In Appendix A we introduce some basic definitions that are needed for our theoretical results. In Appendix B we provide sufficient conditions for Assumption that were mentioned in the main text. In Appendix C and Appendix D we prove the error bounds for PPI and PQI. In Appendix E and Appendix F we present more details of our experimental results.

A Definition of auxiliary MDP and policy projection

First we introduce the definition of an auxiliary MDP $M'$ based on $M$: each state in $M$ has an absorbing action which leads to a self-looping absorbing state. All the other dynamics are preserved.

Rewards are 0 for the absorbing action and unchanged elsewhere. More formally: The auxiliary MDP $M'$ given $M = \langle S, A, R, P, \gamma, \rho \rangle$ is defined as $M' = \langle S', A', R', P', \gamma', \rho' \rangle$, where $S' = S \cup \{s_{abs}\}$, $A' = A \cup \{a_{abs}\}$. $R'$ and $P'$ are the same as $R$ and $P$ for all $(s, a) \in S \times A$.

$R'(s, a)$ if $s = s_{abs}$ or $a = a_{abs}$ is a point mass on 0, and $P'(s, a)$ if $s = s_{abs}$ or $a = a_{abs}$ is a point mass on $s_{abs}$. A data set $D$ generated from distribution $\mu$ on $M$ is also from the distribution $\mu$ on $M'$, since all distributions on $S \times A$ are the same between the two MDPs. This MDP is used only to perform our analysis about the error bounds on the algorithm, and is not needed at all for executing Algorithm 1 and 2. As some of the notations is actually a function of the MDP, we clarify the usage of notation w.r.t. $M/M'$ in the appendix:

1. Policy value functions $V^\pi/F^\pi$ and Bellman operators $T/T^\pi$ correspond to $M'$ unless they have additional subscripts.
2. The definition of $F$, $\Pi$, $T$, $T^\pi$, $\hat{\mu}$ is independent of the change from $M$ to $M'$.
3. $\mu$ is also a distribution over $S' \times A'$. The definition of $\zeta$ will be extended to $S' \times A'$ as follow:

$$
\zeta(s, a) = \begin{cases} 
1 & (\hat{\mu}(s, a) \geq b) \quad s \in S, \ a \in A \\
0 & s = s_{abs} \text{ or } a = a_{abs}
\end{cases}
$$

(That means there is only one version of $\mu$ and $\zeta$ across $M$ and $M'$, instead of like we have $T_{\hat{\mu}}$, and $T_{M'}$ for $M$ and $M'$.)

Recall the definition of semi-norm of any function of state-action pairs. For any function $g : S' \times A' \to \mathbb{R}$, $\nu \in \Delta(S' \times A')$, and $p \geq 1$, define the shorthand $\|g\|_{p, \nu} := (E_{(s, a) \sim \nu} ||g(s, a)||^p)^{1/p}$. With some abuse of notation, later we also use this norm for $\nu \in \Delta(S \times A)$ (specifically, $\mu$) by viewing the probability of $\nu$ on additional $(s, a)$ pairs as zero. Given a policy $\pi$, let $\eta^\pi(s)$ be the marginal distribution of $s_h$ under $\pi$, that is, $\eta^\pi(s) := \text{Pr}[s_h = s | s_0 \sim p, \pi]$. $\eta^\pi(s, a) = \eta^\pi(s) \pi(a | s)$, and $\eta^\pi(s, a) = (1 - \gamma) \sum_{h=0}^\infty \gamma^h \eta^\pi(s, a)$. We also use $P(s, a)$ and $P(\nu)$ to denote the next state distribution given a state action pair or given the current state action distribution.

The norm $\| \cdot \|_{p, \nu}$ are defined over $S' \times A'$. Though for the input space of function $f \in F$ is $S \times A$, the norm can still be well-defined. All of the norm would not need the value of $f(s, a)$ on $s = s_{abs}$ or $a = a_{abs}$, because the distribution does not cover those $(s, a)$, or the $f$ inside of the norm is multiplied by other function that is zero for those $(s, a)$.

We first formally state an obvious result about policy value in $M$ and $M'$.

**Lemma 1.** For any policy $\pi$ that only have non-zero probability for $a \in A$, $v^\pi_{M'} = v^\pi_{M}$. 

**Proof.** By the definition of $M'$, $P$ and $R$ are the same with $M$ over $S \times A$.

$$
v^\pi_{M'} = E_M \left[ \sum_{t=0}^h \gamma^t r_t | s_0 \sim p, \pi \right] = E_M' \left[ \sum_{t=0}^h \gamma^t r_t | s_0 \sim p, \pi \right] = v^\pi_{M'}
$$

For the readability we repeat the Definition here

**Definition 1** ($\zeta$-constrained policy set). Let $\Pi^{\zeta}_{\bar{\epsilon}}$ be the set of policies $\pi : S \to \Delta(A)$ such that

$$
\text{Pr}(\zeta(s, a) = 0 | \pi) \leq \epsilon_\zeta. That is

(1 - \gamma) \sum_{h=0}^\infty \gamma^h E_{s, a \sim \eta^\pi_h} [1(\zeta(s, a) = 0)] \leq \epsilon_\zeta
$$

(4)
Now we introduce another constrained policy set. Different from $\zeta$-constrained policy set which we introduced in Definition [1], this policy set is on $M'$ instead of $M$ and the policy is forced to take action $a_{abs}$ when $\zeta(s, a) = 0$ for all $a$. The reason we introduce this is to help us formally analyze the (lower bound of) performance of the resulting policy. We essentially treat any action taken outside of the support to be $a_{abs}$. Later we will define a projection to achieve that and show results about how the policy value changes after projection.

**Definition 2** (strong $\zeta$-constrained policy set). Let $\Pi_{\zeta SC}^{all}$ be the set of all policies $S' \to \Delta(A')$ such that for $\forall (s, a) \pi(a|s) > 0$ then 1) $\zeta(s, a) > 0$, or 2) $a = a_{abs}$.

Notice that for $\zeta$-constrained policy set we have no requirement for $\pi$ if for any action $\zeta(s, a)$ is zero.

For strong $\zeta$-constrained policy set we enforce $\pi$ to take action $a_{abs}$. The second difference is $\zeta$-constrained policy set requires the condition holds for $s, a$ that is reachable, which means $\eta_H^\zeta(s) > 0$ and $\pi(a|s) > 0$. Here we require the same condition holds for any $s, a$ such that $\pi(a|s) > 0$. In general, this is a stronger definition. However, we can show that for any policy in $\zeta$-constrained policy set, it can be mapped to a policy in strong $\zeta$-constrained policy set, with changing value bounds. Since we only need to change the behavior of policy in the state actions such that the state actions that $\zeta = 0$, the value of policy will not be much different.

Now we define a projection that maps any policy to $\Pi_{\zeta SC}^{all}$.

**Definition 3** ($\zeta$-constrained policy projection). $(\Xi(\pi))(a|s) equals \zeta(s, a)\pi(a|s)$ if $a \in A$, and equals $\sum_{a' \in A'} \pi(a'|s)(1 - \zeta(s, a'))$ if $a = a_{abs}$.

Next we show that the projection of policy will has an equal or smaller value than the original policy.

**Lemma 2.** For any policy $\pi : S' \to \Delta(A')$, $v_{M'}^\pi \geq v_{M'}^\Xi(\pi)$, and $v_{M'}^\pi = v_{M'}^\Xi(\pi)$ if for any $(s, a)$ reachable by $\pi$, $\zeta(s, a) = 1$.

