DECOUPLED Q-CHUNKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Bootstrapping bias problem is a long-standing challenge in temporal-difference (TD) methods in off-policy reinforcement learning (RL). Multi-step return backups can alleviate this issue but require delicate importance sampling to correct their off-policy bias. Recent work has proposed to use chunked critics, which estimate the value of short action sequences ("chunks") rather than individual actions, enabling unbiased multi-step backup. However, extracting policies from chunked critics is challenging: policies must output the entire action chunk open-loop, which can be sub-optimal in environments that require policy reactivity and also challenging to model especially when the chunk length grows. Our key insight is to decouple the chunk length of the critic from that of the policy, allowing the policy to operate over shorter action chunks. We propose a novel algorithm that achieves this by optimizing the policy against a distilled critic for partial action chunks, constructed by optimistically backing up from the original chunked critic to approximate the maximum value achievable when a partial action chunk is extended to a complete one. This design retains the benefits of multi-step value propagation while sidestepping both the open-loop sub-optimality and the difficulty of learning policies over long action chunks. We evaluate our method on challenging, long-horizon offline goal-conditioned benchmarks and show that it reliably outperforms prior methods.

1 INTRODUCTION

A reinforcement learning (RL) agent can in principle solve any task with a well-defined reward function, but training an RL agent from scratch can be sample inefficient. In many practical problems, we have access to an offline dataset of trajectories that serves as a great prior to accelerate learning. Temporal-difference (TD)-based RL algorithms, which learn a value network to perform approximate dynamic programming via value backups, are particularly suitable in this setting because they are designed to handle off-policy data. A well-known yet long-lasting bottleneck, however, is the bootstrapping bias problem (Jaakkola et al., 1993; Sutton et al., 1998; De Asis et al., 2018; Park et al., 2025)—as the value network regresses towards its own estimates, any error compounds across time steps, making accurate value propagation challenging especially in long-horizon, sparse reward tasks.

Multi-step return backups (such as n-step return (Sutton et al., 1998)) can alleviate bootstrapping bias by effectively reducing the time horizon, but naïvely applying them can result in another form of bias that causes the value estimates to be overly conservative/pessimistic. While it is possible to correct such systematic biases with importance sampling (Munos et al., 2016), they often require additional heuristics and truncations to balance a delicate scale between bias and variance which is often tricky to tune. Recent works (Seo & Abbeel, 2024; Li et al., 2025a; Tian et al., 2025; Li et al., 2025b) leverage chunked value functions, which estimate the value of short action sequences ("chunks") rather than a single action. This formulation allows n-step return backup without the pessimistic bias (under the open-loop consistency condition, which we will formalize in Section 4). However, directly optimizing a policy over full action chunks is difficult, particularly as the chunk size grows, and it is still unclear how to best extract a policy from a chunked critic.

In this work, we develop a simple, novel technique to address this challenge. We train a policy to predict a shorter, partial action chunk using the chunked critic that takes in longer, complete action chunks. The key idea enabling this approach is a 'distilled' chunked critic with a chunk size that matches the policy: it optimistically regresses to the original chunked critic to approximate the maximum value that the partial action chunk can achieve after being extended into a full action chunk. Conceptually, while optimization is still performed for the longer, complete action chunks, the policy network is only trained to output the partial action chunk of an optimized complete action chunk. This way, the policy only needs to predict a much shorter action chunk (*e.g.*, in the extreme case, only one action), which often admits a much simpler distribution, while enjoying the value learning benefits from the use of chunked critics.

Our main contributions are two-fold. On the theoretical side, we provide a formal analysis of Q-learning with action chunking, identifying the open-loop value learning bias and characterizing the conditions under which action chunking critic backup is preferable over n-step return backup with a single-step critic. On the empirical side, we propose a novel technique, **Decoupled Q-chunking (DQC)**, that addresses the policy learning challenge in action chunking Q-learning by decoupling the policy chunk size from the critic chunk size. DQC trains a policy to only predict a partial action chunk, significantly reducing the policy learning challenge, while retaining the value learning benefits of the chunked critic. We instantiate this technique as a practical offline RL algorithm that outperforms the previous state-of-the-art method on the hardest set of environments in OGBench (Park et al., 2024a), a challenging, long-horizon goal-conditioned RL benchmark.

2 Related Work

Offline and offline-to-online reinforcement learning methods assume access to an offline dataset to learn a policy without interactions with the environment (offline) (Kumar et al., 2020; Kostrikov et al., 2021; Tarasov et al., 2024) or with as little online interaction with the environment as possible (offline-to-online) (Lee et al., 2022; Ball et al., 2023; Nakamoto et al., 2024). TD-based RL algorithms have been a popular choice for these problem settings as they naturally handle off-policy data while requiring no on-policy rollouts, and they also exhibit good online sample-efficiency (Chen et al., 2021; D'Oro et al., 2022). A large body of literature in these areas has been focusing on tackling the distribution shift challenge by appropriately constraining the policies with respect to the prior offline data, and most of them use the standard 1-step TD backup for Q-learning, which has been known to suffer from the bootstrapping bias problem in the RL literature (Jaakkola et al., 1993; Sutton et al., 1998). To tackle this, recent work (Jeong et al., 2022; Park & Lee, 2024; Park et al., 2025; Li et al., 2025b) has shown that multi-step return backups are effective for improving offline/offline-to-online Q-learning agents. These methods either use a standard single-step critic network (Park et al., 2025) that suffers from the off-policy bias, or use a 'chunked,' multi-step critic network (Li et al., 2025b) that does not have such bias but poses a huge policy learning challenge when the chunk size is too large. Our method brings the best of both worlds—it uses action chunking to avoid the off-policy bias while simultaneously avoiding the policy learning challenge by extracting a simpler policy that predicts a shorter action chunk from the full-chunk-sized critic.

Multi-step return backups are computed with multi-step off-policy rewards that can lead to systematic value underestimation (Sutton et al., 1998; Peng & Williams, 1994; Konidaris et al., 2011; Thomas et al., 2015), and there has been a rich literature (Precup et al., 2000; Munos et al., 2016; Rowland et al., 2020) dedicated to fix these biases via importance sampling (Kloek & Van Dijk, 1978) with truncation (Ionides, 2008). These approaches often require a careful balance between bias and variance that can be tricky to tune. More recently, Seo & Abbeel (2024); Li et al. (2025a); Tian et al. (2025); Li et al. (2025b) group temporally extended sequences of actions as chunks and directly estimate the value of an action chunk rather than a single action. Such a formulation allows the value backup to operate directly in the chunk space, which allows multi-step return backup without the systematic biases from the sub-optimal off-policy data. Despite their empirical success, we still lack a good theoretical understanding of the convergence of TD-learning with 'chunked' critics, as well as when it should be favored over more traditional multi-step returns. Our work lays out the theoretical foundation for Q-learning with critic chunking, and identifies an important yet subtle, often overlooked bias in the TD-backup. We quantify such bias and provide the condition under which TD backup using critic chunking is guaranteed to perform better than the standard n-step return backup with a single-step critic.

See additional discussions for related work in hierarchical reinforcement learning in Appendix I.

3 Preliminaries

Reinforcement learning can be formalized as a Markov decision process, $\mathcal{M} = (\mathcal{S}, \mathcal{A}, T, r, \rho, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $T : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{A}}$ is the transition kernel that defines the next state distribution conditioned on the current state and the current action (*e.g.*, $s' \sim T(\cdot \mid s, a)$), $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function, $\rho \in \Delta_S$ is the initial state distribution, and $\gamma \in [0, 1)$ is the discount factor. We also assume we have access to a prior offline dataset $D = \{(s_0^i, a_0^i, r_0^i, s_1^i, a_1^i, r_1^i, \cdots, s_H^i)\}_{i=1}^{|D|}$ where the goal is to learn a policy, $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$ that maximizes its return, $\eta(\pi) = \mathbb{E}_{s_{t+1}\sim T(\cdot|s_t,a_t),a_t\sim\pi(\cdot|s_t),s_0\sim\rho}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]$, the cumulative discounted sum of rewards that the policy receives in expectation.

Temporal difference learning. Modern value-based reinforcement learning methods often learn a critic network, $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ to approximate the maximum discounted cumulative reward starting from state s and action a, and the critic is often trained using the temporal-difference (TD) loss:

$$L(\phi) = \mathbb{E}_{s,a,s'\sim\mathcal{D}}\left[(Q_\phi(s, a) - r(s, a) - \gamma\bar{Q}(s', a'^\star))^2\right], \tag{1}$$

where $\bar{Q}$ is the target critic that is set to the same critic with its parameters set to an exponential moving average of ϕ, and $a'^\star = \arg\max_{a'} Q(s', a')$ (often approximated by a policy π_θ).

Implicit value learning with implicit maximization loss function. Instead of using $Q(s', a'^\star \sim \pi_\theta(s'))$ as the TD target, we can use what we refer to as an *implicit maximization* loss function f_{imp} to learn a value function $V_\xi(s)$ that approximates the maximum value $Q(s, a^\star)$ (Kostrikov et al., 2021; Hansen-Estruch et al., 2023):

$$L(\xi) = \mathbb{E}_{s,a\sim\mathcal{D}}\left[f^\kappa_{\text{imp}}(\bar{Q}(s, a) - V_\xi(s))\right]. \tag{2}$$

Two popular choices of f^κ_{imp} are (1) expectile: $f^\kappa_{\text{expectile}}(c) = |\kappa - \mathbb{I}_{c<0}|c^2$, and (2) quantile: $f^\kappa_{\text{quantile}}(c) = |\kappa - \mathbb{I}_{c<0}||c|$, for any real value $\kappa \in [0.5, 1)$. At the optimum of $L(\xi)$, $V_\xi(s)$ approximates the κ-expectile/quantile of the distribution of the critic values evaluated at $Q(s, a)$, induced by the data distribution $\mathcal{D}$. With this implicit maximization technique, we no longer need to explicitly find the action a that maximizes $Q(s, a)$ and can use $V_\xi(s)$ as the backup target:

$$L(\phi) = \mathbb{E}_{s,a,s'\sim\mathcal{D}}\left[(Q_\phi(s, a) - r(s, a) - \gamma V_\xi(s'))^2\right]. \tag{3}$$

Multi-step return backup. TD learning can sometimes struggle with long-horizon tasks due to the well-known bootstrapping bias problem, where regressing the value network towards its own potentially inaccurate value estimates amplifies the value estimation errors further. To tackle this challenge, we can instead sample a trajectory segment, $(s_t, a_t, s_{t+1}, \cdots, a_{t+n-1}, s_{t+n})$, to construct an n-step return backup target from states h steps ahead:

$$L_{\text{ns}}(\phi) = \mathbb{E}_{s_t,a_t,\cdots,s_{t+n}}\left[\left(Q_\phi(s_t, a_t) - R_{t:t+n} - \gamma^n\bar{Q}(s_{t+n}, a^\star_{t+n})\right)^2\right], \tag{4}$$

where $a^\star_{t+n} = \arg\max_{a_{t+n}} Q(s_{t+n}, a_{t+n})$, $R_{t:t+n} := \sum_{t'=t}^{t+n-1} \gamma^{t'-t} r(s_{t'}, a_{t'})$. The n-step return value estimate of reduces the effective horizon by a factor of n, alleviating the bootstrapping bias problem. However, such value estimate is always biased towards the off-policy data distribution, and is also commonly referred to as the *uncorrected n-step return estimator* (Fedus et al., 2020; Kozuno et al., 2021). While there are ways to correct this value estimator via importance sampling (Precup et al., 2000; Munos et al., 2016; Rowland et al., 2020), they require additional tricks (*e.g.*, importance ratio truncation) for numerical stability and re-introduce biases into the estimator, ultimately resulting in a delicate trade-off between variances and biases that must be carefully balanced.

Action chunking critic. Alternatively, one may learn an action chunking critic to estimate the value of a short sequence of actions, $a_{t:t+h} := (a_t, a_{t+1}, \cdots, a_{t+h-1})$ (or an *action chunk*) instead: $Q(s_t, a_{t:t+h})$ (Seo & Abbeel, 2024; Li et al., 2025a; Tian et al., 2025; Li et al., 2025b). The TD backup loss for such a critic is naturally multi-step:

$$L_{\text{QC}}(\phi) = \mathbb{E}_{s_{t:t+h+1},a_{t:t+h}}\left[\left(Q_\phi(s_t, a_{t:t+h}) - R_{t:t+h} - \gamma^h\bar{Q}(s_{t+h}, a^\star_{t+h:t+2h})\right)^2\right], \tag{5}$$

where again $a^\star_{t+h:t+2h} = \arg\max_{a_{t+h:t+2h}} Q(s_{t+h}, a_{t+h:t+2h})$. On the one hand, unlike n-step return estimate for single-action critic that is pessimistic, the n-step return estimate (with $n = h$) for the action chunking critic is *unbiased* as long as the action chunk $a_{t:t+h}$ is *independent* of the intermediate states $s_{t+1:t+h+1}$, while enjoying the reduction in effective horizon (Li et al., 2025a;b). On the other hand, action chunking critic implicitly imposes a constraint on the policy that the actions are predicted and executed in chunks. As a result, the policy extracted from the action chunking critic needs to predict the entire action chunk all at once, posing a big learning challenge, especially for environments with complex transition dynamics.

In the following two sections, we offer theoretical insights that characterize the conditions when using action chunking critic is more preferable over n-step return backup with a single critic (Section 4), and develop a practical method that tackles the action chunking policy extraction challenge (Section 5).

4 When should we use action chunking for Q-learning?

In this section, we build a theoretical foundation for Q-learning with action chunking critic functions. We start by formalizing the setup of our analysis in Section 4.1, quantifying the value estimation bias incurred from backing up on non-action chunking data (Theorem 4.4) and the optimality of action chunking policy (Theorem 4.6) in Section 4.2. Using these result, we derive the condition under which we prefer action chunking Q-learning over the standard n-step return learning in Section 4.3. We also include some examples in which the condition holds in Appendix F.5 in the hope of facilitating theoretical analysis of action chunking policy learning in future work.

4.1 Assumptions

To build the foundation of our analysis, we start by describing the trajectory data distribution that we use for Q-learning and the trajectory distribution induced by an open-loop action chunking policy. In particular, we assume that the trajectory data distribution obeys the transition dynamics T:

Assumption 4.1 (Data Distribution Obeys Dynamics)**.** *$\mathcal{D} \in \Delta_{\mathcal{T}}$ is a trajectory distribution generated by rolling out a behavior policy from a distribution of $s_t \sim \mu$. The behavior policy can be non-Markovian (*i.e.*, $\pi_\beta(a_{t+k} \mid s_{t:t+k+1}, a_{t:t+k})$). Each subsequent state is generated obeying the dynamics of the MDP $\mathcal{M}$: $s_{t+k+1} \sim T(\cdot \mid s_{t+k}, a_{t+k}), \forall k \in \{0, 1, \cdots, h-1\}$. The resulting trajectory is $\{s_t, s_{t+1}, \cdots, s_{t+h}, a_t, a_{t+1}, \cdots, a_{t+h}\} \in \mathcal{T} = \mathcal{S}^h \times \mathcal{A}^h$.*

Next, we formally define the open-loop trajectory distribution that we would obtain if we take the same actions in the data and roll them out open-loop in the environment.

Definition 4.2 (Open-loop Trajectory)**.** *From any trajectory distribution $\mathcal{D}$, we can extract an open-loop policy with a horizon of h by marginalizing out all intermediate states. We use $\pi^\circ_{\mathcal{D}} : \mathcal{S} \to \Delta_{\mathcal{A}^h}$ to denote such policy which is formally defined as:*

$$\pi^\circ_{\mathcal{D}}(a_{t:t+h} \mid s_t) := P_{\mathcal{D}}(a_{t:t+h} \mid s_t). \tag{6}$$

By using this open-loop policy to roll-out trajectories in the MDP $\mathcal{M}$, it induces a trajectory distribution $P^\circ_{\mathcal{D}} \in \Delta_{\mathcal{S}^{h+1},\mathcal{A}^h}$ that is generally different from $\mathcal{D}$. We can decompose this open-loop policy step-by-step with the following factorization $\pi^\circ_{\mathcal{D}}(a_{t:t+k} \mid s_t) = \prod_{k=0}^{h-1} \pi^\circ_{\mathcal{D}}(a_{t+k} \mid s_t, a_{t:t+k})$ which allows us to define the induced trajectory distribution $P^\circ_{\mathcal{D}}$ recursively (for $k \in \{1, 2, \cdots, h\}$):

$$\begin{aligned} & P^\circ_{\mathcal{D}}(s_{t+k}, a_{t:t+k} \mid s_t) := \qquad\qquad (7) \\ & \quad P^\circ_{\mathcal{D}}(s_{t+k-1}, a_{t:t+k-1} \mid s_t) T(s_{t+k} \mid s_{t+k-1}, a_{t+k-1}) \pi^\circ_{\mathcal{D}}(a_{t+k} \mid s_t, a_{t:t+k}). \qquad (8) \end{aligned}$$

4.2 Open-loop value bias of action chunking Q-learning

As what we have elucidated in our definition above, replaying the actions from the trajectory data distribution $P_{\mathcal{D}}$ in an open-loop manner, in general, can result in a different trajectory distribution, $P^\circ_{\mathcal{D}}$. This discrepancy between $P^\circ_{\mathcal{D}}$ and $P_{\mathcal{D}}$ has not been carefully analyzed by prior work (*e.g.*, Q-chunking (Li et al., 2025b)) but can play a huge role in the optimal policy that action chunking Q-learning converges to. This is because TD-backup is only unbiased when it is done under the open-loop trajectory distribution $P^\circ_{\mathcal{D}}$. Naïvely running TD-backup on $P_{\mathcal{D}}$ (as done in Li et al. (2025b)) may lead to a *biased Q-target*. We now formalize the discrepancy and analyze such bias.

Definition 4.3 (Open-Loop Consistency)**.** *$\mathcal{D}$ is ε_h-open-loop consistent if for every $s_t \in \mathcal{S}, h' \in \{1, \cdots, h\}$, as long as $s_t \in \mathcal{S}$ has non-zero probability in the data (*i.e.*, $P_{\mathcal{D}}(s_t) > 0$),*

$$D_{\mathrm{TV}}(P^\circ_{\mathcal{D}}(s_{t+h'}, a_{t+h'} \mid s_t) \parallel P_{\mathcal{D}}(s_{t+h'}, a_{t+h'} \mid s_t)) \le \varepsilon_h, \forall h' \in \{1, 2, \cdots, h-1\}, \tag{9}$$

$$D_{\mathrm{TV}}(P^\circ_{\mathcal{D}}(s_{t+h} \mid s_t) \parallel P_{\mathcal{D}}(s_{t+h} \mid s_t)) \le \varepsilon_h. \tag{10}$$

*We say $\mathcal{D}$ is strongly ε_h-open-loop consistent if additionally for $h' \in \{1, 2, \cdots, h\}$, for every $a_{t:t+h} \in \mathcal{A}^h$ with non-zero probability in the data (*i.e.*, $P_{\mathcal{D}}(a_{t:t+h}, s_t) > 0$),*

$$D_{\mathrm{TV}}(T(s_{t+h'} \mid s_t, a_{t:t+h'}) \parallel P_{\mathcal{D}}(s_{t+h'} \mid s_t, a_{t:t+h})) \le \varepsilon_h. \tag{11}$$

Intuitively, $\mathcal{D}$ is ε-open-loop consistent if, when executing the same sequence of actions from it open-loop from s_t, the resulting marginal distribution of the state-action h steps into the future (*i.e.*, s_{t+h}) deviates from the corresponding distribution in the dataset by at most ε in total variation distance. The strong version (Equation (11)) requires the total variation distance bound to hold for every action sequence in the support, whereas the weak version (Equation (9)) only requires the bound to hold in expectation. Having *weak* open-loop consistency of $\mathcal{D}$ is sufficient to show that *behavior* value iteration of an action chunking critic results in a *nominal* value function with a bounded bias from the true value of the open-loop policy $\pi^{\circ}_{\mathcal{D}}$:

Theorem 4.4 (Bias of Action Chunking Critic)**.** *Let $\hat{V}_{\mathrm{ac}} : \mathcal{S} \to [0, 1/(1-\gamma)]$ be a solution of*

$$\hat{V}_{\mathrm{ac}}(s_t) = \mathbb{E}_{s_{t+1:t+h+1}, a_{t:t+h} \sim P_{\mathcal{D}}(\cdot|s_t)} \left[R_{t:t+h} + \gamma^h \hat{V}_{\mathrm{ac}}(s_{t+h}) \right], \tag{12}$$

with $R_{t:t+h} = \sum_{t'=t}^{t+h} \gamma^{t'-t} r(s_{t'}, a_{t'})$ and V_{ac} is the true value of $\pi^{\circ}_{\mathcal{D}} : s_t \mapsto P_{\mathcal{D}}(a_{t:t+h} \mid s_t)$. If $\mathcal{D}$ is ε_h-open-loop consistent, then under $\mathrm{supp}(\mathcal{D})$*,*

$$\left\| V_{\mathrm{ac}} - \hat{V}_{\mathrm{ac}} \right\|_{\infty} \leq \frac{\varepsilon_h \gamma}{(1 - (1-\varepsilon_h)\gamma^h)(1-\gamma)} \leq \frac{\varepsilon_h}{(1-\gamma^h)(1-\gamma)}. \tag{13}$$

The proof of Theorem 4.4 is available in Appendix G.2. We also show this bound is tight in Appendix F.1. A direct consequence of this result is that the true value of the optimal action chunking policy is close to that of the optimal closed-loop policy:

Corollary 4.5 (Optimal Action Chunking Policy)**.** *Let $\pi^\star : \mathcal{S} \to \Delta_{\mathcal{A}}$ be an optimal policy in $\mathcal{M}$ and $\mathcal{D}^\star$ be the data collected by $\pi^\star$. If $\mathcal{D}^\star$ is ε_h-open-loop consistent, then under* $\mathrm{supp}(\mathcal{D}^\star)$*,*

$$\| V^\star_{\mathrm{ac}} - V^\star \|_\infty \leq \left\| \tilde{V}_{\mathrm{ac}} - V^\star \right\|_\infty \leq \frac{\varepsilon_h \gamma}{(1 - (1-\varepsilon_h)\gamma^h)(1-\gamma)} \leq \frac{\varepsilon_h}{(1-\gamma^h)(1-\gamma)}, \tag{14}$$

where $V^\star$ is the value of the optimal policy $\pi^\star$, $V^\star_{\mathrm{ac}}$ is the true value of the optimal action chunking policy, and $\tilde{V}_{\mathrm{ac}}$ is the true value of the action chunking policy from cloning the data $\mathcal{D}^\star$:

$$\tilde{\pi}_{\mathrm{ac}}(a_{t:t+h} \mid s_t) : s_t \mapsto P_{\mathcal{D}^\star}(\cdot \mid s_t). \tag{15}$$

We again show that this bound is tight in Appendix F.2. The proof of Corollary 4.5 (available in Appendix G.4) builds on the observation that the nominal (biased) value of the action chunking critic obtained from behavior value iteration on an optimal data $\mathcal{D}^\star$ (*i.e.*, the data collected from an optimal policy $\pi^\star$) recovers the value of the optimal policy. This allows us to use Theorem 4.4 to show that the value of the action chunking policy obtained by behavior cloning on such optimal data is close to the nominal (biased) value of its critic, and thus close to the optimal value of the closed-loop policy.

Next, we analyze the performance of the action chunking policy obtained by Q-learning. In particular, we analyze the Q-function obtained as a solution of the following equation under $\mathrm{supp}(\mathcal{D})$:

$$\hat{Q}^+_{\mathrm{ac}}(s_t, a_{t:t+h}) = \mathbb{E}_{s_{t+1:t+h+1} \sim P_{\mathcal{D}}(\cdot|s_t, a_{t:t+h})} \left[R_{t:t+h} + \gamma^h \max_{a_{t+h:t+2h}} \hat{Q}^+_{\mathrm{ac}}(s_{t+h}, a_{t+h:t+2h}) \right]. \tag{16}$$

The corresponding action chunking policy is

$$\pi^+_{\mathrm{ac}} : s_t \mapsto \arg\max_{a_{t:t+h}} \hat{Q}^+_{\mathrm{ac}}(s_t, a_{t:t+h}). \tag{17}$$

It turns out that with the weak version of the open-loop consistent condition, the worst case performance of the action chunking policy may be arbitrarily low (see an example in Appendix H). Fortunately, as long as the data $\mathcal{D}$ satisfies the strongly open-loop consistency (Equation (11)), we can show that the learned policy π^+_{ac} is provably near-optimal by combining all the results above together:

Theorem 4.6 (Q-Learning with Action Chunking Policy on Off-policy Data)**.** *If $\mathcal{D}$ is strongly ε_h-open-loop consistent and* $\mathrm{supp}(\mathcal{D}) \supseteq \mathrm{supp}(\mathcal{D}^\star)$*, with $\mathcal{D}^\star$ being the data distribution of an arbitrary optimal policy $\pi^\star$ under $\mathcal{M}$), then the following bound holds under* $\mathrm{supp}(\mathcal{D}^\star)$*:*

$$\| V^+_{\mathrm{ac}} - V^\star \|_\infty \leq \frac{\varepsilon_h \gamma}{1-\gamma} \left[\frac{2}{1 - (1-2\varepsilon_h)\gamma^h} + \frac{1}{1 - (1-\varepsilon_h)\gamma^h} \right] \leq \frac{3\varepsilon_h}{(1-\gamma)(1-\gamma^h)}. \tag{18}$$

where $V^\star$ is the value of an optimal policy under $\mathcal{M}$.

This bound is also tight (as shown in Appendix F.3). The implication of Theorem 4.6 (proof available in Appendix G.6) is that as long as $\mathcal{D}$ satisfies the strongly open-loop consistency condition and contains the behavior in $\mathcal{D}^\star$, Q-learning with action chunking is guaranteed to converge to a near-optimal action chunking policy regardless of how sub-optimal the data $\mathcal{D}$ might be. As we will show in the following section, this is in contrast to n-step return policy where its performance depends on the sub-optimality of the data.

4.3 Comparing to n-step return Q-learning

We now characterize the condition when action chunking Q-learning should be preferred over the standard n-step return backup. We start by introudcing a notion of sub-optimality of the data $\mathcal{D}$:

Definition 4.7 (Sub-optimal data)**.** *$\mathcal{D}$ is δ_n-suboptimal for backup horizon length $n \in \mathbb{N}^+$ if*

$$Q^\star(s_t, a_t) - \mathbb{E}_{P_\mathcal{D}(\cdot|s_t,a_t)}\left[R_{t:t+n} + \gamma^n V^\star(s_{t+n})\right] \geq \delta_h, \forall s_t \in \mathcal{S}, a_t \in \mathcal{A}. \tag{19}$$

Intuitively, δ_n captures how much worse the n-step return policy can get compared to the optimal policy incurred by the backup bias. Under such condition, we can show that the action chunking policy is provably better than the n-step return policy as long as δ_n is large.

Theorem 4.8. *Let $\mathcal{D}$ be strongly ε_h-open-consistent, δ_n-suboptimal, and* $\mathrm{supp}(\mathcal{D}) \supseteq \mathrm{supp}(\mathcal{D}^\star)$. *Let $\pi_n^\star$ be the optimal n-step return policy learned from $\mathcal{D}$, as the solution of*

$$Q_n^\star(s_t, a_t) = \mathbb{E}_{P_\mathcal{D}}\left[R_{t:t+n} + \gamma^n Q_n^\star(s_{t+n}, \pi_n^\star(s_{t+n}))\right], \quad \pi_n^\star : s_t \mapsto \arg\max_{a_t} Q_n^\star(s_t, a_t). \tag{20}$$

As long as $\delta_n > \frac{3\varepsilon_h(1-\gamma^n)}{(1-\gamma)(1-\gamma^h)}$, then from all $s \in \mathrm{supp}(\mathcal{D}^\star)$, the action chunking policy, π_{ac}^+ (Equation (17)), is better than the n-step return policy, π_n (Equation (20)) (i.e., *$V_{\mathrm{ac}}^+(s) > V_n^\star(s)$).*

The proof of Theorem 4.8 is available in Appendix G.9. Notably, for $n = h$, the condition on δ_n and ε_h reduces to $\delta_n > 3\varepsilon_h H$ with effective horizon H (*i.e.*, $H = 1/(1-\gamma)$). As long as $\mathcal{D}$ is more than $O(\varepsilon_h H)$ sub-optimal, the action chunking policy performs provably better than n-step return policy.

4.4 Closed-loop execution of action chunking policy

Under the same strongly ε_h-open-loop consistency assumption, we can guarantee that closed-loop execution of the action chunking policy is also near-optimal. This is based on the intuition that in order for action chunking policy to be near-optimal, the first action in the chunk cannot be too sub-optimal:

Proposition 4.9 (Optimality of Closed-loop Execution of Action Chunking Policy)**.** *Let $V^\bullet$ be the value of the one-step policy, $\pi^\bullet$, defined as the closed-loop execution of the action chunking policy π_{ac}^+ learned from $\mathcal{D}$. That is, for each $s_t \in \mathrm{supp}(P_\mathcal{D}(s_t))$,*

$$\pi^\bullet(s_t) = a_t^+, \quad \text{where } a_{t:t+h}^+ = \pi_{\mathrm{ac}}^+(s_t). \tag{21}$$

If we assume $\mathcal{D}$ and $\mathcal{D}^\star$ are both strongly ε_h-open-loop consistent and $\mathrm{supp}(P_\mathcal{D}(s_t, a_{t:t+h})) \supseteq \mathrm{supp}(P_{\mathcal{D}^\star}(s_t, a_{t:t+h}))$, then under $\mathrm{supp}(\mathcal{D}^\star)$,

$$\|V^\star - V^\bullet\|_\infty \leq \frac{\varepsilon_h \gamma}{(1-\gamma)^2}\left[\frac{2}{1-(1-2\varepsilon_h)\gamma^h} + \frac{1}{1-(1-\varepsilon_h)\gamma^h}\right] \leq \frac{3\varepsilon_h}{(1-\gamma)^2(1-\gamma^h)}. \tag{22}$$

The proof is available in Appendix G.8. This result demonstrates that closed-loop execution is also near-optimal as long as the action chunking policy is near-optimal, though we might have to pay up to a horizon factor H (*i.e.*, $1/(1-\gamma)$) in sub-optimality gap in the worst case. Can we do better than this?

In practical applications, the data distributions that we are dealing with often have more structures. For example, it is common to have a dataset consisting of multiple sources where each data source is collected by either human expert or scripted policy that exhibits a somewhat predictable behavior (*e.g.*, after a robot arm picks up a cube, it will always move up rather than dropping it right away). We formalize this kind of structure as the notion of optimality variability:

Definition 4.10 (Optimality Variability)**.** *We say $\mathcal{D}$ exhibits ϑ_h-variability in optimality conditioned on an event X if*

$$\max_{\mathrm{supp}(P_\mathcal{D}(\cdot|X))}\left[R_{t:t+h} + \gamma^h V^\star(s_{t+h})\right] - \min_{\mathrm{supp}(P_\mathcal{D}(\cdot|X))}\left[R_{t:t+h} + \gamma^h V^\star(s_{t+h})\right] \leq \vartheta_h. \tag{23}$$

See more discussion of this the definition in Appendix J. We can now formalize our results as follows:

Theorem 4.11 (Closed-loop AC Policy under Bounded OV)**.** *Let $\mathcal{D}^\star$ be the data distribution collected by an optimal policy. Assume $\mathcal{D}$ can be decomposed into a mixture of data distributions $\{\mathcal{D}^\star, \mathcal{D}_1, \mathcal{D}_2, \cdots \mathcal{D}_N\}$ such that each data distribution component satisfies Assumption 4.1 and for some $\vartheta_h^L, \vartheta_h^G \geq 0$, they satisfy the following two conditions:*

***1. Locally bounded optimality variability condition**: every $\mathcal{D}_i$ (including $\mathcal{D}^\star$) exhibits ϑ_h^L-bounded variability in optimality conditioned on s_t, a_t for all $(s_t, a_t) \in \mathrm{supp}(P_{\mathcal{D}_i}(s_t, a_t))$, and*

2. Globally bounded optimality variability condition*: $\mathcal{D}$ as a whole exhibits ϑ^G_h-variability in optimality conditioned on $s_t, a_{t:t+h}$ for all $(s_t, a_{t:t+h}) \in \mathrm{supp}(P_\mathcal{D}(s_t, a_{t:t+h}))$.*

Then for all $s_t \in \mathrm{supp}(P_{\mathcal{D}^\star}(s_t))$,

$$V^\star(s_t) - V^\bullet(s_t) \leq \frac{\vartheta^L_h}{1-\gamma} + \frac{\vartheta^G_h + \gamma^h \min(\vartheta^L_h, \vartheta^G_h)}{(1-\gamma)(1-\gamma^h)} \leq \vartheta^L_h H + 2\vartheta^G_h H \bar{H} \tag{24}$$

This bound is also tight up to the exact value (as shown in Appendix F.4). It is worth noting that although the global optimality variability condition looks similar to the strong open-loop consistency condition, they have completely different properties. For instance, a nearly strong open-loop consistent data distribution $\mathcal{D}$ can have unbounded global optimality variability and a data distribution that exhibits zero optimality variability can also have large open-loop inconsistency. The implication of this is that even when the closed-loop execution of an action chunking policy is near-optimal, the same action chunking policy executed in chunks can be very sub-optimal (formalized in Appendix F.4). Furthermore, executing the first action of the original action chunk also brings practical benefits: it removes the need to explicitly train a policy to predict the full action chunk all at once, which is hard when the chunk size grows big. Can we develop a practical method that realizes such potential?

5 DECOUPLED Q-CHUNKING

We propose a new algorithm that enjoys the benefits of value backup speedup of Q-chunking while avoiding the difficulty of learning an open-loop action chunking policy with a large chunk size.

Our core idea is to decouple the chunk size of the critic from that of the policy. In particular, we train a policy $\pi(a_{t:t+h_a} \mid s_t)$ to output an action chunk (with a size of $h_a \ll h$) with the following objective:

$$L(\pi) := -\mathbb{E}_{a_{t:t+h_a} \sim \pi(\cdot|s_t)}[Q_\phi(s, [a_{t:t+h_a}, a^\star_{t+h_a:t+h}])], \tag{25}$$

where $[a_{t:t+h_a}, a^\star_{t+h_a:t+h}]$ represents the concatenation of two partial action chunks (size h_a and size $h - h_a$) into a full action chunk $a_{t:t+h}$ of size h, and $a^\star_{t+h_a:t+h}$ is the best 'second-half' of the action chunk that maximizes the critic value under Q_ϕ:

$$a^\star_{t+h_a:t+h} := \arg\max_{a_{t+h_a:t+h}} Q_\phi(s, [a_{t:t+h_a}, a_{t+h_a:t+h}]). \tag{26}$$

Essentially, we want our policy to predict the partial action chunk (of size h_a) within an optimal action chunk of size h, rather than the entire optimal action chunk. This lowers the policy expressivity requirement and hence the learning challenges associated with it with $h_a < h$.

However, directly optimizing this objective (Equation (25)) does not lead to a novel algorithm because taking the maximization over $a_{t+h_a:t+h}$ seemingly requires us to learn a policy of the original chunk size anyways. To address this issue, we learn a separate partial critic Q^P_ψ, which only takes in the partial action chunk (of size h_a) as input, to approximate the maximum value this partial action chunk can achieve when it is extended to the full action chunk (of size h):

$$Q^P_\psi(s, a_{t:t+h_a}) \approx Q_\phi(s, [a_{t:t+h_a}, a^\star_{t+h_a:t+h}]) \tag{27}$$

To train Q^P_ψ, we can use an *implicit maximization* loss function (as described in Equation (2)):

$$L(\psi) := f^{\kappa_d}_{\mathrm{imp}}(\bar{Q}_\phi(s_t, a_{t:t+h}) - Q^P_\psi(s_t, a_{t:t+h_a})), \tag{28}$$

where $s_t, a_{t:t+h}$ are sampled from $\mathcal{D}$. As a result, the partial critic, Q^P_ψ, is distilled from the original critic via an optimistic regression, where its optimum $Q^\star_\psi(s, a_{t:t+h_a})$ approximates $Q_\phi(s, [a_{t:t+h_a}, a^\star_{t+h_a:t+h}])$ in Equation (25), conveniently removing the need for training a policy to predict the whole optimal action chunk entirely. This allows us to simplify the policy objective as

$$L(\pi) := -\mathbb{E}_{a_{t:t+h_a} \sim \pi(\cdot|s_t)} \left[Q^P_\psi(s, a_{t:t+h_a})\right]. \tag{29}$$

In summary, DQC trains a policy to predict a partial chunk, $a_{t:t+h_a}$ (of size h_a), by hill climbing the value of a partial critic $Q^P_\psi(s, a_{t:t+h_a})$ that is distilled from the original chunked critic $Q_\phi(s, a_{t:t+h})$ via an implicit maximization loss. This allows our policy to fully leverage the chunked critic Q_ϕ (and thus the value speedup benefits associated with Q-chunking) without the need to predict the full action chunk (of size h), mitigating the learning challenge of an action chunking policy.

