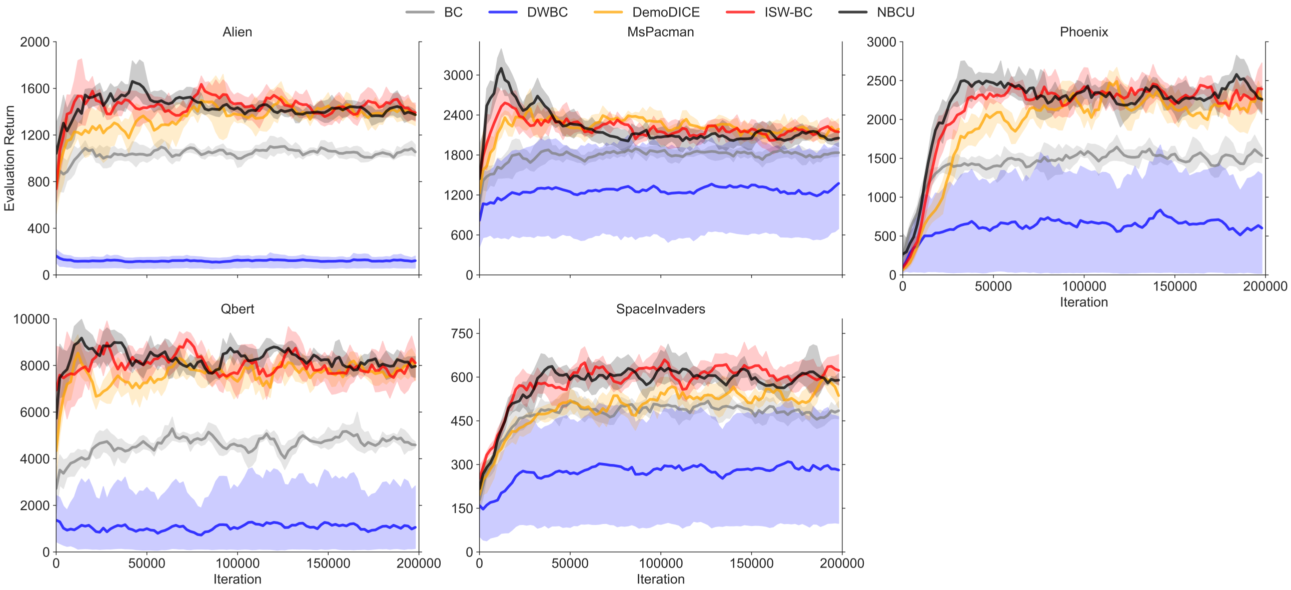

Figure 8. Training curves of algorithms on the Atari games with the full replay setting.