**Proof.** We drop the subscription of $M'$ in this proof for ease of notation. For any given $s$,

$$
\sum_{a \in A} \pi(a|s)Q_{\Xi}(\pi)(s, a) = \sum_{a \in A} \pi(a|s)Q_{\Xi}(\pi)(s, a) \quad (Q^\pi(s, a_{abs} = 0))
$$

$$
\geq \sum_{a \in A} \zeta(s, a)\pi(a|s)Q_{\Xi}(\pi)(s, a)
$$

$$
= \Xi(\pi)(a_{abs}|s)Q_{\Xi}(\pi)(s, a_{abs}) + \sum_{a \in A} \Xi(\pi)(a|s)Q_{\Xi}(\pi)(s, a) \quad (\text{Def of } \Xi)
$$

$$
= \sum_{a \in A'} \Xi(\pi)(a|s)Q_{\Xi}(\pi)(s, a)
$$

$$
= V_{\Xi}(\pi)(s)
$$

The inequality is an equality if for any $a$ s.t. $\pi(a|s) > 0$, $\zeta(s, a) = 1$. By the performance difference lemma [3] [Lemma 6.1]:

$$
v_{\Xi}(\pi) - v_{\pi} = \sum_{h=0}^{\infty} \gamma^h E_{s \sim \eta_H^\pi} \left[ V_{\Xi}(\pi)(s) - \sum_{a \in A'} \pi(a|s)Q_{\Xi}(\pi)(s, a) \right] \leq 0
$$

The inequality is an equality if for any $(s, a)$ s.t. $\eta_H^\pi(s)\pi(a|s) > 0$ for some $h$, $\zeta(s, a) = 1$. In another word for any state-action reachable by $\pi$ $(\eta_H^\pi(s) > 0$ and $\pi(a|s) > 0$ for some $h), \zeta(s, a) = 1$. \hfill \Box

The following results shows for any policy $\pi$ in the $\zeta$-constrained policy set the projection will not change the policy value much.

**Lemma 3.** For any policy $\pi \in \Pi_{\zeta C}^{all}$, $v_M^\pi \leq v_M^{\Xi(\pi)} + \frac{\epsilon_\zeta V_{\max}}{1 - \gamma}$.

**Proof.** Since $\pi$ only takes action in $A$, by Lemma [1] we have that $v_M^\pi = v_M^{\pi_1}$. Since $\pi \in \Pi_{\zeta C}^{all}$, we have that $\Pr (\zeta(s, a) = 0|\pi) \leq \epsilon_\zeta$, which means that:

$$
(1 - \gamma) \sum_{h=0}^{\infty} \gamma^h E_{s \sim \eta_H^\pi} \left[ 1 (\zeta(s, a) = 0) \right] \leq \epsilon_\zeta
$$

(9)
Thus:

\[ v^\Xi(\pi) - v^\pi = \sum_{h=0}^{\infty} \gamma^h \mathbb{E}_{s \sim \eta^h_h} \left[ V^\Xi(\pi)(s) - \sum_{a \in A'} \pi(a|s)Q^\Xi(\pi)(s, a) \right] \]

(10)

\[ = \sum_{h=0}^{\infty} \gamma^h \mathbb{E}_{s \sim \eta^h_h} \left[ V^\Xi(\pi)(s) - \sum_{a \in A'} \pi(a|s)\zeta(s, a)Q^\Xi(\pi)(s, a) \right] \]

(11)

\[ - \sum_{h=0}^{\infty} \gamma^h \mathbb{E}_{s,a \sim \eta^h_h} \left[ \mathbb{I}(\zeta(s, a) = 0)Q^\Xi(\pi)(s, a) \right] \]

(12)

\[ \geq \sum_{h=0}^{\infty} \gamma^h \mathbb{E}_{s \sim \eta^h_h} \left[ V^\Xi(\pi)(s) - \sum_{a \in A'} \pi(a|s)\zeta(s, a)Q^\Xi(\pi)(s, a) \right] \]

(13)

\[ - V_{\max} \sum_{h=0}^{\infty} \gamma^h \mathbb{E}_{s,a \sim \eta^h_h} \left[ \mathbb{I}(\zeta(s, a) = 0) \right] \]

(14)

\[ \geq \sum_{h=0}^{\infty} \gamma^h \mathbb{E}_{s \sim \eta^h_h} \left[ V^\Xi(\pi)(s) - \sum_{a \in A'} \pi(a|s)\zeta(s, a)Q^\Xi(\pi)(s, a) \right] - \frac{V_{\max}\zeta}{1 - \gamma} \]

(15)

\[ = - \frac{V_{\max}\zeta}{1 - \gamma} \]

(16)

The last step follows from the first part in the proof of Lemma 2, \( v^\pi_{M^t} - v^\Xi(\pi)_{M^t} \leq \frac{V_{\max}\zeta}{1 - \gamma}. \)

\[ \square \]

**B Justification of Assumption [1]**

In this section we prove a claim stated in Section 5 about the upper bound on density functions. We are going to prove Assumption [1] holds when the transition density is bounded.

**Lemma 4.** Let \( p(\cdot|s, a) \) be the probability density function of transition distribution: \( \rho(s_0) \leq \sqrt{U} < \infty, p(s_{t+1}|s_t, a_t) \leq \sqrt{U} < \infty \) and \( \forall \pi(a_t|s_t, h) \leq \sqrt{U} < \infty, \) for all \( s_0, s_t, s_{t+1} \in \mathcal{S} \) and \( a \in \mathcal{A}. \)

Then in \( M' \) for any non-stationary policy \( \pi : \mathcal{S'} \times \mathbb{N} \to \Delta(\mathcal{A'}) \) and \( h \geq 0, \eta^h_h(s, a) \leq U \) for any \( s \in \mathcal{S} \) and \( a \in \mathcal{A}. \)

**Proof.** We first prove that \( \eta^h_h(s) \leq \sqrt{U} \) for any non-stationary policy \( \pi. \) For \( h = 0, \eta^0_h(s) = \rho(s) \leq \sqrt{U}. \) For \( h \geq 1 \) and \( s \in \mathcal{S}:

\[ \eta^h_h(s) = \int_{s-1 \in \mathcal{S'}} \sum_{a \in \mathcal{A}'} \eta^{h-1}_{h-1}(s-1)\pi(a_{-1}|s-1, h-1)p(s|s-1, a_{-1})ds_{-1} \]

(17)

\[ = \int_{s-1 \in \mathcal{S}} \sum_{a \in \mathcal{A}} \eta^{h-1}_{h-1}(s)\pi(a_{-1}|s, h-1)p(s|s, a_{-1})ds_{-1} \]

(18)

\[ \leq \mathbb{E} \eta^{h-1}_{h-1} \times \pi(h-1) \left[ p(s|s-1, a_{-1}) \right] \]

(19)

\[ \leq \sqrt{U} \]

(20)

The first step follows from the inductive definition of \( \eta^h_h(s). \) The second step follows from that \( s_{abs} \) is absorbing state and \( a_{abs} \) only leads to absorbing state. The third step follows from transition density \( p(s|s-1, a_{-1}) \) is non-negative. The last step follows from that the transition density \( p(s|s-1, a_{-1}) \) is the same between \( M \) and \( M' \) for \( s, s-1 \in \mathcal{S}, a_{-1} \in \mathcal{A}, \) and \( p(s|s-1, a_{-1}) \) in \( M \) is upper bounded by \( U. \) Finally, the joint density function over \( s \) and \( a, \eta^h_h(s, a) = \eta^h_h(s)\pi(a|s, h) \) is bounded by \( U, \) and we finished the proof. \( \square \)