Algorithm 1 Decoupled Q-chunking (DQC).

```
Given: $D, Q_\phi(s_t, a_{t:t+h}), Q^P_\psi(s_t, a_{t:t+h_a}), V_\xi(s_t), \pi_\beta(a_{t:t+h_a} \mid s_t)$
1. Agent Update:
$(s_{t:t+h+1}, a_{t:t+h}, r_{t:t+h}) \sim D.$    ▷ sample trajectory chunk from the offline dataset
Optimize $Q_\phi$ with $L(\phi) = \left(Q_\phi(s_t, a_{t:t+h}) - \sum_{k=0}^{h-1} \gamma^k r_{t+k} - \gamma^h \bar{V}_\xi(s_{t+h})\right)^2.$
Optimize $Q^P_\psi$ with $L(\psi) = f^{\kappa_{\text{d}}}_{\text{expectile}}\left(\bar{Q}_\phi(s_t, a_{t:t+h}) - Q^P_\psi(s_t, a_{t:t+h_a})\right).$
Optimize $V_\xi$ with $L(\xi) = f^{\kappa_{\text{b}}}_{\text{quantile}}(\bar{Q}^P_\psi(s_t, a^\beta_{t:t+h_a}) - V_\xi(s_t)), a^\beta_{t:t+h_a} \sim \pi_\beta(\cdot \mid s_t)$
2. Policy Extration:
$a^1_{t:t+h_a}, a^2_{t:t+h_a}, \cdots, a^N_{t:t+h_a} \sim \pi_\beta(\cdot \mid s_t)$    ▷ sample $N$ actions from behavior policy
$a^\star_{t:t+h_a} \leftarrow \arg\max_{\{a^i_{t:t+h_a}\}_{i=1}^N} Q^P_\psi(s_t, a_{t:t+h_a})$    ▷ take the action with the highest Q-value
```

Practical considerations for offline RL. Finally, we describe several implementation details that we find to work well in the offline RL setting, which our experiments primarily focus on. Our implementation draws inspirations from a prior method, IDQL (Hansen-Estruch et al., 2023).

We first train a behavior cloning flow policy π_β using a standard flow-matching objective (Liu et al., 2022) on the offline dataset D. Then, we approximate the policy optimization objective in DQC (Equation (29)) using best-of-N sampling without explicitly modeling π:

$$a^\star_{t:t+h_a} \leftarrow \arg\max_{\{a^i_{t:t+h_a}\}_{i=1}^N} Q^P_\psi(s_t, a_{t:t+h_a}), \quad \text{where } a^1_{t:t+h_a}, \cdots, a^N_{t:t+h_a} \sim \pi_\beta(\cdot \mid s_t). \tag{30}$$

where $a^\star_{t:t+h_a}$ is output of the policy that we extract from Q^P_ψ for state s_t. Essentially, this sampling procedure is a test-time approximation of the objective in Equation (29), where it outputs action (chunk) that maximizes Q^P_ψ, subject to the behavior prior, as modeled by π_β.

For TD learning of Q_ϕ, directly computing the TD backup target from either $Q_{\bar\phi}$ or $Q^P_{\bar\psi}$ is computationally expensive, as either requires samples from the current policy, which is approximated via the best-of-N sampling procedure as described above. Instead, we use the implicit value backup (Kostrikov et al., 2021) (*i.e.*, as described in Equation (2)) to approximate the target:

$$L(\xi) = f^{\kappa_b}_{\text{quantile}}(\bar{Q}^P_\psi(s_t, a^\beta_{t:t+h_a} - V_\xi(s_t)), \quad a^\beta_{t:t+h_a} \sim \pi_\beta(\cdot \mid s_t) \tag{31}$$

where we pick the quantile regression loss as the implicit maximization loss function. This is because the Q-value obtained from best-of-N sampling can be seen as the largest order statistic of a random batch (of size N) of the behavior Q-values (*i.e.*, $\{Q(s, a^i)\}_{i=1}^N, a^i \sim \pi_\beta(\cdot \mid s)$). Such statistic estimates the behavior Q-value distribution's $\frac{N}{1-N}$-quantile, which is the same as $V_\xi(s)$ at the optimum of $L(\xi)$ if we set $\kappa_b = \frac{N}{1-N}$. In practice, we use a larger κ_b for numerical stability.

Finally, we pick the expectile regression loss for training the distilled partial critic Q^P_ψ because prior work has found it to work the best among all implicit maximization loss functions (Hansen-Estruch et al., 2023). A summary of the algorithm is available in Algorithm 1.

6 Experimental Setup

We conduct experiments to evaluate the benefits of decoupling the policy chunk size and the critic chunk size on OGBench (Park et al., 2024a)—a challenging long-horizon, goal-conditioned offline RL benchmark consisting of a diverse set of environments (from manipulation to locomotion). In particular, we use the more difficult environments introduced by Park et al. (2025) (Figure 6), where multi-step return backups are crucial. These environments require highly complex, long-horizon reasoning, and serve as an ideal testbed for our algorithm, which improves upon n-step returns and Q-chunking. We now describe our main comparisons, starting with direct ablation baselines:

DQC-naïve is a naïve attempt at decoupling the critic chunk size from the policy chunk size, where it takes the QC policy to predict full action chunks of size h but only execute the first h_a actions.

QC (Li et al., 2025b) uses a single critic that has the same chunk length as that of the policy (*i.e.*, $h = h_a$). This baseline tests whether having *decoupled* chunk sizes is important.

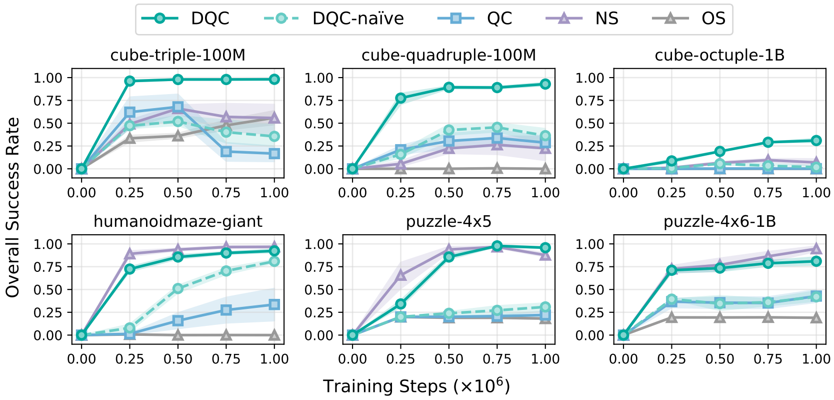


Figure 1: **Offline goal-conditioned RL results.** Our method (*DQC*) uses a *decoupled* critic and policy chunk sizes, which allows it to outperform our baselines by a large margin on `cube-*` and competitive on others. *QC*: Q-chunking (Li et al., 2025b); *DQC-naïve*: QC but only executing a partial action chunk open-loop; *NS*: n-step return backup; *OS*: 1-step TD-backup.

NS: n-step return TD backup. This baseline uses a single one-step critic (*i.e.*, $Q(s_t, a_t)$). Compared to DQC with $h = n$ and $h_a = 1$, this baseline tests whether using a chunked critic is important.

OS: Standard 1-step TD backup. This is the same as NS but with $n = 1$.

Beyond the ablation baselines, we also consider the following strong goal-conditioned baselines:

FBC/HFBC: Goal-conditioned and hierarchical goal-conditioned flow behavior cloning baselines considered in Park et al. (2025).

IQL/HIQL (Kostrikov et al., 2021; Park et al., 2023): These are strong goal-conditioned RL methods that train a goal-conditioned value function with implicit value backups and extract a flat (IQL) or hierarchical (HIQL) policy from the value function.

SHARSA (Park et al., 2025): The previous state-of-the-art method on the long-horizon environments that we evaluate on. The method uses a combination of n-step return and bi-level hierarchical policies.

In our ablation study, we also consider an additional baseline, **QC-NS**, that uses the idea of decoupled policy chunking and critic chunking ($h_a < h$), but without using a distilled critic. This baseline simply uses n-step return targets to directly train a critic with a chunk size of h_a without implicit maximization (Equation (28)). The performance of this baseline helps determine how important it is to learn a separate distilled critic for partial action chunks with implicit maximization. For all our main results, we run 3 seeds and report the means and the 95% confidence intervals.

Task	FBC	HFBC	IQL	HIQL	SHARSA	OS	NS	QC	DQC-naïve	DQC
`cube-triple-100M`	$53_{[48,57]}$	$57_{[54,61]}$	$64_{[59,68]}$	$36_{[27,45]}$	$82_{[78,88]}$	$56_{[48,64]}$	$56_{[37,71]}$	$17_{[8,25]}$	$36_{[24,49]}$	$\mathbf{98}_{[\mathbf{97,99}]}$
`cube-quadruple-100M`	$32_{[30,33]}$	$38_{[34,41]}$	$53_{[53,53]}$	$24_{[18,30]}$	$67_{[62,74]}$	$0_{[0,0]}$	$22_{[9,36]}$	$29_{[22,36]}$	$36_{[28,44]}$	$\mathbf{93}_{[\mathbf{91,95}]}$
`cube-octuple-1B`	$0_{[0,0]}$	$28_{[27,28]}$	$0_{[0,0]}$	$18_{[14,21]}$	$\mathbf{33}_{[\mathbf{30,35}]}$	$0_{[0,0]}$	$7_{[3,11]}$	$0_{[0,0]}$	$2_{[0,4]}$	$\mathbf{31}_{[\mathbf{29,33}]}$
`humanoidmaze-giant`	$1_{[0,3]}$	$4_{[2,5]}$	$4_{[2,6]}$	$24_{[20,28]}$	$18_{[13,25]}$	$0_{[0,0]}$	$\mathbf{97}_{[\mathbf{95,98}]}$	$34_{[16,51]}$	$81_{[79,83]}$	$92_{[90,94]}$
`puzzle-4x5`	$0_{[0,0]}$	$0_{[0,0]}$	$20_{[20,20]}$	$0_{[0,0]}$	$1_{[0,2]}$	$18_{[17,19]}$	$88_{[86,90]}$	$22_{[20,26]}$	$31_{[26,35]}$	$\mathbf{96}_{[\mathbf{95,97}]}$
`puzzle-4x6-1B`	$0_{[0,0]}$	$5_{[3,5]}$	$7_{[2,13]}$	$10_{[3,17]}$	$62_{[57,71]}$	$19_{[19,20]}$	$\mathbf{95}_{[\mathbf{92,98}]}$	$43_{[36,50]}$	$42_{[37,48]}$	$81_{[77,86]}$

Table 1: **Comparisons with prior methods.** Our method outperforms SHARSA (Park et al., 2025) (the previous state-of-the-art method on this benchmark) on most tasks except `cube-octuple` where our performance is on par with SHARSA. In contrast, our n-step return baseline (NS), Q-chunking baseline (QC), and naïvely executing partial action chunks from QC (naïve DQC) all fail to outperform SHARSA on `cube-*`.

7 Results

In this section, we present our experimental results to answer the following three questions:

(Q1) Does DQC improve upon n-step return, Q-chunking? Figure 1 compares DQC (ours) to both n-step and QC across six challenging long-horizon GCRL tasks, with our method performing on par or better across the board. Table 1 shows DQC also consistently outperforms the previous state-of-the-art method on this benchmark, SHARSA (Park et al., 2025), on all environments. For each environment, we tune DQC (ours), QC, NS, OS (see the tuning range in Table 8) and pick the

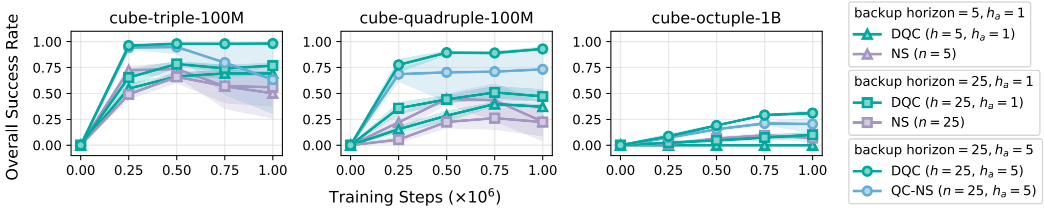


Figure 2: **Distilled critic ablations.** Each group in the legend contains DQC and its non-distilled counterpart with the same configuration (*i.e.*, same backup horizon and same policy chunk size). Our method (DQC) performs on par or better than the non-distilled counterpart across all configurations.

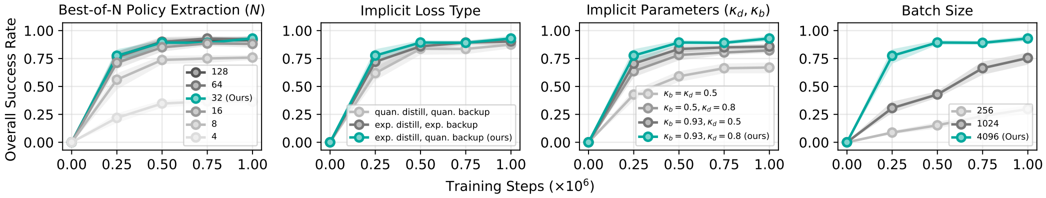


Figure 3: **Hyperparameter sensitivity analysis on `cube-quadruple-100M`.** *Best-of-N*: the number of action samples drawn from $\pi_\beta(\cdot \mid s)$ during policy evaluation; *Implicit loss type*: the implicit maximization loss function used for distillation and value backup; *Batch size*: the number of examples used in each gradient step.

best configuration (Table 7) for hyperparameters used in Figure 1 and Table 1. For all baselines from prior work (SHARSA, HIQL, IQL, HFBC, FBC), we directly use their tuned hyperparameters and run with the same batch size (*i.e.*, 4096) as used in our method and other baselines. See the complete table for all combinations of h, n, h_a in Appendix A.

(Q2) Is training a separate distilled critic Q_ψ^P necessary? In Figure 2, we compare DQC to DQC without using the distilled critic across three different (h, h_a) configurations: $(h = 25, h_a = 5)$, $(h = 25, h_a = 1)$, and $(h = 5, h_a = 1)$. For configurations with $h_a = 1$, the baseline without using the distilled critic is the same as the n-step return baseline (with $n = h$) and for the configuration with $h_a = 5$, it is the same as combining Q-chunking and n-step return. Across three configurations, DQC performs on par or better than its non-distilled counterpart. This highlights that the use of a separate distilled critic for the partial action chunk is necessary for the effectiveness of DQC.

(Q3) How sensitive is DQC to its hyperparameters? Figure 3 shows that our method is not sensitive to the implicit backup method (quantile or expectile), and somewhat sensitive to the implicit parameters κ_b, κ_d. In particular, DQC is still reasonably effective as long as some form of optimism is employed (*i.e.*, either $\kappa_b \neq 0.5$ or $\kappa_d \neq 0.5$). Using no optimism ($\kappa_b = \kappa_d = 0.5$) results in a big performance drop. The other important hyperparameters are N in best-of-N policy extraction and the batch size. Having large enough batch size (*i.e.*, 4096) and N (*e.g.*, $N = 32$) is crucial for good performance, though a larger N ($N = 128$) does not lead to better performance.

8 Discussion

We provide a theoretical foundation for action chunking Q-learning and demonstrate how to effectively extract policies from chunked critics. Theoretically, we provide a formal analysis of action chunking Q-learning, identifying the TD backup bias that arises from *open-loop inconsistency* and characterizing the conditions under which action chunking Q-learning is preferred over n-step return learning. Empirically, we develop a novel technique that enables effective policy extraction from chunked critics with long action chunks, *scaling up action chunking Q-learning* to much harder environments. Together, these contributions advance the goal of tackling bootstrapping bias in TD-learning. Several challenges remain, indicating promising avenues for future research. Our method still inherits the open-loop value bias identified in Theorem 4.4, and developing techniques to actively correct for this bias could further improve performance. Moreover, our method relies on a fixed policy action chunk size h_a and critic action chunk size h across all states, even though the optimal action chunk size may vary by state. Developing practical methods that can support flexible, state-dependent chunk sizes would be a natural next step.

Reproducibility Statement

To facilitate future research, we include our source code as part of the supplementary materials, along with example scripts for both our method and our baselines. We describe our environments in Appendix D and hyperparameters in Appendix E. For our theoretical results, we fully state our assumption in Assumption 4.1 and provide complete proofs in Appendix G.

References

Anurag Ajay, Aviral Kumar, Pulkit Agrawal, Sergey Levine, and Ofir Nachum. OPAL: Offline primitive discovery for accelerating offline reinforcement learning. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=V69LGwJ0lIN`.

Pierre-Luc Bacon, Jean Harb, and Doina Precup. The option-critic architecture. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Akhil Bagaria and George Konidaris. Option discovery using deep skill chaining. In *International Conference on Learning Representations*, 2019.

Akhil Bagaria, Ben Abbatematteo, Omer Gottesman, Matt Corsaro, Sreehari Rammohan, and George Konidaris. Effectively learning initiation sets in hierarchical reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Philip J Ball, Laura Smith, Ilya Kostrikov, and Sergey Levine. Efficient online reinforcement learning with offline data. In *International Conference on Machine Learning*, pp. 1577–1594. PMLR, 2023.

Boyuan Chen, Chuning Zhu, Pulkit Agrawal, Kaiqing Zhang, and Abhishek Gupta. Self-supervised reinforcement learning that transfers using random features. *Advances in Neural Information Processing Systems*, 36, 2024.

Xinyue Chen, Che Wang, Zijian Zhou, and Keith Ross. Randomized ensembled double q-learning: Learning fast without a model. *arXiv preprint arXiv:2101.05982*, 2021.

Nuttapong Chentanez, Andrew Barto, and Satinder Singh. Intrinsically motivated reinforcement learning. *Advances in neural information processing systems*, 17, 2004.

Imre Csiszár. On information-type measure of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

Christian Daniel, Gerhard Neumann, Oliver Kroemer, and Jan Peters. Hierarchical relative entropy policy search. *Journal of Machine Learning Research*, 17(93):1–50, 2016.

Peter Dayan and Geoffrey E Hinton. Feudal reinforcement learning. *Advances in neural information processing systems*, 5, 1992.

Kristopher De Asis, J Hernandez-Garcia, G Holland, and Richard Sutton. Multi-step reinforcement learning: A unifying algorithm. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Anita de Mello Koch, Akhil Bagaria, Bingnan Huo, Zhiyuan Zhou, Cameron Allen, and George Konidaris. Learning transferable sub-goals by hypothesizing generalizing features. 2025.

Thomas G Dietterich. Hierarchical reinforcement learning with the maxq value function decomposition. *Journal of artificial intelligence research*, 13:227–303, 2000.

Pierluca D'Oro, Max Schwarzer, Evgenii Nikishin, Pierre-Luc Bacon, Marc G Bellemare, and Aaron Courville. Sample-efficient reinforcement learning by breaking the replay ratio barrier. In *Deep Reinforcement Learning Workshop NeurIPS 2022*, 2022.

Ishan P Durugkar, Clemens Rosenbaum, Stefan Dernbach, and Sridhar Mahadevan. Deep reinforcement learning with macro-actions. *arXiv preprint arXiv:1606.04615*, 2016.

William Fedus, Prajit Ramachandran, Rishabh Agarwal, Yoshua Bengio, Hugo Larochelle, Mark Rowland, and Will Dabney. Revisiting fundamentals of experience replay. In *International conference on machine learning*, pp. 3061–3071. PMLR, 2020.

Roy Fox, Sanjay Krishnan, Ion Stoica, and Ken Goldberg. Multi-level discovery of deep options. *arXiv preprint arXiv:1703.08294*, 2017.

Kevin Frans, Seohong Park, Pieter Abbeel, and Sergey Levine. Unsupervised zero-shot reinforcement learning via functional reward encodings. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 13927–13942. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/frans24a.html.

Jonas Gehring, Gabriel Synnaeve, Andreas Krause, and Nicolas Usunier. Hierarchical skills for efficient exploration. *Advances in Neural Information Processing Systems*, 34:11553–11564, 2021.

Philippe Hansen-Estruch, Ilya Kostrikov, Michael Janner, Jakub Grudzien Kuba, and Sergey Levine. Idql: Implicit q-learning as an actor-critic method with diffusion policies. *arXiv preprint arXiv:2304.10573*, 2023.

Hao Hu, Yiqin Yang, Jianing Ye, Ziqing Mai, and Chongjie Zhang. Unsupervised behavior extraction via random intent priors. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=4vGVQVz5KG.

Edward L Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008.

Tommi Jaakkola, Michael Jordan, and Satinder Singh. Convergence of stochastic iterative dynamic programming algorithms. *Advances in neural information processing systems*, 6, 1993.

Jihwan Jeong, Xiaoyu Wang, Michael Gimelfarb, Hyunwoo Kim, Baher Abdulhai, and Scott Sanner. Conservative bayesian model-based value expansion for offline policy optimization. *arXiv preprint arXiv:2210.03802*, 2022.

Teun Kloek and Herman K Van Dijk. Bayesian estimates of equation system parameters: an application of integration by monte carlo. *Econometrica: Journal of the Econometric Society*, pp. 1–19, 1978.

George Konidaris, Scott Niekum, and Philip S Thomas. TD$_\gamma$: Re-evaluating complex backups in temporal difference learning. *Advances in Neural Information Processing Systems*, 24, 2011.

George Dimitri Konidaris. *Autonomous robot skill acquisition*. University of Massachusetts Amherst, 2011.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit Q-learning. *arXiv preprint arXiv:2110.06169*, 2021.

Tadashi Kozuno, Yunhao Tang, Mark Rowland, Rémi Munos, Steven Kapturowski, Will Dabney, Michal Valko, and David Abel. Revisiting Peng's Q (λ) for modern reinforcement learning. In *International Conference on Machine Learning*, pp. 5794–5804. PMLR, 2021.

Tejas D Kulkarni, Karthik Narasimhan, Ardavan Saeedi, and Josh Tenenbaum. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. *Advances in neural information processing systems*, 29, 2016.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020.

Seunghyun Lee, Younggyo Seo, Kimin Lee, Pieter Abbeel, and Jinwoo Shin. Offline-to-online reinforcement learning via balanced replay and pessimistic Q-ensemble. In *Conference on Robot Learning*, pp. 1702–1712. PMLR, 2022.

Ge Li, Dong Tian, Hongyi Zhou, Xinkai Jiang, Rudolf Lioutikov, and Gerhard Neumann. TOP-ERL: Transformer-based off-policy episodic reinforcement learning. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=N4NhVN30ph.

Qiyang Li, Zhiyuan Zhou, and Sergey Levine. Reinforcement learning with action chunking. *arXiv preprint arXiv:2507.07969*, 2025b.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Amy McGovern and Richard S Sutton. Macro-actions in reinforcement learning: An empirical analysis. 1998.

Ishai Menache, Shie Mannor, and Nahum Shimkin. Q-cut—dynamic discovery of sub-goals in reinforcement learning. In *Machine Learning: ECML 2002: 13th European Conference on Machine Learning Helsinki, Finland, August 19–23, 2002 Proceedings 13*, pp. 295–306. Springer, 2002.

Josh Merel, Leonard Hasenclever, Alexandre Galashov, Arun Ahuja, Vu Pham, Greg Wayne, Yee Whye Teh, and Nicolas Heess. Neural probabilistic motor primitives for humanoid control. *arXiv preprint arXiv:1811.11711*, 2018.

Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc Bellemare. Safe and efficient off-policy reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

Ofir Nachum, Shixiang Shane Gu, Honglak Lee, and Sergey Levine. Data-efficient hierarchical reinforcement learning. *Advances in neural information processing systems*, 31, 2018.

Mitsuhiko Nakamoto, Simon Zhai, Anikait Singh, Max Sobol Mark, Yi Ma, Chelsea Finn, Aviral Kumar, and Sergey Levine. Cal-QL: Calibrated offline RL pre-training for efficient online fine-tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

Soroush Nasiriany, Tian Gao, Ajay Mandlekar, and Yuke Zhu. Learning and retrieval from prior data for skill-based imitation learning. In *Conference on Robot Learning*, 2022.

Alexandros Paraschos, Christian Daniel, Jan R Peters, and Gerhard Neumann. Probabilistic movement primitives. *Advances in neural information processing systems*, 26, 2013.

Kwanyoung Park and Youngwoon Lee. Model-based offline reinforcement learning with lower expectile q-learning. *arXiv preprint arXiv:2407.00699*, 2024.

Seohong Park, Dibya Ghosh, Benjamin Eysenbach, and Sergey Levine. HIQL: Offline goal-conditioned RL with latent states as actions. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=cLQCCtVDuW.

Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. Ogbench: Benchmarking offline goal-conditioned rl. *ArXiv*, 2024a.

Seohong Park, Tobias Kreiman, and Sergey Levine. Foundation policies with hilbert representations. In *Forty-first International Conference on Machine Learning*, 2024b. URL https://openreview.net/forum?id=LhNsSaAKub.

Seohong Park, Kevin Frans, Deepinder Mann, Benjamin Eysenbach, Aviral Kumar, and Sergey Levine. Horizon reduction makes RL scalable. *arXiv preprint arXiv:2506.04168*, 2025.

Jing Peng and Ronald J Williams. Incremental multi-step Q-learning. In *Machine Learning Proceedings 1994*, pp. 226–232. Elsevier, 1994.

Xue Bin Peng, Glen Berseth, KangKang Yin, and Michiel Van De Panne. Deeploco: Dynamic locomotion skills using hierarchical deep reinforcement learning. *Acm transactions on graphics (tog)*, 36(4):1–13, 2017.

Karl Pertsch, Youngwoon Lee, and Joseph Lim. Accelerating reinforcement learning with learned skill priors. In *Conference on robot learning*, pp. 188–204. PMLR, 2021.

Doina Precup, Richard S Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. In *ICML*, volume 2000, pp. 759–766. Citeseer, 2000.

Martin Riedmiller, Roland Hafner, Thomas Lampe, Michael Neunert, Jonas Degrave, Tom Wiele, Vlad Mnih, Nicolas Heess, and Jost Tobias Springenberg. Learning by playing solving sparse reward tasks from scratch. In *International conference on machine learning*, pp. 4344–4353. PMLR, 2018.

Mark Rowland, Will Dabney, and Rémi Munos. Adaptive trade-offs in off-policy learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 34–44. PMLR, 2020.

Younggyo Seo and Pieter Abbeel. Reinforcement learning with action sequence for data-efficient robot learning. 2024.

Younggyo Seo, Jafar Uruç, and Stephen James. Continuous control with coarse-to-fine reinforcement learning. In *8th Annual Conference on Robot Learning*, 2024. URL https://openreview.net/forum?id=WjDR48cL3O.

Tanmay Shankar and Abhinav Gupta. Learning robot skills with temporal variational inference. In *International Conference on Machine Learning*, pp. 8624–8633. PMLR, 2020.

Özgür Şimşek and Andrew G. Barto. Betweenness centrality as a basis for forming skills. Workingpaper, University of Massachusetts Amherst, April 2007.

Aravind Srinivas, Ramnandan Krishnamurthy, Peeyush Kumar, and Balaraman Ravindran. Option discovery in hierarchical reinforcement learning using spatio-temporal clustering. *arXiv preprint arXiv:1605.05359*, 2016.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Richard S Sutton, Doina Precup, and Satinder Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

Denis Tarasov, Vladislav Kurenkov, Alexander Nikulin, and Sergey Kolesnikov. Revisiting the minimalist approach to offline reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Philip S Thomas, Scott Niekum, Georgios Theocharous, and George Konidaris. Policy evaluation using the Ω-return. *Advances in Neural Information Processing Systems*, 28, 2015.

Dong Tian, Ge Li, Hongyi Zhou, Onur Celik, and Gerhard Neumann. Chunking the critic: A transformer-based soft actor-critic with N-step returns. *arXiv preprint arXiv:2503.03660*, 2025.

Ahmed Touati, Jérémy Rapin, and Yann Ollivier. Does zero-shot reinforcement learning exist? In *The Eleventh International Conference on Learning Representations*, 2022.

Alexander Vezhnevets, Volodymyr Mnih, Simon Osindero, Alex Graves, Oriol Vinyals, John Agapiou, et al. Strategic attentive writer for learning macro-actions. *Advances in neural information processing systems*, 29, 2016.

Alexander Sasha Vezhnevets, Simon Osindero, Tom Schaul, Nicolas Heess, Max Jaderberg, David Silver, and Koray Kavukcuoglu. Feudal networks for hierarchical reinforcement learning. In *International conference on machine learning*, pp. 3540–3549. PMLR, 2017.

Max Wilcoxson, Qiyang Li, Kevin Frans, and Sergey Levine. Leveraging skills from unlabeled prior data for efficient online exploration. *arXiv preprint arXiv:2410.18076*, 2024.

Yihong Wu. Lecture notes on information-theoretic methods for high-dimensional statistics. *Lecture Notes for ECE598YW (UIUC)*, 16:15, 2017.

Kevin Xie, Homanga Bharadhwaj, Danijar Hafner, Animesh Garg, and Florian Shkurti. Latent skill planning for exploration and transfer. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jXe91kq3jAq.

Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.

A Full results

Table 2 reports the performance of our method (DQC) and baselines for all hyperparameter configurations. All of them use the same hyperparameters in Table 5 with the only exception that SHARSA handles goal-sampling for training behavior cloning policies slightly differently as we discuss in more details in Appendix E. We also include the full batch size sensitivity analysis in Figure 4.

Task	**OS**	**DQC**	**DQC-naïve**	**QC**	**NS**	**DQC**	**DQC-naïve**	**NS**	**DQC**	**DQC-naïve**	**QC-NS**	**QC**	**SHARSA**	**HIQL**	**IQL**	**FBC**	**HFBC**
		$(h=5, h_a=1)$	$(h=5, h_a=1)$	$(h_a=5)$	$(n=5)$	$(h=25, h_a=1)$	$(h=25, h_a=1)$	$(n=25)$	$(h=25, h_a=5)$	$(h=25, h_a=5)$	$(n=25, h_a=5)$	$(h_a=25)$					
`cube-triple-100M`	$56_{[48,64]}$	$69_{[68,70]}$	$15_{[8,20]}$	$16_{[8,23]}$	$50_{[26,72]}$	$77_{[75,79]}$	$21_{[14,29]}$	$56_{[38,71]}$	$\mathbf{98}_{[\mathbf{97,99}]}$	$36_{[24,49]}$	$63_{[31,96]}$	$27_{[18,37]}$	$82_{[78,88]}$	$36_{[27,45]}$	$64_{[59,68]}$	$53_{[48,57]}$	$57_{[54,61]}$
`cube-quadruple-100M`	$0_{[0,0]}$	$37_{[36,39]}$	$36_{[28,44]}$	$37_{[30,44]}$	$23_{[4,41]}$	$47_{[42,53]}$	$9_{[0,20]}$	$22_{[9,36]}$	$\mathbf{93}_{[\mathbf{91,95}]}$	$18_{[2,37]}$	$73_{[44,89]}$	$7_{[1,15]}$	$67_{[62,74]}$	$24_{[18,30]}$	$53_{[53,53]}$	$32_{[30,33]}$	$38_{[34,41]}$
`cube-octuple-1B`	$0_{[0,0]}$	$0_{[0,0]}$	$0_{[0,0]}$	$0_{[0,0]}$	$0_{[0,0]}$	$10_{[8,13]}$	$1_{[0,2]}$	$7_{[3,11]}$	$\mathbf{31}_{[\mathbf{29,33}]}$	$2_{[0,4]}$	$20_{[12,26]}$	$0_{[0,1]}$	$\mathbf{33}_{[\mathbf{30,35}]}$	$18_{[14,21]}$	$0_{[0,0]}$	$0_{[0,0]}$	$28_{[27,28]}$
`humanoidmaze-giant`	$0_{[0,0]}$	$0_{[0,0]}$	$81_{[79,83]}$	$49_{[46,52]}$	$1_{[0,1]}$	$92_{[91,94]}$	$18_{[17,19]}$	$\mathbf{97}_{[\mathbf{95,98}]}$	$51_{[47,54]}$	$0_{[0,1]}$	$66_{[64,68]}$	$0_{[0,0]}$	$18_{[13,25]}$	$24_{[20,28]}$	$4_{[2,6]}$	$1_{[0,3]}$	$4_{[2,5]}$
`puzzle-4x5`	$18_{[17,19]}$	$19_{[19,20]}$	$20_{[20,20]}$	$20_{[20,20]}$	$66_{[61,71]}$	$90_{[87,94]}$	$30_{[25,34]}$	$88_{[86,90]}$	$\mathbf{96}_{[\mathbf{95,97}]}$	$31_{[26,35]}$	$\mathbf{96}_{[\mathbf{95,97}]}$	$28_{[25,32]}$	$1_{[0,2]}$	$0_{[0,0]}$	$20_{[20,20]}$	$0_{[0,0]}$	$0_{[0,0]}$
`puzzle-4x6-1B`	$19_{[19,20]}$	$35_{[33,37]}$	$25_{[23,28]}$	$26_{[24,28]}$	$54_{[46,61]}$	$81_{[77,86]}$	$41_{[36,47]}$	$\mathbf{95}_{[\mathbf{92,97}]}$	$75_{[71,79]}$	$42_{[37,49]}$	$\mathbf{98}_{[\mathbf{97,99}]}$	$45_{[41,51]}$	$62_{[57,71]}$	$10_{[3,17]}$	$7_{[2,13]}$	$0_{[0,0]}$	$5_{[3,5]}$

Table 2: **Complete results for all configurations.** All means and 95% bootstrapped confidence intervals are computed over 6 seeds. ($\star$) indicates that we take the results from the original paper (Park et al., 2025), where we take the results with larger 10M-sized datasets for `humanoidmaze-giant` (originally 4M) and `puzzle-4x5` (originally 3M). For **QC** ($h = 25$), we use $\kappa_b = 0.93$ for `cube-*`, $\kappa_b = 0.9$ on `humanoidmaze-giant` and `puzzle-4x5`, $\kappa_b = 0.7$ on `puzzle-4x6` (same as **QC** with $h = 5$). For **QC-NS**, we use the same implicit parameters as **DQC**. For **NS** ($n = 5$), we use $\kappa_b = 0.93$ on `cube-*`, $\kappa_b = 0.7$ on `humanoidmaze-giant` and `puzzle-4x5`, $\kappa_b = 0.5$ on `puzzle-4x6` (same as **NS** with $n = 25$).

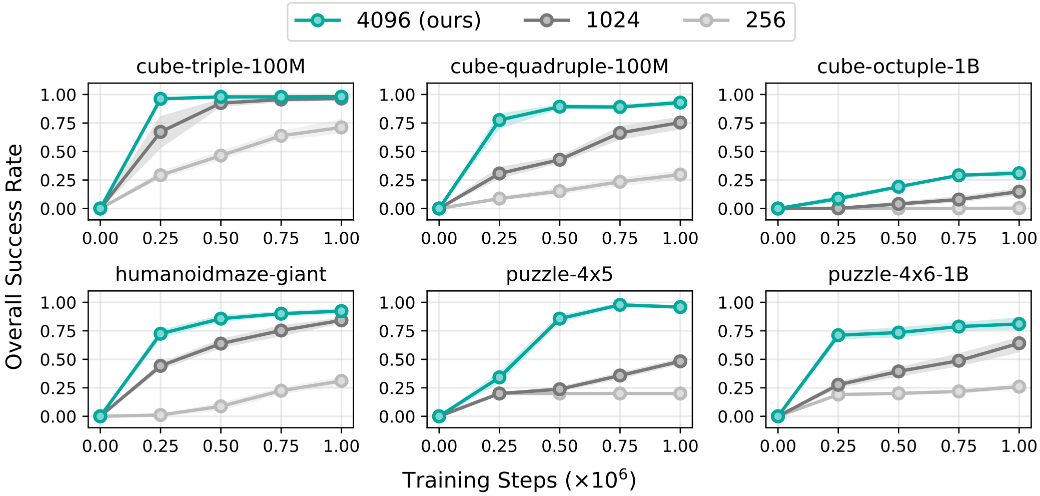


Figure 4: **Batch size sensitivity.** Large batch size is crucial for DQC's performance especially on hard tasks.