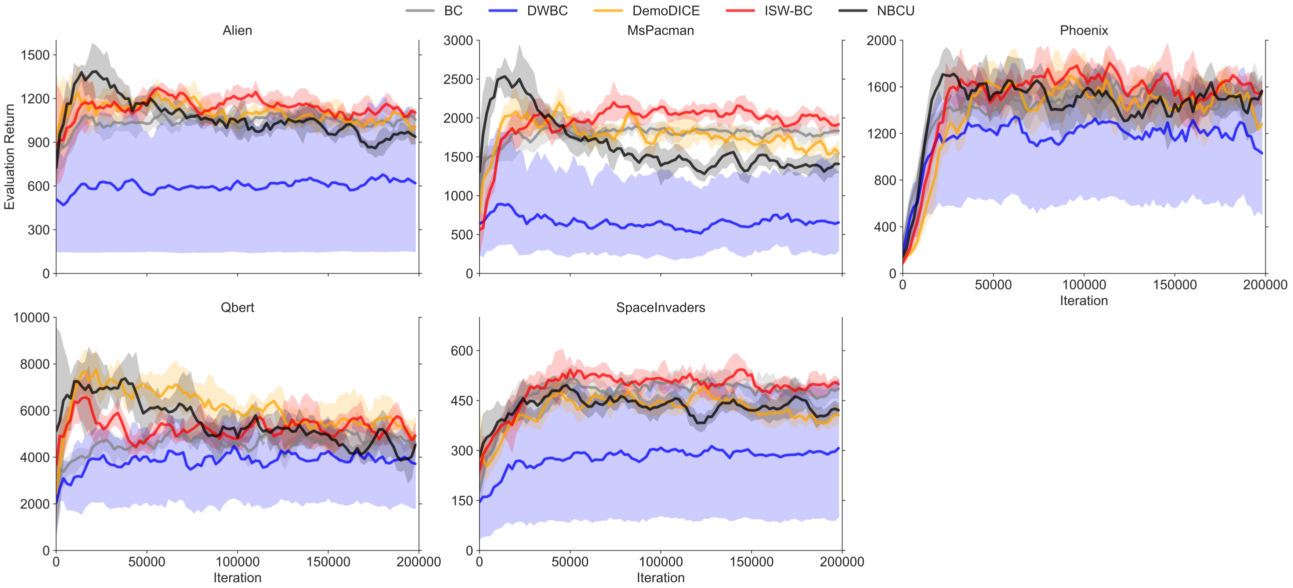

Figure 9. Training curves of algorithms on the Atari games with the noisy expert setting.

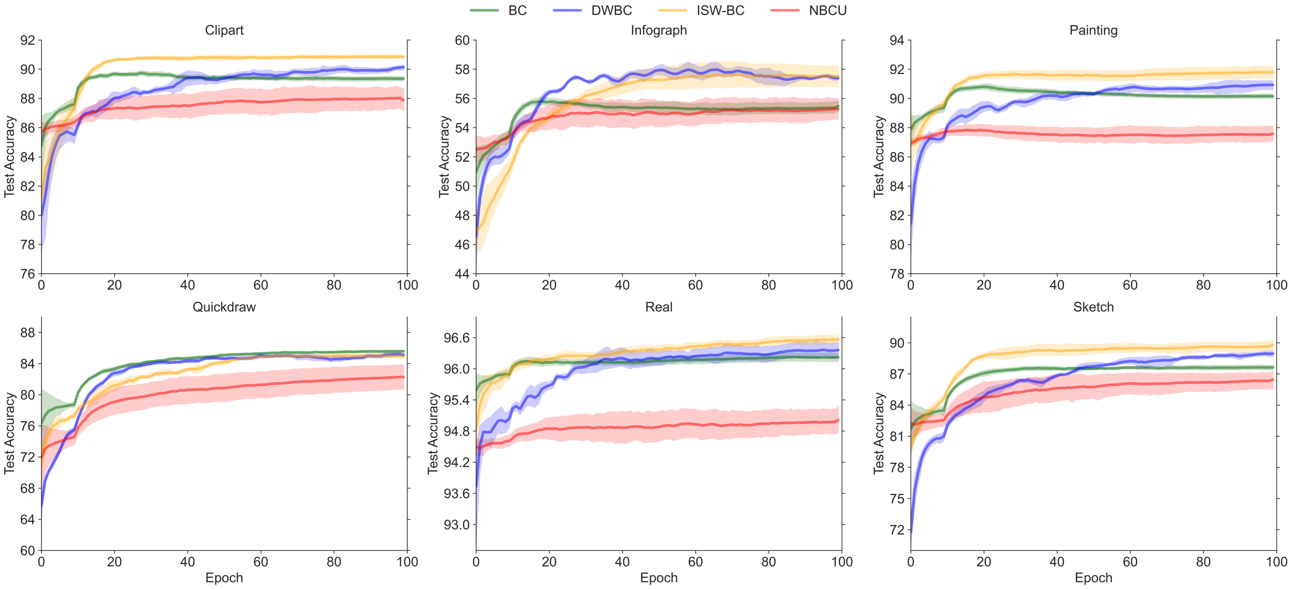

Figure 10. Training curves of algorithms on the object recognition task using the DomainNet dataset.