For the convenience of notation later we use *admissible distribution* to refer to state-action distributions introduced by non-stationary policy \( \pi \) in \( M'. \) This definition is from [1]:

**Definition 4 (Admissible distributions).** We say a distribution or its density function \( \nu \in \Delta(\mathcal{S'} \times \mathcal{A'}) \) is admissible in MDP \( M', \) if there exists \( h \geq 0, \) and a (non-stationary) policy \( \pi : \mathcal{S'} \times \mathbb{N} \to \Delta(\mathcal{A'}), \) such that \( \nu(s, a) = \eta^h_h(s, a). \)
C Proofs for Policy Iteration Guarantees

In this section we are going to prove the result of Theorem 1 using the definition of the strong ζ-constrained policy set. At a high level, the proof is done in two steps. First we prove similar result to Theorem 1 for any policy in the strong ζ-constrained policy set: an upper bound of \( v^\pi_M - v^\pi_t \), where \( \pi \) can be any policy in the strong ζ-constrained policy set and \( \pi_t \) is the output of the algorithm (Theorem 2, formally stated in Appendix C.4). Then we are going to show that for any policy \( \pi \) in the ζ-constrained policy set after a projection \( \Xi \) it is in the strong ζ-constrained policy set and \( v^\pi_M \leq v^{\Xi(\pi)}_M + \frac{V_{\max} \varepsilon}{1-\gamma} \). Then we can provide the upper bound for \( v^\pi_M - v^\pi_t \) for any \( \pi \) in ζ-constrained policy set.

The proof of Theorem 2 (the \( \Pi_{SC}^{\pi_t} \), version of Theorem 1, formally stated in Appendix C.4) goes as follow. First, we show the fixed point of \( T_\varepsilon \) is \( Q^{\Xi(\pi)} \) for any policy \( \pi \), indicating the inner loop of policy evaluation step is actually evaluating \( \pi_t = \Xi(\hat{\pi}_t) \). We prove this result formally in Lemma 6.

To bound the gap between \( \pi_t \) and any policy \( \tilde{\pi} \) in the ζ-constrained policy set, we use the contraction property of \( T_\varepsilon \) to recursively decompose it into a discounted summation over policy improvement gap \( Q^{\pi_{t+1}} - Q^{\pi_t} \). \( \tilde{\pi} \) in the ζ-constrained policy set is needed because the operator \( T_\varepsilon \) constrains the backup on the support set of \( \zeta \).

Next, we bound the policy improvement gap in Lemma 12
\[
Q^{\pi_{t+1}} - Q^{\pi_t} \geq -\mathcal{O}(||\zeta(Q^{\pi_t} - f_{t,K})||_{1,\nu})
\]
for some admissible distribution \( \nu \) related to \( \pi_{t+1} \). The fact that we only need to measure the error on the support set of \( \zeta \) is important. It follows from the fact that both \( \pi_{t+1} \) and \( \pi_t \) only takes action on the support set of \( \zeta \) except \( a_{obs} \) which gives us a constant value. This allows us to change the measure from arbitrary distribution \( \nu \) to data distribution \( \mu \), without needing concentratability.

The rest of proof is to upper bound \( ||\zeta(Q^{\pi_t} - f_{t,K})||_{1,\nu} \) using contraction and concentration inequalities. First, \( ||\zeta(Q^{\pi_t} - f_{t,K})||_{1,\nu} \) is upper bounded by \( C||f_{t,K} - T_\varepsilon f_{t,K}||_{2,\mu}/(1-\gamma) \) in Lemma 9 using a standard contraction analysis technique. Notice that here we can change the measure to \( \mu \) with cost \( C \) to allow us to apply concentration inequality. Then Lemma 8 bounds \( ||f_{t,K} - T_\varepsilon f_{t,K}||_{2,\mu} \) by a function of sample size \( n \) and completeness error \( \epsilon_F \) using Bernstein’s inequality.

While writing the proof, we will first introduce the fixed point of \( T_\varepsilon \) is \( Q^{\Xi(\pi)} \) in section C.1. We prove the upper bound of the policy evaluation error \( ||\zeta(Q^{\pi_t} - f_{t,K})||_{1,\nu} \) in section C.2 and the policy improvement step in section C.3. After we proved the main theorem, we will prove when we can bound the value gap with the optimal value in Corollary 1 as we showed in the main text.

C.1 Fixed point property

In Algorithm 1, the output policy is \( \hat{\pi}_{t+1} \). However, we will show that is actually equivalent with the following algorithm,

**Algorithm 3 Pessimistic Policy Iteration (PPI, repeat Algorithm 1)**

Input: \( D, \mathcal{F}, \Pi, \hat{\mu}, b \)
Output: \( \hat{\pi}_{T+1} \)

Initialize \( \pi_0 \in \Pi \).
for \( t = 0 \) to \( T - 1 \) do
  Initialize \( f_{t,0} \in \mathcal{F} \)
  for \( k = 0 \) to \( K \) do
    // Policy Evaluation
    \( f_{t,k+1} \leftarrow \arg \min_{f \in \mathcal{F}} \mathcal{L}_D(f, f_{t,k}; \pi_t) \)
  end for
  // Policy Improvement
  \( \hat{\pi}_{t+1} \leftarrow \arg \max_{\pi \in \Pi} \mathbb{E}_D[\mathbb{E}_\pi [\zeta(s,a) f_{t,K}(s,a)]] \)
  \( \pi_{t+1} \leftarrow \Xi(\hat{\pi}_{t+1}) \)
end for
The output policy is still \( \hat{\pi}_{t+1} \), and we know that \( v_{\hat{\pi}_{t+1}} \geq v^{\pi_{t+1}} \). So if we can lower bound \( v^{\pi_{t+1}} \) we immediately have the lower bound on \( v^{\hat{\pi}_{t+1}} \). The only difference in algorithm is we change the policy evaluation operator from \( T_{\pi}^\xi \) to \( T_{\hat{\pi}}^\xi \), where \( \pi_t \) is the projection of \( \hat{\pi}_t \). The following result shows these two operators are actually the same. For the ease of notation, we refer to Algorithm 3 in our analysis.

**Lemma 5.** For any policy \( \pi : S' \to \Delta(A') \), \( T^\pi_\xi = T^\Xi_\xi \).

**Proof.** We only need to prove for any \( f \), \( T^\pi_\xi f = T^\Xi_\xi f \). For any \( a \in A \),

\[
(T^\pi_\xi f)(s, a) = r(s, a) + \gamma \mathbb{E} \left[ \sum_{a' \in A} \pi(a'|s') \zeta(s', a') f(s', a') \right] = r(s, a) + \gamma \mathbb{E} \left[ \sum_{a' \in A} \pi(a'|s') \zeta^2(s', a') f(s', a') \right] = r(s, a) + \gamma \mathbb{E}_{a'} \left[ \sum_{a' \in A} \Xi(\pi)(a'|s') \zeta(s', a') Q^\Xi(s', a') \right] = (T^\Xi_\xi f)(s, a)
\]

For \( a = a_{abs} \), \( (T^\pi_\xi f)(s, a) = 0 = (T^\Xi_\xi f)(s, a) \). ☐

The next result is a key insight about \( T^\pi_\xi \)'s behavior in \( M' \) that guide our analysis.