B Additional empirical analysis

To gain more insights of the role of the implicit parameters κ_b and κ_d in DQC, we plot the average value of V_ξ, Q_ϕ and Q_ψ^P over the course of training for each task in Figure 5.

C Computation resource

All our experiments are run NVIDIA RTX-A5000 GPU. On average, each 1M-training-step experiment takes about 8-10 hours (depending on the method). To reproduce our main results (*e.g.*, Table 2), we estimate it would take around $\underbrace{10}_{\text{hours per single run}} \times \underbrace{14}_{\text{\# of methods}} \times \underbrace{6}_{\text{\# of tasks}} \times \underbrace{6}_{\text{\# of seeds}} =$ 5 040 GPU hours. Reproducing our sensitivity analysis in Figure 3 and Figure 4 would take another extra $\underbrace{10}_{\text{hours per single run}} \times \underbrace{22}_{\text{\# of analysis curves}} \times \underbrace{6}_{\text{\# of seeds}} = 1\,320$ GPU hours. We also report the training speed and the parameter count for both our method and all our baselines in Table 3.

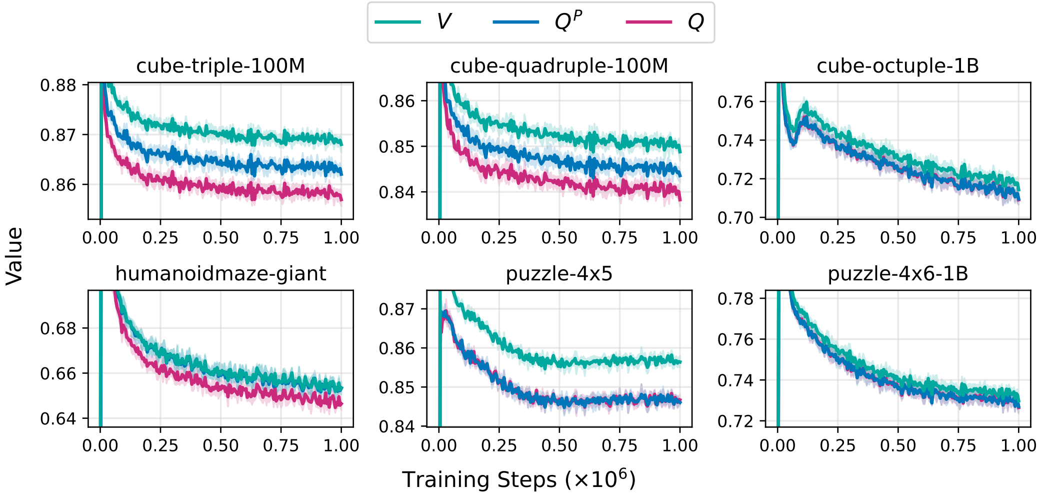


Figure 5: **Value of V_ξ, Q_ϕ, Q^P_ψ over the course of training of our method, DQC.** For `cube-triple` and `cube-quadruple`, DQC uses $\kappa_b = 0.93, \kappa_d = 0.8$. This is reflected as the value gap between V, Q^P and Q. The partial critic Q^P optimistically distills from the full critic Q and the value V optimistically backs up from Q^P. For `cube-octuple` and `puzzle-4x5`, we use $\kappa_d = 0.5$, which causes Q^P to closely track Q. For `humanoidmaze-giant`, DQC uses $\kappa_b = 0.5$ and $\kappa_d = 0.8$ which make V closely tracks Q^P and Q^P optimistically distills from Q. Finally, for `puzzle-4x6`, we use $\kappa_b = \kappa_d = 0.5$ which causes all value functions to output a similar value.

	DQC	**QC**	**NS / OS**	**SHARSA**	**HIQL**	**IQL**	**HFBC**	**FBC**
`training speed (sec/step)`	0.0271	0.0203	0.0200	0.0235	0.0401	0.0243	0.0101	0.0066
`parameter count`	26 218 528	19 507 230	19 384 330	22 677 526	22 605 853	19 390 474	6 490 129	3 237 893

Table 3: **Training speed and parameter count for each method on `cube-quadruple-100M`.**

D Environments and Datasets

To evaluate our method, we consider 8 goal-conditioned environments in OGBench with varying difficulties (Figure 6). The dataset size, episode length, and the action dimension for each environment is available in Table 4. We describe each of the environments and the datasets we use as follows.

Environment `cube-*`: We consider three cube environments (`cube-triple`, `cube-quadruple`, `cube-octuple`). As the names suggest, the goal of these environments involve using a robot arm to manipulate 3/4/8 cubes from some initial configuration to some specified goal configuration. We use the same five evaluation tasks used in OGBench (Park et al., 2024a) for `cube-triple` and `cube-quadruple` and the same five evaluation tasks used in Park et al. (2025) for `cube-octuple`. We refer the environment detail to the corresponding references.

Environment	**Dataset Size**	**Episode Length**	**Action Dim.** (A)
`cube-triple-100M`	100M	1000	5
`cube-quadruple-100M`	100M	1000	5
`cube-octuple-1B`	1B	1500	5
`humanoidmaze-giant`	4M (default)	4000	21
`puzzle-4x5`	3M (default)	1000	5
`puzzle-4x6-1B`	1B	1000	5

Table 4: **Environment metadata.** For both `humanoidmaze-giant` and `puzzle-4x5`, we use the default dataset that is released in the original OGBench benchmark (Park et al., 2024a). For the other environments, we use larger datasets as we find them to be essential for achieving good performances on these environments.

Environment humanoidmaze-*: We also consider the hardest locomotion environment available in OGBench. The goal of the environment is to control and navigate a humanoid agent from some initial location to some specified goal location in a 16 × 12 maze. This environment also has the longest episode length (4000, more than twice as long as the second longest episode length as used in cube-octuple). We refer the environment detail to Park et al. (2024a).

Environment puzzle-*: Finally, we consider two environments that involve solving a combinatorial puzzle with a robot arm. The puzzle consists of a board of 4×5 or 4×6 buttons, organized as a regular grid (4 rows and 5 or 6 columns). Each button has a binary state. Whenever the end-effector of the arm touches a button, the button and all its adjacent four buttons (three or two if the button is on the edge of the grid or in the corner) flip its binary state. The goal of the environment is to transform the board from some initial state to some specified goal state. We refer the environment detail to Park et al. (2025).

At the test-time/evaluation-time, the goal-conditioned agent is tested on five evaluation tasks for each of the six environments we consider. The overall success rate is the average over 5 tasks with 50 evaluation trials each.

Datasets. We use play datasets for all cube-* and puzzle-* environments and navigate dataset for humanoidmaze-*. We use the original datasets available for humanoidmaze-giant and puzzle-4x5 because they are sufficient for solving the environments. Using larger datasets on these environments do not help differentiating among different methods/baselines. For each of the other environments, we use the largest dataset available from Park et al. (2025) as we find it to be necessary to solve these environments (or achieve non-trivial performance on the hardest cube-octuple environment).

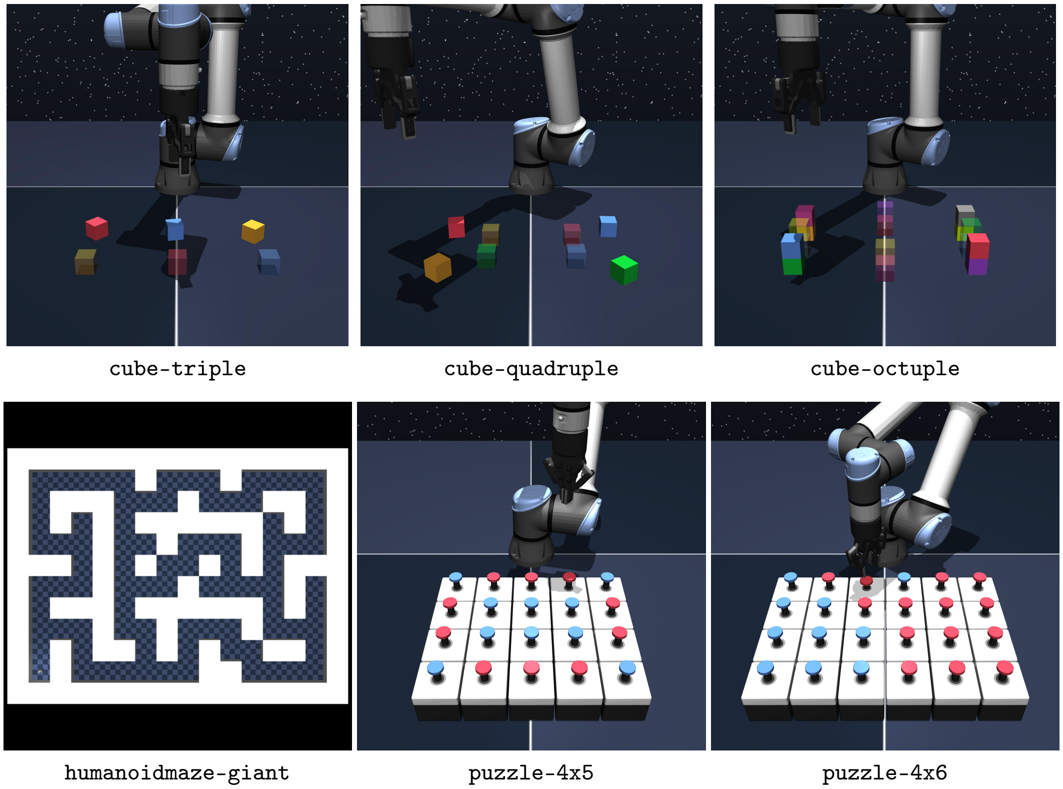


Figure 6: **Visualization of environments**.

E HYPERPARAMETERS AND IMPLEMENTATION DETAILS

Hyperparameters. Table 5 describes the common hyperparameters used in all our experiments. Table 7 describe the environment-specific hyperparameters and Table 8 describes the range of hyperparameters we use for tuning each method.

Parameter	Value
Batch size	4096
Discount factor (γ)	0.999
Optimizer	Adam
Learning rate	3×10^{-4}
Target network update rate (λ)	5×10^{-3}
Critic ensemble size (K)	2
Critic target	$\min(Q_1, Q_2)$ for `cube-*` $(Q_1 + Q_2)/2$ for `puzzle-*` and `humanoid-*`
Implicit Backup Quantile (κ_b)	0.9
Value loss type	binary cross entropy
Best-of-N sampling (N)	32
Number of flow steps	10
Number of training steps	10^6
Network width	1024
Network depth	4 hidden layers
Value goal sampling ($w^{\text{v}}_{\text{cur}}, w^{\text{v}}_{\text{geom}}, w^{\text{v}}_{\text{traj}}, w^{\text{v}}_{\text{rand}}$)	$(0.2, 0, 0.5, 0.3)$
Actor goal sampling ($w^{\text{p}}_{\text{cur}}, w^{\text{p}}_{\text{geom}}, w^{\text{p}}_{\text{traj}}, w^{\text{p}}_{\text{rand}}$)	DQC/QC/NS/OS: π_β is not goal-conditioned SHARSA (`cube`): $(0, 1, 0, 0)$ SHARSA (`puzzle`): $(0, 0, 1, 0)$ SHARSA (`humanoidmaze`): $(0, 0, 1, 0)$

Table 5: **Common hyperparameters.** For the GCRL goal-sampling distribution we follow the same hyperparameters used in Park et al. (2025).

Goal-conditioned RL implementation details. While we have described in the main body of the paper how DQC works as a general RL algorithm, we have not touched on how DQC and similarly all our baselines works with the goal-condition RL (GCRL) setting. We consider the setting where we have access to an oracle goal representation $\Psi : \mathcal{S} \to \mathcal{G}$ where $\mathcal{G}$ is the goal space (see Table 6 for the oracle goal representation description for each environment). The goal-conditioned reward function $r : (s, g) \mapsto \mathbb{I}_{\Psi(s)=g}$ is a binary reward function where its output is 1 if the goal g is reached by the current state s. We can treat g as part of an extended state $\tilde{s} = [s, g] \in \tilde{S} = \mathcal{S} \times \mathcal{G}$ and learn value functions (*e.g.*, $Q_\phi(\tilde{s}, a)$) normally with such extended state.

Environment	Goal Representation (Ψ)	Goal Domain ($\mathcal{G}$)
`cube-triple`	(x, y, z) of three cubes (rel. to center)	$\mathbb{R}^9$
`cube-quadruple`	(x, y, z) of four cubes (rel. to center)	$\mathbb{R}^{12}$
`cube-octuple`	(x, y, z) of eight cubes (rel. to center)	$\mathbb{R}^{24}$
`humanoidmaze-giant`	(x, y) of the humanoid	$\mathbb{R}^2$
`puzzle-4x5`	the binary state for each button	$\{0, 1\}^{20}$
`puzzle-4x6`	the binary state for each button	$\{0, 1\}^{24}$

Table 6: **Oracle goal representation description for each environment.** Following Park et al. (2025), we assume access to an oracle goal representation for each environment. More detailed definition of these oracle goal representations is available in OGBench (Park et al., 2024a).

A common trick in the GCRL setting is to use goal relabeling. That is, during training for each (s, a) pair in the training batch, a goal g is sampled from some distribution (*i.e.*, $p^{\mathcal{D}}(\cdot \mid s, a)$) and the reward of the transition is relabeled with the goal-conditioned reward function. Following Park et al. (2025),

Environment	DQC $(h, h_a, \kappa_b, \kappa_d)$	DQC-naïve (h, h_a, κ_b)	QC $(h = h_a, \kappa_b)$	NS (n, κ_b)	OS (κ_b)	SHARSA (n)	HIQL (h, κ, α)	IQL (α)	HFBC (h)
`cube-triple-100M`	$(25, 5, 0.93, 0.8)$	$(25, 5, 0.93)$	$(5, 0.93)$	$(25, 0.5)$	0.5	25	$(25, 0.5, 10)$	3	25
`cube-quadruple-100M`	$(25, 5, 0.93, 0.8)$	$(5, 1, 0.93)$	$(5, 0.93)$	$(25, 0.5)$	0.7	25	$(25, 0.5, 10)$	3	25
`cube-octuple-1B`	$(25, 5, 0.93, 0.5)$	$(25, 5, 0.93)$	$(25, 0.93)$	$(25, 0.97)$	0.7	25	$(50, 0.5, 10)$	10	50
`humanoidmaze-giant`	$(25, 1, 0.5, 0.8)$	$(5, 1, 0.9)$	$(5, 0.9)$	$(25, 0.7)$	0.5	50	$(50, 0.5, 3)$	0.3	50
`puzzle-4x5`	$(25, 5, 0.9, 0.5)$	$(25, 5, 0.9)$	$(5, 0.9)$	$(25, 0.7)$	0.7	50	$(25, 0.7, 3)$	1	25
`puzzle-4x6-1B`	$(25, 1, 0.7, 0.5)$	$(25, 5, 0.7)$	$(5, 0.7)$	$(25, 0.5)$	0.7	50	$(25, 0.7, 3)$	1	25

Table 7: **Environment-specific hyperparameters for `DQC`, `QC`, `NS`, `OS`, and `SHARSA`** . For `SHARSA`, we follow the hyperparameters in the original paper (Park et al., 2025).

Environment	Backup Quantile (κ_b)	Distillation Expectile (κ_d)	Backup horizon (h) or (n)	Policy Chunk Size (h_a)
`cube-*`	$\{0.5, 0.7, 0.9, 0.93, 0.95, 0.97, 0.99\}$	$\{0.5, 0.8\}$	$\{5, 25\}$	$\{1, 5, 25\}$
`{humanoidmaze/puzzle}-*`	$\{0.5, 0.7, 0.9\}$	$\{0.5, 0.8\}$	$\{5, 25\}$	$\{1, 5, 25\}$

Table 8: **Hyperparameter tuning range for all methods**. For **NS**, we only tune κ_b and n because the policy chunk size is always 1 and there is no distilled critic. Similarly, for **QC**, we only tune κ_b and $h = h_a$ because the policy chunk size is the same as the crtici chunk size and there is no distilled critic. For **OS**, we only tune κ_b.

the goal distribution $P^g(\cdot \mid s, a) : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{G}}$ is a mixture of four distributions, conditioned on the training state-action example:

$$P^g = w_{\text{cur}} P^g_{\text{cur}} + w_{\text{geom}} P^g_{\text{geom}} + w_{\text{traj}} P^g_{\text{traj}} + w_{\text{rand}} P^g_{\text{rand}}, \tag{32}$$

where

1. $P^g_{\text{cur}}(\cdot \mid s, a) = \delta_{\Psi(s)}$: the goal is the same as the current state;
2. $P^g_{\text{geom}}(\cdot \mid s, a)$: geometric distribution over the future states in the same trajectory that (s, a) is from;
3. $P^g_{\text{traj}}(\cdot \mid s, a)$: uniform distribution over the future states in the same trajectory that (s, a) is from; and finally
4. $P^g_{\text{rand}}(\cdot \mid s, a) = \Psi(\mathcal{U}_{\mathcal{D}(s)})$: uniform distribution over the dataset ($\mathcal{D}(s)$ is the distribution of states in the dataset).

and $w_{\text{cur}}, w_{\text{geom}}, w_{\text{traj}}, w_{\text{rand}} > 0$ are the corresponding weights for each of the distribution components with $w_{\text{cur}} + w_{\text{geom}} + w_{\text{traj}} + w_{\text{rand}} = 1$.

In practice, it has been found to be beneficial to use a separate set of goal sampling weights for TD backup (Park et al., 2024a) (*i.e.*, $(w^{\text{v}}_{\text{cur}}, w^{\text{v}}_{\text{geom}}, w^{\text{v}}_{\text{traj}}, w^{\text{v}}_{\text{rand}})$) and for policy learning (*i.e.*, $(w^{\text{p}}_{\text{cur}}, w^{\text{p}}_{\text{geom}}, w^{\text{p}}_{\text{traj}}, w^{\text{p}}_{\text{rand}})$). However, in our implementation of DQC/QC/NS/OS, we do not train a goal-conditioned policy, as our policy extraction is done entirely at test-time by best-of-N sampling from an *unconditional* (*i.e.*, not goal-conditioned) behavior policy π_β. In particular, we use an unconditioned flow policy $\pi_\beta(\cdot \mid s)$ that is parameterized by a velocity field $v_\beta : \mathcal{S} \times \mathbb{R}^A \times [0, 1] \to \mathbb{R}^A$ that is trained with the standard flow-matching objective:

$$L_{\text{FM}}(\beta) = \mathbb{E}_{u \sim \mathcal{U}[0,1], z \sim \mathcal{N}, (s,a) \sim \mathcal{D}} \left[\| v_\beta(s, (1-u)z + ua, u) - a + z \|_2^2 \right] \tag{33}$$

For SHARSA, we use the official implementation where both flow policies (high-level and low-level) are goal-conditioned (and thus are trained with the goal distribution mixture specified by $w^{\text{p}}_{\text{cur}}, w^{\text{p}}_{\text{geom}}, w^{\text{p}}_{\text{traj}}, w^{\text{p}}_{\text{rand}}$). The goal sampling distribution for training the value networks (for all methods) and the goal sampling distribution for the policy networks (for SHARSA only) are provided in Table 5.

F Lower-bound analyses

F.1 AC Value Bias (Proof in Appendix G.3)

Theorem F.1 (Worst-case AC Value Bias). *For any $\gamma \in [0, 1), \varepsilon_h \in [0, 1/2]$, there exists an MDP $\mathcal{M}$ and a weakly ε_h-open-loop consistent $\mathcal{D}$ such that for some $s \in \operatorname{supp}(P_{\mathcal{D}}(s_t))$,*

$$V_{\text{ac}}(s) - \hat{V}_{\text{ac}}(s) = \pm \frac{\gamma \varepsilon_h}{(1-\gamma)(1-(1-\varepsilon_h)\gamma^h)}. \tag{34}$$

F.2 Optimality Gap for Action Chunking Policy (Proof in Appendix G.5)

Corollary F.2 (Worse-case Optimality Gap for Action Chunking Policy). *For any $\gamma \in [0, 1), \varepsilon_h \in [0, 1/2]$, there exists an MDP $\mathcal{M}$ whose optimal policy $\pi^\star$ induces a data distribution $\mathcal{D}^\star$ that is weakly ε_h-open-loop consistent, such that for some $s \in \operatorname{supp}(P_{\mathcal{D}^\star}(s_t))$,*

$$V^\star(s) - V^\star_{\text{ac}}(s) = \frac{\gamma \varepsilon_h}{(1-\gamma)(1-(1-\varepsilon_h)\gamma^h)}. \tag{35}$$

F.3 Q-learning with Action Chunking Policy (Proof in Appendix G.7)

Theorem F.3 (Worst-case Analysis of Q-Learning with Action Chunking Policy on Off-policy Data). *For any $\varepsilon_h \in (0, 1/5)$, $\gamma \in (0, 1)$, $c_1 \in (0, \varepsilon_h/2)$, and $c_2 \in (0, 2\varepsilon_h\gamma)$, there exists an MDP $\mathcal{M}$ and strongly ε_h-open-loop consistent data distribution $\mathcal{D}$ and $\mathcal{D}^\star$ with $\operatorname{supp}(P_{\mathcal{D}}(s_t, a_{t:t+h})) \supseteq \operatorname{supp}(P_{\mathcal{D}^\star}(s_t, a_{t:t+h}))$, such that for some $s \in \operatorname{supp}(P_{\mathcal{D}^\star}(s_t))$,*

$$V^\star(s) - V^+_{\text{ac}}(s) = \frac{2\varepsilon_h\gamma - c_2}{(1-\gamma)(1-(1-2\varepsilon_h)\gamma^h)} + \frac{\varepsilon_h\gamma}{(1-\gamma)(1-(1-\varepsilon_h-c_1)\gamma^h)}, \tag{36}$$

where $V^\star$ is the value of an optimal policy and V^+_{ac} is the true *value of π^+_{ac}. As $c_1, c_2 \to 0$,*

$$V^\star(s) - V^+_{\text{ac}}(s) \to \frac{\varepsilon_h\gamma}{1-\gamma}\left[\frac{2}{1-(1-2\varepsilon_h)\gamma^h} + \frac{1}{1-(1-\varepsilon_h)\gamma^h}\right]. \tag{37}$$

F.4 Closed-loop AC Policy under BOV (Proof in Appendix G.11)

Theorem F.4 (Worst-case Closed-loop AC Policy under BOV). *For any $\gamma \in (0, 1)$, $\vartheta_h^G, \vartheta_h^L \in \left(0, \frac{\gamma-\gamma^h}{4(1-\gamma)}\right]$, $c \in \left[0, \frac{\gamma-\gamma^h}{4(1-\gamma^h)}\right)$, $\sigma \in \left(0, \frac{\min(\vartheta_h^G, \vartheta_h^L)}{1-\gamma}\right)$, there exists $\mathcal{M}$ and $\mathcal{D}$ satisfying the mixture assumption in Theorem 4.11 such that there exists $s_t \in \operatorname{supp}(P_{\mathcal{D}^\star}(s_t))$, where*

$$V^\star(s_t) - V^\bullet(s_t) = \frac{\vartheta_h^L}{1-\gamma} + \frac{\vartheta_h^G + \gamma^h \min(\vartheta_h^L, \vartheta_h^G)}{(1-\gamma)(1-\gamma^h)} - \sigma, \quad V^\star(s_t) - V^+_{\text{ac}}(s_t) \geq \frac{c}{1-\gamma} \tag{38}$$

F.5 ε-deterministic dynamics is weakly open-loop consistent

To provide some intuitions on what this open-loop consistency implies, we discuss a concrete family of MDPs where any data distribution from these MDPs are (weakly) ε_h-open-loop consistent (Proposition F.6, with proof available in Appendix G.12).

Definition F.5 (Near-deterministic Dynamics). *A transition dynamics T is ε-deterministic if there exists a deterministic transition dynamics represented by function $f : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ and another arbitrary transition dynamics $\tilde{T} : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}$, and T is a combination of f and $\tilde{T}$:*

$$T(s' \mid s, a) = (1-\varepsilon)\delta_{f(s,a)}(s') + \varepsilon \tilde{T}(s' \mid s, a), \forall s, s' \in \mathcal{S}, a \in \mathcal{A}. \tag{39}$$

Proposition F.6 (Deterministic Dynamics are Weakly Open-loop Consistent). *If a transition dynamics $\mathcal{M}$ is ε-deterministic, then any data $\mathcal{D}$ collected from $\mathcal{M}$ is weakly ε_h-open-loop consistent with respect to $\mathcal{M}$ for any $h \in \mathbb{N}^+$ as long as $\varepsilon_h \geq 3(1-(1-\varepsilon)^{h-1})$.*

An ε-deterministic dynamics acts like a deterministic one most of the time (with $1 - \varepsilon$ probability) and a non-deterministic one occasionally (with ε probability). This bounded stochasticity allows the results of taking an action sequence (of length h) open-loop to be deterministically determined in the event that the deterministic dynamics is 'triggered' (with a joint $(1 - \varepsilon)^{h-1}$ probability across h time steps). It is clear that under such event, there is no gap between the 'replayed' open-loop data $P_{\mathcal{D}}^{\circ}$ and the original data distribution $P_{\mathcal{D}}$, and as result there is also no value estimation bias under this event, and thus intuitively we can bound the value estimation error by a function of the probability that the stochastic dynamics is 'triggered' (*i.e.*, with $1 - (1 - \varepsilon)^{h-1}$ probability).

G Proofs of main results

G.1 Utility Lemmata

Lemma G.1 (Mean value theorem for conditional probabilities)**.** *Let $P_1, P_2 \in \Delta_{\mathcal{X}\times\mathcal{Y}}$ and $P(x,y) := \hat{\alpha}(y)P_1(x,y) + (1-\hat{\alpha}(y))P_2(x,y)$ and there exists $\alpha > 0$ such that $\hat{\alpha}(y) \le \alpha, \forall y \in \mathcal{Y}$. Then, there exists $y \in \mathcal{Y}$ and $\tilde{\alpha} \le \alpha$ such that*

$$P(\cdot \mid y) = \tilde{\alpha}P_1(\cdot \mid y) + (1-\tilde{\alpha})P_2(\cdot \mid y) \tag{40}$$

Proof.

$$\begin{aligned}\frac{P(x,y)}{P(y)} &= \frac{\hat{\alpha}(y)P_1(y)P_1(x \mid y) + (1-\hat{\alpha}(y))P_2(x \mid y)}{\hat{\alpha}(y)P_1(y) + (1-\hat{\alpha}(y))P_2(y)} \\ &= \beta(y)P_1(x \mid y) + (1-\beta(y))P_2(x \mid y)\end{aligned} \tag{41}$$

where $\beta(y) := \frac{\hat{\alpha}(y)P_1(y)}{\hat{\alpha}(y)P_1(y)+(1-\hat{\alpha}(y))P_2(y)}$. We now prove $\exists y \in \mathcal{Y}, \tilde{\alpha} \le \alpha$ for Equation (40) to hold by contradiction.

We first assume $\tilde{\alpha} = \beta(y) > \alpha, \forall y \in \mathcal{Y}$. Now, substitute $\beta(y)$ in and integrate both side by y to obtain

$$\hat{\alpha}(y)P_1(y) > \alpha\hat{\alpha}(y)P_1(y) + \alpha(1-\hat{\alpha}(y))P_2(y) \tag{42}$$

$$\hat{\alpha}(y) > \alpha\hat{\alpha}(y) + \alpha - \alpha\hat{\alpha}(y) = \alpha, \tag{43}$$

which is a contradiction to the condition $\hat{\alpha}(y) \le \alpha$.

Therefore, there must exist $y \in \mathcal{Y}$ with $\tilde{\alpha} \le \alpha$ such that Equation (40) holds. □

Lemma G.2 (Expectation difference for bounded function and TV)**.** *For two distributions $P, Q \in \Delta_{\mathcal{X}}$ and two bounded functions $f, g : \mathcal{X} \to [0,1]$, if the TV distance between P and Q is no larger than ε and $\|f-g\|_\infty \le \delta$ under $\operatorname{supp}(P) \cap \operatorname{supp}(Q)$, then*

$$|\mathbb{E}_{x\sim P}[f(x)] - \mathbb{E}_{x\sim Q}[g(x)]| \le (1-\varepsilon)\delta + \varepsilon. \tag{44}$$

Proof. Let's decompose the probability mass of P and Q in terms of $d_P, d_{PQ}, d_Q : \mathcal{X} \to \mathbb{R}$ as the following:

$$P(x) = d_P(x) + d_{PQ}(x), \tag{45}$$

$$Q(x) = d_{PQ}(x) + d_Q(x). \tag{46}$$

The $\int d_P(x)\mathrm{d}x$ maximizing solution is

$$d_P(x) = \max(P(x), Q(x)) - Q(x) \tag{47}$$

$$d_Q(x) = \max(P(x), Q(x)) - P(x) \tag{48}$$

$$d_{PQ}(x) = P(x) + Q(x) - \max(P(x), Q(x)). \tag{49}$$

It is clear that under this decomposition,

$$\int d_P(x)\mathrm{d}x = \int d_Q(x)\mathrm{d}x = \hat{\varepsilon} \le \varepsilon, \tag{50}$$

$$\int d_{PQ}(x)\mathrm{d}x = 1 - \hat{\varepsilon} \ge 1 - \varepsilon. \tag{51}$$

Now we are ready to bound the expectation difference:

$$
\begin{aligned}
&|\mathbb{E}_{x\sim P}[f(x)] - \mathbb{E}_{x\sim Q}[g(x)]| \\
&= \left|\left(\int d_P(x)f(x)\mathrm{d}x - \int d_Q(x)g(x)\mathrm{d}x\right) + \left(\int d_{PQ}(x)(f(x)-g(x))\mathrm{d}x\right)\right| \\
&\le \left|\int d_P(x)f(x)\mathrm{d}x - \int d_Q(x)g(x)\mathrm{d}x\right| + \left|\int d_{PQ}(x)(f(x)-g(x))\mathrm{d}x\right| \\
&\le \max\left(\sup_x f(x)\int d_P(x)\mathrm{d}x - \inf_x g(x)\int d_Q(x)\mathrm{d}x, \sup_x g(x)\int d_Q(x)\mathrm{d}x - \inf_x f(x)\int d_P(x)\mathrm{d}x\right) \\
&\qquad + \left|\left(\sup_{x:d_{PQ}(x)>0}|f(x)-g(x)|\right)\int d_{PQ}(x)\mathrm{d}x\right| \\
&\le \hat{\varepsilon} + \left(\sup_{x\in\mathrm{supp}(P)\cap\mathrm{supp}(Q)}|f(x)-g(x)|\right)(1-\hat{\varepsilon}) \\
&= \hat{\varepsilon} + \|f-g\|_\infty(1-\hat{\varepsilon}) \\
&\le \hat{\varepsilon}(1-\delta) + \delta \\
&= (1-\varepsilon)\delta + \varepsilon
\end{aligned}
\tag{52}
$$

as desired. □

Lemma G.3 (Total variation under event conditioning). *For two random variables $X \in \Delta_{\mathcal{X}}$ and $Y \in \Delta_{\mathcal{Y}}$ and any $y \in \mathcal{Y}$,*

$$D_{\mathrm{TV}}(P(X \mid Y=y) \parallel P(X)) \le 1 - P(Y=y) \tag{53}$$

Proof. Let $p = P(Y=y)$

$$
\begin{aligned}
&D_{\mathrm{TV}}(P(X \mid Y=y) \parallel P(X)) \\
&= \frac{1}{2}\int |P(x) - P(x \mid y)|\mathrm{d}x \\
&= \frac{1}{2}\int |P(x \mid Y=y)(P(Y=y)-1) + P(x \mid Y\neq y)P(Y\neq y)|\mathrm{d}x \\
&= \frac{1-p}{2}\int |(P(x \mid Y\neq y) - P(x \mid Y=y))|\mathrm{d}x \\
&= (1-p)D_{\mathrm{TV}}(P(X \mid Y=y) \parallel P(X \mid Y\neq y)) \\
&\le 1-p
\end{aligned}
\tag{54}
$$

□

Lemma G.4 (Data Processing Inequality for f-divergence (Csiszár, 1967)). *For two random variables $A, B \in \Delta_{\mathcal{X}}$ and a deterministic function $f : \mathcal{X} \to \mathcal{Y}$, and $C := g(A), D := g(B)$*

$$D_f(P_A \parallel P_B) \ge D_f(P_C \parallel P_D). \tag{55}$$

Since TV-distance is a f-divergence with $f = |x-1|$, we have

$$D_{\mathrm{TV}}(P_A \parallel P_B) \ge D_{\mathrm{TV}}(P_C \parallel P_D). \tag{56}$$

Proof from Wu (2017).

$$
\begin{aligned}
D_f(P_A \parallel P_B) &= \mathbb{E}_{x\sim P_B}\left[f(P_A(x)/P_B(x))\right] \\
&= \mathbb{E}_{P_{BD}}\left[f(P_{AC}/P_{BD})\right] \\
&= \mathbb{E}_{(x,y)\sim P_D}\left[\mathbb{E}_{P_{B|D}}\left[f(P_{AC}(x,y)/P_{BD}(x,y))\right]\right] \\
&\ge \mathbb{E}_{y\sim P_D}\left[f\left(\mathbb{E}_{x\sim P_{B|D=y}}\left[P_{AC}(x,y)/P_{BD}(x,y)\right]\right)\right] \\
&= \mathbb{E}_{y\sim P_D}\left[f\left(\mathbb{E}_{x\sim P_{B|D=y}}\left[P_C(y)/P_D(y)\right]\right)\right] \\
&= D_f(P_C \parallel P_D).
\end{aligned}
\tag{57}
$$

□

G.2 Proof of Theorem 4.4

Theorem 4.4 (Bias of Action Chunking Critic)**.** *Let* $\hat{V}_{\text{ac}} : \mathcal{S} \to [0, 1/(1-\gamma)]$ *be a solution of*

$$\hat{V}_{\text{ac}}(s_t) = \mathbb{E}_{s_{t+1:t+h+1}, a_{t:t+h} \sim P_{\mathcal{D}}(\cdot|s_t)} \left[R_{t:t+h} + \gamma^h \hat{V}_{\text{ac}}(s_{t+h}) \right], \tag{12}$$

with $R_{t:t+h} = \sum_{t'=t}^{t+h} \gamma^{t'-t} r(s_{t'}, a_{t'})$ *and* V_{ac} *is the true value of* $\pi^\circ_{\mathcal{D}} : s_t \mapsto P_{\mathcal{D}}(a_{t:t+h} \mid s_t)$*. If* $\mathcal{D}$ *is* ε_h*-open-loop consistent, then under* $\text{supp}(\mathcal{D})$*,*

$$\left\| V_{\text{ac}} - \hat{V}_{\text{ac}} \right\|_\infty \leq \frac{\varepsilon_h \gamma}{(1 - (1 - \varepsilon_h)\gamma^h)(1-\gamma)} \leq \frac{\varepsilon_h}{(1-\gamma^h)(1-\gamma)}. \tag{13}$$

Proof. Since $\mathcal{D}$ is $\varepsilon_{h'}$-open-loop consistent in state-action for $h' < h$, the state-action distribution leading up to step h admits the following bound:

$$D_{\text{TV}}(P_{\mathcal{D}}(s_{t+h}, a_{t+h} \mid s_t) \parallel P^\circ_{\mathcal{D}}(s_{t+h}, a_{t+h} \mid s_t)) \leq \varepsilon_h \tag{58}$$

Let $R_{t:t+h} = \sum_{k=0}^{h-1} \gamma^k r(s_{t+k}, a_{t+k})$ be the h-step reward distribution. Then the difference in h-step reward is bounded by

$$\begin{aligned}
&\left| \mathbb{E}_{P_{\mathcal{D}}(\cdot|s_t)}[R_{t:t+h}] - \mathbb{E}_{P^\circ_{\mathcal{D}}(\cdot|s_t)}[R_{t:t+h}] \right| \\
&\leq \sum_{h'=1}^{h-1} \left[\gamma^{h'} \mathbb{E}_{P_{\mathcal{D}}(s_{t+h'}, a_{t+h'}|s_t)}[r(s_{t+h'}, a_{t+h'})] - \mathbb{E}_{P^\circ_{\mathcal{D}}(s_{t+h'}, a_{t+h'}|s_t)}[r(s_{t+h'}, a_{t+h'})] \right] \\
&\leq \sum_{h'=1}^{h-1} \gamma^{h'} \varepsilon_h.
\end{aligned} \tag{59}$$

where the first inequality uses Lemma G.2 and the fact that TV distance is bounded (Equation (58)).