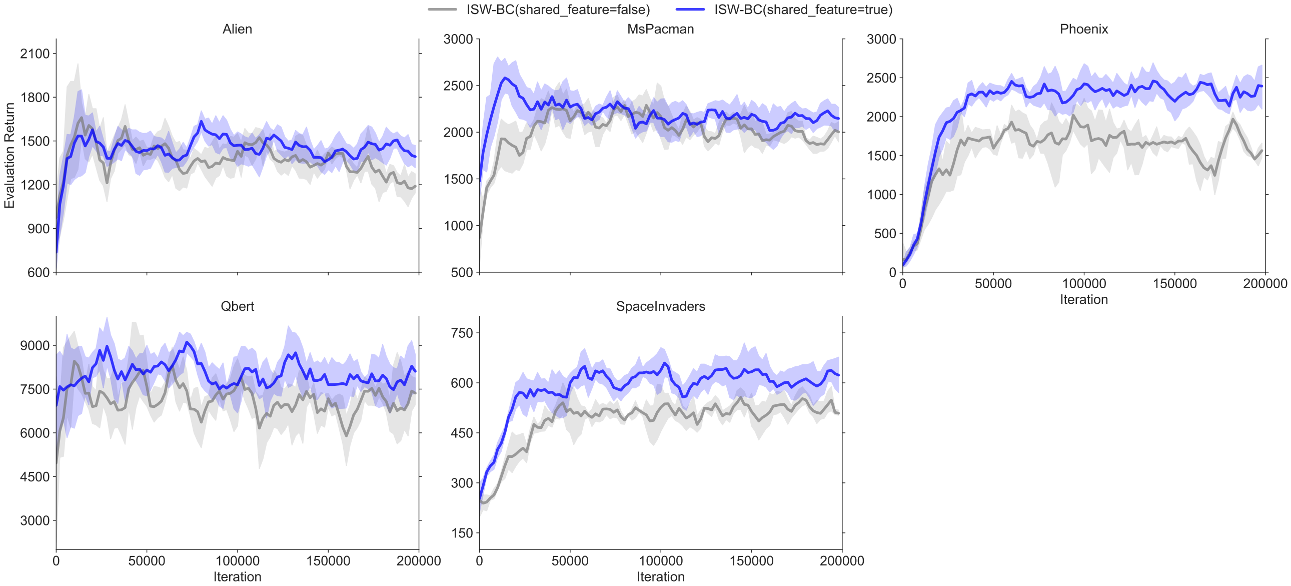

Figure 11. Training curves of ISW-BC on the Atari games in the full replay setting. We test the performance with different feature extractors of the discriminator.