**Lemma 6.** For any policy \( \pi : S' \to \Delta(A') \), the fixed point solution of \( T^\pi_\xi \) is equal to \( Q^\Xi_\xi \) on \( S \times A \).

**Proof.** By definition \( Q^\Xi_\xi \) is the fixed point of the standard Bellman evaluation operator on \( M' \): \( T^\Xi_\xi \). So for any \( (s, a) \in S \times A \):

\[
Q^\Xi_\xi(s, a) = (T^\Xi_\xi Q^\Xi_\xi)(s, a) = r(s, a) + \gamma \mathbb{E}_{a'} \left[ \sum_{a' \in A'} \Xi(\pi)(a'|s') Q^\Xi_\xi(s', a') \right] = r(s, a) + \gamma \mathbb{E}_{a'} \left[ \Xi(\pi)(a_{abs}|s') Q^\Xi_\xi(s', a_{abs}) + \sum_{a' \in A} \Xi(\pi)(a'|s') Q^\Xi_\xi(s', a') \right] = r(s, a) + \gamma \mathbb{E}_{a'} \left[ \sum_{a' \in A} \Xi(\pi)(a'|s') Q^\Xi_\xi(s', a') \right] = r(s, a) + \gamma \mathbb{E}_{a'} \left[ \sum_{a' \in A} \pi(a'|s') \zeta(s', a') Q^\Xi_\xi(s', a') \right] = (T^\Xi_\xi Q^\Xi_\xi)(s, a)
\]

So we proved that \( Q^\Xi_\xi \) is also the fixed-point solution of \( T^\pi_\xi \) constrained on \( S \times A \). ☐

An obvious consequences of these two lemmas is that the fixed point solution of \( T^\pi_\xi = T^\Xi_\xi \) equals \( Q^\Xi_\xi \) on \( S \times A \).
C.2 Proofs for policy evaluation step

We start with an useful result of the expected loss of the solution from empirical loss minimization, by applying a concentration inequality.

**Lemma 7.** Given \( \pi \in \Xi(\Pi) \) and Assumption 3, let \( g^*_f = \arg \min_{g \in \mathcal{F}} \|g - T^\pi f\|_{2,\mu} \). Then \( \|g^*_f - T^\pi f\|_{2,\mu} \leq \epsilon_F \). The dataset \( D \) is generated i.i.d. from \( M \) as follows: \((s, a) \sim \mu, r = R(s, a)\), \( s' \sim P(s, a) \). Define \( \mathcal{L}_\mu(f; f', \pi) = \mathbb{E}_D[|\mathcal{L}_D(f; f', \pi)|] \). We have that \( \forall f \in \mathcal{F} \), with probability at least \( 1 - \delta \),

\[
\mathcal{L}_\mu(T^\pi f; f, \pi) - \mathcal{L}_\mu(g^*_f; f, \pi) \leq \frac{121V^2}{3n} \ln \left( \frac{|\mathcal{F}| |\Pi|}{\delta} \right) + \sqrt{\frac{64V^2}{n} \ln \left( \frac{|\mathcal{F}| |\Pi|}{\delta} \right) \epsilon_F}
\]

where \( T^\pi f = \arg \min_{g \in \mathcal{F}} \mathcal{L}_D(g; f, \pi) \).

**Proof.** This proof is similar with the proof of Lemma 16 in \( \Xi \), and we adapt it to the \( \zeta \)-constrained Bellman evaluation operator \( T^\pi_\zeta \). First, there is no difference in \( \mathcal{L}_D \) and \( \mathcal{L}_\mu \) between \( M \) and \( M' \), and the right hand side is also the same constant for \( M \) and \( M' \). The distribution of \( D \) in \( M \) and \( M' \) are the same, since \( \mu \) does not cover \( s_{\text{abs}} \) and \( a_{\text{abs}} \). So we are going to prove the inequality for \( M \), and thus this bound holds for \( M' \) too.

For the simplicity of notations, let \( V^\pi_f(s) = \sum_{a \in A} \pi(a|s) \zeta(s, a) f(s, a) \). Fix any \( f, g \in \mathcal{F} \), and define

\[
X(g, f, g^*_f) := (g(s, a) - r - \gamma V^\pi_f(s'))^2 - (g^*_f(s, a) - r - \gamma V^\pi_f(s'))^2.
\]

Plugging each \((s, a, r, s') \in D\) into \( X(g, f, g^*_f) \), we get i.i.d. variables \( X_1(g, f, g^*_f), X_2(g, f, g^*_f), \ldots, X_n(g, f, g^*_f) \). It is easy to see that

\[
\frac{1}{n} \sum_{i=1}^{n} X_i(g, f, g^*_f) = \mathcal{L}_D(g; f, \pi) - \mathcal{L}_D(g^*_f; f, \pi).
\]

By the definition of \( \mathcal{L}_\mu \), it is also easy to show that

\[
\mathcal{L}_\mu(g; f, \pi) = \|g - T^\pi f\|_{2,\mu}^2 + \mathbb{E}_{s,a \sim \mu} \left[ \mathbb{V}_{r,s'} \left( r + \gamma \sum_{a' \in A} \pi(a'|s') \zeta(s', a') f(s', a') \right) \right],
\]

where \( \mathbb{V}_{r,s'} \) is the variance over conditional distribution of \( r \) and \( s' \) given \((s, a)\). Notice that the second part does not depends on \( g \). Then

\[
\mathcal{L}_\mu(g; f, \pi) - \mathcal{L}_\mu(T^\pi_\zeta f; f, \pi) = \|g - T^\pi_\zeta f\|_{2,\mu}^2.
\]

Then we bound the variance of \( X \):

\[
\mathbb{V}[X(g, f, g^*_f)] \leq \mathbb{E}[X(g, f, g^*_f)^2] = \mathbb{E}_\mu \left( (g(s, a) - r - \gamma V_f(s'))^2 - (g^*_f(s, a) - r - \gamma V_f(s'))^2 \right) \quad \text{(Definition of } X) \]

\[
\leq 4V^2 \mathbb{E}_\mu \left[ (g(s, a) - g^*_f(s, a))^2 \right] \leq 4V^2 \|g - g^*_f\|_{2,\mu}^2 \leq 8V^2 \mathbb{E}[X(g, f, g^*_f)] + 2\epsilon_F.
\]
The last step holds because
\[ \|g - g_f^*\|^2_{2,\mu} \]
\[ \leq 2 \left( \|g - T^\pi_{\zeta,D} f\|^2_{2,\mu} + \|T^\pi_{\zeta,D} f - g_f^*\|^2_{2,\mu} \right) \]
\[ = 2 \left( \|g - T^\pi_{\zeta,D} f\|^2_{2,\mu} - \|T^\pi_{\zeta,D} f - g^*_f\|^2_{2,\mu} + 2\|T^\pi_{\zeta,D} f - g^*_f\|^2_{2,\mu} \right) \]
\[ = 2 \left[ (\mathcal{L}_\mu(g; f, \pi) - \mathcal{L}_\mu(T^\pi_{\zeta,D} f; f, \pi)) - (\mathcal{L}_\mu(g^*_f; f, \pi) - \mathcal{L}_\mu(T^\pi_{\zeta,D} f; f, \pi)) \right] + 2\|T^\pi_{\zeta,D} f - g^*_f\|^2_{2,\mu} \]