Since $\mathcal{D}$ is ε_h-open-loop consistent for h in state, we have

$$D_{\text{TV}}(P_{\mathcal{D}}(s_{t+h} \mid s_t) \parallel P^\circ_{\mathcal{D}}(s_{t+h} \mid s_t)) \leq \varepsilon_h, \tag{60}$$

which can then be used to bound the estimation error using Lemma G.2:

$$\begin{aligned}
&\left| \mathbb{E}_{s_{t+h} \sim P_{\mathcal{D}}(s_{t+h}|s_t)} \left[\hat{V}_{\text{ac}}(s_{t+h}) \right] - \mathbb{E}_{s_{t+h} \sim P^\circ_{\mathcal{D}}(s_{t+h}|s_t)} \left[V_{\text{ac}}(s_{t+h}) \right] \right| \\
&\leq \frac{\varepsilon_h}{1-\gamma} + (1 - \varepsilon_h) \sup_{s_{t+h} \in \text{supp}(P_{\mathcal{D}}(s_{t+h}|s_t))} \left[|\hat{V}_{\text{ac}}(s_{t+h}) - V_{\text{ac}}(s_{t+h})| \right]
\end{aligned} \tag{61}$$

For all $s_t \in \text{supp}(P_{\mathcal{D}}(s_t))$,

$$\begin{aligned}
&\left| \hat{V}_{\text{ac}}(s_t) - V_{\text{ac}}(s_t) \right| \\
&\leq \left| \mathbb{E}_{P_{\mathcal{D}}(\cdot|s_t)}[R_{t:t+h}] - \mathbb{E}_{P^\circ_{\mathcal{D}}(\cdot|s_t)}[R_{t:t+h}] \right| \\
&\quad + \gamma^h \left| \mathbb{E}_{s_{t+h} \sim P_{\mathcal{D}}(s_{t+h}|s_t)} \left[\hat{V}_{\text{ac}}(s_{t+h}) \right] - \mathbb{E}_{s_{t+h} \sim P^\circ_{\mathcal{D}}(s_{t+h}|s_t)} \left[V_{\text{ac}}(s_{t+h}) \right] \right| \\
&\leq \sum_{h'=0}^{h-1} \left[\gamma^{h'} \varepsilon_h \right] + \frac{\gamma^h \varepsilon_h}{1-\gamma} + \gamma^h (1 - \varepsilon_h) \sup_{s_{t+h} \in \text{supp}(P_{\mathcal{D}}(s_{t+h}|s_t))} \left[|\hat{V}_{\text{ac}}(s_{t+h}) - V_{\text{ac}}(s_{t+h})| \right].
\end{aligned} \tag{62}$$

Since the support of $s_{t+h} \mid s_t$ is a subset of the support for s_t by Assumption 4.1, we can recursively apply the inequality to obtain,

$$\begin{aligned}
\left| \hat{V}_{\text{ac}}(s_t) - V_{\text{ac}}(s_t) \right| &\leq \frac{1}{1 - (1-\varepsilon_h)\gamma^h} \left(\sum_{h'=1}^{h-1} \left[\gamma^{h'} \varepsilon_h \right] + \frac{\gamma^h \varepsilon_h}{1-\gamma} \right) \\
&= \frac{\gamma \varepsilon_h}{(1-\gamma)(1 - (1-\varepsilon_h)\gamma^h)},
\end{aligned} \tag{63}$$

as desired. □

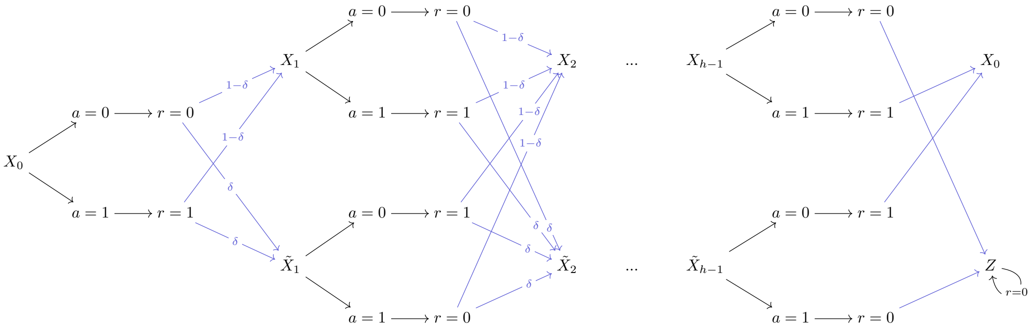


Figure 7: **A $2h$-state MDP that is constructed to meet the upper-bound in Theorem 4.4.** The data distribution $\mathcal{D}$ that achieves such upper-bound is collected by the optimal policy: $\pi(X_i) = 1, \pi(\tilde{X}_i) = 0$.

G.3 Proof of Theorem F.1

Theorem F.1 (Worst-case AC Value Bias). *For any $\gamma \in [0,1)$, $\varepsilon_h \in [0,1/2]$, there exists an MDP $\mathcal{M}$ and a weakly ε_h-open-loop consistent $\mathcal{D}$ such that for some $s \in \mathrm{supp}(P_{\mathcal{D}}(s_t))$,*

$$V_{\mathrm{ac}}(s) - \hat{V}_{\mathrm{ac}}(s) = \pm \frac{\gamma\varepsilon_h}{(1-\gamma)(1-(1-\varepsilon_h)\gamma^h)}. \tag{34}$$

Proof. Let $\delta \in [0,1]$ be any value that satisfies $\varepsilon_h = 2\delta(1-\delta)$. δ must exist because $\varepsilon_h \in [0,1/2]$. Let us define a MDP that has $S = 2h$ states, $\mathcal{S} = \{X_0, X_1, \tilde{X}_1, \cdots, X_{h-1}, \tilde{X}_{h-1}, Z\}$, and $A = 2$ actions, $\mathcal{A} = \{0,1\}$, and the following transition function T and reward function r (see a diagram in Figure 7):

$$\begin{aligned}
&T(\tilde{X}_{i+1} \mid X_i, a) = T(\tilde{X}_{i+1} \mid \tilde{X}_i, a) = \delta, && \forall a \in \{0,1\}, i \in \{1, \cdots, h-2\}\\
&T(X_{i+1} \mid X_i, a) = T(X_{i+1} \mid \tilde{X}_i, a) = 1-\delta, && \forall a \in \{0,1\}, i \in \{0, \cdots, h-2\}\\
&T(Z \mid \tilde{X}_{h-1}, a=1) = T(Z \mid X_{h-1}, a=0) = 1 &&\\
&T(X_0 \mid \tilde{X}_{h-1}, a=0) = T(X_0 \mid X_{h-1}, a=1) = 1 &&\\
&r(\tilde{X}_i, a=0) = r(X_i, a=1) = 1, && \forall i \in \{0, \cdots, h-1\}\\
&r(\tilde{X}_i, a=1) = r(X_i, a=0) = 0, && \forall i \in \{0, \cdots, h-1\}\\
&r(Z, a=1) = r(Z, a=0) = 0 &&\\
&T(Z \mid Z, a=0) = T(Z \mid Z, a=1) = 1 &&
\end{aligned} \tag{64}$$

Now, we assume that the data $\mathcal{D}$ is collected by the optimal closed-loop policy where

$$\pi(X_i) = 1, \pi(\tilde{X}_i) = 0. \tag{65}$$

First, we check $\mathcal{D}$ is ε_h-open-loop consistent.

We can show that by computing the distribution for $P_{\mathcal{D}}(s_{t+i}, a_{t+i} \mid s_t = X_0)$ and $P^{\circ}_{\mathcal{D}}(s_{t+i}, a_{t+i} \mid s_t = X_0)$ as follows:

$$\begin{aligned}
&\begin{bmatrix} P_{\mathcal{D}}(s_{t+i} = \tilde{X}_i, a_{t+i} = 0 \mid X_0) & P_{\mathcal{D}}(s_{t+i} = \tilde{X}_i, a_{t+i} = 1 \mid X_0) \\ P_{\mathcal{D}}(s_{t+i} = X_i, a_{t+i} = 0 \mid X_0) & P_{\mathcal{D}}(s_{t+i} = X_i, a_{t+i} = 1 \mid X_0) \end{bmatrix} = \begin{bmatrix} \delta & 0 \\ 0 & 1-\delta \end{bmatrix}\\
&\begin{bmatrix} P^{\circ}_{\mathcal{D}}(s_{t+i} = \tilde{X}_i, a_{t+i} = 0 \mid X_0) & P^{\circ}_{\mathcal{D}}(s_{t+i} = \tilde{X}_i, a_{t+i} = 1 \mid X_0) \\ P^{\circ}_{\mathcal{D}}(s_{t+i} = X_i, a_{t+i} = 0 \mid X_0) & P^{\circ}_{\mathcal{D}}(s_{t+i} = X_i, a_{t+i} = 1 \mid X_0) \end{bmatrix} = \begin{bmatrix} \delta^2 & (1-\delta)\delta \\ \delta(1-\delta) & (1-\delta)^2 \end{bmatrix}
\end{aligned} \tag{66}$$

From the calculation above, it is clear that

$$D_{\mathrm{TV}}(P^{\circ}_{\mathcal{D}}(s_{t+i}, a_{t+i} \mid s_t) \parallel P_{\mathcal{D}}(s_{t+i}, a_{t+i} \mid s_t)) = \varepsilon_h, \quad \forall i \in \{1, 2, \cdots, h-1\}. \tag{67}$$

From the computed values of $P^{\circ}_{\mathcal{D}}(s_{t+h-1}, a_{t+h-1} \mid s_t)$ and $P_{\mathcal{D}}(s_{t+h-1}, a_{t+h-1} \mid s_t)$, we can derive

$$\begin{aligned}
&P_{\mathcal{D}}(s_{t+h} = Z \mid s_t = X_0) = 0,\\
&P^{\circ}_{\mathcal{D}}(s_{t+h} = Z \mid s_t = X_0) = 2(1-\delta)\delta = \varepsilon_h.
\end{aligned} \tag{68}$$

From the calculation above, it is clear that

$$D_{\mathrm{TV}}(P_{\mathcal{D}}^{\circ}(s_{t+h} \mid s_t) \parallel P_{\mathcal{D}}(s_{t+h} \mid s_t)) = \varepsilon_h. \tag{69}$$

Up to now, we have checked that $\mathcal{D}$ is ε_h-open-loop consistent. Now, we are left with analyzing $\hat{V}_{\mathrm{ac}}$ and V_{ac}. With some calculations, we can obtain the following:

$$\begin{aligned}\mathbb{E}_{P_{\mathcal{D}}^{\circ}}\left[R_{t:t+h}\right] &= 1 + \frac{(1-\varepsilon_h)(\gamma - \gamma^h)}{1-\gamma},\\ \hat{V}_{\mathrm{ac}}(X_0) &= \frac{1}{1-\gamma},\\ V_{\mathrm{ac}}(Z) &= 0.\end{aligned} \tag{70}$$

Now, we are ready to compute $V_{\mathrm{ac}}(X_0)$:

$$\begin{aligned}V_{\mathrm{ac}}(X_0) &= \frac{(1-\gamma^h) - \varepsilon_h(\gamma - \gamma^h)}{(1-\gamma)} + \gamma^h\left[(1-\varepsilon_h)V_{\mathrm{ac}}(X_0) + \varepsilon_h V_{\mathrm{ac}}(Z)\right]\\ &= \frac{1-\gamma^h - \varepsilon_h(\gamma-\gamma^h)}{(1-\gamma)(1-\gamma^h(1-\varepsilon_h))}\end{aligned} \tag{71}$$

Finally, with $X_0 \in \mathrm{supp}(\mathcal{D})$, we obtain the desired value difference

$$\hat{V}_{\mathrm{ac}}(X_0) - V_{\mathrm{ac}}(X_0) = \frac{\varepsilon_h \gamma}{(1-\gamma)(1-\gamma^h(1-\varepsilon_h))}. \tag{72}$$

By symmetry, we can flip the reward value (*i.e.*, $0 \to 1$ and $1 \to 0$) to construct the example such that

$$V_{\mathrm{ac}}(X_0) - \hat{V}_{\mathrm{ac}}(X_0) = \frac{\varepsilon_h \gamma}{(1-\gamma)(1-\gamma^h(1-\varepsilon_h))}. \tag{73}$$

□

G.4 Proof of Corollary 4.5

Corollary 4.5 (Optimal Action Chunking Policy)**.** *Let $\pi^\star : \mathcal{S} \to \Delta_{\mathcal{A}}$ be an optimal policy in $\mathcal{M}$ and $\mathcal{D}^\star$ be the data collected by $\pi^\star$. If $\mathcal{D}^\star$ is ε_h-open-loop consistent, then under* $\mathrm{supp}(\mathcal{D}^\star)$*,*

$$\|V^\star_{\mathrm{ac}} - V^\star\|_\infty \le \left\|\tilde{V}_{\mathrm{ac}} - V^\star\right\|_\infty \le \frac{\varepsilon_h \gamma}{(1-(1-\varepsilon_h)\gamma^h)(1-\gamma)} \le \frac{\varepsilon_h}{(1-\gamma^h)(1-\gamma)}, \tag{14}$$

where $V^\star$ is the value of the optimal policy $\pi^\star$, $V^\star_{\mathrm{ac}}$ is the true value of the optimal action chunking policy, and $\tilde{V}_{\mathrm{ac}}$ is the true value of the action chunking policy from cloning the data $\mathcal{D}^\star$:

$$\tilde{\pi}_{\mathrm{ac}}(a_{t:t+h} \mid s_t) : s_t \mapsto P_{\mathcal{D}^\star}(\cdot \mid s_t). \tag{15}$$

Proof. Let $\hat{V}_{\mathrm{ac}}$ be the fixed point of the following equation:

$$\hat{V}_{\mathrm{ac}}(s_t) = \mathbb{E}_{s_{t+1:t+h+1}, a_{t:t+h} \sim P_{\mathcal{D}^\star}(\cdot|s_t)} \left[R_{t:t+h} + \gamma^h \hat{V}_{\mathrm{ac}}(s_{t+h})\right] \tag{74}$$

where again $R_{t:t+h} = \sum_{t'=t}^{t+h} \gamma^{t'-t} r(s_{t'}, a_{t'})$. The value of the optimal policy is the fixed point of the following equation:

$$\begin{aligned} V^\star(s_t) &= \mathbb{E}_{s_{t+1}, a_t \sim P_{\mathcal{D}^\star}(\cdot|s_t)} \left[r(s_t, a_t) + \gamma V^\star(s_{t+1})\right] \\ &= \mathbb{E}_{s_{t:t+2}, a_{t:t+1} \sim P_{\mathcal{D}^\star}(\cdot|s_t)} \left[r(s_t, a_t) + \gamma r(s_{t+1}, a_{t+1}) + \gamma V^\star(s_{t+2})\right] \\ &\dots \\ &= \mathbb{E}_{s_{t+1:t+h+1}, a_{t:t+h} \sim P_{\mathcal{D}^\star}(\cdot|s_t)} \left[R_{t:t+h} + \gamma^h V^\star(s_{t+h})\right] \end{aligned} \tag{75}$$

which is equivalent to fixed-point equation for $\hat{V}_{\mathrm{ac}}$. Therefore $\hat{V}_{\mathrm{ac}} = V^\star$. By Theorem 4.4, we know that the true value V_{ac} of the action chunking policy $\tilde{\pi}_{\mathrm{ac}}$ that clones $\mathcal{D}^\star$ is close to $\hat{V}_{\mathrm{ac}}$. More specifically, for all $s_t \in \mathrm{supp}(\mathcal{D}^\star)$,

$$\left|\hat{V}_{\mathrm{ac}}(s_t) - \tilde{V}_{\mathrm{ac}}(s_t)\right| \le \frac{\gamma \varepsilon_h}{(1-\gamma)(1-(1-\varepsilon_h)\gamma^h)}, \tag{76}$$

which means that

$$V^\star(s_t) - \tilde{V}_{\mathrm{ac}}(s_t) \le \frac{\gamma \varepsilon_h}{(1-\gamma)(1-(1-\varepsilon_h)\gamma^h)}, \tag{77}$$

where we can remove the absolute value operator because $V^\star(s_t)$ is by definition always at least as large as $\tilde{V}_{\mathrm{ac}}(s_t)$. Since the optimal action chunking policy, by definition, attains equally good or better values (over $\mathcal{S}$) represented by V_{ac}, and the optimal policy $\pi^\star$ also attains equally good or better value (*i.e.*, $V^\star$) compared to that of the optimal action chunking policy $\pi^\star_{\mathrm{ac}}$ (*i.e.*, $V^\star_{\mathrm{ac}}$), the following inequality holds for all $s_t \in \mathrm{supp}(\mathcal{D}^\star)$:

$$V^\star(s_t) \ge V^\star_{\mathrm{ac}}(s_t) \ge \tilde{V}_{\mathrm{ac}}(s_t). \tag{78}$$

Therefore,

$$V^\star_{\mathrm{ac}}(s_t) - V^\star(s_t) \le \tilde{V}_{\mathrm{ac}}(s_t) - V^\star(s_t) \le \frac{\gamma \varepsilon_h}{(1-\gamma)(1-(1-\varepsilon_h)\gamma^h)}, \tag{79}$$

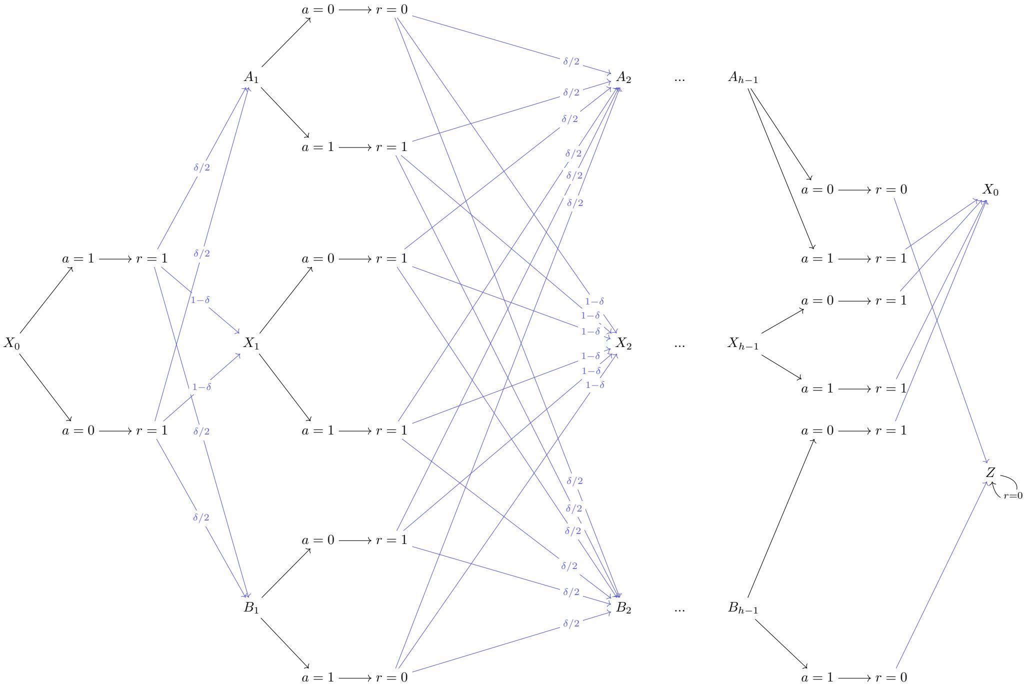
as desired. □

G.5 Proof of Corollary F.2

Corollary F.2 (Worse-case Optimality Gap for Action Chunking Policy)**.** *For any $\gamma \in [0, 1), \varepsilon_h \in [0, 1/2]$, there exists an MDP $\mathcal{M}$ whose optimal policy $\pi^\star$ induces a data distribution $\mathcal{D}^\star$ that is weakly ε_h-open-loop consistent, such that for some $s \in \operatorname{supp}(P_{\mathcal{D}^\star}(s_t))$,*

$$V^\star(s) - V^\star_{\mathrm{ac}}(s) = \frac{\gamma \varepsilon_h}{(1-\gamma)(1-(1-\varepsilon_h)\gamma^h)}. \tag{35}$$

Proof. To show this, we need a slightly more complicated MDP (compared to the $2h$-state MDP we use in the proof Appendix G.3). The MDP we construct for this proof is a $(3h-1)$-state MDP as illustrated in Figure 8.



X_0
$a=1 \longrightarrow r=1$
$a=0 \longrightarrow r=1$
$\delta/2$
$1-\delta$
A_1
X_1
B_1
$a=0 \longrightarrow r=0$
$a=1 \longrightarrow r=1$
$a=0 \longrightarrow r=1$
$a=1 \longrightarrow r=1$
$a=0 \longrightarrow r=1$
$a=1 \longrightarrow r=0$
A_2
X_2
B_2
...
A_{h-1}
X_{h-1}
B_{h-1}
$a=0 \longrightarrow r=0$
$a=1 \longrightarrow r=1$
$a=0 \longrightarrow r=1$
$a=1 \longrightarrow r=1$
$a=0 \longrightarrow r=1$
$a=1 \longrightarrow r=0$
X_0
Z
$r=0$

Figure 8: **A $(3h-1)$-state MDP that is constructed to meet the upper-bound in Corollary 4.5.**

The optimal policy we pick is described as the following:

$$\begin{aligned}
\pi^\star(a = 0 \mid X_i) &= 1/2 \\
\pi^\star(a = 1 \mid X_i) &= 1/2 \\
\pi^\star(a = 1 \mid A_i) &= 1 \\
\pi^\star(a = 0 \mid B_i) &= 1/2
\end{aligned} \tag{80}$$

This induces the following state distribution,

$$\begin{aligned}
&P_{\mathcal{D}^\star}(s_{t+i} = A_i \mid s_t = X_0) = P_{\mathcal{D}^\star}(s_{t+i} = B_i \mid s_t = X_0) \\
&\quad = P^\circ_{\mathcal{D}^\star}(s_{t+i} = A_i \mid s_t = X_0) = P^\circ_{\mathcal{D}^\star}(s_{t+i} = B_i \mid s_t = X_0) = \delta/2, \\
&P_{\mathcal{D}^\star}(s_{t+i} = X_i \mid s_t = X_0) = P^\circ_{\mathcal{D}^\star}(s_{t+i} = X_i \mid s_t = X_0) = 1 - \delta,
\end{aligned} \tag{81}$$

and a fully factorized distribution for the action chunk,

$$P^\circ_{\mathcal{D}^\star}(a_{t+i} = 0 \mid s_t) = P^\circ_{\mathcal{D}^\star}(a_{t+i} = 0 \mid s_t, a_{t:t+i}) = \frac{1}{2}(\delta_{a=0} + \delta_{a=1}). \tag{82}$$

Now, we derive the condition on δ when the optimal data $\mathcal{D}^\star$ is ε_h-open-loop consistent. We start by calculating the TV distance discrepancy for the future state-action distribution:

$$\begin{aligned}&D_{\mathrm{TV}}(P^{\mathrm{open}}_{\mathcal{D}^\star}(s_{t+i}, a_{t+i} \mid s_t) \parallel P_{\mathcal{D}^\star}(s_{t+i}, a_{t+i} \mid s_t))\\ &= \frac{1}{2}\left\|\begin{bmatrix} 0 & \delta/2 \\ (1-\delta)/2 & (1-\delta)/2 \\ \delta/2 & 0 \end{bmatrix} - \begin{bmatrix} \delta/4 & \delta/4 \\ (1-\delta)/2 & (1-\delta)/2 \\ \delta/4 & \delta/4 \end{bmatrix}\right\|_{1,1} \\ &= \delta/2.\end{aligned} \tag{83}$$

In the second line of the equations above, each row in the matrix corresponds to a distinct action $a_{t+i} \in \{0, 1\}$ and each row in the matrix corresponds to a distinct state $s_{t+i} \in \{A_i, X_i, B_i\}$.

Next, we calculate the TV distance discrepancy for s_{t+h}:

$$\begin{aligned}&D_{\mathrm{TV}}(P^{\mathrm{open}}_{\mathcal{D}^\star}(s_{t+h} \mid s_t) \parallel P_{\mathcal{D}^\star}(s_{t+h} \mid s_t))\\ &= \frac{1}{2}\left\|[1 \quad 0] - [1-\delta/2 \quad \delta/2]\right\|_1 \\ &= \delta/2.\end{aligned} \tag{84}$$

In the second line of the equations above, each element in the vector corresponds to a distinct state $s_{t+h} \in \{X_0, Z\}$. Up to now, we have concluded that $\mathcal{D}^\star$ is $(\delta/2)$-open-loop consistent.

Due to the symmetric structure of this MDP, it is clear that any action chunking policy $\pi_{\mathrm{ac}}(X_0) = a_{t:t+h}$ with $a_{t:t+h} \in \{0, 1\}$ is optimal and achieve the following value:

$$\begin{aligned}V^\star_{\mathrm{ac}}(X_0) &= 1 + (1-\delta/2)\left[\frac{\gamma - \gamma^h}{1-\gamma} + \gamma^h V^\star_{\mathrm{ac}}(X_0)\right] \\ &= \frac{(1-\gamma) + (1-\delta/2)(\gamma - \gamma^h)}{(1-\gamma)(1-(1-\delta/2)\gamma^h)}.\end{aligned} \tag{85}$$

The optimal closed-loop policy can achieve the maximum possible return

$$V^\star(X_0) = \frac{1}{1-\gamma}. \tag{86}$$

Therefore, with $\varepsilon_h = \delta/2$, the optimality gap achieved by this $(3h-1)$-state MDP is

$$V^\star(X_0) - V^\star_{\mathrm{ac}}(X_0) = \frac{\varepsilon_h \gamma}{(1-\gamma)(1-(1-\varepsilon_h)\gamma^h)}, \tag{87}$$

as desired.

□

G.6 Proof of Theorem 4.6

Theorem 4.6 (Q-Learning with Action Chunking Policy on Off-policy Data)**.** *If $\mathcal{D}$ is strongly ε_h-open-loop consistent and $\operatorname{supp}(\mathcal{D}) \supseteq \operatorname{supp}(\mathcal{D}^\star)$, with $\mathcal{D}^\star$ being the data distribution of an arbitrary optimal policy $\pi^\star$ under $\mathcal{M}$), then the following bound holds under $\operatorname{supp}(\mathcal{D}^\star)$:*

$$\|V^+_{\text{ac}} - V^\star\|_\infty \le \frac{\varepsilon_h \gamma}{1-\gamma}\left[\frac{2}{1-(1-2\varepsilon_h)\gamma^h} + \frac{1}{1-(1-\varepsilon_h)\gamma^h}\right] \le \frac{3\varepsilon_h}{(1-\gamma)(1-\gamma^h)}. \tag{18}$$

where $V^\star$ is the value of an optimal policy under $\mathcal{M}$.

Proof of Theorem 4.6. We start by constructing a bound between $\hat{Q}^+_{\text{ac}}$ and $Q^\star_{\text{ac}}$, the solution of the following bellman equation:

$$Q^\star_{\text{ac}}(s_t, a_{t:t+h}) = \mathbb{E}_{s_{t+1:t+h+1} \sim P^\circ_{\mathcal{D}}(\cdot \mid s_t, a_{t:t+h})}\left[R_{t:t+h} + \gamma^h \max_{a_{t+h:t+2h}} Q^\star_{\text{ac}}(s_{t+h}, a_{t+h:t+2h})\right]. \tag{88}$$

Intuitively, $Q^\star_{\text{ac}}$ is the Q-function of the optimal action chunking policy $\pi^\star_{\text{ac}}$ that can be learned from $\mathcal{D}$. Because $\operatorname{supp}(\mathcal{D}) \supseteq \operatorname{supp}(\mathcal{D}^\star)$, $\pi^\star_{\text{ac}}$ is at least as good as $\tilde{\pi}_{\text{ac}}$, the action chunking policy obtained by behavior cloning $\mathcal{D}^\star$. Bounding the difference between $\hat{Q}^+_{\text{ac}}$ and $Q^\star_{\text{ac}}$ allows us to leverage the bound in Corollary 4.5 to form a bound between $\hat{V}^+_{\text{ac}}$ and $V^\star$.

Since $\mathcal{D}$ is strongly ε_h-open-loop consistent,

$$D_{\text{TV}}(T(s_{t+h'} \mid s_t, a_{t:t+h'}) \,\|\, P_{\mathcal{D}}(s_{t+h'} \mid s_t, a_{t:t+h})) \le \varepsilon_h, \forall h' \in \{1, \cdots, h-1\}. \tag{89}$$

Since $\mathcal{D}^\star$ is also strongly ε_h-open-loop consistent,

$$D_{\text{TV}}(T(s_{t+h'} \mid s_t, a_{t:t+h'}) \,\|\, P_{\mathcal{D}^\star}(s_{t+h'} \mid s_t, a_{t:t+h})) \le \varepsilon_h, \forall h' \in \{1, \cdots, h-1\}. \tag{90}$$

Using the transitive property of TV distance, we have

$$D_{\text{TV}}(P_{\mathcal{D}}(s_{t+h'} \mid s_t, a_{t:t+h}) \,\|\, P_{\mathcal{D}^\star}(s_{t+h'} \mid s_t, a_{t:t+h})) \le 2\varepsilon_h, \forall h' \in \{1, \cdots, h-1\}. \tag{91}$$

Now, for the h-step reward, we have

$$\begin{aligned}
&\left|\mathbb{E}_{P_{\mathcal{D}}(\cdot \mid s_t, a_{t:t+h})}\left[R_{t:t+h}\right] - \mathbb{E}_{P_{\mathcal{D}^\star}(\cdot \mid s_t, a_{t:t+h})}\left[R_{t:t+h}\right]\right| \\
&\le \sum_{h'=1}^{h-1}\left[\gamma^{h'} D_{\text{TV}}(P_{\mathcal{D}}(s_{t+h'} \mid s_t, a_{t:t+h}) \,\|\, P_{\mathcal{D}^\star}(s_{t+h'} \mid s_t, a_{t:t+h}))\right] \\
&\le \frac{2(\gamma - \gamma^h)\varepsilon_h}{1-\gamma}.
\end{aligned} \tag{92}$$

Similarly, for the value h-step into the future, we can use Lemma G.2 to obtain the following bound:

$$\begin{aligned}
&\left|\mathbb{E}_{s_{t+h} \sim P_{\mathcal{D}}(s_{t+h} \mid s_t)}\left[V^\star(s_{t+h})\right] - \mathbb{E}_{s_{t+h} \sim P_{\mathcal{D}^\star}(s_{t+h} \mid s_t)}\left[\hat{V}^+_{\text{ac}}(s_{t+h})\right]\right| \\
&\le 2\varepsilon_h + (1 - 2\varepsilon_h) \sup_{s_{t+h} \in \mathcal{D}^\star}\left|V^\star(s_{t+h}) - \hat{V}^+_{\text{ac}}(s_{t+h})\right|.
\end{aligned} \tag{93}$$

We define $Q^\star(s_t, a_{t:t+h})$ to be

$$Q^\star(s_t, a_{t:t+h}) := \mathbb{E}_{P_{\mathcal{D}^\star}(\cdot \mid s_t, a_{t:t+h})}\left[R_{t:t+h} + \gamma^h V^\star(s_{t+h})\right]. \tag{94}$$

It is clear that

$$V^\star(s_t) = \mathbb{E}_{a_{t:t+h} \sim P_{\mathcal{D}^\star}}\left[Q^\star(s_t, a_{t:t+h})\right]. \tag{95}$$

Combining the bound for the h-step reward and the bound on the value for s_{t+h}, for all $s_t, a_{t:t+h} \in \mathrm{supp}(P_{\mathcal{D}^\star}(s_t, a_{t:t+h}))$,

$$\begin{aligned}
&\Delta(s_t, a_{t:t+h}) = Q^\star(s_t, a_{t:t+h}) - \hat{Q}^+_{\text{ac}}(s_t, a_{t:t+h}) \\
&\leq 2\varepsilon_h \gamma^h + \frac{2(\gamma - \gamma^h)\varepsilon_h}{1-\gamma} + (1 - 2\varepsilon_h)\gamma^h \left(V^\star(s_{t+h}) - \hat{V}^+_{\text{ac}}(s_{t+h}) \right) \\
&\leq \frac{2\varepsilon_h \gamma}{1-\gamma} + (1 - 2\varepsilon_h)\gamma^h \left(\mathbb{E}_{P_{\mathcal{D}^\star}} \left[Q^\star(s_{t+h}, a_{t+h:t+2h}) \right] - \sup_{a_{t+h:t+2h}} \hat{Q}^+_{\text{ac}}(s_{t+h}, a_{t+h:t+2h}) \right) \\
&\leq \frac{2\varepsilon_h \gamma}{1-\gamma} + (1 - 2\varepsilon_h)\gamma^h \left(\mathbb{E}_{P_{\mathcal{D}^\star}} \left[\hat{Q}^+_{\text{ac}}(s_{t+h}, a_{t+h:t+2h}) + \Delta(s_{t+h}, a_{t+h:t+2h}) \right] - \sup_{a_{t+h:t+2h}} \hat{Q}^+_{\text{ac}}(s_{t+h}, a_{t+h:t+2h}) \right) \\
&\leq \frac{2\varepsilon_h \gamma}{1-\gamma} + (1 - 2\varepsilon_h)\gamma^h \sup_{s_{t+h}, a_{t+h:t+2h}} \left[\Delta(s_{t+h}, a_{t+h:t+2h}) \right],
\end{aligned} \tag{96}$$

which can be recursively expanded to obtain

$$V^\star(s_t) - \hat{V}^+_{\text{ac}}(s_t) \leq \frac{2\varepsilon_h \gamma}{(1-\gamma)(1 - (1 - 2\varepsilon_h)\gamma^h)}. \tag{97}$$

By Theorem 4.4, for all $s_t \in \mathrm{supp}(\mathcal{D})$,

$$\left| \hat{V}^+_{\text{ac}}(s_t) - V^+_{\text{ac}}(s_t) \right| \leq \frac{\varepsilon_h \gamma}{(1-\gamma)(1 - (1 - \varepsilon_h)\gamma^h)}. \tag{98}$$

Combining the two inequalities above, for all $s_t \in \mathrm{supp}(\mathcal{D}^\star)$,

$$V^\star(s_t) - V^+_{\text{ac}}(s_t) \leq \frac{\varepsilon_h \gamma}{1-\gamma} \left[\frac{2}{1 - (1 - 2\varepsilon_h)\gamma^h} + \frac{1}{1 - (1 - \varepsilon_h)\gamma^h} \right]. \tag{99}$$

□

G.7 Proof of Theorem F.3

Theorem F.3 (Worst-case Analysis of Q-Learning with Action Chunking Policy on Off-policy Data)**.** *For any $\varepsilon_h \in (0, 1/5)$, $\gamma \in (0, 1)$, $c_1 \in (0, \varepsilon_h/2)$, and $c_2 \in (0, 2\varepsilon_h\gamma)$, there exists an MDP $\mathcal{M}$ and strongly ε_h-open-loop consistent data distribution $\mathcal{D}$ and $\mathcal{D}^\star$ with $\mathrm{supp}(P_{\mathcal{D}}(s_t, a_{t:t+h})) \supseteq \mathrm{supp}(P_{\mathcal{D}^\star}(s_t, a_{t:t+h}))$, such that for some $s \in \mathrm{supp}(P_{\mathcal{D}^\star}(s_t))$,*

$$V^\star(s) - V^+_{\mathrm{ac}}(s) = \frac{2\varepsilon_h\gamma - c_2}{(1-\gamma)(1-(1-2\varepsilon_h)\gamma^h)} + \frac{\varepsilon_h\gamma}{(1-\gamma)(1-(1-\varepsilon_h-c_1)\gamma^h)}, \tag{36}$$

where $V^\star$ is the value of an optimal policy and V^+_{ac} is the true *value of π^+_{ac}. As $c_1, c_2 \to 0$,*

$$V^\star(s) - V^+_{\mathrm{ac}}(s) \to \frac{\varepsilon_h\gamma}{1-\gamma}\left[\frac{2}{1-(1-2\varepsilon_h)\gamma^h} + \frac{1}{1-(1-\varepsilon_h)\gamma^h}\right]. \tag{37}$$

The examples in the following proof of Theorem F.3 (available in Appendix G.7) provide insights on the factor of 3 in $V^\star - V^+_{\mathrm{ac}} \leq 3\varepsilon_h H\bar{H}$ (with $H = 1/(1-\gamma)$, $\bar{H} = 1/(1-\gamma^h)$) is necessary. In particular, the worse case can be roughly seen as a combination of the two main results that we have presented so far:

1. $V^\star - V^\star_{\mathrm{ac}} \approx \varepsilon_h H\bar{H}$ (Corollary 4.5, Corollary F.2): the optimal action chunking policy is $(\varepsilon_h H^2)$-suboptimal due to its inability to react to environment stochasticity, quantified by the strongly-ε_h open-loop consistency of $\mathcal{D}^\star$.
2. $V^\star_{\mathrm{ac}} - \hat{V}^+_{\mathrm{ac}} \approx \varepsilon_h H\bar{H}$ (a transformation of Theorem 4.4 and Theorem F.1 on the optimal action chunking policy $\pi^\star_{\mathrm{ac}}$): the value *under-estimation* bias can incur another factor of $\varepsilon_h H\bar{H}$ bringing up the sub-optimality of $\hat{V}^+_{\mathrm{ac}}$ to at most $2\varepsilon_h H\bar{H}$, and finally,
3. $\hat{V}^+_{\mathrm{ac}} - V^+_{\mathrm{ac}} \approx \varepsilon_h H\bar{H}$ (Theorem 4.4, Theorem F.1): the action chunking value function may prefer an *overestimated* action chunking policy π^+_{ac} where its actual value is again $\varepsilon_h H\bar{H}$ from its estimated value, resulting in a total sub-optimality of $3\varepsilon_h H\bar{H}$.