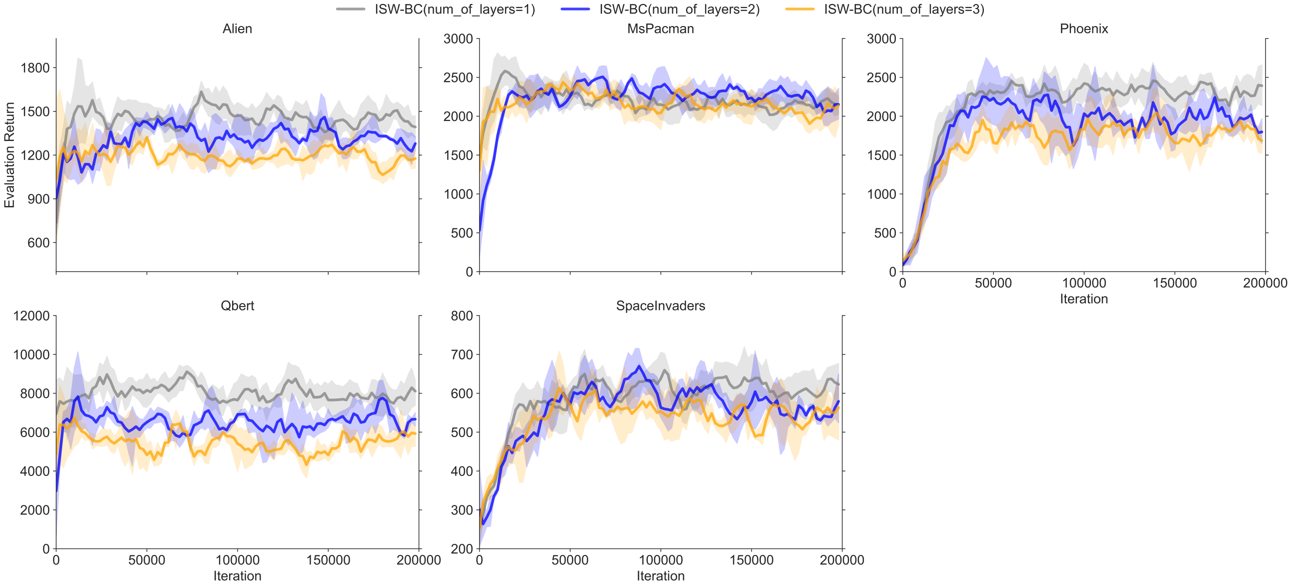

Figure 12. Training curves of ISW-BC on the Atari games in the full replay setting. We test the performance with different number of layers for the discriminator network.

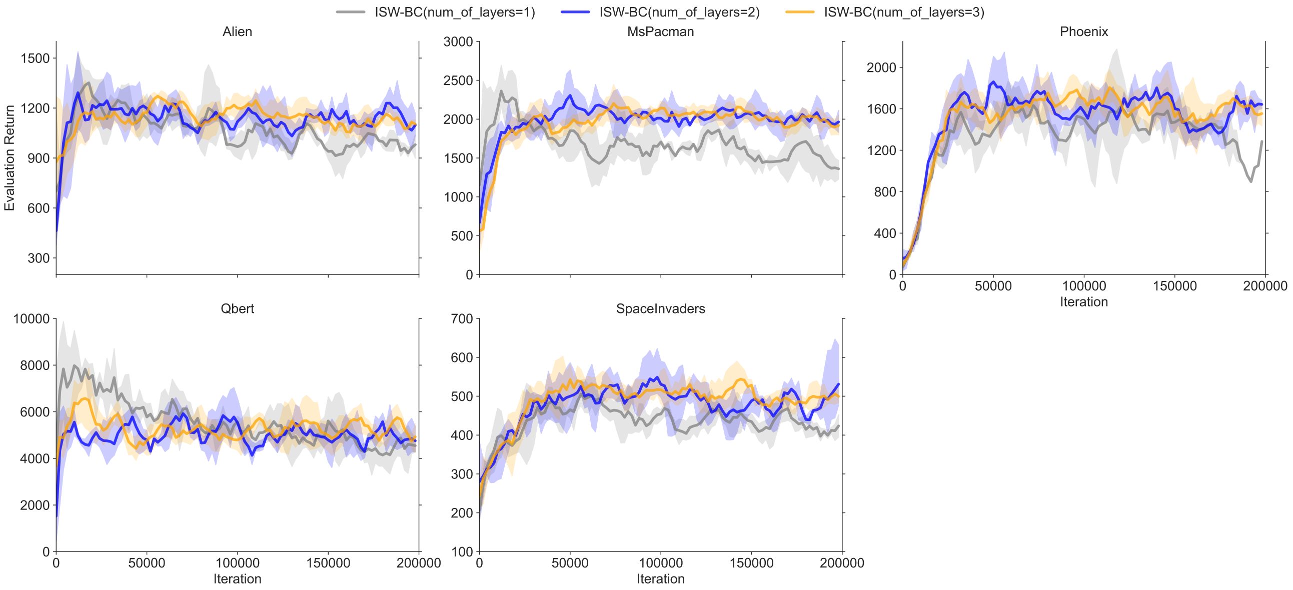

Figure 13. Training curves of ISW-BC on the Atari games with the noisy expert setting. We test the performance with different number of layers for the discriminator network.