(Equation (35))

Next, we apply (one-sided) Bernstein’s inequality and union bound over all \( f \in \mathcal{F}, g \in \mathcal{F}, \) and \( \pi \in \Xi(\Pi). \) With probability at least \( 1 - \delta, \) we have
\[ \mathbb{E}[X(g, f, g_f^*)] - \frac{1}{n} \sum_{i=1}^n X_i(f, f, g_f^*) \leq \sqrt{\frac{2V^2_n[gX(g, f, g_f^*)] \ln \frac{\|f\|_\Pi}{\delta}}{n} + \frac{4V^2_n \ln \frac{\|f\|_\Pi}{\delta}}{3n}} \]
\[ = \sqrt{\frac{32V^2_n \left( \mathbb{E}[X(g, f, g_f^*)] + 2\epsilon_f \right) \ln \frac{\|f\|_\Pi}{\delta}}{n} + \frac{8V^2_n \ln \frac{\|f\|_\Pi}{\delta}}{3n}}. \]

Since \( T^\pi_{\zeta,D} f \) minimizes \( \mathcal{L}_D(\cdot; f, \pi), \) it also minimizes \( \frac{1}{n} \sum_{i=1}^n X_i(\cdot, f, g_f^*). \) This is because the two objectives only differ by a constant \( \mathcal{L}_D(g_f^*; f, \pi). \) Hence,
\[ \frac{1}{n} \sum_{i=1}^n X_i(T^\pi_{\zeta,D} f, f, g_f^*) \leq \frac{1}{n} \sum_{i=1}^n X_i(g_f^*, f, g_f^*) = 0. \]

Then,
\[ \mathbb{E}[X(T^\pi_{\zeta,D} f, f, g_f^*)] \leq 0 + \sqrt{\frac{32V^2_n \left( \mathbb{E}[X(T^\pi_{\zeta,D} f, f, g_f^*)] + 2\epsilon_f \right) \ln \frac{\|f\|_\Pi}{\delta}}{n} + \frac{8V^2_n \ln \frac{\|f\|_\Pi}{\delta}}{3n}}. \]

Solving for the quadratic formula,
\[ \mathbb{E}[X(T^\pi_{\zeta,D} f, f, g_f^*)] \leq \frac{48 \left( \frac{8V^2_n \ln \frac{\|f\|_\Pi}{\delta}}{3n} \right)^2}{3n} + \frac{64V^2_n \ln \frac{\|f\|_\Pi}{\delta}}{n} \epsilon_f + \frac{56V^2_n \ln \frac{\|f\|_\Pi}{\delta}}{3n} \epsilon_f \]
\[ \leq \frac{(56 + 32\sqrt{3})V^2_n \ln \frac{\|f\|_\Pi}{\delta}}{3n} \epsilon_f \]
\[ \leq \sqrt{\frac{56 \ln \frac{\|f\|_\Pi}{\delta}}{\delta} \epsilon_f} \]
\[ \leq \frac{112V^2_n \ln \frac{\|f\|_\Pi}{\delta}}{3n} + \sqrt{\frac{64V^2_n \ln \frac{\|f\|_\Pi}{\delta}}{n} \epsilon_f}. \]

Noticing that \( \mathbb{E}[X(T^\pi_{\zeta,D} f, f, g_f^*)] = \mathcal{L}_\mu(T^\pi_{\zeta,D} f; f, \pi) - \mathcal{L}_\mu(g_f^*; f, \pi), \) we complete the proof.

\[ \Box \]

**Lemma 8** (Policy Evaluation Accuracy). For any \( t, k \geq 1 \) and \( \pi_t, f_{t,k} \) and \( f_{t,k-1} \) from Algorithm 7
\[ \|f_{t,k} - T^\pi_{\zeta,D} f_{t,k-1}\|^2_{2,\mu} \leq \epsilon_1 \]
where \( \epsilon_1 = \frac{208V^2_n \ln \frac{\|f\|_\Pi}{\delta}}{3n} + 2\epsilon_f. \)
Proof.

\[
\begin{align*}
\|f_{t,k} - T_{\pi_t} f_{t,k-1}\|_{2,\mu}^2 &= \mathcal{L}_\mu(f_{t,k}; f_{t,k-1}, \pi_t) - \mathcal{L}_\mu(f_{t,k}; f_{t,k-1}, \pi_t) \\
&= \left(\mathcal{L}_\mu(f_{t,k}; f_{t,k-1}, \pi_t) - \mathcal{L}_\mu(g_{f_{t,k-1}}; f_{t,k-1}, \pi_t) \right) - \left(\mathcal{L}_\mu(T_{\pi_t} f_{t,k-1}; f_{t,k-1}, \pi_t) - \mathcal{L}_\mu(g_{f_{t,k-1}}; f_{t,k-1}, \pi_t) \right) \\
&\leq \frac{112V_2^2 \ln \frac{|\mathcal{X}|}{\delta}}{3n} + \sqrt{64V_2^2 \ln \frac{|\mathcal{X}|}{\delta}} \epsilon_F + \|g_{f_{t,k-1}} - T_{\pi_t} f_{t,k-1}\|_{2,\mu} \\
&\leq \frac{112V_2^2 \ln \frac{|\mathcal{X}|}{\delta}}{3n} + \sqrt{64V_2^2 \ln \frac{|\mathcal{X}|}{\delta}} \epsilon_F + \epsilon_F & (\text{Definition of } g_{f_{t,k-1}} \text{ and Assumption 3}) \\
&\leq \frac{112V_2^2 \ln \frac{|\mathcal{X}|}{\delta}}{3n} + \frac{32V_2^2 \ln \frac{|\mathcal{X}|}{\delta}}{n} + \epsilon_F + \epsilon_F = \epsilon_1 & (\sqrt{2ab} \leq a + b)
\end{align*}
\]

From this lemma to the proof of main theorem, we are going to condition on the fact that the event in Assumption 2 holds. In the proof of the main theorem we will impose the union bound on all failures.

**Lemma 9.** For any admissible distribution \(\nu\) on \(S' \times A'\), and any \(\pi_t\) from Algorithm 2,

\[
\|\zeta(s,a) (f_{t,K}(s,a) - Q_{\pi_t}(s,a))\|_{1,\nu} \leq C \left(\sqrt{\epsilon_1 + \frac{V_{\max} \epsilon_\mu}{1 - \gamma}} + \gamma K V_{\max} \right)
\]

where \(\epsilon_1\) is defined in Lemma 8

(Although \(f_{t,K}\) is only defined on \(S \times A\), \(\zeta\) is always zero for any other \((s,a)\). Thus the all values used in the proof are well-defined. Later, when it is necessary for proof, we define the value of \(f_{t,K}\) outside of \(S \times A\) to be zero. In the algorithm, we will never need to query the value of \(f_{t,K}\) outside of \(S \times A\).)