Our construction (in the proof of Theorem F.3) directly builds on the above insights by using a 2-part MDP where the first part corresponds to an $(\varepsilon_h H\bar{H})$-underestimated action chunking policy that has a $(\varepsilon_h H\bar{H})$-optimality gap from the optimal closed-loop policy and the second part corresponds to an $(\varepsilon_h H\bar{H})$-overestimated action chunking policy that has a $(3\varepsilon_h H\bar{H})$-optimality gap that is preferred by the value function.

Before we start our main proof, we first introduce a Lemma that helps simplifies the inequalities.

Lemma G.5 (Optimality gap comparator)**.** *For any $\tilde{\gamma} \in [0, 1)$ and $0 < \varepsilon_1 < \varepsilon_2 < 1$,*

$$\frac{\varepsilon_1}{1-(1-\varepsilon_1)\tilde{\gamma}} < \frac{\varepsilon_2}{1-(1-\varepsilon_2)\tilde{\gamma}}. \tag{100}$$

Proof.

$$\begin{aligned} 0 &< (1-\gamma)(\varepsilon_2 - \varepsilon_1) \\ &= \varepsilon_2 - \varepsilon_2\tilde{\gamma} - \varepsilon_1 + \varepsilon_1\tilde{\gamma} \\ &= \varepsilon_2 - \varepsilon_2\tilde{\gamma} + \varepsilon_1\varepsilon_2\tilde{\gamma} - \varepsilon_1 + \varepsilon_1\tilde{\gamma} - \varepsilon_1\varepsilon_2\tilde{\gamma} \\ &= \varepsilon_2(1-(1-\varepsilon_1)\tilde{\gamma}) - \varepsilon_1(1-(1-\varepsilon_2)\tilde{\gamma}) \end{aligned} \tag{101}$$

Since $1-(1-\varepsilon_1)\tilde{\gamma} > 0$ and $1-(1-\varepsilon_2)\tilde{\gamma} > 0$, we can divide both sides by $(1-(1-\varepsilon_1)\tilde{\gamma})(1-(1-\varepsilon_2)\tilde{\gamma})$ to get

$$0 < \frac{\varepsilon_2}{1-(1-\varepsilon_2)\tilde{\gamma}} - \frac{\varepsilon_1}{1-(1-\varepsilon_1)\tilde{\gamma}}, \tag{102}$$

as desired.

□

Now, we begin the main proof as follows.

Proof of Theorem F.3. We prove by constructing the following $(2h+4)$-state MDP where the agent can take any of the three actions $\{0, 1, 2\}$ at each state (see a diagram in Figure 9).

Notations: we start by introducing some abbreviations for all action chunks that appear in this proof:

$$\begin{aligned}
a^{\star}_{t:t+h} &= (0, 0, 0, \cdots, 0) \\
a^{\diamond}_{t:t+h} &= (0, 1, 0, \cdots, 0) \\
a^{\bullet}_{t:t+h} &= (0, 2, 0, \cdots, 0) \\
a^{\triangle}_{t:t+h} &= (1, 1, 1, \cdots, 1) \\
a^{\circ}_{t:t+h} &= (1, 0, 1, \cdots, 1) \\
a^{\times}_{t:t+h} &= (1, 2, 1, \cdots, 1)
\end{aligned} \tag{103}$$

The first three action chunks $a^{\star}_{t:t+h}, a^{\diamond}_{t:t+h}, a^{\bullet}_{t:t+h}$ are only possible in the top branch and the last three action chunks $a^{\triangle}_{t:t+h}, a^{\circ}_{t:t+h}, a^{\times}_{t:t+h}$ are only possible in the bottom branch because the first action in the action chunk deterministically divides it into the two branches.

Among these action chunks, it is clear by inspection that $\pi_{\text{ac}}(X_0) = (0, 0, \cdots, 0)$ is the optimal action chunking policy, and thus we directly use '$\star$' to denote $a^{\star}_{t:t+h} = (0, 0, \cdots, 0)$. $a^{\triangle}_{t:t+h}$ is also of great importance: as we will show later, $\pi^{+}_{\text{ac}}(X_0) = a^{\triangle}_{t:t+h}$. The *actual* values and *nominal/estimated* values for these action chunks are $(V^{\star}_{\text{ac}}, V^{\diamond}_{\text{ac}}, V^{\bullet}_{\text{ac}}, V^{\triangle}_{\text{ac}}, V^{\circ}_{\text{ac}}, V^{\times}_{\text{ac}})$ and $(\hat{V}^{\star}_{\text{ac}}, \hat{V}^{\diamond}_{\text{ac}}, \hat{V}^{\bullet}_{\text{ac}}, \hat{V}^{\triangle}_{\text{ac}}, \hat{V}^{\circ}_{\text{ac}}, \hat{V}^{\times}_{\text{ac}})$ respectively. Much of the focus of this proof is to calculate the optimality gap, which is the difference between the optimal closed-loop value and the action chunking policy value (either estimated or actual):

$$\texttt{actual optimality gap:} \quad V^{\star}(X_0) - V^{[\cdot]}_{\text{ac}}(X_0) \tag{104}$$

$$\texttt{nominal optimality gap:} \quad V^{\star}(X_0) - \hat{V}^{[\cdot]}_{\text{ac}}(X_0) \tag{105}$$

High-level proof sketch: The MDP contains two branches: a top branch where (as we will show) both the optimal policy $\pi^{\star}$ and the optimal action chunking policy $\pi^{\star}_{\text{ac}}$ take, and a bottom branch where (as we will also show) the learned action chunking policy π^{+}_{ac} takes. The key idea of the construction is that for the top branch, we have

$$V^{\star}(X_0) - \hat{V}^{\star}_{\text{ac}}(X_0) \approx \frac{2\varepsilon_h \gamma}{(1-\gamma)(1-(1-2\varepsilon_h)\gamma^h)}, \tag{106}$$

and for the bottom branch, we have

$$\hat{V}^{\star}_{\text{ac}}(X_0) < \hat{V}^{+}_{\text{ac}}(X_0) \approx V^{+}_{\text{ac}}(X_0) + \frac{\varepsilon_h \gamma}{(1-\gamma)(1-(1-\varepsilon_h)\gamma^h)}. \tag{107}$$

Combining these two together gives

$$V^{\star}(X_0) - V^{+}_{\text{ac}}(X_0) \approx \frac{\varepsilon_h \gamma}{1-\gamma}\left[\frac{2}{1-(1-2\varepsilon_h)\gamma^h} + \frac{1}{1-(1-\varepsilon_h)\gamma^h}\right]. \tag{108}$$

We use '$\approx$' because the equalities are not strictly achievable but (as we will show) can be made arbitrarily close.

The proof can be roughly divided into the following steps (we use '$\approx$' to help illustrate the high-level idea below and use more precise argument in the actual proof):

1. `MDP description`: we formally describe the transition dynamics T and the reward function r for each state-action pair for both the top and the bottom branches.

2. `Strong` ε_h`-open-loop consistency of` $\mathcal{D}^{\star}$: we then check the strong open-loop consistency assumption for $\mathcal{D}^{\star}$.

3. `Data distribution` $\mathcal{D}_{\text{top}}$ `for the top branch`: we use a mixture data distribution from two policies to construct $\mathcal{D}_{\text{top}}$.

4. Strong ε_h-open-loop consistency of $\mathcal{D}_{\text{top}}$: we then check that the constructed data distribution of the top branch satisfies the strongly open-loop consistency assumption. Note that we can do so separately for the top and the bottom because these two distributions have non-overlapping support in $a_{t:t+h}$.

5. The optimality gap and value estimation error for the top branch: we prove that $V^\star(X_0) - V^\star_{\text{ac}}(X_0) = \frac{\varepsilon_h \gamma}{(1-\gamma)(1-(1-2\varepsilon_h)\gamma^h)}$ and $V^\star(X_0) - \hat{V}^\star_{\text{ac}}(X_0) = \frac{2\varepsilon_h \gamma}{(1-\gamma)(1-(1-2\varepsilon_h)\gamma^h)}$ and the other two possible action chunks $a^\diamond_{t:t+h} = (0, 1, 0, \cdots)$ and $a^\bullet_{t:t+h} = (0, 2, 0, \cdots)$ both admit lower estimated values compared to $a^\star_{t:t+h}$: $\hat{V}^\diamond_{\text{ac}}(X_0) < \hat{V}^\star_{\text{ac}}(X_0)$ and $\hat{V}^\bullet_{\text{ac}}(X_0) < \hat{V}^\star_{\text{ac}}(X_0)$.

6. Data distribution $\mathcal{D}_{\text{bottom}}$ for the bottom branch: we again use a mixture data distribution from two different policies to construct $\mathcal{D}_{\text{bottom}}$.

7. Strong ε_h-open-loop consistency of $\mathcal{D}_{\text{bottom}}$: we then check that the constructed data distribution of the bottom branch satisfies the strongly open-loop consistency assumption.

8. The optimality gap and value estimation error for the bottom branch: we prove that $V^\star(X_0) - \hat{V}^\triangle_{\text{ac}}(X_0) \approx \frac{2\varepsilon_h \gamma}{(1-\gamma)(1-(1-2\varepsilon_h)\gamma^h)}$ and $\hat{V}^\triangle_{\text{ac}}(X_0) - V^\triangle_{\text{ac}}(X_0) = \frac{\varepsilon_h \gamma}{(1-\gamma)(1-(1-\varepsilon_h)\gamma^h)}$, and the other two possible action chunks $a^\circ_{t:t+h} = (1, 0, 0, \cdots)$ and $a^\times_{t:t+h} = (1, 2, 0, \cdots)$ both admit lower estimated values compared to $a^\triangle_{t:t+h}$: $\hat{V}^\circ_{\text{ac}}(X_0) < \hat{V}^\triangle_{\text{ac}}(X_0)$ and $\hat{V}^\times_{\text{ac}}(X_0) < \hat{V}^\triangle_{\text{ac}}(X_0)$. Moreover $a^\star_{t:t+h}$ also admits a lower estimated value compared to $a^\triangle_{t:t+h}$: $\hat{V}^\star_{\text{ac}}(X_0) < \hat{V}^\triangle_{\text{ac}}(X_0)$ which proves $\pi^+_{\text{ac}}(X_0) = a^\triangle_{t:t+h}$ and thus concluding our proof: $V^\star(X_0) - V^+_{\text{ac}}(X_0) \approx \frac{2\varepsilon_h \gamma}{(1-\gamma)(1-(1-2\varepsilon_h)\gamma^h)} + \frac{\varepsilon_h \gamma}{(1-\gamma)(1-(1-\varepsilon_h)\gamma^h)}$.

Now we begin our proof as follows.

Step 1. MDP description (Figure 9).

The transition function T of the MDP is defined as follows (from left to right):

$$
\begin{aligned}
T(Z \mid Z, a) = T(G \mid G, a) &= 1, \quad \forall a, \\
T(Z \mid s, a = 2) &= 1, \quad \forall a, \forall s : s \neq G \\
T(X_1 \mid X_0, a = 0) &= 1 - 2\varepsilon_h \\
T(\tilde{X}_1 \mid X_0, a = 0) &= \varepsilon_h \\
T(C \mid X_0, a = 0) &= \varepsilon_h \\
T(Y_1 \mid X_0, a = 1) &= 1 - \varepsilon_h - c_1 \\
T(\tilde{Y}_1 \mid X_0, a = 1) &= \varepsilon_h \\
T(G \mid X_0, a = 1) &= c_1 \\
T(X_2 \mid X_1, a = 0) &= 1 \\
T(X_2 \mid \tilde{X}_1, a = 1) &= 1 \\
T(X_2 \mid C, a = 1) &= 1 \\
T(Z \mid X_1, a = 1) &= 1 \\
T(Z \mid C, a = 0) &= 1 \\
T(G \mid \tilde{X}_1, a = 0) &= 1 \\
T(Y_2 \mid Y_1, a = 1) &= 1 \\
T(Y_2 \mid \tilde{Y}_1, a = 0) &= 1 \\
T(Z \mid Y_1, a = 0) &= 1 \\
T(Z \mid \tilde{Y}_1, a = 1) &= 1 \\
T(X_{i+1} \mid X_i, a \in \{0, 1\}) = T(Y_{i+1} \mid Y_i, a \in \{0, 1\}) &= 1, \quad \forall i \in \{2, \cdots, h-2\} \\
T(X_0 \mid X_{h-1}, a \in \{0, 1\}) = T(Y_0 \mid Y_{h-1}, a \in \{0, 1\}) &= 1
\end{aligned}
\tag{109}
$$

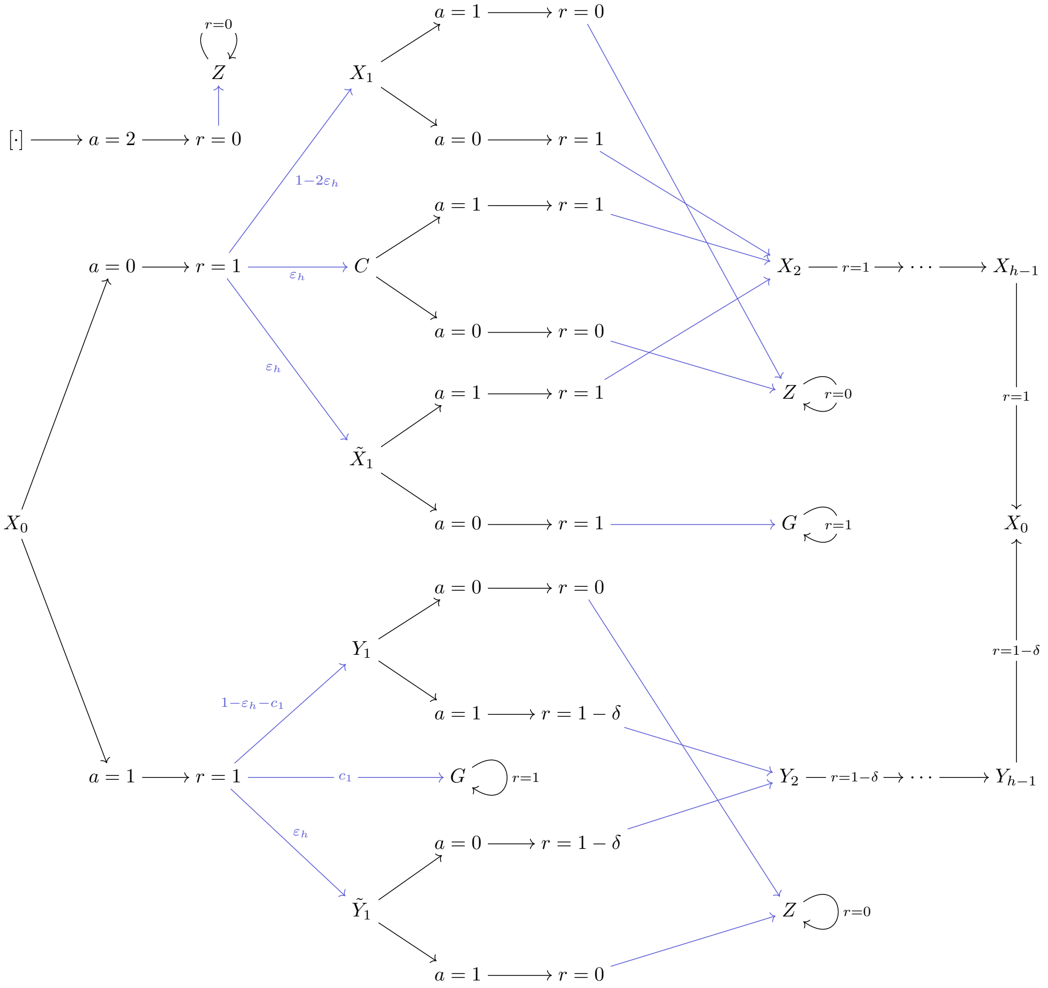


Figure 9: **A** $(2h + 4)$**-state MDP that is constructed to illustrate the MDP constructed to meet the exact upper-bound optimality gap in Theorem 4.6.** We redraw the same states Z, G, X_0 in multiple locations in the diagram above for better clarity.

The reward function is defined as

$$\begin{aligned}
r(Z, a) &= 0, \quad \forall a \\
r(G, a) &= 1, \quad \forall a \\
r(s, a = 2) &= 0, \quad \forall s : s \neq G \\
r(X_0, a = 0) = r(X_0, a = 1) &= 1, \\
r(C, a = 1) = r(X_1, a = 0) = r(\tilde{X}_1, a \in \{0, 1\}) &= 1, \\
r(C, a = 0) = r(X_1, a = 1) &= 0, \\
r(Y_1, a = 1) = r(\tilde{Y}_1, a = 0) &= 1 - \delta, \\
r(Y_1, a = 0) = r(\tilde{Y}_1, a = 1) &= 0, \\
r(X_i, a \in \{0, 1\}) &= 1, \quad \forall i \in \{2, \cdots, h - 1\} \\
r(Y_i, a \in \{0, 1\}) &= 1 - \delta, \quad \forall i \in \{2, \cdots, h - 1\}
\end{aligned} \tag{110}$$

Notably, there are some special states:

- State Z: a self-looping “black hole” state that always gets 0 reward at each time step and thus has a constant value of 0.

- State G: a self-looping "black hole" state that always gets 1 reward at each time step and thus has a constant value of $1/(1-\gamma)$.
- State X_0: the special state that branches out based on the action taken. The agent periodically visit this state every h steps unless it has been trapped in either Z or G. As we proceed in the proof, we will encounter factors in the form of $\frac{1}{1-b\gamma^h}$ in the calculation of the optimality gap. These factors come from the agent looping around and revisiting X_0 with b-probability each cycle.

These two absorbing states are important because their values sit at the boundary of the value range of our value function $V(s) \in [0, 1/(1-\gamma)]$. Shifting the reaching probability from Z to G or the other way around results in the biggest possible difference in the policy value. Our construction hinges on the constructing $\mathcal{D}$ such that

1. $P_{\mathcal{D}}(\cdot \mid s_t, \pi^\star(s_t))$ and $T(\cdot \mid s_t, \pi^\star(s_t))$ differs by only ε_h (in TV distance as required by the strongly open-loop consistency assumption) where precisely ε_h probability mass is moved from reaching state Z to reaching state G. This causes the $\hat{V}^\star_{\text{ac}}$ to precisely underestimates the value of $V^\star_{\text{ac}}$ by $\frac{\varepsilon_h\gamma^h}{(1-\gamma)(1-(1-2\varepsilon_h)\gamma^h)}$. It is worth noting that we cannot make the $2\varepsilon_h$ in the denominator ε_h because $V^\star_{\text{ac}}$ needs to simultaneously maintain a value gap with $V^\star$. If we were to construct an example where $\hat{V}^\star_{\text{ac}}(X_0) - V^\star_{\text{ac}}(X_0) = V^\star_{\text{ac}}$ be $\frac{\varepsilon_h\gamma^h}{(1-\gamma)(1-(1-\varepsilon_h)\gamma^h)}$, it would enforce $V^\star_{\text{ac}}(X_0) = V^\star(X_0)$ because there would be no probability mass left to create the gap between $V^\star$ and $V^\star_{\text{ac}}$. With an extra ε_h in the denominator, we can also make the optimality gap of $V^\star_{\text{ac}}$ precisely $\frac{\varepsilon_h\gamma^h}{(1-\gamma)(1-(1-2\varepsilon_h)\gamma^h)}$, bringing the combined value gap (between $V^\star$ and $\hat{V}^\star_{\text{ac}}$) up to $\frac{2\varepsilon_h\gamma^h}{(1-\gamma)(1-(1-2\varepsilon_h)\gamma^h)}$.
2. $P_{\mathcal{D}}(\cdot \mid s_t, \pi^+(s_t))$ and $T(\cdot \mid s_t, \pi^+(s_t))$ differs by only ε_h (again in TV distance as required by the strongly open-loop consistency assumption) where precisely ε_h probability mass is moved from reaching state G to reaching state Z. This causes the $\hat{V}^+_{\text{ac}}$ to precisely overestimates the value of V^+_{ac} by $\frac{\varepsilon_h\gamma^h}{(1-\gamma)(1-(1-\varepsilon_h)\gamma^h)}$.

We use a special action $a=2$ where upon taking the action the agent immediately transitions to Z and receives a reward of 0 (except in G). As we will see soon, this action is useful for constructing a data distribution with an easily 'controllable' probability of reaching Z for the top branch and an easily 'controllable' probability of reaching G for the bottom branch. Before we start constructing $\mathcal{D}$, we first check the condition that $\mathcal{D}^\star$ is strongly ε_h-open-loop consistent.

`Step 2. Strong` ε_h`-open-loop consistency of` $\mathcal{D}^\star$: It is clear that one possible $\pi^\star$ that achieves $1/(1-\gamma)$ value is

$$\begin{aligned} \pi^\star(X_i) &= 0 \\ \pi^\star(C) &= 1 \\ \pi^\star(\tilde{X}) &= 0 \end{aligned} \tag{111}$$

We can easily check that $\mathcal{D}^\star$ collected by $\pi^\star$ is strongly ε_h-open-loop consistent by observing that the only path that $\pi^\star$ outputs $(0,1,0,0,\cdots)$ has ε_h probability, which causes the state distribution of s_{t+1} to differ by at most ε_h under the TV distance (subject to $a=(0,1,0,0,\cdots)$ or $a=(0,0,0,\cdots)$ conditioning). This concludes that $\mathcal{D}^\star$ generated by $\pi^\star$ above is strongly ε_h-open-loop consistent.

Now, depending on the first action a_t, the MDP can be decomposed into two parts: *the top* $(a=0)$ and *the bottom* $(a=1)$. We construct the data distribution for each branch and analyze the *actual* and *nominal* optimality gap for each branch in the following steps.

`Step 3. Data distribution` $\mathcal{D}_{\text{top}}$ `for the top branch`: we use a mixture of the following two policies to construct a strongly ε_h-open-loop consistent $\mathcal{D}_{\text{top}}$.

Policy π^1_{top}:

$$\begin{aligned} \pi^1_{\text{top}}(X_0) = \pi^1_{\text{top}}(C) = \pi^1_{\text{top}}(Z) = 0, \\ \pi^1_{\text{top}}(X_1) = \pi^1_{\text{top}}(\tilde{X}_1) = 2. \end{aligned} \tag{112}$$

π^1_{top} always take $a = 2$ unless it is in state X_0, C or Z where it always takes $a = 0$. It is clear that this policy only produces two possible action chunks: $a_{t:t+h} = (0, 0, \cdots, 0)$ or $a_{t:t+h} = a^{\bullet}_{t:t+h} := (0, 2, 0, \cdots)$. We note that the $a_{t:t+h}$ policy always leads to state Z:

$$P_{\mathcal{D}_{\pi^1_{\text{top}}}}(s_{t+h} = Z \mid s_t, a_{t:t+h} = 0) = 1. \tag{113}$$

Policy π^2_{top}:

$$\begin{aligned} \pi^2_{\text{top}}(X_0) &= 0, \\ \pi^2_{\text{top}}(\tilde{X}_1) = \pi^2_{\text{top}}(\tilde{C}) &= 1, \\ \pi^2_{\text{top}}(a = 0 \mid X_1) &= 1 - \delta_G, \\ \pi^2_{\text{top}}(a = 1 \mid X_1) &= \delta_G, \end{aligned} \tag{114}$$

with some $\delta_G \in (0, 1)$. π^2_{top} can also only produce two possible action chunks: $a_{t:t+h} = (0, 0, \cdots, 0)$ or $a_{t:t+h} = a^{\diamond}_{t:t+h} := (0, 1, 0, \cdots, 0)$.

The distribution of s_{t+h} conditioned on $a_{t:t+h} = 0$ is

$$\begin{aligned} P_{\mathcal{D}_{\pi^2_{\text{top}}}}(s_{t+h} = Z \mid s_t, a_{t:t+h} = 0) &= 0, \\ P_{\mathcal{D}_{\pi^2_{\text{top}}}}(s_{t+h} = G \mid s_t, a_{t:t+h} = 0) &= 0, \\ P_{\mathcal{D}_{\pi^2_{\text{top}}}}(s_{t+h} = X_0 \mid s_t, a_{t:t+h} = 0) &= 1. \end{aligned} \tag{115}$$

Mixing π^1_{top} *and* π^2_{top}*:* Let $\mathcal{D}_{\text{top}}$ be a mixture of $\mathcal{D}_{\pi^1_{\text{top}}}$ and $\mathcal{D}_{\pi^2_{\text{top}}}$:

$$P_{\mathcal{D}_{\text{top}}} = (1 - \varsigma) P_{\mathcal{D}^1_{\text{top}}} + \varsigma P_{\mathcal{D}^2_{\text{top}}}, \tag{116}$$

where

$$\varsigma = \frac{1}{2(1 - \delta_G) + 1}. \tag{117}$$

It is clear that $0 < \varsigma < 1$ (because $\delta_G \in (0, 1)$), making it valid mixing ratio.

We now compute the marginal state distribution of the mixture by first analyzing the action probability:

$$\begin{aligned} P_{\mathcal{D}^1_{\text{top}}}(a^{\star}_{t:t+h} \mid s_t) &= \varepsilon_h, \\ P_{\mathcal{D}^2_{\text{top}}}(a^{\star}_{t:t+h} \mid s_t) &= (1 - 2\varepsilon_h)(1 - \delta_G). \end{aligned} \tag{118}$$

The state marginals are then

$$\begin{aligned} P_{\mathcal{D}_{\text{top}}}(s_{t+h} = Z \mid s_t, a^{\star}_{t:t+h}) &= \frac{P_{\mathcal{D}_{\text{top}}}(s_{t+h} = Z, a^{\star}_{t:t+h} \mid s_t)}{P_{\mathcal{D}_{\text{top}}}(a^{\star}_{t:t+h} \mid s_t)} \\ &= \frac{(1 - \varsigma) P_{\mathcal{D}^1_{\text{top}}}(a^{\star}_{t:t+h} \mid s_t)}{(1 - \varsigma) P_{\mathcal{D}^1_{\text{top}}}(a^{\star}_{t:t+h} \mid s_t) + \varsigma P_{\mathcal{D}^2_{\text{top}}}(a^{\star}_{t:t+h} \mid s_t)} \\ &= \frac{\varepsilon_h(1 - \varsigma)}{\varepsilon_h(1 - \varsigma) + (1 - 2\varepsilon_h)(1 - \delta_G)\varsigma} \\ &= 2\varepsilon_h. \end{aligned} \tag{119}$$

Therefore,

$$\begin{aligned} P_{\mathcal{D}_{\text{top}}}(s_{t+h} = Z \mid s_t, a^{\star}_{t:t+h}) &= 2\varepsilon_h, \\ P_{\mathcal{D}_{\text{top}}}(s_{t+h} = X_0 \mid s_t, a^{\star}_{t:t+h}) &= 1 - 2\varepsilon_h, \\ P_{\mathcal{D}_{\text{top}}}(s_{t+h} = G \mid s_t, a^{\star}_{t:t+h}) &= 0. \end{aligned} \tag{120}$$

Step 4. Strong ε_h-open-loop consistency of $\mathcal{D}_{\text{top}}$: Now, we check for strong open-loop consistency for the three possible action chunks on the top branch:

$$\begin{aligned} a^{\star}_{t:t+h} &= (0, 0, 0, \cdots) \\ a^{\diamond}_{t:t+h} &= (0, 1, 0, \cdots) \\ a^{\bullet}_{t:t+h} &= (0, 2, 0, \cdots) \end{aligned} \tag{121}$$

For $a^{\star}_{t:t+h} = 0$*,* we can compute open-loop marginal state distribution as follows:

$$\begin{aligned} T(s_{t+h} = Z \mid s_t, a^{\star}_{t:t+h}) &= \varepsilon_h, \\ T(s_{t+h} = X_0 \mid s_t, a^{\star}_{t:t+h}) &= 1 - 2\varepsilon_h, \\ T(s_{t+h} = G \mid s_t, a^{\star}_{t:t+h}) &= \varepsilon_h. \end{aligned} \tag{122}$$

Combining this with the data distribution calculated in Equation (120), it is clear that

$$D_{\mathrm{TV}}\big(T(s_{t+h} \mid s_t, a_{t:t+h} = 0) \,\big\|\, P_{\mathcal{D}_{\mathrm{top}}}(s_{t+h} \mid s_t, a_{t:t+h} = 0)\big) = \varepsilon_h. \tag{123}$$

We can repeat the same procedure to show that

$$D_{\mathrm{TV}}\big(T(s_{t+h'} \mid s_t, a_{t:t+h'} = 0) \,\big\|\, P_{\mathcal{D}_{\mathrm{top}}}(s_{t+h'} \mid s_t, a_{t:t+h} = 0)\big) = \varepsilon_h, \quad \forall h' \in \{1, \cdots, h-1\} \tag{124}$$

because the only difference in these distributions is that they occupy $s_{t+h'} = X_{h'}$ with $2\varepsilon_h$ probability instead of $s_{t+h} = X_0$ with $2\varepsilon_h$ probability.

For $a^{\bullet}_{t:t+h} = (0, 2, 0, \cdots)$, it is clear that

$$D_{\mathrm{TV}}\big(T(s_{t+h'} \mid s_t, a_{t:t+h} = a^{\bullet}_{t:t+h}) \,\big\|\, P_{\mathcal{D}_{\mathrm{top}}}(s_{t+h'} \mid s_t, a_{t:t+h} = a^{\bullet}_{t:t+h})\big) = \varepsilon_h \tag{125}$$

holds for any $h' \in \{1, 2, \cdots, h\}$ since the only difference between these two distributions is the ε_h-probability path (*i.e.*, $X_0 \to C \to Z$ where the probability is under $T(\cdot \mid s_t, a^{\bullet}_{t:t+h})$).

For $a^{\diamond}_{t:t+h} = (0, 1, 0, \cdots)$, we first compute the marginal state distributions:

$$\begin{aligned} P_{\mathcal{D}_{\mathrm{top}}}(s_{t+h} = Z \mid s_t, a^{\diamond}_{t:t+h}) &= \frac{(1 - 2\varepsilon_h)\delta_G}{2\varepsilon_h + (1 - 2\varepsilon_h)\delta_G}, \\ P_{\mathcal{D}_{\mathrm{top}}}(s_{t+h} = X_0 \mid s_t, a^{\diamond}_{t:t+h}) &= \frac{2\varepsilon_h}{2\varepsilon_h + (1 - 2\varepsilon_h)\delta_G}, \\ P_{\mathcal{D}_{\mathrm{top}}}(s_{t+h} = G \mid s_t, a^{\diamond}_{t:t+h}) &= 0. \\ P_{\mathcal{D}_{\mathrm{top}}}(s_{t+1} = X_1 \mid s_t, a^{\diamond}_{t:t+h}) &= \frac{(1 - 2\varepsilon_h)\delta_G}{2\varepsilon_h + (1 - 2\varepsilon_h)\delta_G}. \\ P_{\mathcal{D}_{\mathrm{top}}}(s_{t+1} = \tilde{X}_1 \mid s_t, a^{\diamond}_{t:t+h}) &= \frac{\varepsilon_h}{2\varepsilon_h + (1 - 2\varepsilon_h)\delta_G}. \\ P_{\mathcal{D}_{\mathrm{top}}}(s_{t+1} = C \mid s_t, a^{\diamond}_{t:t+h}) &= \frac{\varepsilon_h}{2\varepsilon_h + (1 - 2\varepsilon_h)\delta_G}. \end{aligned} \tag{126}$$

We can also compute the open-loop marginal state distribution as follows:

$$\begin{aligned} T(s_{t+h} = Z \mid s_t, a^{\diamond}_{t:t+h}) &= 1 - 2\varepsilon_h \\ T(s_{t+h} = X_0 \mid s_t, a^{\diamond}_{t:t+h}) &= 2\varepsilon_h \\ T(s_{t+h} = G \mid s_t, a^{\diamond}_{t:t+h}) &= 0. \\ T(s_{t+1} = X_1 \mid s_t, a^{\diamond}_{t:t+h}) &= 1 - 2\varepsilon_h. \\ T(s_{t+1} = \tilde{X}_1 \mid s_t, a^{\diamond}_{t:t+h}) &= \varepsilon_h. \\ T(s_{t+1} = C \mid s_t, a^{\diamond}_{t:t+h}) &= \varepsilon_h. \end{aligned} \tag{127}$$

Let c_3 be any value that satisfies $c_3 \in (0, \varepsilon_h/2)$, we can set

$$\delta_G = \frac{\varepsilon_h(1 - 2\varepsilon_h - 2c_3)}{(\varepsilon_h + c_3)(1 - 2\varepsilon_h)}, \tag{128}$$

such that

$$\begin{aligned} P_{\mathcal{D}_{\mathrm{top}}}(s_{t+h} = Z \mid s_t, a^{\diamond}_{t:t+h}) &= 1 - 2\varepsilon_h - 2c_3, \\ P_{\mathcal{D}_{\mathrm{top}}}(s_{t+h} = X_0 \mid s_t, a^{\diamond}_{t:t+h}) &= 2\varepsilon_h + 2c_3, \\ P_{\mathcal{D}_{\mathrm{top}}}(s_{t+h} = G \mid s_t, a^{\diamond}_{t:t+h}) &= 0, \\ P_{\mathcal{D}_{\mathrm{top}}}(s_{t+1} = X_1 \mid s_t, a^{\diamond}_{t:t+h}) &= 1 - 2\varepsilon_h - 2c_3, \\ P_{\mathcal{D}_{\mathrm{top}}}(s_{t+1} = \tilde{X}_1 \mid s_t, a^{\diamond}_{t:t+h}) &= \varepsilon_h + c_3, \\ P_{\mathcal{D}_{\mathrm{top}}}(s_{t+1} = C \mid s_t, a^{\diamond}_{t:t+h}) &= \varepsilon_h + c_3. \end{aligned} \tag{129}$$

It is easy to check that $0 < \delta_G < 1$ (a valid probability) because in Equation (128), each term in the numerator has a larger term in the denominator (*i.e.*, $\varepsilon_h < \varepsilon_h + c_3$ and $1 - 2\varepsilon_h - 2c_3 < 1 - 2\varepsilon_h$).

Now, for all $h' \in \{1, 2, \cdots, h\}$, using the values calculated in Equation (127) and Equation (129), we have

$$D_{\mathrm{TV}}\big(T(s_{t+h'} \mid s_t, a_{t:t+h'} = a^{\diamond}_{t:t+h'}) \,\big\|\, P_{\mathcal{D}_{\mathrm{top}}}(s_{t+h'} \mid s_t, a_{t:t+h} = a^{\diamond}_{t:t+h})\big) = 2c_3. \tag{130}$$

Since $c_3 < \varepsilon_h/2$, the strong open-loop consistency assumption holds for $a^{\diamond}_{t:t+h}$ as well.

`Step 5. The optimality gap and value estimation error for the top branch:` Now we can compute the optimality gap for the estimated value for $a^{\diamond}_{t:t+h}$:

$$V^\star(X_0) - \hat{V}^{\diamond}_{\mathrm{ac}}(X_0) = \frac{(1 - 2\varepsilon_h - 2c_3)\gamma}{(1-\gamma)(1 - 2(\varepsilon_h + c_3)\gamma^h)}, \tag{131}$$

where the h-step reward suboptimality gap is a sole result of the reaching Z with $(1 - 2\varepsilon_h - 2c_3)$ probability (and hence the $(1 - 2\varepsilon_h - 2c_3)\gamma$ term in the numerator), and the h-step distribution gap is reflected in the $(1 - 2(\varepsilon_h + c_3)\gamma^h)$ term at bottom because the probability of reaching X_0 after h steps is $2(\varepsilon_h + c_3)$.

Similarly, we can compute the optimality gap for $V^\star_{\mathrm{ac}}$ and $\hat{V}^\star_{\mathrm{ac}}$:

$$\begin{aligned} V^\star(X_0) - V^\star_{\mathrm{ac}}(X_0) &= \varepsilon_h \frac{\gamma - \gamma^h}{1-\gamma} + \frac{\varepsilon_h \gamma^h}{1-\gamma} + \gamma^h(1 - 2\varepsilon_h)(V^\star - \hat{V}_{\mathrm{ac}}) \\ &= \frac{\varepsilon_h \gamma}{(1-\gamma)(1 - (1 - 2\varepsilon_h)\gamma^h)}, \end{aligned} \tag{132}$$

$$\begin{aligned} V^\star(X_0) - \hat{V}^\star_{\mathrm{ac}}(X_0) &= \frac{2\varepsilon_h(\gamma - \gamma^h)}{1-\gamma} + \frac{2\varepsilon_h \gamma^h}{1-\gamma}\gamma^h(1 - 2\varepsilon_h)(V^\star - \hat{V}_{\mathrm{ac}}) \\ &= \frac{2\varepsilon_h \gamma}{(1-\gamma)(1 - (1 - 2\varepsilon_h)\gamma^h)}. \end{aligned} \tag{133}$$

Now, we observe that

$$1 - 2\varepsilon_h - 2c_3 > 1 - 3\varepsilon_h > 2\varepsilon_h, \tag{134}$$

where the first inequality is due to $c_3 \in (0, \varepsilon_h/2)$ and the second inequality is due to $\varepsilon_h \in (0, 1/5)$ in our assumption.

This allows us to lower-bound the estimated optimality gap for $a^{\diamond}_{t:t+h}$ as follows:

$$\begin{aligned} V^\star(X_0) - \hat{V}^{\diamond}_{\mathrm{ac}}(X_0) &= \frac{(1 - 2\varepsilon_h - 2c_3)\gamma}{(1-\gamma)(1 - 2(\varepsilon_h + c_3)\gamma^h)} \\ &> \frac{2\varepsilon_h \gamma}{(1-\gamma)(1 - (1 - 2\varepsilon_h)\gamma^h)} \\ &= V^\star(X_0) - \hat{V}^\star_{\mathrm{ac}}(X_0), \end{aligned} \tag{135}$$

where the inequality is obtained by triggering Lemma G.5 (*e.g.*, by setting $\varepsilon_1 = 2\varepsilon_h, \varepsilon_2 = (1 - 2\varepsilon_h - 2c_3), \tilde{\gamma} = \gamma^h$). The bound above rules out the possibility of $a^{\diamond}_{t:t+h}$ being picked by $\hat{\pi}^+_{\mathrm{ac}}$ because it has a lower estimated value compared to $a^\star_{t:t+h}$.

Finally, for $a^{\bullet}_{t:t+h}$, since it is correlated with $s_{t+h} = Z$ and receives no reward except the first step in $\mathcal{D}_{\mathrm{top}}$, the estimated value is just 1, being trivially smaller than $\hat{V}^\star_{\mathrm{ac}}(X_0)$ and would never get picked by $\hat{\pi}^+_{\mathrm{ac}}$.

Up to now, we have finished our data distribution construction and analysis for the top branch. We summarize the key intermediate results as the remark below:

Remark G.6 (Intermediate results from Step 1-4)**.** *The optimal action chunk is $a^\star_{t:t+h}$ and the estimated values for the two other possible action chunks $a^{\bullet}_{t:t+h}$, $a^{\diamond}_{t:t+h}$ are smaller than that of $a^\star_{t:t+h}$:*

$$\hat{V}^{\bullet}_{\mathrm{ac}}(X_0) < \hat{V}^{\diamond}_{\mathrm{ac}}(X_0) < \hat{V}^\star_{\mathrm{ac}}(X_0) = V^\star(X_0) - \frac{2\varepsilon_h \gamma}{(1-\gamma)(1 - (1 - 2\varepsilon_h)\gamma^h)}. \tag{136}$$

In addition, both $\mathcal{D}_{\mathrm{top}}$ and $\mathcal{D}^\star$ are strongly ε_h-open-loop consistent.

Next, we move on to the bottom branch.

Step 6. Data distribution $\mathcal{D}_{\text{bottom}}$ for the bottom branch: For the bottom, we again use two policies.

Policy π^1_{bottom}*:*

$$\begin{aligned} \pi^1_{\text{bottom}}(X_0) = \pi^1_{\text{bottom}}(G) = \pi^1_{\text{bottom}}(Z) = 1, \\ \pi^1_{\text{bottom}}(Y_1) = \pi^1_{\text{bottom}}(\tilde{Y}_1) = 2. \end{aligned} \tag{137}$$

π^1_{bottom} takes $a = 1$ at X_0 and G and Z, and takes $a = 2$ otherwise (at Y_1, $\tilde{Y}_1$). It is clear that this policy only produces two possible action chunks: $a^{\triangle}_{t:t+h} = (1, 1, 1, \cdots)$ or $a^{\times}_{t:t+h} = (1, 2, 1, \cdots)$.

Policy π^2_{bottom}*:*

$$\begin{aligned} \pi^2_{\text{bottom}}(X_0) &= 1, \\ \pi^2_{\text{bottom}}(a = 0 \mid Y_1) &= \delta_Z, \\ \pi^2_{\text{bottom}}(a = 1 \mid Y_1) &= 1 - \delta_Z, \\ \pi^2_{\text{bottom}}(\tilde{Y}_1) &= 0, \\ \pi^2_{\text{bottom}}(Y_i) &= 1, \quad \forall i \in \{2, \cdots, h-1\}, \end{aligned} \tag{138}$$

where $\delta_Z \in (0, 1)$ and we shall specify the exact value of δ_Z shortly.

π^2_{bottom} takes $a = 1$ when it is at Y_i and takes either $a = 0$ (with δ_Z probability) or $a = 1$ (with $1 - \delta_Z$ probability) when it is at $\tilde{Y}_1$. It is clear that this policy only produces two possible action chunks: $a^{\triangle}_{t:t+h} = (1, 1, 1, \cdots)$ or $a^{\circ}_{t:t+h} = (1, 0, 1, \cdots)$.

Now, we observe that the marginal state distributions for both policies conditioned on $a^{\triangle}_{t:t+h}$ are independent of c_1 and δ_Z because the action chunk only appears when π^1_{bottom} reaches G and when π^2_{bottom} reaches X_0. More specifically,

$$P_{\mathcal{D}^1_{\text{bottom}}}(s_{t+1} = G \mid s_t, a^{\triangle}_{t:t+h}) = P_{\mathcal{D}^1_{\text{bottom}}}(s_{t+h} = G \mid s_t, a^{\triangle}_{t:t+h}) = 1, \tag{139}$$

$$P_{\mathcal{D}^2_{\text{bottom}}}(s_{t+i} = X_i \mid s_t, a^{\triangle}_{t:t+h}) = P_{\mathcal{D}^2_{\text{bottom}}}(s_{t+h} = X_0 \mid s_t, a^{\triangle}_{t:t+h}) = 1, \forall i \in \{1, \cdots, h-1\}. \tag{140}$$

We can now mix $\mathcal{D}^1_{\text{bottom}}$ and $\mathcal{D}^2_{\text{bottom}}$ with an appropriate ratio to control the state marginals for $s_{t:t+h} = G$ and $s_{t:t+h} = X_0$ arbitrarily ($s_{t:t+h} = Z$ stays at 0 because none of the policies take/have taken $a^{\triangle}_{t:t+h}$ when they reach Z).

Mixing π^1_{bottom} *and* π^2_{bottom}*:* Let $\mathcal{D}_{\text{bottom}}$ be a mixture of $\mathcal{D}^1_{\text{bottom}}$ and $\mathcal{D}^2_{\text{bottom}}$:

$$P_{\mathcal{D}_{\text{bottom}}} = (1 - \vartheta) P_{\mathcal{D}^1_{\text{bottom}}} + \vartheta P_{\mathcal{D}^2_{\text{bottom}}}, \tag{141}$$

where we set the mixing ratio to be

$$\vartheta = \frac{c_1}{c_1 + (1 - \delta_Z)(\varepsilon_h + c_1)}. \tag{142}$$

This mixing ratio helps the calculations to be simpler later on.

We can now compute the marginal state distribution of the mixture. We start by analyzing the action probability:

$$\begin{aligned} P_{\mathcal{D}^1_{\text{bottom}}}(a^{\triangle}_{t:t+h} \mid s_t) &= c_1, \\ P_{\mathcal{D}^2_{\text{bottom}}}(a^{\triangle}_{t:t+h} \mid s_t) &= (1 - \varepsilon_h - c_1)(1 - \delta_Z). \end{aligned} \tag{143}$$

The state marginal is then

$$\begin{aligned}
P_{\mathcal{D}_{\text{bottom}}}(s_{t+h} = X_0 \mid s_t, a^{\triangle}_{t:t+h}) &= \frac{P_{\mathcal{D}_{\text{bottom}}}(s_{t+h} = X_0, a^{\triangle}_{t:t+h} \mid s_t)}{P_{\mathcal{D}_{\text{bottom}}}(a^{\triangle}_{t:t+h} \mid s_t)} \\
&= \frac{\vartheta P_{\mathcal{D}^2_{\text{bottom}}}(a^{\triangle}_{t:t+h} \mid s_t)}{(1-\vartheta) P_{\mathcal{D}^1_{\text{bottom}}}(a^{\triangle}_{t:t+h} \mid s_t) + \vartheta P_{\mathcal{D}^2_{\text{bottom}}}(a^{\triangle}_{t:t+h} \mid s_t)} \\
&= \frac{(1-\varepsilon_h - c_1)(1-\delta_Z)\vartheta}{c_1(1-\vartheta) + (1-\varepsilon_h - c_1)(1-\delta_Z)\vartheta} \\
&= 1 - \varepsilon_h - c_1.
\end{aligned} \tag{144}$$

We can use it to deduce the rest of the marginals as follows:

$$\begin{aligned}
P_{\mathcal{D}_{\text{bottom}}}(s_{t+h} = G \mid s_t, a^{\triangle}_{t:t+h}) &= \varepsilon_h + c_1, \quad \forall h' \in \{1, \cdots, h-1\}, \\
P_{\mathcal{D}_{\text{bottom}}}(s_{t+h} = X_0 \mid s_t, a^{\triangle}_{t:t+h}) &= 1 - \varepsilon_h - c_1, \\
P_{\mathcal{D}_{\text{bottom}}}(s_{t+h} = Z \mid s_t, a^{\triangle}_{t:t+h}) &= 0, \\
P_{\mathcal{D}_{\text{bottom}}}(s_{t+h'} = Y_{h'} \mid s_t, a^{\triangle}_{t:t+h}) &= 1 - \varepsilon_h - c_1, \quad \forall h' \in \{1, \cdots, h-2\}, \\
P_{\mathcal{D}_{\text{bottom}}}(s_{t+1} = \tilde{Y}_1 \mid s_t, a^{\triangle}_{t:t+h}) &= 0.
\end{aligned} \tag{145}$$

Up to now, we have established $\mathcal{D}_{\text{bottom}}$ and we are ready to check the strong open-loop consistency.

Step 7. Strong ε_h-open-loop consistency of $\mathcal{D}_{\text{bottom}}$:

For $a^{\triangle}_{t:t+h} = (1, 1, \cdots)$, we can compute the open-loop marginals as follows:

$$\begin{aligned}
T(s_{t+h'} = G \mid s_t, a^{\triangle}_{t:t+h}) &= c_1, \quad \forall h' \in \{1, \cdots, h-1\}, \\
T(s_{t+h} = X_0 \mid s_t, a^{\triangle}_{t:t+h}) &= 1 - \varepsilon_h - c_1, \\
T(s_{t+h} = Z \mid s_t, a^{\triangle}_{t:t+h}) &= \varepsilon_h. \\
T(s_{t+h'} = Y_{h'} \mid s_t, a^{\triangle}_{t:t+h}) &= 1 - \varepsilon_h - c_1, \quad \forall h' \in \{1, \cdots, h-2\} \\
T(s_{t+1} = \tilde{Y}_1 \mid s_t, a^{\triangle}_{t:t+h}) &= \varepsilon_h.
\end{aligned} \tag{146}$$

Combining it with the marginals calculated in Equation (145), it is clear that for all $h' \in \{1, \cdots, h-1\}$,

$$D_{\text{TV}}\big(T(s_{t+h'} \mid s_t, a_{t:t+h'} = a^{+}_{t:t+h'}) \,\big\|\, P_{\mathcal{D}_{\text{bottom}}}(s_{t+h'} \mid s_t, a_{t:t+h} = a^{+}_{t:t+h})\big) = \varepsilon_h, \tag{147}$$

satisfying the open-loop consistency.

For $a^{\times}_{t:t+h} = (1, 2, 1, \cdots)$, the data and open-loop state marginals are

$$\begin{aligned}
P_{\mathcal{D}_{\text{bottom}}}(s_{t+h} = Z \mid s_t, a^{\times}_{t:t+h}) &= 1, \\
P_{\mathcal{D}_{\text{bottom}}}(s_{t+1} = Y_1 \mid s_t, a^{\times}_{t:t+h}) &= \frac{1 - \varepsilon_h - c_1}{1 - c_1}, \\
P_{\mathcal{D}_{\text{bottom}}}(s_{t+1} = \tilde{Y}_1 \mid s_t, a^{\times}_{t:t+h}) &= \frac{\varepsilon_h}{1 - c_1}, \\
T(s_{t+h} = Z \mid s_t, a^{\times}_{t:t+h}) &= 1 - c_1, \\
T(s_{t+h} = G \mid s_t, a^{\times}_{t:t+h}) &= c_1, \\
T(s_{t+1} = Y_1 \mid s_t, a^{\times}_{t:t+h}) &= 1 - \varepsilon_h - c_1, \\
T(s_{t+1} = \tilde{Y}_1 \mid s_t, a^{\times}_{t:t+h}) &= \varepsilon_h, \\
T(s_{t+1} = G \mid s_t, a^{\times}_{t:t+h}) &= c_1.
\end{aligned} \tag{148}$$

This allows us to bound the TV distance for all $h' \in \{1, \cdots, h-1\}$ as

$$D_{\text{TV}}\big(T(s_{t+h'} \mid s_t, a_{t:t+h'} = a^{\times}_{t:t+h'}) \,\big\|\, P_{\mathcal{D}_{\text{bottom}}}(s_{t+h'} \mid s_t, a_{t:t+h} = a^{\times}_{t:t+h})\big) \leq \frac{c_1}{1 - c_1}. \tag{149}$$

Since $c_1 < \varepsilon_h/2 < 1/10$,

$$\frac{c_1}{1-c_1} < \frac{10}{9}c_1 < 5\varepsilon_h/9 < \varepsilon_h, \tag{150}$$

satisfying the strong open-loop consistency assumption.

For $a^\circ_{t:t+h} = (1, 0, 1, \cdots)$, we first compute the state marginals in $\mathcal{D}_{\text{bottom}}$ as follows:

$$\begin{aligned}
P_{\mathcal{D}_{\text{bottom}}}(s_{t+h} = Z \mid s_t, a^\circ_{t:t+h}) &= \frac{(1-\varepsilon_h - c_1)\delta_Z}{\varepsilon_h + (1-\varepsilon_h - c_1)\delta_Z}, \\
P_{\mathcal{D}_{\text{bottom}}}(s_{t+h} = X_0 \mid s_t, a^\circ_{t:t+h}) &= \frac{\varepsilon_h}{\varepsilon_h + (1-\varepsilon_h - c_1)\delta_Z}, \\
P_{\mathcal{D}_{\text{bottom}}}(s_{t+1} = Y_1 \mid s_t, a^\circ_{t:t+h}) &= \frac{(1-\varepsilon_h - c_1)\delta_Z}{\varepsilon_h + (1-\varepsilon_h - c_1)\delta_Z}. \\
P_{\mathcal{D}_{\text{bottom}}}(s_{t+1} = \tilde{Y}_1 \mid s_t, a^\circ_{t:t+h}) &= \frac{\varepsilon_h}{\varepsilon_h + (1-\varepsilon_h - c_1)\delta_Z}.
\end{aligned} \tag{151}$$

We can also compute the open-loop marginal state distribution as follows:

$$\begin{aligned}
T(s_{t+h} = Z \mid s_t, a^\circ_{t:t+h}) &= 1 - \varepsilon_h - c_1, \\
T(s_{t+h} = X_0 \mid s_t, a^\circ_{t:t+h}) &= \varepsilon_h, \\
T(s_{t+h} = G \mid s_t, a^\circ_{t:t+h}) &= c_1, \\
T(s_{t+1} = Y_1 \mid s_t, a^\circ_{t:t+h}) &= 1 - \varepsilon_h - c_1, \\
T(s_{t+1} = \tilde{Y}_1 \mid s_t, a^\circ_{t:t+h}) &= \varepsilon_h, \\
T(s_{t+1} = G \mid s_t, a^\circ_{t:t+h}) &= c_1.
\end{aligned} \tag{152}$$

Let $c_4 \in (c_1, \varepsilon_h)$, and we set

$$\delta_Z = \frac{\varepsilon_h(1-\varepsilon_h - c_4)}{(\varepsilon_h + c_4)(1-\varepsilon_h - c_1)}. \tag{153}$$

Then, we have

$$\begin{aligned}
P_{\mathcal{D}_{\text{bottom}}}(s_{t+h} = Z \mid s_t, a^\circ_{t:t+h}) &= 1 - \varepsilon_h - c_4, \\
P_{\mathcal{D}_{\text{bottom}}}(s_{t+h} = X_0 \mid s_t, a^\circ_{t:t+h}) &= \varepsilon_h + c_4, \\
P_{\mathcal{D}_{\text{bottom}}}(s_{t+h} = G \mid s_t, a^\circ_{t:t+h}) &= 0, \\
P_{\mathcal{D}_{\text{bottom}}}(s_{t+1} = Y_1 \mid s_t, a^\circ_{t:t+h}) &= 1 - \varepsilon_h - c_4, \\
P_{\mathcal{D}_{\text{bottom}}}(s_{t+1} = \tilde{Y}_1 \mid s_t, a^\circ_{t:t+h}) &= \varepsilon_h + c_4.
\end{aligned} \tag{154}$$

The TV distance is then

$$D_{\text{TV}}\big(T(s_{t+h'} \mid s_t, a_{t:t+h'} = a^\circ_{t:t+h'}) \,\big\|\, P_{\mathcal{D}_{\text{bottom}}}(s_{t+h'} \mid s_t, a_{t:t+h} = a^\circ_{t:t+h})\big) = c_4. \tag{155}$$

Since $c_4 < \varepsilon_h$, the strong open-loop consistency is also satisfied for $a^\circ_{t:t+h}$.

Up to now, we have checked that all three possible action chunks in the bottom branch satisfy the strong open-loop consistency assumption. Since $\mathcal{D}_{\text{top}}$ and $\mathcal{D}_{\text{bottom}}$ have non-overlapping supports for $a_{t:t+h}$, and they are both strongly ε_h-open-loop consistent on their own, we can construct $\mathcal{D}$ as

$$P_{\mathcal{D}}(\cdot \mid s_t) = (1-\varrho)P_{\mathcal{D}_{\text{top}}}(\cdot \mid s_t) + \varrho P_{\mathcal{D}_{\text{bottom}}}(\cdot \mid s_t), \tag{156}$$

for any $\varrho \in (0, 1)$, and conclude that

Remark G.7 (Intermediate result from Step 5-7)**.** *$\mathcal{D}$ is strongly ε_h-open-loop consistent.*

Up to now, we have constructed and checked both $\mathcal{D}$ and $\mathcal{D}^\star$ are strongly ε_h-open-loop consistent.

As the final step, we calculate the optimality gap and value estimation error for these action chunks.

`Step 8. The optimality gap and value estimation error for the bottom branch:`

We first note that similar to $a^{\bullet}_{t:t+h}$, $a^{\times}_{t:t+h}$ is correlated with $s_{t+h} = Z$ and always receives 0 reward except the first step in $\mathcal{D}$. Thus, the estimated value $\hat{V}^{\times}$ is just 1, being trivially smaller than $\hat{V}^{\star}_{\text{ac}}$ and would never get picked by $\hat{\pi}^{\triangle}_{\text{ac}}$. The only top contenders are $a^{+}_{t:t+h}$, $a^{\circ}_{t:t+h}$ and $a^{\star}_{t:t+h}$ (which we already analyzed in Step 5 above).

We start with $a^{\circ}_{t:t+h}$ where we can compute optimality gap as follows:

$$V^\star(X_0) - \hat{V}^{\circ}_{\text{ac}}(X_0) = \frac{(1-\varepsilon_h - c_4)\gamma + \delta(1-\gamma) + (\varepsilon_h + c_4)\delta(\gamma - \gamma^h)}{(1-\gamma)(1-(\varepsilon_h + c_4)\gamma^h)}. \tag{157}$$

Now, observe that

$$\varepsilon_h + c_4 < 2\varepsilon_h < 1 - 2\varepsilon_h, \tag{158}$$

where again the last inequality comes from the fact that $\varepsilon_h < 1/4$.

We can now lower-bound the optimality gap as follows:

$$\begin{aligned} V^\star(X_0) - \hat{V}^{\circ}_{\text{ac}}(X_0) &> \frac{2\varepsilon_h\gamma + \delta(1-\gamma) + (\varepsilon_h + c_4)\delta(\gamma - \gamma^h)}{(1-\gamma)(1-(1-2\varepsilon_h)\gamma^h)} \\ &> \frac{2\varepsilon_h\gamma}{(1-\gamma)(1-(1-2\varepsilon_h)\gamma^h)} \\ &= V^\star(X_0) - \hat{V}^{\star}_{\text{ac}}(X_0). \end{aligned} \tag{159}$$

where the first inequality is obtained by triggering Lemma G.5 (*e.g.*, by setting $\varepsilon_1 = 2\varepsilon_h, \varepsilon_2 = (1-\varepsilon_h - c_4), \tilde{\gamma} = \gamma^h$).

With this lower-bound, we can conclude that $a^{\circ}_{t:t+h}$ would not be picked by π^{+}_{ac} as well because $\hat{V}^{\circ}_{\text{ac}}(X_0) < \hat{V}^{\star}_{\text{ac}}(X_0)$.

Up to now, we have eliminated both $a^{\circ}_{t:t+h}$ and $a^{\times}_{t:t+h}$ (for the possibility of being picked by π^{+}_{ac}) and the only remaining contender left is $a^{\triangle}_{t:t+h}$.

We can also compute the estimated and the actual values for $a_{t:t+h} = a^{\triangle}_{t:t+h} = 1$ in terms of their optimality gaps:

$$V^\star(X_0) - \hat{V}^{\triangle}_{\text{ac}}(X_0) = \frac{\delta(1-\varepsilon_h - c_1)\gamma}{(1-\gamma)(1-(1-\varepsilon_h - c_1)\gamma^h)}, \tag{160}$$

$$V^\star(X_0) - V^{\triangle}_{\text{ac}}(X_0) = \frac{[\delta(1-\varepsilon_h - c_1) + \varepsilon_h]\,\gamma}{(1-\gamma)(1-(1-\varepsilon_h - c_1)\gamma^h)}. \tag{161}$$

Let

$$\delta = \frac{2\varepsilon_h\gamma - c_2}{(1-\varepsilon_h - c_1)\gamma}\,\frac{1-(1-\varepsilon_h - c_1)\gamma^h}{1-(1-2\varepsilon_h)\gamma^h}. \tag{162}$$

We first check $1-\delta$ is a valid reward value (within $[0,1]$):

$$\begin{aligned} \delta &< \frac{2\varepsilon_h}{1-\varepsilon_h - c_1}\,\frac{1-(1-\varepsilon_h - c_1)\gamma^h}{1-(1-2\varepsilon_h)\gamma^h} \\ &< \frac{2\varepsilon_h}{1-2\varepsilon_h}\,\frac{1-(1-2\varepsilon_h)\gamma^h}{1-(1-2\varepsilon_h)\gamma^h} \\ &= \frac{2\varepsilon_h}{1-2\varepsilon_h} \\ &\leq 1, \end{aligned} \tag{163}$$

where the first inequality is because $c_2 > 0$, the second inequality is due to $c_1 < \varepsilon_h$, and the final inequality is due to $\varepsilon_h < 1/4$.

It is also clear that $\delta > 0$ because all terms are positive in the fraction (Equation (162)).

Next, we substitute δ in to obtain

$$V^\star(X_0) - \hat{V}^{\triangle}_{\text{ac}}(X_0) = \frac{2\varepsilon_h\gamma - c_2}{(1-\gamma)(1-(1-2\varepsilon_h)\gamma^h)}, \tag{164}$$

$$V^\star(X_0) - V^{\triangle}_{\text{ac}}(X_0) = \frac{2\varepsilon_h\gamma - c_2}{(1-\gamma)(1-(1-2\varepsilon_h)\gamma^h)} + \frac{\varepsilon_h\gamma}{(1-\gamma)(1-(1-\varepsilon_h-c_1)\gamma^h)}, \tag{165}$$

where intuitively the second term in $V^\star(X_0) - V^{\triangle}_{\text{ac}}(X_0)$ is due to the fact that from $P_{\mathcal{D}}(\cdot \mid s_t, a^{\triangle}_{t:t+h})$ to $T(\cdot \mid s_t, a^{\triangle}_{t:t+h})$, there is a shift in ε_h probability mass from $s_{t:t+h} = (X_0, G, \cdots)$ to $s_{t:t+h} = (X_0, \tilde{Y}_1, Z, \cdots)$ incurring an additional $\frac{\varepsilon_h\gamma}{1-\gamma}$ suboptimality in terms of the h-step reward, and then amplified by the value recursion by an additional factor of $\frac{1}{1-(1-\varepsilon_h-c_1)\gamma^h}$ (where $1-\varepsilon_h-c_1$ is the probability that $a^{\triangle}_{t:t+h}$ reaches X_0 for the value recursion to occur).

Since $c_2 > 0$, we can now show that $a^{\triangle}_{t:t+h}$ achieves the highest estimated value among six possible action chunks:

$$V^\star - \hat{V}^{\triangle}_{\text{ac}} < V^\star - \hat{V}^{\star}_{\text{ac}} = \frac{2\varepsilon_h\gamma}{(1-\gamma)(1-(1-2\varepsilon_h)\gamma^h)}, \tag{166}$$

which means that $\pi^+_{\text{ac}}(X_0) = a^{\triangle}_{t:t+h} = (1, 1, \cdots)$, or equivalently $\hat{V}^{\triangle}_{\text{ac}} = \hat{V}^{+}_{\text{ac}}$!

Finally, putting everything together, we have

$$V^\star(X_0) - V^{+}_{\text{ac}}(X_0) = \frac{2\varepsilon_h\gamma - c_2}{(1-\gamma)(1-(1-2\varepsilon_h)\gamma^h)} + \frac{\varepsilon_h\gamma}{(1-\gamma)(1-(1-\varepsilon_h-c_1)\gamma^h)}, \tag{167}$$

as desired. □

G.8 Proof of Proposition 4.9

Proposition 4.9 (Optimality of Closed-loop Execution of Action Chunking Policy)**.** *Let $V^\bullet$ be the value of the one-step policy, $\pi^\bullet$, defined as the closed-loop execution of the action chunking policy π^+_{ac} learned from $\mathcal{D}$. That is, for each $s_t \in \operatorname{supp}(P_{\mathcal{D}}(s_t))$,*

$$\pi^\bullet(s_t) = a^+_t, \quad \textit{where } a^+_{t:t+h} = \pi^+_{\text{ac}}(s_t). \tag{21}$$

If we assume $\mathcal{D}$ and $\mathcal{D}^\star$ are both strongly ε_h-open-loop consistent and $\operatorname{supp}(P_{\mathcal{D}}(s_t, a_{t:t+h})) \supseteq \operatorname{supp}(P_{\mathcal{D}^\star}(s_t, a_{t:t+h}))$, then under $\operatorname{supp}(\mathcal{D}^\star)$,

$$\|V^\star - V^\bullet\|_\infty \leq \frac{\varepsilon_h\gamma}{(1-\gamma)^2}\left[\frac{2}{1-(1-2\varepsilon_h)\gamma^h} + \frac{1}{1-(1-\varepsilon_h)\gamma^h}\right] \leq \frac{3\varepsilon_h}{(1-\gamma)^2(1-\gamma^h)}. \tag{22}$$

Proof. We observe that

$$\begin{aligned} V^+_{\text{ac}}(s_t) &= Q^+_{\text{ac}}(s_t, a^+_{t:t+h}) \\ &\leq Q^\star(s_t, a^+_t). \end{aligned} \tag{168}$$

Combining this with Theorem 4.6, we get

$$Q^\star(s_t, a^+_t) \geq V^\star(s_t) - \Delta, \tag{169}$$

where $\Delta = \frac{\varepsilon_h\gamma}{1-\gamma}\left[\frac{2}{1-(1-2\varepsilon_h)\gamma^h} + \frac{1}{1-(1-\varepsilon_h)\gamma^h}\right]$.

Now, we can bound $V^\bullet$ as follows:

$$\begin{aligned} V^\star(s_t) - V^\bullet(s_t) &\leq Q^\star(s_t, a^+_t) - Q^\bullet(s_t, a^+_t) + \Delta \\ &\leq \gamma\mathbb{E}_{T(\cdot \mid s_t, a^+_t)}\left[V^\star(s_{t+1}) - V^\bullet(s_{t+1})\right] + \Delta \\ &\leq \frac{\varepsilon_h\gamma}{(1-\gamma)^2}\left[\frac{2}{1-(1-2\varepsilon_h)\gamma^h} + \frac{1}{1-(1-\varepsilon_h)\gamma^h}\right]. \end{aligned} \tag{170}$$

□

G.9 Proof of Theorem 4.8

Theorem 4.8. *Let $\mathcal{D}$ be strongly ε_h-open-consistent, δ_n-suboptimal, and* $\mathrm{supp}(\mathcal{D}) \supseteq \mathrm{supp}(\mathcal{D}^\star)$. *Let $\pi_n^\star$ be the optimal n-step return policy learned from $\mathcal{D}$, as the solution of*

$$Q_n^\star(s_t, a_t) = \mathbb{E}_{P_\mathcal{D}}\left[R_{t:t+n} + \gamma^n Q_n^\star(s_{t+n}, \pi_n^\star(s_{t+n}))\right], \quad \pi_n^\star : s_t \mapsto \arg\max_{a_t} Q_n^\star(s_t, a_t). \tag{20}$$

As long as $\delta_n > \frac{3\varepsilon_h(1-\gamma^n)}{(1-\gamma)(1-\gamma^h)}$, then from all $s \in \mathrm{supp}(\mathcal{D}^\star)$, the action chunking policy, π_{ac}^+ (Equation (17)), is better than the n-step return policy, π_n (Equation (20)) (i.e., $V_{\mathrm{ac}}^+(s) > V_n^\star(s)$).

To prove Theorem 4.8, we first prove the following helper Lemma G.8 to quantify sub-optimality for n-step return policy.

Lemma G.8. *Let $Q_n^\star$ be the solution of the uncorrected n-step return backup equation:*

$$Q_n^\star(s_t, a_t) = \mathbb{E}_{P_\mathcal{D}(\cdot|s_t,a_t)}\left[R_{t:t+n} + \gamma^n \max_{a_{t+n}} Q_n^\star(s_{t+n}, a_{t+n})\right] \tag{171}$$

The following inequality holds as long as $\mathcal{D}$ is δ_n-suboptimal:

$$Q^\star(s_t, a_t) \geq Q_n^\star(s_t, a_t) + \frac{\delta_n}{1-\gamma^n}, \forall s_t \in \mathcal{S}, a_t \in \mathcal{A} \tag{172}$$

where $Q^\star$ is the Q-function of the optimal policy in $\mathcal{M}$. For the n-step return policy

$$\pi_n^\star : s_t \mapsto \arg\max_{a_t} Q_n^\star(s_t, a_t), \tag{173}$$

its corresponding value admits a similar bound:

$$V^\star(s_t) \geq V_n^\star(s_t) + \frac{\delta_n}{1-\gamma^n}, \forall s_t \tag{174}$$

Proof. Using the definition of suboptimal data (Definition 4.7), we have

$$\begin{aligned} Q_n^\star(s_t, a_t) &= \mathbb{E}_{P_\mathcal{D}(\cdot|s_t,a_t)}\left[R_{t:t+n} + \gamma^n \max_{a_{t+n}} Q_n^\star(s_{t+n}, a_{t+n})\right] \\ &\leq Q^\star(s_t, a_t) - \delta_n + \gamma^h \mathbb{E}_{P_\mathcal{D}(\cdot|s_t,a_t)}\left[\max_{a_{t+n}} Q_n^\star(s_{t+n}, a_{t+n}) - V^\star(s_{t+h})\right] \end{aligned} \tag{175}$$

Rearranging the inequality above yields

$$Q_n^\star(s_t, a_t) - Q^\star(s_t, a_t) \leq -\delta_n + \gamma^n \mathbb{E}_{P_\mathcal{D}(\cdot|s_t)}[V_n^\star(s_{t+n}) - V^\star(s_{t+n})], \forall s_t \in \mathcal{S}, a_t \in \mathcal{A} \tag{176}$$

By recursively applying the inequality above, we have

$$Q^\star(s_t, a_t) \geq Q_n^\star(s_t, a_t) + \frac{\delta_n}{1-\gamma^n}, \forall s_t \in \mathcal{S}, a_t \in \mathcal{A} \tag{177}$$

By choosing $a_t^\star = \pi_n^\star(s_t)$, we see that

$$\begin{aligned} V^\star(s_t) &\geq Q^\star(s_t, a_t) \\ &\geq Q_n^\star(s_t, a_t^\star) + \frac{\delta_n}{1-\gamma^n} \\ &= V_n^\star(s_t) + \frac{\delta_n}{1-\gamma^n} \end{aligned} \tag{178}$$

□

Now we are ready to prove the main Theorem 4.8.