**Proof.** For any \(k \geq 1\) and any distribution \(\nu\) on \(S' \times A'\):

\[
\|\zeta(f_{t,k} - Q_{\pi_t})\|_{1,\nu} \leq \|\zeta(f_{t,k} - T_{\pi_t} f_{t,k-1})\|_{1,\nu} + \|\zeta(T_{\pi_t} f_{t,k-1} - T_{\pi_t} Q_{\pi_t})\|_{1,\nu} \\
\leq C \|f_{t,k} - T_{\pi_t} f_{t,k-1}\|_{1,\mu} + \|T_{\pi_t} f_{t,k-1} - T_{\pi_t} Q_{\pi_t}\|_{1,\nu} \\
\leq C \left(\|f_{t,k} - T_{\pi_t} f_{t,k-1}\|_{1,\mu} + V_{\max} \epsilon_\mu\right) + \|T_{\pi_t} f_{t,k-1} - T_{\pi_t} Q_{\pi_t}\|_{1,\nu} \quad (\text{Jensen’s inequality}) \\
\leq C(\sqrt{\epsilon_1} + V_{\max} \epsilon_\mu) + \|T_{\pi_t} f_{t,k-1} - T_{\pi_t} Q_{\pi_t}\|_{1,\nu} \quad (\text{Lemma 8})
\]

\[
= C(\sqrt{\epsilon_1} + V_{\max} \epsilon_\mu) + E_{\nu} \left[\gamma E_{P(\nu)} \sum_{a' \in A} \pi_t(a'|s') \zeta(s',a') (f_{t,k-1}(s',a') - Q_{\pi_t}(s',a'))\right] \\
= C(\sqrt{\epsilon_1} + V_{\max} \epsilon_\mu) + E_{\nu} \left[\gamma E_{P(\nu) \times \pi_t} \zeta(s',a') (f_{t,k-1}(s',a') - Q_{\pi_t}(s',a'))\right] \\
\leq C(\sqrt{\epsilon_1} + V_{\max} \epsilon_\mu) + C(\sqrt{\epsilon_1} + V_{\max} \epsilon_\mu) + \gamma \|\zeta(f_{t,k-1} - Q_{\pi_t})\|_{1,P(\nu) \times \pi}
\]
Equation (42) holds since for all \((s, a)\) s.t. \(\zeta(s, a) > 0, \nu(s, a) \leq U \leq \frac{v}{\epsilon} \mu(s, a) = C \mu(s, a)\).

Equation (43) holds since the total variation distance between \(\mu\) and \(\mu\) is bounded by \(\epsilon\) and the Bellman error is bounded in \([-V_{\text{max}}, V_{\text{max}}]\). Equation (44) follows from \(\pi_t \in \Pi_{\text{off}}\). So if \(\zeta(s, a) = 0, \pi(a(s) = 0\) for all \(a \in A\). Equation (45) holds since \(\zeta(\cdot, a_{\text{abs}}) = 0\). The next equation follows from that \(\zeta = \zeta^2\).

Note that this holds for any admissible distribution \(\nu\) on \(S' \times A'\) and and \(k\), as well as \(c_1\) does not depend on \(k\). Repeating this for \(k\) from \(K\) to \(1\) we will have that

\[
\|\zeta(s, a) (f_t, K(s, a) - Q^*_{\nu}(s, a))\|_{1,\nu} \leq \frac{1 - \gamma^K}{1 - \gamma} C (\sqrt{\epsilon_1} + V_{\text{max}} \epsilon_{\mu}) + \gamma^K V_{\text{max}}
\]

\[
< \frac{C (\sqrt{\epsilon_1} + V_{\text{max}} \epsilon_{\mu})}{1 - \gamma} + \gamma^K V_{\text{max}}
\]

C.3 Proofs for policy improvement step

Lemma 10 (Concentration of Policy Improvement Loss). For any \(f \in \mathcal{F}\), with probability at least \(1 - \delta\),

\[
\left\| \mathbb{E}_{\pi} [\zeta(s, a) f(s, a)] - \max_{a \in A} \zeta(s, a) f(s, a) \right\|_{1,\mu} \leq \epsilon_{\Pi} + 2V_{\text{max}} \sqrt{\frac{\ln(|\mathcal{F}| \Pi / \delta)}{2n}}
\]

where \(\pi_f = \arg \max_{\pi \in \Pi} \mathbb{E}_D [\mathbb{E}_{\pi} [\zeta(s, a) f(s, a)]]\).

Proof. Fixed \(f\), define \(X(s; \pi) = \max_{a \in A} \zeta(s, a) f(s, a) - \mathbb{E}_\pi [\zeta(s, a) f(s, a)]\). Notice that by definition \(X(s; \pi)\) is always non-negative, and \(\pi_f = \arg \max_{\pi \in \Pi} \mathbb{E}_D [\mathbb{E}_\pi [\zeta(s, a) f(s, a)]] = \arg \min_{\pi \in \Pi} \mathbb{E}_D [X(s; \pi)]\).

Only in this proof, let \(\pi_f\) be:

\[
\arg \min_{\pi \in \Pi} \mathbb{E}_\mu [X(s; \pi)] = \arg \min_{\pi \in \Pi} \left\| \mathbb{E}_\pi [\zeta(s, a) f(s, a)] - \max_{a \in A} \zeta(s, a) f(s, a) \right\|_{1,\mu}.
\]

\(X(s; \pi) \in [0, V_{\text{max}}]\). By Hoeffding’s inequality and union bound over all \(\pi \in \Pi, f \in \mathcal{F}\), with probability at least \(1 - \delta\) for any \(f\) and \(\pi \neq \pi_f\),

\[
\mathbb{E}_\mu [X(s; \pi)] - \mathbb{E}_D [X(s; \pi)] \leq V_{\text{max}} \sqrt{\frac{\ln(|\mathcal{F}| \Pi / \delta)}{2n}}
\]

for \(\pi = \pi_f\)

\[
\mathbb{E}_D [X(s; \pi)] - \mathbb{E}_\mu [X(s; \pi)] \leq V_{\text{max}} \sqrt{\frac{\ln(|\mathcal{F}| \Pi / \delta)}{2n}}
\]

If \(\pi_f = \pi_f\), then \(\mathbb{E}_\mu [X(s; \pi_f)] \leq \epsilon_{\Pi}\). Otherwise,

\[
\mathbb{E}_\mu [X(s; \pi_f)] \leq \mathbb{E}_D [X(s; \pi_f)] + V_{\text{max}} \sqrt{\frac{\ln(|\mathcal{F}| \Pi / \delta)}{2n}}
\]

\[
\leq \mathbb{E}_D [X(s; \pi_f)] + V_{\text{max}} \sqrt{\frac{\ln(|\mathcal{F}| \Pi / \delta)}{2n}}
\]

\[
\leq \mathbb{E}_\mu [X(s; \pi_f)] + 2V_{\text{max}} \sqrt{\frac{\ln(|\mathcal{F}| \Pi / \delta)}{2n}}
\]

\[
= \min_{\pi \in \Pi} \left\| \mathbb{E}_\pi [\zeta(s, a) f(s, a)] - \max_{a \in A} \zeta(s, a) f(s, a) \right\|_{1,\mu} + 2V_{\text{max}} \sqrt{\frac{\ln(|\mathcal{F}| \Pi / \delta)}{2n}}
\]

\[
= \epsilon_{\Pi} + 2V_{\text{max}} \sqrt{\frac{\ln(|\mathcal{F}| \Pi / \delta)}{2n}}
\]
By Lemma 10:

\[ \text{Proof.} \quad \text{Recall that } \pi_{t+1} = \hat{\xi}(\pi_{t+1}). \text{ So } \pi_{t+1}(a|s) = \hat{\pi}_{t+1}(a|s) \text{ for all } a \text{ such that } \zeta(s,a) = 1. \text{ Then} \]