Proof of Theorem 4.8. From Lemma G.8 and Theorem 4.6, we have

$$V_n^\star(s) + \frac{\delta_n}{1-\gamma^n} \leq V^\star(s) \leq V_{\mathrm{ac}}^+(s) + \frac{\varepsilon_h \gamma}{1-\gamma}\left[\frac{2}{1-(1-2\varepsilon_h)\gamma^h} + \frac{1}{1-(1-\varepsilon_h)\gamma^h}\right]. \tag{179}$$

Rearranging the terms give

$$V_{\mathrm{ac}}^+(s) - V_n^\star(s) \geq \frac{\delta_n}{1-\gamma^n} - \frac{\varepsilon_h \gamma}{1-\gamma}\left[\frac{2}{1-(1-2\varepsilon_h)\gamma^h} + \frac{1}{1-(1-\varepsilon_h)\gamma^h}\right]. \tag{180}$$

□

G.10 Proof Theorem 4.11

Theorem 4.11 (Closed-loop AC Policy under Bounded OV)**.** *Let $\mathcal{D}^\star$ be the data distribution collected by an optimal policy. Assume $\mathcal{D}$ can be decomposed into a mixture of data distributions $\{\mathcal{D}^\star, \mathcal{D}_1, \mathcal{D}_2, \cdots \mathcal{D}_N\}$ such that each data distribution component satisfies Assumption 4.1 and for some $\vartheta_h^L, \vartheta_h^G \geq 0$, they satisfy the following two conditions:*

*1. **Locally bounded optimality variability condition**: every $\mathcal{D}_i$ (including $\mathcal{D}^\star$) exhibits ϑ_h^L-bounded variability in optimality conditioned on s_t, a_t for all $(s_t, a_t) \in \mathrm{supp}(P_{\mathcal{D}_i}(s_t, a_t))$, and*

*2. **Globally bounded optimality variability condition**: $\mathcal{D}$ as a whole exhibits ϑ_h^G-variability in optimality conditioned on $s_t, a_{t:t+h}$ for all $(s_t, a_{t:t+h}) \in \mathrm{supp}(P_{\mathcal{D}}(s_t, a_{t:t+h}))$.*

Then for all $s_t \in \mathrm{supp}(P_{\mathcal{D}^\star}(s_t))$,

$$V^\star(s_t) - V^\bullet(s_t) \leq \frac{\vartheta_h^L}{1-\gamma} + \frac{\vartheta_h^G + \gamma^h \min(\vartheta_h^L, \vartheta_h^G)}{(1-\gamma)(1-\gamma^h)} \leq \vartheta_h^L H + 2\vartheta_h^G H \bar{H} \tag{24}$$

The proof of Theorem 4.11 below is made possible by observing that $V^\star(s_t) - \hat{V}_{\mathrm{ac}}^+$ and $\hat{V}_{\mathrm{ac}}^+(s_t) - Q^\star(s_t, a_t^+)$ are bounded by $\vartheta_h^G/(1-\gamma^h)$ and $\vartheta_h^L/(1-\gamma^h)$ respectively. Combining this two bounds naïvely already allows us to derive a relatively loose bound $V^\star(s_t) - Q^\star(s_t, a_t^+) \leq (\vartheta_h^L + \vartheta_h^G)/(1-\gamma^h)$ which leads to $V^\star(s_t) - V^\bullet(s_t) \leq (\vartheta_h^L + \vartheta_h^G)/(1-\gamma^h)/(1-\gamma)$. To obtain the tight bound in Theorem 4.11, we leverage a key insight that the amount of overestimation in V_{ac}^+ can *never exceed* $\vartheta_h^L + \frac{\vartheta^G}{1-\gamma^h}$ as otherwise the nominal value of the action chunking policy h-step into the future, $\hat{V}_{\mathrm{ac}}^+(s_{t+h})$, would have an optimality gap higher than $\vartheta_h^G/(1-\gamma^h)$, which is impossible under the global optimality variability condition. Forming this tight bound is important because it effectively shaves off a factor of $\bar{H} = 1/(1-\gamma^h)$ from the ϑ_h^L term (the stronger local condition) and only bumps up a factor of ≈ 2 to the ϑ_h^G term (the weaker global condition).

Proof of Theorem 4.11. Consider any $s_t \in \mathrm{supp}(P_{\mathcal{D}^\star}(s_t))$. Let $a_{t:t+h}^+ = \pi_{\mathrm{ac}}^+(s_t)$ and

$$a_{t:t+h}^\circ := \arg\max_{a_{t:t+h} \in \mathrm{supp}(P_{\mathcal{D}^\star}(a_{t:t+h}|s_t))} \left[\mathbb{E}_{P_{\mathcal{D}^\star}(\cdot|s_t, a_{t:t+h})}\left[R_{t:t+h} + \gamma^h V^\star(s_{t+h})\right]\right]. \tag{181}$$

We first observe that

$$\mathbb{E}_{P_{\mathcal{D}^\star}(\cdot|s_t, a_{t:t+h}^\circ)}\left[R_{t:t+h} + \gamma^h V^\star(s_{t+h})\right] \geq V^\star(s_t), \tag{182}$$

because

$$V^\star(s_t) = \mathbb{E}_{a_{t:t+h} \sim P_{\mathcal{D}^\star}(\cdot|s_t)}\left[\mathbb{E}_{P_{\mathcal{D}^\star}(\cdot|s_t, a_{t:t+h})}\left[R_{t:t+h} + \gamma^h V^\star(s_{t+h})\right]\right], \tag{183}$$

and the maximum value of a random variable is no less than its expectation.

Let

$$\tilde{Q}_{\min}(s_t, a_{t:t+h}^\circ) := \min_{\mathrm{supp}(P_{\mathcal{D}}(\cdot|s_t, a_{t:t+h}^\circ))} \left[R_{t:t+h} + V^\star(s_{t+h})\right], \tag{184}$$

$$\tilde{Q}_{\max}(s_t, a_{t:t+h}^\circ) := \max_{\mathrm{supp}(P_{\mathcal{D}}(\cdot|s_t, a_{t:t+h}^\circ))} \left[R_{t:t+h} + V^\star(s_{t+h})\right]. \tag{185}$$

Since $\mathcal{D}$ exhibits ϑ_h^G-variability in optimality, we have

$$\tilde{Q}_{\min}(s_t, a_{t:t+h}^\circ) \geq \tilde{Q}_{\max}(s_t, a_{t:t+h}^\circ) - \vartheta_h^G. \tag{186}$$

$$\begin{aligned}
&V^\star(s_t) - Q^\star(s_t, a_t^+) \\
&= V^\star(s_t) - \hat{V}_{\mathrm{ac}}^+(s_t) + \hat{V}_{\mathrm{ac}}^+(s_t) - Q^\star(s_t, a_t^+) \\
&= V^\star(s_t) - \hat{Q}_{\mathrm{ac}}^+(s_t, a_{t:t+h}^+) + \hat{Q}_{\mathrm{ac}}^+(s_t, a_{t:t+h}^+) - Q^\star(s_t, a_t^+) \\
&\leq V^\star(s_t) - \hat{Q}_{\mathrm{ac}}^+(s_t, a_{t:t+h}^\circ) + \vartheta_h^L + \gamma^h \mathbb{E}_{P_{\mathcal{D}}(\cdot|s_t, a_{t:t+h}^+)}\left[\hat{V}_{\mathrm{ac}}^+(s_{t+h}) - V^\star(s_{t+h})\right] \\
&= V^\star(s_t) - \hat{Q}_{\mathrm{ac}}^+(s_t, a_{t:t+h}^\circ) + \vartheta_h^L + \gamma^h \mathbb{E}_{P_{\mathcal{D}}(\cdot|s_t, a_{t:t+h}^+)}\left[\hat{V}_{\mathrm{ac}}^+(s_{t+h}) - Q^\star(s_{t+h}, a_{t+h}^+)\right] - \\
&\qquad \gamma^h \mathbb{E}_{P_{\mathcal{D}}(\cdot|s_t, a_{t:t+h}^+)}\left[V^\star(s_{t+h}) - Q^\star(s_{t+h}, a_{t+h}^+)\right].
\end{aligned} \tag{187}$$

We can use it to lower-bound $\hat{V}^+_{\text{ac}}(s_t)$ as follows:

$$\begin{aligned}
\hat{V}^+_{\text{ac}}(s_t) &= \hat{Q}^+_{\text{ac}}(s_t, a^+_{t:t+h}) \\
&\geq \hat{Q}^+_{\text{ac}}(s_t, a^\circ_{t:t+h}) \\
&= \mathbb{E}_{P_{\mathcal{D}}(\cdot|s_t, a^\circ_{t:t+h})}\left[R_{t:t+h} + \gamma^h \hat{V}^+_{\text{ac}}(s_{t+h})\right] \\
&= \mathbb{E}_{P_{\mathcal{D}}(\cdot|s_t, a^\circ_{t:t+h})}\left[R_{t:t+h} + \gamma^h V^\star(s_{t+h})\right] + \mathbb{E}_{P_{\mathcal{D}}(\cdot|s_t, a^\circ_{t:t+h})}\left[\gamma^h(\hat{V}^+_{\text{ac}}(s_{t+h}) - V^\star(s_{t+h}))\right] \\
&\geq \tilde{Q}_{\text{min}}(s_t, a^\circ_{t:t+h}) + \mathbb{E}_{P_{\mathcal{D}}(\cdot|s_t, a^\circ_{t:t+h})}\left[\gamma^h(\hat{V}^+_{\text{ac}}(s_{t+h}) - V^\star(s_{t+h}))\right] \\
&\geq \tilde{Q}_{\text{max}}(s_t, a^\circ_{t:t+h}) - \vartheta^G_h + \mathbb{E}_{P_{\mathcal{D}}(\cdot|s_t, a^\circ_{t:t+h})}\left[\gamma^h(\hat{V}^+_{\text{ac}}(s_{t+h}) - V^\star(s_{t+h}))\right] \\
&\geq \mathbb{E}_{P_{\mathcal{D}^\star}(\cdot|s_t, a^\circ_{t:t+h})}\left[R_{t:t+h} + \gamma^h V^\star(s_{t+h})\right] - \vartheta^G_h + \gamma^h \mathbb{E}_{P_{\mathcal{D}}(\cdot|s_t, a^\circ_{t:t+h})}\left[(\hat{V}^+_{\text{ac}}(s_{t+h}) - V^\star(s_{t+h}))\right] \\
&\geq V^\star(s_t) - \vartheta^G_h + \gamma^h \mathbb{E}_{P_{\mathcal{D}}(\cdot|s_t, a^\circ_{t:t+h})}\left[(\hat{V}^+_{\text{ac}}(s_{t+h}) - V^\star(s_{t+h}))\right] \\
&\geq V^\star(s_t) - \frac{\vartheta^G_h}{1 - \gamma^h}.
\end{aligned} \tag{188}$$

Let $\mathbb{M}^+ = \{\tilde{\mathcal{D}}_1, \cdots \tilde{\mathcal{D}}_{M^+}\}$ be all data distributions from $\{\mathcal{D}^\star, \mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_N\}$ where $(s_t, a^+_{t:t+h})$ is in the support. Let $\tilde{\mathcal{D}}^+$ be any mixture of $\mathbb{M}$ where each mixture component has non-zero weight:

$$P_{\tilde{\mathcal{D}}^+} = \sum_{i=1}^{M} w_i P_{\tilde{\mathcal{D}}_i}, \tag{189}$$

where $w_i > 0, \sum_i w_i = 1$.

Let

$$\begin{aligned}
\tilde{Q}^\star_{\text{min}}(s_t, a_t) &:= \min_{\text{supp}(P_{\mathcal{D}^\star}(\cdot|s_t, a_t))} \left[R_{t:t+h} + V^\star(s_{t+h})\right], & (190) \\
\tilde{Q}^\star_{\text{max}}(s_t, a_t) &:= \max_{\text{supp}(P_{\mathcal{D}^\star}(\cdot|s_t, a_t))} \left[R_{t:t+h} + V^\star(s_{t+h})\right], & (191) \\
\tilde{Q}^i_{\text{min}}(s_t, a_t) &:= \min_{\text{supp}(P_{\mathcal{D}^i}(\cdot|s_t, a_t))} \left[R_{t:t+h} + V^\star(s_{t+h})\right], & (192) \\
\tilde{Q}^i_{\text{max}}(s_t, a_t) &:= \max_{\text{supp}(P_{\mathcal{D}^i}(\cdot|s_t, a_t))} \left[R_{t:t+h} + V^\star(s_{t+h})\right], & (193) \\
\tilde{Q}^+_{\text{max}}(s_t, a^+_t) &:= \max_{\text{supp}(P_{\tilde{\mathcal{D}}^+}(\cdot|s_t, a^+_t))} \left[R_{t:t+h} + V^\star(s_{t+h})\right], & (194) \\
\tilde{Q}^+_{\text{max}}(s_t, a^+_{t:t+h}) &:= \max_{\text{supp}(P_{\tilde{\mathcal{D}}^+}(\cdot|s_t, a^+_{t:t+h}))} \left[R_{t:t+h} + V^\star(s_{t+h})\right]. & (195)
\end{aligned}$$

The minimum and the maximum is over the remaining trajectory conditioned on s_t, a_t or $s_t, a_{t:t+h}$ that is still in the support of the corresponding data distribution.

From the ϑ^L_h-bounded variability in optimality and the Assumption 4.1 of each data mixture, we observe that

$$\begin{aligned}
Q^\star(s_t, a_t) &\geq \tilde{Q}^i_{\text{min}}(s_t, a_t) \geq \tilde{Q}^i_{\text{max}}(s_t, a_t) - \vartheta^L_h, \quad \forall i \in \{1, 2, \cdots, N\} & (196) \\
Q^\star(s_t, a_t) &\geq \tilde{Q}^\star_{\text{min}}(s_t, a_t) \geq \tilde{Q}^\star_{\text{max}}(s_t, a_t) - \vartheta^L_h. & (197)
\end{aligned}$$

We can then derive that

$$\begin{aligned}
\tilde{Q}^+_{\text{max}}(s_t, a^+_t) &= \max(\tilde{Q}^\star_{\text{max}}(s_t, a^+_t), \tilde{Q}^1_{\text{max}}(s_t, a^+_t), \cdots, \tilde{Q}^N_{\text{max}}(s_t, a^+_t)) \\
&\leq Q^\star(s_t, a_t) + \vartheta^L_h.
\end{aligned} \tag{198}$$

With this, we can now upper-bound $\hat{V}^+_{\text{ac}}(s_t)$ as follows:

$$\begin{aligned}
\hat{V}^+_{\text{ac}}(s_t) &= \hat{Q}^+_{\text{ac}}(s_t, a^+_{t:t+h}) \\
&= \mathbb{E}_{P_{\mathcal{D}}(\cdot|s_t, a^+_{t:t+h})}\left[R_{t:t+h} + \gamma^h \hat{V}^+_{\text{ac}}(s_{t+h})\right] \\
&= \mathbb{E}_{P_{\tilde{\mathcal{D}}^+}(\cdot|s_t, a^+_{t:t+h})}\left[R_{t:t+h} + \gamma^h \hat{V}^+_{\text{ac}}(s_{t+h})\right] \\
&= \mathbb{E}_{P_{\tilde{\mathcal{D}}^+}(\cdot|s_t, a^+_{t:t+h})}\left[R_{t:t+h} + \gamma^h V^\star(s_{t+h})\right] + \gamma^h \mathbb{E}_{P_{\tilde{\mathcal{D}}^+}(\cdot|s_t, a^+_{t:t+h})}\left[\hat{V}^+_{\text{ac}}(s_{t+h}) - V^\star(s_{t+h})\right] \\
&\leq \tilde{Q}^+_{\max}(s_t, a^+_{t:t+h}) + \gamma^h \mathbb{E}_{P_{\mathcal{D}}(\cdot|s_t, a^+_{t:t+h})}\left[\hat{V}^+_{\text{ac}}(s_{t+h}) - V^\star(s_{t+h})\right] \\
&\leq \tilde{Q}^+_{\max}(s_t, a^+_t) + \gamma^h \mathbb{E}_{P_{\mathcal{D}}(\cdot|s_t, a^+_{t:t+h})}\left[\hat{V}^+_{\text{ac}}(s_{t+h}) - V^\star(s_{t+h})\right] \\
&\leq Q^\star(s_t, a^+_t) + \vartheta^L_h + \gamma^h \mathbb{E}_{P_{\mathcal{D}}(\cdot|s_t, a^+_{t:t+h})}\left[\hat{V}^+_{\text{ac}}(s_{t+h}) - V^\star(s_{t+h})\right].
\end{aligned} \tag{199}$$

Let

$$\Delta(s_t) := V^\star(s_t) - Q^\star(s_t, a^+_t). \tag{200}$$

$$\hat{\Delta}(s_t) := \hat{V}^+_{\text{ac}}(s_t) - Q^\star(s_t, a^+_t). \tag{201}$$

From the inequalities above, we have

$$\hat{\Delta}(s_t) \leq \vartheta^L_h + \gamma^h \sup_{s_{t+h}}\left[\hat{\Delta}(s_{t+h}) - \Delta(s_{t+h})\right], \tag{202}$$

$$0 \leq \Delta(s_t) \leq \frac{\vartheta^G_h}{1-\gamma^h} + \hat{\Delta}(s_t), \tag{203}$$

$$\hat{\Delta}(s_t) - \Delta(s_t) \leq \min\left\{\frac{\vartheta^G_h}{1-\gamma^h}, \hat{\Delta}(s_t)\right\}. \tag{204}$$

The minimum operator allows us to obtain two upper-bounds on Δ:

$$\Delta(s_t) \leq \vartheta^L_h + \frac{(1+\gamma^h)\vartheta^G_h}{1-\gamma^h}, \tag{205}$$

$$\Delta(s_t) \leq \frac{\vartheta^G_h}{1-\gamma^h} + \hat{\Delta}(s_t) \leq \frac{\vartheta^L_h + \vartheta^G_h}{1-\gamma^h}. \tag{206}$$

Finally, combining these two upper-bounds together and recursively applying the inequality yields our desired results:

$$V^\star(s_t) - Q^\bullet(s_t, a^+_t) \leq \frac{\vartheta^L_h}{1-\gamma} + \frac{\vartheta^L_h}{(1-\gamma)(1-\gamma^h)} + \frac{\gamma^h \min(\vartheta^G_h, \vartheta^L_h)}{(1-\gamma)(1-\gamma^h)}. \tag{207}$$

□

G.11 Proof of Theorem F.4

Theorem F.4 (Worst-case Closed-loop AC Policy under BOV)**.** *For any $\gamma \in (0,1)$, $\vartheta_h^G, \vartheta_h^L \in \left(0, \frac{\gamma-\gamma^h}{4(1-\gamma)}\right]$, $c \in \left[0, \frac{\gamma-\gamma^h}{4(1-\gamma^h)}\right)$, $\sigma \in \left(0, \frac{\min(\vartheta_h^G, \vartheta_h^L)}{1-\gamma}\right)$, there exists $\mathcal{M}$ and $\mathcal{D}$ satisfying the mixture assumption in Theorem 4.11 such that there exists $s_t \in \operatorname{supp}(P_{\mathcal{D}^\star}(s_t))$, where*

$$V^\star(s_t) - V^\bullet(s_t) = \frac{\vartheta_h^L}{1-\gamma} + \frac{\vartheta_h^G + \gamma^h \min(\vartheta_h^L, \vartheta_h^G)}{(1-\gamma)(1-\gamma^h)} - \sigma, \quad V^\star(s_t) - V^+_{\mathrm{ac}}(s_t) \geq \frac{c}{1-\gamma} \tag{38}$$

To show that our upper-bound is achievable, we need to carefully design both the MDP and the data distribution. For clarity of the proof, we divide up the construction into two parts. The first part (Lemma G.9) focuses on designing part of the MDP and two data distributions $\mathcal{D}^\star$ and $\mathcal{D}^\diamond$ such that any action chunk that has a value bigger than $V^\star - \frac{\vartheta_h^G}{1-\gamma^h}$ is preferred over the action chunks in $\mathcal{D}^\star$ and $\mathcal{D}^\diamond$. The second part (Lemma G.10) focuses on constructing the remaining MDP and the $\mathcal{D}^\triangle$ that contains the action chunk that π^+_{ac} picks where $\hat{V}^+_{\mathrm{ac}}$ overestimates the value of this action chunk by $\vartheta_h^L + \frac{\gamma^h \min(\vartheta^L, \vartheta_h^G)}{1-\gamma^h}$. Finally, we assemble these two results (combining $\mathcal{D}^\star$, $\mathcal{D}^\diamond$, $\mathcal{D}^\triangle$) to show that the MDP and the mixture data achieve our upper-bound *exactly*.

Lemma G.9 ("The Castle")**.** *For $\delta \in (0,1)$, $\vartheta_h^G < \frac{\gamma-\gamma^h}{2(1-\gamma)}$, consider a 2-state, 2-action MDP in Figure 10. Let there be two data distributions, $\mathcal{D}^\star$ and $\mathcal{D}^\diamond$. $\mathcal{D}^\star$ is collected by the following optimal closed-loop policy from X and Y:*

$$\pi^\star(X) = 0, \pi^\star(Y) = 1. \tag{208}$$

$\mathcal{D}^\diamond$ is collected by the following optimal closed-loop policy from X and Y:

$$\pi^\diamond(X) = 1, \pi^\diamond(Y) = 0. \tag{209}$$

Let $\mathcal{D}$ be a mixture of $\mathcal{D}^\star$ and $\mathcal{D}^\diamond$ with

$$P_{\mathcal{D}} = (1-\varsigma)P_{\mathcal{D}^\star} + \varsigma P_{\mathcal{D}^\diamond}. \tag{210}$$

There exists $c_1 \in (0, 1/2)$ such that

1. *$\mathcal{D}^\star$ and $\mathcal{D}^\diamond$ both individually exhibits 0-variability in optimality conditioned on s_t, a_t for all $s_t, a_t \in \operatorname{supp}(P_{\mathcal{D}}(s_t, a_t))$,*
2. *$\mathcal{D}$ exhibits ϑ_h^G-variability in optimality conditioned on $s_t, a_{t:t+h}$ for all $s_t, a_{t:t+h} \in \operatorname{supp}(P_{\mathcal{D}}(s_t, a_{t:t+h}))$,*

and

$$\hat{V}^+_{\mathrm{ac}}(X) = \hat{V}^+_{\mathrm{ac}}(Y) = \frac{1-\gamma+\varsigma(\gamma-\gamma^h)}{2(1-\gamma^h)(1-\gamma)} - \frac{\varsigma\vartheta_h^G}{1-\gamma^h}. \tag{211}$$

Proof. Set

$$c_1 = \frac{(1-\gamma)\vartheta_h^G}{\gamma-\gamma^h}. \tag{212}$$

We first check whether $c_1 \in (0, 1/2)$. For the upper-bound, it is clear that $c_1 < 1/2$ because $\vartheta_h^G < \frac{\gamma-\gamma^h}{2(1-\gamma)}$. For the lower-bound, $c > 0$ because all terms in the fraction are positive.

We now check the two optimality variability conditions. The first (local) one is trivial because $\pi^\diamond$ always receives $r = 1/2 - c_1$ and $\pi^\star$ always receives $r = 1/2$, and the optimal value for X and Y are both $V^\star(X) = V^\star(Y) = \frac{1}{2(1-\gamma)}$.

Next, we check the second (global) condition by analyzing all possible states and action chunks in $\mathcal{D}$. We observe that for any $a_{t:t+h}$ that starts with $a_t = 0$, we have

$$\tilde{Q}_{\min}(X, a_{t:t+h}) = \frac{1-2c_1(\gamma-\gamma^h)}{2(1-\gamma)}, \tag{213}$$

$$\tilde{Q}_{\max}(X, a_{t:t+h}) = \frac{1}{2(1-\gamma)}, \tag{214}$$

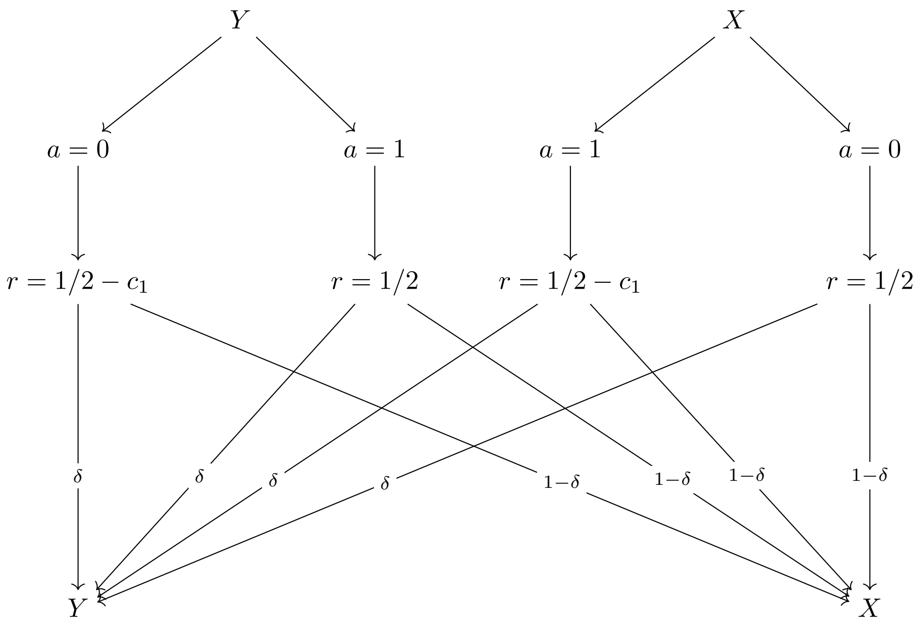


Figure 10: **MDP construction Part 1 for Theorem F.4 (“the castle”).** This diagram describes state X and Y and how actions $a = 0$ and $a = 1$ transition between them. The main purpose of this construction is to make $\hat{V}^+_{\mathrm{ac}}(X)$ underestimate $V^\star$ by exactly $\vartheta^G_h/(1-\gamma^h)$. This allows the action chunk that appears in the second part of the construction to be preferred (by π^+_{ac}) over the action chunks that start with with $a = 0$ or $a = 1$.

which gives

$$\tilde{Q}_{\max}(X, a_{t:t+h}) - \tilde{Q}_{\min}(X, a_{t:t+h}) = \vartheta^G_h. \tag{215}$$

By symmetry, we also have

$$\tilde{Q}_{\max}(Y, a_{t:t+h}) - \tilde{Q}_{\min}(Y, a_{t:t+h}) = \vartheta^G_h. \tag{216}$$

for all $a_{t:t+h}$ that starts with $a_t = 1$.

Now, for any $a_{t:t+h}$ that starts with $a_t = 1$, we have

$$\tilde{Q}_{\min}(X, a_{t:t+h}) = \frac{\gamma - 2c_1(\gamma - \gamma^h)}{2(1-\gamma)}, \tag{217}$$

$$\tilde{Q}_{\max}(X, a_{t:t+h}) = \frac{\gamma}{2(1-\gamma)}, \tag{218}$$

which admits the same gap as the case when $a_t = 0$. The same also holds for Y with $a_t = 1$. Thus, $\mathcal{D}$ exhibits ϑ^G_h-variability in optimality conditioned on $s_t, a_{t:t+h}$ for all $s_t, a_{t:t+h} \in \mathrm{supp}(P_{\mathcal{D}}(s_t, a_{t:t+h}))$.

Finally, we check for the value,

$$\begin{aligned}
\hat{V}^+_{\mathrm{ac}}(X) = \hat{V}^+_{\mathrm{ac}}(Y) &= (1-\varsigma)/2 + \varsigma(1/2 + (1-2c_1)\frac{\gamma-\gamma^h}{2(1-\gamma)}) \\
&= \frac{1}{1-\gamma^h}\left[1/2 + \varsigma\frac{(1-2c_1)(\gamma-\gamma^h)}{2(1-\gamma)}\right] \\
&= \frac{1}{2(1-\gamma^h)}\left[1 + \varsigma\frac{\gamma-\gamma^h-2(1-\gamma)\vartheta^G_h}{1-\gamma}\right] \\
&= \frac{1-\gamma+\varsigma(\gamma-\gamma^h)}{2(1-\gamma^h)(1-\gamma)} - \frac{\varsigma\vartheta^G_h}{1-\gamma^h},
\end{aligned} \tag{219}$$

as desired. □

Lemma G.10 ("The Flower"). *Assume $\vartheta_h^G \in \left(0, \frac{1-\gamma^h}{8}\right]$, $\vartheta_h^L \in \left(0, \frac{\gamma-\gamma^h}{4(1-\gamma)}\right]$, $\gamma \in (0,1)$, and Consider a 5-state, 3-action MDP in Figure 11 building on top of the transitions that already in Figure 10. Let $\mathcal{D}^{\triangle}$ be a data distribution induced by a cycling, time-dependent (with a time cycle length of h) policy $\pi^{\triangle}$ (we use the subscript to indicate the time step from 0 to $h-1$):*

$$\pi_0^{\triangle}(s_t = X) = \pi_0^{\triangle}(s_t = \tilde{X}) = 2, \tag{220}$$

$$\pi_0^{\triangle}(s_t = Y) = 3 \tag{221}$$

$$\pi_k^{\triangle}(s_{t+k} = \tilde{C}) = \pi_k^{\triangle}(s_{t+k} = \tilde{D}) = 2, \quad \forall k \in \{1, 2, \cdots, h-2\}, \tag{222}$$

$$\pi_k^{\triangle}(s_{t+h-1} = \tilde{C}) = \pi_k^{\triangle}(s_{t+h-1} = \tilde{D}) = 0, \tag{223}$$

$$\pi_k^{\triangle}(s_{t+k} = X) = 0, \quad \forall k \in \{1, 2, \cdots, h-1\}, \tag{224}$$

$$\pi_k^{\triangle}(s_{t+k} = Y) = 1, \quad \forall k \in \{1, 2, \cdots, h-1\}. \tag{225}$$

Let $\hat{V}_{\text{ac}}^{+}$ be the nominal value of the action chunking policy π_{ac}^{+} learned from $\mathcal{D}^{\triangle}$ and let

$$\Delta = \vartheta_h^L + \frac{\vartheta_h^G}{1-\gamma^h} + \frac{\gamma^h \min(\vartheta_h^G, \vartheta_h^L)}{1-\gamma^h}. \tag{226}$$

For any $c \in \left[0, \frac{\gamma-\gamma^h}{4(1-\gamma^h)}\right)$, there exists some $0 < c_2 \le 1/2$, $0 < c_3 \le 1/2$, $\delta, \delta_2 \in (0,1)$, such that for every $0 < \tilde{\Delta} < \min\left(\Delta, \frac{2\vartheta_h^G}{1-\gamma^h}\right)$,

1. *$\mathcal{D}^{\triangle}$ exhibits 0-variability in optimality conditioned on $s_t, a_{t:t+h}$ for all $s_t, a_{t:t+h} \in \text{supp}(P_{\mathcal{D}^{\triangle}}(s_t, a_{t:t+h}))$,*
2. *$\mathcal{D}^{\triangle}$ exhibits ϑ_h^L-variability in optimality conditioned on s_t, a_t for all $s_t, a_t \in \text{supp}(P_{\mathcal{D}^{\triangle}}(s_t, a_t))$,*

and

$$\hat{V}_{\text{ac}}^{+}(X) = \frac{1}{2(1-\gamma)} - \frac{\vartheta_h^G}{1-\gamma^h} + \tilde{\Delta}, \tag{227}$$

$$V^{\star}(X) - V^{\bullet}(X) = \frac{\Delta - \tilde{\Delta}}{1-\gamma}, \tag{228}$$

$$V^{\star}(X) - V_{\text{ac}}^{+}(X) \ge \frac{c}{1-\gamma}, \tag{229}$$

$$V^{\star}(X) - V_{\text{ac}}^{\star}(X) \ge \frac{c}{1-\gamma}. \tag{230}$$

Proof. Without the loss of generality, we assume we always start from state X. Due to symmetry, the same analysis applies to state Y (with the first action being $a_t = 3$ rather than $a_t = 2$).

Due to cycling nature of the data collection policy, we observe that all action chunks starting from X are in the form of $a_{t:t+h} = (2, \underbrace{\cdots}_{\text{0's and 1's}})$ or $a_{t:t+h} = (2, 2, \cdots, 2, 0)$. These two possibilities correspond to two different paths that the data collection policy takes:

- $a^{\circ}_{t:t+h} = (2, \underbrace{\cdots}_{\text{0's and 1's}})$: Stay in either X or Y. The agent going on this path receives a constant reward of $1/2$ except the first step where it receives a reward of $(1-c_2)/2$.
- $a^{\triangle}_{t:t+h} = (2, 2, \cdots, 2, 0)$: Visit $\tilde{C}$ and then stays there for $h-1$ until it goes out with $a = 0$ to visit $\tilde{X}$. The agent going on this path receives a constant reward of $(1+c_3)/2$ except the first step where it receives a reward of $(1-c_2)/2$.

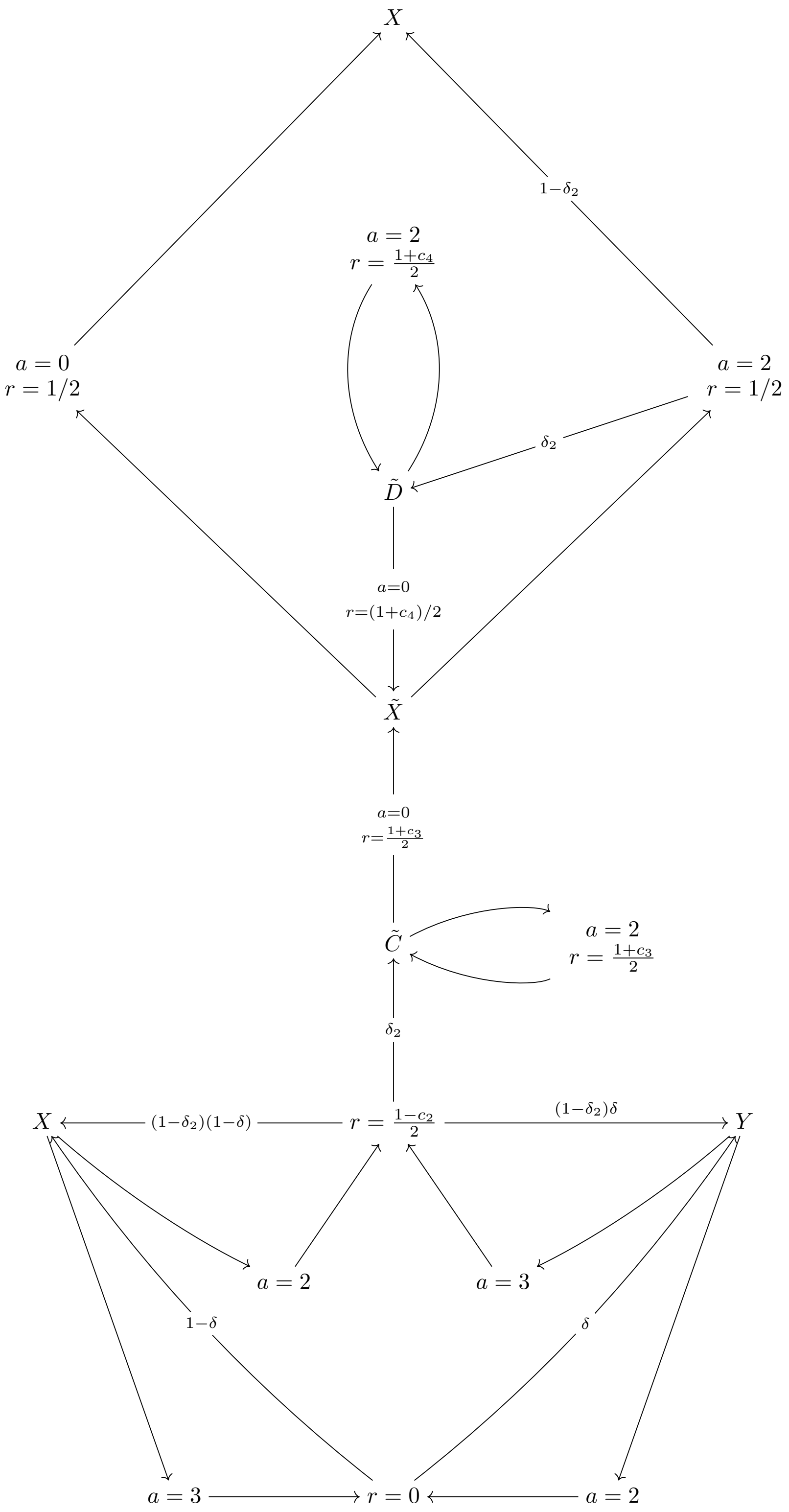


Figure 11: **MDP construction Part 2 for Theorem F.4 (“the flower”).** This diagram describes the remaining states $\tilde{C}$, $\tilde{D}$ and $\tilde{X}$, and what actions $a = 2$ and $a = 3$ do in state X and Y. The main purpose of this construction is to make $\hat{V}^{+}_{\text{ac}}(X)$ overestimate the optimal value of the action chunks that π^{+}_{ac}, $Q^{\star}(X, a^{+}_{t})$, by exactly $\vartheta^{L}_{h} + \gamma^{h} \min(\vartheta^{L}_{h}, \vartheta^{G}_{h})/(1 - \gamma^{h})$.

Similarly, all action chunks starting from $\tilde{X}$ are in the form of $a_{t:t+h} = (2, \underbrace{\cdots}_{\text{0's and 1's}})$ or $a_{t:t+h} = (2, 2, \cdots, 2, 0)$. These two possibilities correspond to two different paths that the data collection policy takes:

- $a^{\circ}_{t:t+h} = (2, \underbrace{\cdots}_{\text{0's and 1's}})$: Stay in either X or Y. The agent going on this path receives a constant reward of $1/2$.
- $a^{\triangle}_{t:t+h} = (2, 2, \cdots, 2, 0)$: Visit $\tilde{C}$ and then stays there for $h-1$ until it goes out with $a = 0$ to visit $\tilde{X}$. The agent going on this path receives a constant reward of $(1+c_4)/2$ except the first step where it receives a reward of $1/2$.

Now, we divide up the problem into two cases depending on the relative values of ϑ^L_h and ϑ^G_h.

1. Case $\vartheta^L_h \geq \vartheta^G_h$:

Set

$$c_2 = 2\left[\vartheta^L_h + \frac{(1+\gamma^h)\vartheta^G_h}{1-\gamma^h}\right] - 2\tilde{\Delta} > 0, \tag{231}$$

$$c_3 = \frac{2(1-\gamma)\vartheta^L_h}{\gamma - \gamma^h} > 0, \tag{232}$$

$$c_4 = \frac{2(1-\gamma)\vartheta^G_h}{\gamma - \gamma^h} > 0. \tag{233}$$

Next, we check that $c_2, c_3, c_4 \leq 1/2$.

We first observe that

$$(1-\gamma)(1-\gamma^h) - 2(\gamma - \gamma^h) = 1 - 3\gamma + \gamma^h(\gamma+1) \leq 1 - 3\gamma + \gamma(\gamma+1) = (1-\gamma)^2 \geq 0. \tag{234}$$

Dividing both sides by $8(1-\gamma)$ yields

$$\frac{1-\gamma^h}{8} \geq \frac{\gamma - \gamma^h}{4(1-\gamma)} \geq \vartheta^L_h \geq \vartheta^G_h. \tag{235}$$

Now, using the inequality above, we have

$$\begin{aligned} c_2 &= 2\left[\vartheta^L_h + \frac{(1+\gamma^h)\vartheta^G_h}{1-\gamma^h}\right] - \tilde{2}\Delta \\ &\leq 2\left[\vartheta^L_h + \frac{(1+\gamma^h)\vartheta^L_h}{1-\gamma^h}\right] \\ &\leq \frac{4\vartheta^L_h}{1-\gamma^h} \\ &\leq 1/2. \end{aligned} \tag{236}$$

Furthermore,

$$c_4 \leq c_3 = \frac{2(1-\gamma)\vartheta^L_h}{\gamma - \gamma^h} \leq 1/2. \tag{237}$$

Next, we check the data distribution $\mathcal{D}^{\triangle}$ satisfies both optimality variability conditions. We first note that we only need to check for $s_t \in \{X, \tilde{X}\}$ because all other states are out of the support due to the cycling nature of the data collection policies. The first (global) optimality condition is trivial because the h-step reward received is deterministic conditioned on $a_{t:t+h} \in \{a^{\circ}_{t:t+h}, a^{\triangle}_{t:t+h}\}$, and the optimal value of $V^\star(s_{t+h})$ is always $\frac{1}{2(1-\gamma)}$. This leads to 0-variability in optimality conditioned on $s_t, a_{t:t+h}$. For the second (local) optimality condition, we check the difference in optimality for two paths from $s_t, a_t = 2$ for both $s_t = X$ and $s_t = \tilde{X}$.

For $s_t = X$, the optimality gap is

$$c_3 \frac{\gamma - \gamma^h}{2(1-\gamma^h)} = \vartheta_h^L. \tag{238}$$

For $s_t = \tilde{X}$, the optimality gap is

$$c_4 \frac{\gamma - \gamma^h}{2(1-\gamma^h)} = \vartheta_h^G \leq \vartheta_h^L. \tag{239}$$

This concludes that the second (local) optimality condition is also satisfied.

Next, we first analyze which action chunk π_{ac}^+ prefers by computing $\hat{Q}_{\text{ac}}^+$'s:

$$\hat{Q}_{\text{ac}}^+(X, a_{t:t+h}^{\circ}) = \frac{1}{2}\left[(1-c_2) + \frac{\gamma-\gamma^h}{1-\gamma}\right] + \gamma^h \hat{V}_{\text{ac}}^+(X), \tag{240}$$

$$\hat{Q}_{\text{ac}}^+(X, a_{t:t+h}^{\triangle}) = \frac{1}{2}\left[(1-c_2) + (1+c_3)\frac{\gamma-\gamma^h}{1-\gamma}\right] + \gamma^h \hat{V}_{\text{ac}}^+(\tilde{X}), \tag{241}$$

$$\hat{Q}_{\text{ac}}^+(\tilde{X}, a_{t:t+h}^{\circ}) = \frac{1}{2}\left[\frac{1-\gamma^h}{1-\gamma}\right] + \gamma^h \hat{V}_{\text{ac}}^+(X), \tag{242}$$

$$\hat{Q}_{\text{ac}}^+(\tilde{X}, a_{t:t+h}^{\triangle}) = \frac{1}{2}\left[1 + (1+c_4)\frac{\gamma-\gamma^h}{1-\gamma}\right] + \gamma^h \hat{V}_{\text{ac}}^+(\tilde{X}). \tag{243}$$

We first observe that

$$\begin{aligned}\hat{Q}_{\text{ac}}^+(\tilde{X}, a_{t:t+h}^{\triangle}) - \hat{Q}_{\text{ac}}^+(X, a_{t:t+h}^{\triangle}) &= \frac{1}{2}\left[c_2 - (c_3 - c_4)\frac{\gamma-\gamma^h}{1-\gamma}\right] \\ &= \vartheta_h^L + \frac{(1+\gamma^h)\vartheta_h^G}{1-\gamma^h} - \vartheta_h^L + \vartheta_h^G - \tilde{\Delta} \\ &= \frac{2\vartheta_h^G}{1-\gamma^h} - \tilde{\Delta} \\ &> 0.\end{aligned} \tag{244}$$

Also,

$$\hat{Q}_{\text{ac}}^+(\tilde{X}, a_{t:t+h}^{\circ}) - \hat{Q}_{\text{ac}}^+(X, a_{t:t+h}^{\circ}) = c_2 > 0 \tag{245}$$

Therefore,

$$\begin{aligned}\hat{V}_{\text{ac}}^+(X) &= \max(\hat{Q}_{\text{ac}}^+(X, a_{t:t+h}^{\circ}), \hat{Q}_{\text{ac}}^+(X, a_{t:t+h}^{\triangle})) \\ &< \max(\hat{Q}_{\text{ac}}^+(\tilde{X}, a_{t:t+h}^{\circ}), \hat{Q}_{\text{ac}}^+(\tilde{X}, a_{t:t+h}^{\triangle})) \\ &= \hat{V}_{\text{ac}}^+(\tilde{X}).\end{aligned} \tag{246}$$

Now, we can compare the values for the action chunks for X and $\tilde{X}$:

$$\hat{Q}_{\text{ac}}^+(X, a_{t:t+h}^{\triangle}) - \hat{Q}_{\text{ac}}^+(X, a_{t:t+h}^{\circ}) = c_3 \frac{\gamma-\gamma^h}{2(1-\gamma)} + \gamma^h(\hat{V}_{\text{ac}}^+(\tilde{X}) - \hat{V}_{\text{ac}}^+(X)) > 0, \tag{247}$$

$$\hat{Q}_{\text{ac}}^+(\tilde{X}, a_{t:t+h}^{\triangle}) - \hat{Q}_{\text{ac}}^+(\tilde{X}, a_{t:t+h}^{\circ}) = c_4 \frac{\gamma-\gamma^h}{2(1-\gamma)} + \gamma^h(\hat{V}_{\text{ac}}^+(\tilde{X}) - \hat{V}_{\text{ac}}^+(X)) > 0, \tag{248}$$

since $c_3, c_4 > 0$ and $h > 1, 0 < \gamma < 1$ (and thus $\frac{\gamma-\gamma^h}{1-\gamma} > 0$).

This concludes that $\pi_{\text{ac}}^+(X) = \pi_{\text{ac}}^+(\tilde{X}) = a_{t:t+h}^{\triangle} = (2, 2, \cdots, 2, 0)$ and thus

$$\hat{V}_{\text{ac}}^+(\tilde{X}) = \frac{1-\gamma+(\gamma-\gamma^h)(1+c_4)}{2(1-\gamma^h)(1-\gamma)}, \tag{249}$$

and

$$\begin{aligned}\hat{V}^{+}_{\mathrm{ac}}(X) &= \frac{1}{2}\left[(1-c_2) + (1+c_3)\frac{\gamma-\gamma^h}{1-\gamma}\right] + \frac{\gamma^h}{1-\gamma^h}\hat{V}^{+}_{\mathrm{ac}}(\tilde{X}) \\ &= \frac{1}{2(1-\gamma)} - \frac{\vartheta^G_h}{1-\gamma^h} + \frac{\tilde{\Delta}}{2}.\end{aligned} \tag{250}$$

We can now compute the remaining values as follows:

$$V^{\star}(X) = \frac{1}{2(1-\gamma)}, \tag{251}$$

$$Q^{\star}(X, a=2) = \frac{(1-c_2)(1-\gamma)+\gamma}{2(1-\gamma)}, \tag{252}$$

$$Q^{\bullet}(X, a=2) = \frac{1-c_2}{2(1-\gamma)}. \tag{253}$$

Substituting the value of c_2 yields

$$V^{\star}(X) - V^{\bullet}(X) = \frac{\vartheta^L_h}{1-\gamma} + \frac{(1+\gamma^h)\vartheta^G_h}{(1-\gamma)(1-\gamma^h)} - \frac{\tilde{\Delta}}{2(1-\gamma)}. \tag{254}$$

2. Case $\vartheta^L_h < \vartheta^G_h$:

Set

$$\Delta = 2\left[\frac{\vartheta^L_h + \vartheta^G_h}{1-\gamma^h}\right] \tag{255}$$

$$c_2 = 2\left[\frac{\vartheta^L_h + \vartheta^G_h}{1-\gamma^h}\right] - \tilde{\Delta} > 0, \tag{256}$$

$$c_3 = c_4 = \frac{2(1-\gamma)\vartheta^L_h}{\gamma-\gamma^h} > 0 \tag{257}$$

where again $\tilde{\Delta}$ is any value that satisfies $0 < \tilde{\Delta} \le \Delta$.

From the definitions above and the value range of ϑ^G_h ($\vartheta^G_h \le \frac{1-\gamma^h}{4}$), it is clear that

$$c_3 = c_4 < c_2 \le \frac{4\vartheta^G_h}{1-\gamma^h} \le \frac{2(1-\gamma)}{\gamma-\gamma^h} \le 1/2. \tag{258}$$

Next, we check the data distribution $\mathcal{D}^{\triangle}$ satisfies both optimality variability conditions. With the same argument as the previous case, we can quickly conclude that the global optimality condition is satisfied. We just need to show the remaining local optimality condition. We repeat the procedure from the previous case.

For $s_t = X$, the local optimality gap is

$$c_3\frac{\gamma-\gamma^h}{2(1-\gamma^h)} = \vartheta^L_h. \tag{259}$$

For $s_t = \tilde{X}$ the local optimality gap is the same because $c_4 = c_3$:

$$c_4\frac{\gamma-\gamma^h}{2(1-\gamma^h)} = \vartheta^L_h. \tag{260}$$

This concludes that the second (local) optimality condition is also satisfied for the second case.

Now, we can follow the same procedure as the previous case to show that $\hat{Q}^{+}_{\mathrm{ac}}(X, a^{\triangle}_{t:t+h}) - \hat{Q}^{+}_{\mathrm{ac}}(X, a^{\circ}_{t:t+h}) > 0$ and $\hat{Q}^{+}_{\mathrm{ac}}(\tilde{X}, a^{\triangle}_{t:t+h}) - \hat{Q}^{+}_{\mathrm{ac}}(\tilde{X}, a^{\circ}_{t:t+h}) > 0$.

This concludes that $\pi^{+}_{\mathrm{ac}}(X) = \pi^{+}_{\mathrm{ac}}(\tilde{X}) = a^{\triangle}_{t:t+h} = (2, 2, \cdots, 2, 0)$, and thus

$$\hat{V}^{+}_{\mathrm{ac}}(\tilde{X}) = \frac{1}{2}\left[\frac{1-\gamma+(1+c_3)(\gamma-\gamma^h)}{(1-\gamma)(1-\gamma^h)}\right], \tag{261}$$

and

$$\begin{aligned}\hat{V}^{+}_{\mathrm{ac}}(X) &= \frac{1}{2}\left[(1-c_2)+(1+c_3)\frac{\gamma-\gamma^h}{1-\gamma}\right] + \frac{\gamma^h}{1-\gamma^h}\hat{V}^{+}_{\mathrm{ac}}(\tilde{X}) \\ &= \frac{1}{2(1-\gamma)} - \frac{\vartheta^G_h}{1-\gamma^h} + \frac{\tilde{\Delta}}{2}.\end{aligned} \tag{262}$$

Repeating the same procedure as the previous case, we obtain

$$V^\star(X) - Q^\star(X, a=2) = \frac{\vartheta^L_h + \vartheta^G_h}{1-\gamma^h} - \tilde{\Delta}, \tag{263}$$

resulting in an optimality of

$$V^\star(X) - V^\bullet(X) = \frac{\vartheta^L_h + \vartheta^G_h}{(1-\gamma)(1-\gamma^h)} - \frac{\tilde{\Delta}}{1-\gamma}. \tag{264}$$

3. Sub-optimality of V^{+}_{ac}:

Finally, we can use a pretty crude upper-bound on the actual value of the action chunking policy π^{+}_{ac} (reparameterizing $\tilde{\delta}_2 = 1-(1-\delta_2)^h$):

$$V^{+}_{\mathrm{ac}}(X) \leq (1-\tilde{\delta}_2)\left[(1-c_2)/2 + \frac{\delta(\gamma-\gamma^h)}{2(1-\gamma)} + \gamma^h V^{+}_{\mathrm{ac}}(X)\right] + \frac{\tilde{\delta}_2}{1-\gamma} \tag{265}$$

$$\leq \frac{1-\tilde{\delta}_2}{2(1-\gamma^h)(1-\gamma)}\left[1-\gamma+\delta(\gamma-\gamma^h)\right] + \frac{\tilde{\delta}_2}{1-\gamma}. \tag{266}$$

Set $\delta = 1/2$, we have

$$V^{+}_{\mathrm{ac}}(X) \leq \frac{1-\tilde{\delta}_2}{2(1-\gamma^h)(1-\gamma)}\left[1-\gamma/2-\gamma^h/2\right] + \frac{\tilde{\delta}_2}{1-\gamma}. \tag{267}$$

We set

$$\delta_2 = 1 - \left[1 - \frac{\gamma-\gamma^h-4c(1-\gamma^h)}{2-3\gamma^h+\gamma}\right]^{1/h}, \tag{268}$$

which results in

$$\tilde{\delta}_2 = \frac{\gamma-\gamma^h-4c(1-\gamma^h)}{2-3\gamma^h+\gamma}. \tag{269}$$

It is clear that $0 < \delta_2 < 1$ because $c < \frac{\gamma-\gamma^h}{4(1-\gamma^h)}$ and $\frac{\gamma-\gamma^h}{2-3\gamma^h+\gamma} < 1$.

Substituting $\tilde{\delta}_2$ in the bound of $V^{+}_{\mathrm{ac}}(X)$ above, we obtain

$$V^\star(X) - V^{+}_{\mathrm{ac}}(X) \geq \frac{c}{1-\gamma}. \tag{270}$$

□

Proof of Theorem F.4. Let

$$\Delta = \vartheta^L_h + \frac{\vartheta^G_h}{1-\gamma^h} + \frac{\gamma^h \min(\vartheta^G_h, \vartheta^L_h)}{1-\gamma^h}. \tag{271}$$

Consider the 5-state, 3-action MDP constructed in Lemma G.9 and Lemma G.10 and a data distribution consisting of a mixture of three data distributions $\mathcal{D}^\star, \mathcal{D}^\diamond$ (from Lemma G.9) and $\mathcal{D}^\triangle$ (from Lemma G.10):

$$P_{\mathcal{D}} = \alpha(1-\varsigma)P_{\mathcal{D}^\star} + \varsigma P_{\mathcal{D}^\diamond} + (1-\alpha)P_{\mathcal{D}^\triangle}. \tag{272}$$

We set α to be any value between 0 and 1 (non-inclusive) and set ς as any positive value such that

$$\varsigma < \frac{(\gamma - \gamma^h) - 2\vartheta_h^G(1-\gamma) + 2\tilde{\Delta}(1-\gamma)(1-\gamma^h)}{(\gamma-\gamma^h) - 2\vartheta_h^G(1-\gamma)}, \tag{273}$$

where $\tilde{\Delta} = \sigma(1-\gamma) < \min(\vartheta_h^L, \vartheta_h^G) < \min(\Delta, \frac{2\vartheta_h^G}{1-\gamma^h})$ (satisfying the condition for $\tilde{\Delta}$ in Lemma G.10).

The numerator and the denominator are both positive:

$$(\gamma-\gamma^h) - 2\vartheta_h^G(1-\gamma) + 2\tilde{\Delta}(1-\gamma)(1-\gamma^h) > (\gamma-\gamma^h) - 2\vartheta_h^G(1-\gamma) > 0, \tag{274}$$

meaning such ς always exists.

Substituting the inequality to the result of Lemma G.9 results in

$$\frac{1-\gamma+\varsigma(\gamma-\gamma^h)}{2(1-\gamma^h)(1-\gamma)} - \frac{\varsigma\vartheta_h^G}{1-\gamma^h} < \frac{1}{2(1-\gamma)} - \frac{\vartheta_h^G}{1-\gamma^h} + \tilde{\Delta}, \tag{275}$$

which shows that π_{ac}^+ will always prefer $a_{t:t+h}^\triangle$ over action chunks in $\mathcal{D}^\star$ and $\mathcal{D}^\diamond$.

This means that the value $\hat{V}_{\text{ac}}^+$ and the action chunking policy π_{ac}^+ we learn from $\mathcal{D}$ coincides with these of $\mathcal{D}^\triangle$, allowing us to directly use the results of Lemma G.10.

Thus, we can conclude that

$$V^\star(s_t) - V_{\text{ac}}^+(s_t) \geq \frac{c}{1-\gamma}, \tag{276}$$

and

$$V^\star(X) - V^\bullet(X) = \frac{\Delta - \tilde{\Delta}}{1-\gamma} = \frac{\vartheta_h^L}{1-\gamma} + \frac{\vartheta_h^G}{(1-\gamma)(1-\gamma^h)} + \frac{\gamma^h \min(\vartheta_h^L, \vartheta_h^G)}{(1-\gamma)(1-\gamma^h)} - \sigma, \tag{277}$$

as desired.

□

G.12 Proof of Proposition F.6

Proposition F.6 (Deterministic Dynamics are Weakly Open-loop Consistent). *If a transition dynamics $\mathcal{M}$ is ε-deterministic, then any data $\mathcal{D}$ collected from $\mathcal{M}$ is weakly ε_h-open-loop consistent with respect to $\mathcal{M}$ for any $h \in \mathbb{N}^+$ as long as $\varepsilon_h \geq 3(1-(1-\varepsilon)^{h-1})$.*

Proof. Since T is ε-deterministic, it can be represented as $T(\cdot \mid s, a) = (1-\varepsilon)\delta_{f(s,a)} + \varepsilon\tilde{T}(\cdot \mid s, a)$ for some $f : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ and $\tilde{T} : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}$. Let $f(s, a_1, \cdots, a_h) = f(\cdots f(f(s, a_1), a_2) \cdots a_h)$

Let $I \in \{0, 1\}$ a binary indicator variable that is 1 if and only if

$$s_{t+k+1} = f(s_{t+k}, a_{t+k}), \forall k \in \{0, 1, 2, \cdots, h-1\} \tag{278}$$

Intuitively $I = 1$ when the trajectory is generated deterministically until but not including the last state s_h in the trajectory chunk.

From the fact that T is ε-deterministic, we know that

$$P_{\mathcal{D}}(I_h = 1) \geq (1-\varepsilon)^{h-1} \tag{279}$$

We also have

$$P_{\mathcal{D}}(a_{t:t+h} \mid s_t) = P_{\mathcal{D}}(I_h = 1)P_{\mathcal{D}}(a_{t:t+h} \mid s_t, I_h = 1) + P_{\mathcal{D}}(I_h = 0)P_{\mathcal{D}}(a_{t:t+h} \mid s_t, I = 0) \tag{280}$$

Then we have

$$D_{\mathrm{TV}}(P_{\mathcal{D}}(a_{t:t+h} \mid s_t) \parallel P_{\mathcal{D}}(a_{t:t+h} \mid s_t, I_h = 1)) \leq (1-(1-\varepsilon)^{h-1}) \tag{281}$$

If we transform each distribution of $a_{t:t+h}$ deterministically by $f(s_t, \cdot)$, by data processing inequality (DPI; Lemma G.4), we have

$$D_{\mathrm{TV}}\left(\mathbb{E}_{a_{t:t+h} \sim P_{\mathcal{D}}(\cdot \mid s_t)}\left[\delta_{f(s_t, a_{t:t+h})}\right] \parallel \mathbb{E}_{a_{t:t+h} \sim P_{\mathcal{D}}(\cdot \mid s_t, I_h = 1)}\left[\delta_{f(s_t, a_{t:t+h})}\right]\right) \leq (1-(1-\varepsilon)^{h-1}) \tag{282}$$

Similarly, we have

$$D_{\mathrm{TV}}(P_{\mathcal{D}}(a_{t:t+h+1} \mid s_t) \parallel P_{\mathcal{D}}(a_{t:t+h+1} \mid s_t, I_{h+1} = 1)) \leq (1-(1-\varepsilon)^{h}) \tag{283}$$

which can be also deterministically transformed by taking $a_{t:t+h+1} \mapsto (f(s_t, \cdot), a_{t+h})$ (again with DPI, Lemma G.4) to obtain

$$\begin{aligned} &D_{\mathrm{TV}}\Big(\mathbb{E}_{a_{t:t+h} \sim P_{\mathcal{D}}(\cdot \mid s_t)}\left[\pi^{\circ}_{\mathcal{D}}(a_{t+h} \mid s_t, a_{t:t+h})\mathbb{I}_{f(s_t, a_{t:t+h})}\right] \parallel \\ &\mathbb{E}_{a_{t:t+h} \sim P_{\mathcal{D}}(\cdot \mid s_t, I_{h+1} = 1)}\left[\pi^{\circ}_{\mathcal{D}}(a_{t+h} \mid s_t, a_{t:t+h}, I_{h+1} = 1)\mathbb{I}_{f(s_t, a_{t:t+h})}\right]\Big) \leq (1-(1-\varepsilon)^{h}) \end{aligned} \tag{284}$$

Now, if we analyze the distribution of s_{t+h} subject to the open-loop execution of the action sequence from $P_{\mathcal{D}}(\cdot \mid s_t)$ and break it up into the deterministic and the non-deterministic case, we get

$$\begin{aligned} \mathbb{E}_{a_{t:t+h} \sim P_{\mathcal{D}}(\cdot \mid s_t)}\left[T_{a_{t:t+h}}(\cdot \mid s_t)\right] = P_T(I = 1)\mathbb{E}_{a_{t:t+h} \sim P_{\mathcal{D}}(\cdot \mid s_t)}\left[\delta_{f(s_t, a_{t:t+h})}\right] + \\ P_T(I = 0)\mathbb{E}_{a_{t:t+h} \sim P_{\mathcal{D}}(\cdot \mid s_t)}\left[T_{a_{t:t+h}}(\cdot \mid s_t, I_h = 0)\right] \end{aligned} \tag{285}$$

Note that $P_T(I = 1)$ denotes the probability that an open-loop executed trajectory using $a_{t:t+h} \sim P_{\mathcal{D}}(\cdot \mid s_t)$ is deterministic. This is different from $P_{\mathcal{D}}(I_h = 1)$ because the latter is based on $P_{\mathcal{D}}(s_{t:t+h+1}, a_{t:t+h})$ whereas $P_T(I_h = 1)$ is based on the open-loop trajectory distribution: $P_{\mathcal{D}}(\cdot \mid s_t) \prod_{k=0}^{h-1} T(s_{t+k} \mid s_t, a_{t:t+k})$. They both admit the same lower bound of $2(1-(1-\varepsilon)^{h-1})$.

Therefore,

$$D_{\mathrm{TV}}\left(\mathbb{E}_{a_{t:t+h} \sim P_{\mathcal{D}}(\cdot \mid s_t)}\left[T_{a_{t:t+h}}(\cdot \mid s_t)\right] \parallel \mathbb{E}_{a_{t:t+h} \sim P_{\mathcal{D}}(\cdot \mid s_t)}\left[\delta_{f(s_t, a_{t:t+h})}\right]\right) \leq (1-(1-\varepsilon)^{h-1}) \tag{286}$$

Similarly for the state-action case, we can multiply both side by the same conditional distribution $\pi^{\circ}_{\mathcal{D}}(a_{t+h} \mid s_t, a_{t:t+h})$ which preserves the TV bound. For the left-hand side, we have

$$P^{\circ}_{\mathcal{D}}(s_{t+h}, a_{t+h} \mid s_t) = \mathbb{E}_{a_{t:t+h} \sim P_{\mathcal{D}}(\cdot \mid s_t)}\left[\pi^{\circ}_{\mathcal{D}}(a_{t+h} \mid s_t, a_{t:t+h})T_{a_{t:t+h}}(s_{t+h} \mid s_t)\right] \tag{287}$$

Therefore, we get

$$D_{\text{TV}}\big(P^{\circ}_{\mathcal{D}}(s_{t+h}, a_{t+h} \mid s_t) \parallel \mathbb{E}_{a_{t:t+h}\sim P_{\mathcal{D}}(\cdot|s_t)} \left[\pi^{\circ}_{\mathcal{D}}(a_{t+h} \mid s_t, a_{t:t+h})\mathbb{I}_{f(s_t,a_{t:t+h})}\right]\big) \\ \leq (1-(1-\varepsilon)^{h-1}) \tag{288}$$

We also have

$$P_{\mathcal{D}}(s_{t+h} \mid s_t) = (1-\varepsilon)^{h-1}P_{\mathcal{D}}(s_{t+h} \mid s_t, I=1) + (1-(1-\varepsilon)^{h-1})P_{\mathcal{D}}(s_{t+h} \mid s_t, I_h = 0) \tag{289}$$

Similarly, we have

$$D_{\text{TV}}(P_{\mathcal{D}}(s_{t+h} \mid s_t) \parallel P_{\mathcal{D}}(s_{t+h} \mid s_t, I_h = 1)) \\ = D_{\text{TV}}\big(P_{\mathcal{D}}(s_{t+h} \mid s_t) \parallel \mathbb{E}_{a_{t:t+h}\sim P_{\mathcal{D}}(\cdot|s_t,I_h=1)} \left[\delta_{f(s_t,a_{t:t+h})}\right]\big) \leq (1-(1-\varepsilon)^{h-1}) \tag{290}$$

For state-action, we can also get

$$P_{\mathcal{D}}(s_{t+h}, a_{t+h} \mid s_t) = (1-\varepsilon)^{h}P_{\mathcal{D}}(s_{t+h}, a_{t+h} \mid s_t, I_{h+1} = 1) \\ + (1-(1-\varepsilon)^{h})P_{\mathcal{D}}(s_{t+h}, a_{t+h} \mid s_t, I_{h+1} = 0) \tag{291}$$

which can be turned into the TV distance bound:

$$D_{\text{TV}}(P_{\mathcal{D}}(s_{t+h}, a_{t+h} \mid s_t) \parallel P_{\mathcal{D}}(s_{t+h}, a_{t+h} \mid s_t, I_{h+1} = 1)) \\ = D_{\text{TV}}\Big(P_{\mathcal{D}}(s_{t+h}, a_{t+h} \mid s_t) \parallel \\ \mathbb{E}_{a_{t:t+h}\sim P_{\mathcal{D}}(\cdot|s_t,I_{h+1}=1)} \left[\pi^{\circ}_{\mathcal{D}}(a_{t+h} \mid s_t, a_{t:t+h}, I_{h+1} = 1)\mathbb{I}_{f(s_t,a_{t:t+h})}\right]\Big) \\ \leq (1-(1-\varepsilon)^{h}) \tag{292}$$

Connecting all three total variation inequality (Equations (282), (286) and (290)) together, we get

$$D_{\text{TV}}\big(P_{\mathcal{D}}(s_{t+h} \mid s_t) \parallel \mathbb{E}_{a_{t:t+h}\sim P_{\mathcal{D}}(\cdot|s_t)} \left[T_{a_{t:t+h}}(\cdot \mid s_t)\right]\big) \leq 3(1-(1-\varepsilon)^{h-1}) \leq \varepsilon_h \tag{293}$$

Connecting all three total variable inequality for state-action (Equations (284), (287) and (292)) together, we get

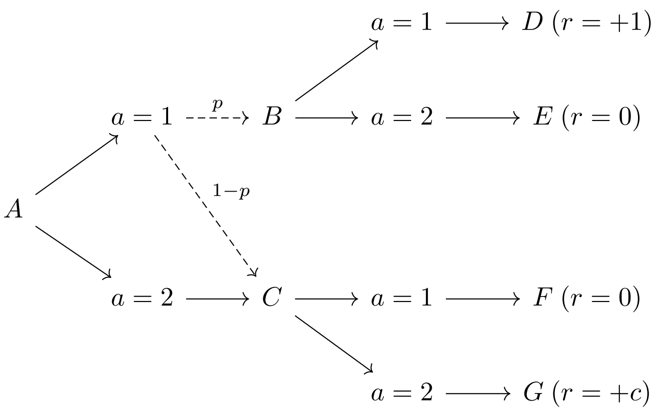
$$D_{\text{TV}}(P^{\circ}_{\mathcal{D}}(s_{t+h-1}, a_{t+h-1} \mid s_t) \parallel P_{\mathcal{D}}(s_{t+h}, a_{t+h} \mid s_t)) \leq 3 - 2(1-\varepsilon)^{h-1} - (1-\varepsilon)^{h-2} \\ \leq 3(1-(1-\varepsilon)^{h-1}) \\ \leq \varepsilon_h \tag{294}$$

Therefore, $\mathcal{D}$ is ε_h-open-loop consistent as desired. □

H A PATHOLOGICAL FAILURE OF ACTION CHUNKING POLICIES WITHOUT THE STRONG OPEN-LOOP CONSISTENCY ASSUMPTION

In this section, we show an example where the optimal action chunking policy defined in Equation (17) can be highly suboptimal in the absence of the strong open-loop consistency condition.

We define an MDP as follows. Let $\mathcal{S} = \{A, B, C, D, E, F, G\}$ and $\mathcal{A} = \{1, 2\}$. Define the transition dynamics and reward function as shown in the diagram below:



where $p, c \in (0, 1)$ are real numbers and dotted lines denote stochastic transitions. For simplicity, assume that the MDP has a length-2 finite horizon with $\gamma = 1$, and the reward function depends only on states ($r(A) = r(B) = r(C) = r(E) = r(F) = 0$, $r(D) = 1$, and $r(G) = c$). Assume that the dataset is collected by a policy $\pi_{\mathcal{D}}$ defined as $\pi_{\mathcal{D}}(A) = 1$ (with probability 0.5) or 2 (with probability 0.5), $\pi_{\mathcal{D}}(B) = 1$ (with probability 1), and $\pi_{\mathcal{D}}(C) = 2$ (with probability 1).

Then, we have the following:

$$P_{\mathcal{D}}(A, (1, 1)) = D, \ R(A, (1, 1)) = 1, \tag{295}$$

$$P_{\mathcal{D}}(A, (1, 2)) = G, \ R(A, (1, 2)) = c, \tag{296}$$

$$P_{\mathcal{D}}(A, (2, 2)) = G, \ R(A, (2, 2)) = c, \tag{297}$$

where we denote action chunks as a tuple and slightly abuse notation to denote deterministic outputs of $P_{\mathcal{D}}(\cdot \mid s_0, a_{0:2})$ (*e.g.*, $P_{\mathcal{D}}(A, (1, 1)) = D$ indicates that all length-2 trajectories in $\mathcal{D}$ from state A with $a_0 = a_1 = 1$ have $s_2 = D$ with probability 1). From this, we can compute $\hat{Q}^+_{\text{ac}}$ as follows:

$$\hat{Q}^+_{\text{ac}}(A, (1, 1)) = 1, \tag{298}$$

$$\hat{Q}^+_{\text{ac}}(A, (1, 2)) = c, \tag{299}$$

$$\hat{Q}^+_{\text{ac}}(A, (2, 2)) = c. \tag{300}$$

Then, assuming the missing data has a Q-value of 0 (*i.e.*, $\hat{Q}^+_{\text{ac}}(A, (2, 1)) = 0$), the optimal action chunking policy is defined as $\hat{\pi}^+_{\text{ac}}(A) = (1, 1)$ (Equation (17)).

The true value of this action chunking policy is p. However, if p is small enough and c is large enough, the optimal strategy in this MDP is to always choose $(a_0, a_1) = (2, 2)$, in which case the agent receives a constant return of c. The suboptimality in this example is therefore $c - p$, which can be made arbitrarily close to 1 (the maximum possible regret in any finite, length-2 sparse-reward MDP with a terminal reward bounded by $[0, 1]$). This shows a pathological failure of an action chunking policy without the strong open-loop consistency assumption.

I Additional Related Work on Hierarchical Reinforcement Learning

Hierarchical reinforcement learning methods (Dayan & Hinton, 1992; Dietterich, 2000; Peng et al., 2017; Riedmiller et al., 2018; Shankar & Gupta, 2020; Pertsch et al., 2021; Gehring et al., 2021; Xie et al., 2021) solve tasks by typically leveraging a bi-level structure: a set of low-level/skill policies that directly interact with the environment and a high-level policy that selects among low-level policies. The low-level policies can also be learned via online RL (Kulkarni et al., 2016; Vezhnevets et al., 2016; 2017; Nachum et al., 2018) or offline pre-training on a prior dataset (Paraschos et al., 2013; Merel et al., 2018; Ajay et al., 2021; Pertsch et al., 2021; Touati et al., 2022; Nasiriany et al., 2022; Hu et al., 2023; Frans et al., 2024; Chen et al., 2024; Park et al., 2024b). In the options framework, these low-level policies are often additionally associated with initiation and termination conditions that specify when and for how long these actions can be used (Sutton et al., 1999; Menache et al., 2002; Chentanez et al., 2004; Şimşek & Barto, 2007; Konidaris, 2011; Daniel et al., 2016; Srinivas et al., 2016; Fox et al., 2017; Bacon et al., 2017; Bagaria & Konidaris, 2019; Bagaria et al., 2024; de Mello Koch et al., 2025). A long-lasting challenge in HRL is optimization stability because the high-level policy needs to optimize for an objective that is shaped by the constantly changing low-level policies (Nachum et al., 2018). Prior work (Ajay et al., 2021; Pertsch et al., 2021; Wilcoxson et al., 2024) avoided this by first pre-train low-level policies and then keep them frozen during the optimization of the high-level policy. Macro-actions (McGovern & Sutton, 1998; Durugkar et al., 2016), or action chunking (Zhao et al., 2023) is another form of temporally extended action, a special case of the low-level policies often considered in HRL, options literature, where a short horizon of actions are predicted all at once and executed in open loop. Such approach collapses the bi-level structure, conveniently side stepping optimization instability, and when combined with Q-learning, has shown great empirical successes in offline-to-online RL setting (Seo et al., 2024; Li et al., 2025b). Action chunking policies need to predict multiple actions open-loop, which can be difficult to learn and sacrifice reactivity. Our approach regains policy reactivity by predicting and executing only a partial action chunk, while still learning with the fully chunked critic for TD-backup. This design preserves the value propagation benefits of chunked critic without relying on fully open-loop action chunking policies, allowing our approach to work well on a wider range of tasks.

J Intuition behind optimality variability

In this section, we provide more intuition on the definition of optimality variability. With Definition 4.10, if we pick X to be the current state and the current action (*i.e.*, s_t, a_t), a bounded optimality variability subject to such conditioning means that as long as we observe the initial action (*e.g.*, picking up the cube), the optimality of the outcome after h-steps does not vary too much (*e.g.*, does not misdrop the object that fails the task immediately). It turns out that if (1) the data distribution is a mixture of a bunch of data sources where the optimality variability conditioned on the *current actions* is bounded within each data source, and additionally (2) the optimality variability conditioned on the *current action chunks* is bounded globally across mixture, we can form a much stronger bound on the optimality of $\pi^\bullet$. It is worth noting that the second optimality variability condition is *much weaker* than the first one because it is conditioned on the event where we observe the state s_t and the entire action chunk $a_{t:t+h}$ (rather than the first action a_t). For example, for data mixture where each pair of data distributions has non-overlapping support on the action chunks, the second condition is trivially implied by the first condition.