\[ E_{\nu} \left[ E_{\pi_{t+1}} [\zeta(s,a) f_{t,K}(s,a)] - E_{\pi} [\zeta(s,a) f_{t,K}(s,a)] \right] \geq -C \left( \epsilon_1 + V_{max} \epsilon_\mu + 2V_{max} \sqrt{\frac{\ln(|\mathcal{F}|)}{2n}} \right) \]

\[ \text{Lemma 11. For any admissible distribution } \nu \text{ on } S', \text{ any policy } \pi : S' \rightarrow \Delta(A'), \]

\[ E_{\nu} \left[ E_{\pi_{t+1}} [\zeta(s,a) f_{t,K}(s,a)] - E_{\pi} [\zeta(s,a) f_{t,K}(s,a)] \right] \geq \]

\[ = E_{\nu} \left[ E_{\hat{\pi}_{t+1}} [\zeta(s,a) f_{t,K}(s,a)] - E_{\pi} [\zeta(s,a) f_{t,K}(s,a)] \right] \]

\[ = E_{\nu} \left[ E_{\hat{\pi}_{t+1}} [\zeta(s,a) f_{t,K}(s,a)] - \max_{a \in A} \zeta(s,a) f_{t,K}(s,a) + \max_{a \in A} \zeta(s,a) f_{t,K}(s,a) - E_{\pi} [\zeta(s,a) f_{t,K}(s,a)] \right] \]

\[ \geq E_{\nu} \left[ E_{\hat{\pi}_{t+1}} [\zeta(s,a) f_{t,K}(s,a)] - \max_{a \in A} \zeta(s,a) f_{t,K}(s,a) \right] \]

\[ \geq -E_{\nu} \left[ E_{\hat{\pi}_{t+1}} [\zeta(s,a) f_{t,K}(s,a)] - \max_{a \in A} \zeta(s,a) f_{t,K}(s,a) \right] \]

\[ = - \left\| E_{\hat{\pi}_{t+1}} [\zeta(s,a) f_{t,K}(s,a)] - \max_{a \in A} \zeta(s,a) f_{t,K}(s,a) \right\|_{1,\mu} \]

\[ \geq -C \left\| E_{\hat{\pi}_{t+1}} [\zeta(s,a) f_{t,K}(s,a)] - \max_{a \in A} \zeta(s,a) f_{t,K}(s,a) \right\|_{1,\hat{\mu}} \]

The last step follows from that \( \zeta(s,a) = 1 \Rightarrow \hat{\mu}(s,a) \geq b \Rightarrow \hat{\nu}(s) \geq b \Rightarrow -\nu(s) \geq -U \geq -C\hat{\nu}(s) \), and for all other \((s,a)\) the term inside of norm is zero. Since the total variation distance between \( \hat{\mu} \) and \( \mu \) is bounded by \( \epsilon_\mu \)

\[ \left\| E_{\hat{\pi}_{t+1}} [\zeta(s,a) f_{t,K}(s,a)] - \max_{a \in A} \zeta(s,a) f_{t,K}(s,a) \right\|_{1,\hat{\mu}} \]

\[ \leq \left\| E_{\hat{\pi}_{t+1}} [\zeta(s,a) f_{t,K}(s,a)] - \max_{a \in A} \zeta(s,a) f_{t,K}(s,a) \right\|_{1,\mu} + V_{max} \epsilon_\mu \]

By Lemma 10,

\[ \left\| E_{\hat{\pi}_{t+1}} [\zeta(s,a) f_{t,K}(s,a)] - \max_{a \in A} \zeta(s,a) f_{t,K}(s,a) \right\|_{1,\hat{\mu}} \leq \epsilon_\Pi + 2V_{max} \sqrt{\frac{\ln(|\mathcal{F}|)}{2n}} \]

Then we finished the proof by plug this into the last equation.\]

\[ \text{Lemma 12. For any } (s,a) \in S' \times A', \text{ and any } \pi_t, \pi_{t+1} \text{ in Algorithm}[7] \]

\[ Q^{\pi_{t+1}}(s,a) - Q^{\pi_t}(s,a) \geq -2C\sqrt{\epsilon_1} + 3V_{max} C \epsilon_\mu \frac{1}{(1 - \gamma)^2} - \epsilon_2 + 2\gamma K V_{max} \]

where \( \epsilon_1 \) is defined in Lemma 8, \( \epsilon_2 = C \left( \epsilon_\Pi + 2V_{max} \sqrt{\frac{\ln(|\mathcal{F}|)}{2n}} \right) \).
\[ V^{\pi^{t+1}}(s') - V^{\pi^t}(s') \]  
\[ = \sum_{h=1}^{\infty} \gamma^{h-1} \mathbb{E}_{z \sim \eta_h} \left[ \sum_{a \in A'} (\pi_{t+1}(a|z)Q^{\pi_t}(z,a) - \pi_t(a|z)Q^{\pi_t}(z,a)) \right] \]  
\[ = \sum_{h=1}^{\infty} \gamma^{h-1} \mathbb{E}_{z \sim \eta_h} \left[ \sum_{a \in A'} (1 - \zeta(z,a)) (\pi_{t+1}(a|z)Q^{\pi_t}(z,a) - \pi_t(a|z)Q^{\pi_t}(z,a)) + \zeta(z,a) (\pi_{t+1}(a|z)Q^{\pi_t}(z,a) - \pi_t(a|z)Q^{\pi_t}(z,a)) \right] \]  

Because \( \pi_t, \pi_{t+1} \in \Pi_{\text{alt}}, \zeta(z,a) = 0 \) means either \( \pi_t(a|z) = \pi_{t+1}(a|z) = 0 \) or \( a = a_{\text{abs}} \). So the first term is zero. Then:

\[ V^{\pi^{t+1}}(s') - V^{\pi^t}(s') \]  
\[ = \sum_{h=1}^{\infty} \gamma^{h-1} \mathbb{E}_{z \sim \eta_h} \left[ \sum_{a \in A'} \zeta(z,a) (\pi_{t+1}(a|z)Q^{\pi_t}(z,a) - \pi_t(a|z)Q^{\pi_t}(z,a)) \right] \]  
\[ = \sum_{h=1}^{\infty} \gamma^{h-1} \mathbb{E}_{z \sim \eta_h} \left[ \sum_{a \in A} \zeta(z,a) (\pi_{t+1}(a|z)Q^{\pi_t}(z,a) - \pi_t(a|z)Q^{\pi_t}(z,a)) \right] \]  
\[ = \sum_{h=1}^{\infty} \gamma^{h-1} \mathbb{E}_{z \sim \eta_h} \left[ \sum_{a \in A} \zeta(z,a) (\pi_{t+1}(a|z)f_{t,K}(z,a) - \pi_t(a|z)f_{t,K}(z,a)) + \zeta(z,a) (\pi_{t+1}(a|z)f_{t,K}(z,a) - \pi_t(a|z)Q^{\pi_t}(z,a)) \right] \]  

Equation 75 follows from \( Q^{\pi}(s,a_{\text{abs}}) = 0 \) for any \( \pi \) and \( s \). By Lemma 11 for any \( h \),

\[ \mathbb{E}_{z \sim \eta_h} \left[ \sum_{a \in A} \zeta(z,a) (\pi_{t+1}(a|z)f_{t,K}(z,a) - \pi_t(a|z)f_{t,K}(z,a)) \right] \]  
\[ = \mathbb{E}_{z \sim \eta_h} \left[ \mathbb{E}_{\pi_t+1} [\zeta(s,a)f_{t,K}(s,a)] - \mathbb{E}_{\pi_t} [\zeta(s,a)f_{t,K}(s,a)] \right] \geq -\epsilon_2 - CV_{\text{max}}\epsilon_\mu \]
Then

\[
V^{\pi_t+1}(s') - V^{\tilde{\pi}}(s') 
\]  
\[
\geq \sum_{h=1}^{\infty} \gamma^{h-1} \mathbb{E}_{z \sim \eta_t^{\pi_t+1}} \left[ \sum_{a \in A} \zeta(z,a) (\pi_t+1(a|z)Q^{\pi_t}(z,a) - \pi_{t+1}(a|z)f_{t,K}(z,a)) \right] - \frac{\epsilon_2 + CV_{\max} \epsilon_\mu}{1 - \gamma} 
\]  
\[
\geq - \sum_{h=1}^{\infty} \gamma^{h-1} \left( \|\zeta(z,a)(Q^{\pi_t}(z,a) - f_{t,K}(z,a))\|_{1,\eta_t^{\pi_t+1}} \right) - \frac{\epsilon_2 + CV_{\max} \epsilon_\mu}{1 - \gamma} 
\]  
\[
\geq - \sum_{h=1}^{\infty} \gamma^{h-1} \left( \|\zeta(z,a)(Q^{\pi_t}(z,a) - f_{t,K}(z,a))\|_{2,\eta_t^{\pi_t+1}} \right) - \frac{\epsilon_2 + CV_{\max} \epsilon_\mu}{1 - \gamma} 
\]  
\[
\geq - \frac{2C \sqrt{\epsilon_1} + V_{\max} \epsilon_\mu}{(1 - \gamma)^2} - 2\gamma^K V_{\max} \frac{\epsilon_2 + CV_{\max} \epsilon_\mu}{1 - \gamma} 
\]  
\[
\geq - \frac{2C \sqrt{\epsilon_1} + 3CV_{\max} \epsilon_\mu}{(1 - \gamma)^2} - 2\gamma^K V_{\max} \frac{\epsilon_2}{1 - \gamma} 
\]  

Equation [87] follows from Jensen’s inequality. Since this holds for any \(s'\), we proved that for any \((s,a),\)

\[
[Q^{\pi_t+1}(s,a) - Q^{\tilde{\pi}}(s,a)] 
\]  
\[
= \gamma \mathbb{E}_{s' \sim \mu_t} [V^{\pi_t+1}(s') - V^{\tilde{\pi}}(s')] 
\]  
\[
\geq - \frac{2C \sqrt{\epsilon_1} + V_{\max} \epsilon_\mu}{(1 - \gamma)^2} - 2\gamma^K V_{\max} \frac{\epsilon_2 + CV_{\max} \epsilon_\mu}{1 - \gamma} 
\]  
\[
\geq - \frac{2C \sqrt{\epsilon_1} + 3CV_{\max} \epsilon_\mu}{(1 - \gamma)^2} - 2\gamma^K V_{\max} \frac{\epsilon_2}{1 - \gamma} 
\]

\[\square\]

### C.4 Proof of main theorems

**Theorem 2.** Given an MDP \(M = \langle S, A, R, P, \gamma, p \rangle\), a dataset \(D = \{(s,a,r,s')\}\) with \(n\) samples that are drawn i.i.d. from \(\mu \times R \times P\), and a finite Q-function classes \(\mathcal{F}\) and a finite policy class \(\Pi\) satisfying Assumption 3 and \(\overline{\pi} = \Xi(\bar{\pi})\) from Algorithm 7 satisfies that with probability at least \(1 - 3\delta\),

\[
v^{\overline{\pi}} - v^{\pi_t} \leq \frac{4C}{(1 - \gamma)^3} \left( \sqrt{419V_{\max}^2 \ln \frac{||\mathcal{F}||}{\delta}} + 2\sqrt{c_{\mathcal{F}}} \right) + \frac{6CV_{\max} \epsilon_\mu}{(1 - \gamma)^3} + \frac{2C \epsilon_{\Pi} + 3\gamma^{K-1}V_{\max}}{(1 - \gamma)^2}
\]

for any policy \(\overline{\pi} \in \Pi_{\text{SC}}^{\text{all}}\).

**Proof.** For simplicity of the notation, let \(\epsilon_1 = \frac{208V_{\max} \ln \frac{||\mathcal{F}||}{\delta}}{3n} + 2\epsilon_\mathcal{F}, \epsilon_2 = 2C \sqrt{\epsilon_1} + 3CV_{\max} \epsilon_\mu\) and \(\epsilon_3 = \frac{2C \epsilon_{\Pi} + 3\gamma^{K-1}V_{\max}}{1 - \gamma}\). We start by proving a stronger result. For any \(\overline{\pi} \in \Pi_{\text{SC}}^{\text{all}}\), we will upper bound \(\mathbb{E}_{s' \sim \mu_t}[V^{\overline{\pi}} - V^{\pi_t}]\) for any admissible
distribution $\nu$ over $\mathcal{S}'$ which will naturally be an upper bound for $\nu^\pi - a^\pi_t$

\[
E_{\nu} [V^\pi - V^{\pi_{t+1}}] \\
= E_{\nu} \left[ V^\pi(s) - \sum_{a \in A'} \pi_{t+1}(a|s)Q^\pi(s, a) + \sum_{a \in A'} \pi_{t+1}(a|s)Q^\pi(s, a) - V^{\pi_{t+1}}(s) \right] \\
= E_{\nu} \left[ V^\pi(s) - \sum_{a \in A'} \pi_{t+1}(a|s)Q^\pi(s, a) + \sum_{a \in A'} \pi_{t+1}(a|s) (Q^\pi(s, a) - Q^{\pi_{t+1}}(s, a)) \right] \\
\leq E_{\nu} \sum_{a \in A'} \tilde{\pi}(a|s)Q^\pi(s, a) - \pi_{t+1}(a|s)Q^\pi(s, a) + \epsilon_3 \\
(\text{Lemma} [12]) \\
= E_{\nu} \sum_{a \in A'} \zeta(s, a)[\tilde{\pi}(a|s)Q^\pi(s, a) - \pi_{t+1}(a|s)Q^\pi(s, a)] + \epsilon_3 \\
= E_{\nu} \left[ E_{\pi} \left[ \zeta(s, a)Q^\pi(s, a) - \pi_{t+1}(a|s)Q^\pi(s, a) \right] - \pi_{t+1}(a|s)Q^\pi(s, a) \right] + \epsilon_3 \\
+ E_{\pi_{t+1}} \left[ \zeta(s, a)f_1(s, a) - \pi_{t+1}(a|s)Q^\pi(s, a) \right] + \epsilon_3 \\
\leq E_{\nu} \left[ E_{\pi} \left[ \zeta(s, a)Q^\pi(s, a) - \pi_{t+1}(a|s)Q^\pi(s, a) \right] - \pi_{t+1}(a|s)Q^\pi(s, a) \right] + \epsilon_3 \\
+ \sqrt{\epsilon_1 + CV_{\max} + \gamma K_{\max} + \epsilon_3} \quad (\text{Lemma} [9]) \\
\leq E_{\nu} \left[ E_{\pi} \left[ \zeta(s, a)Q^\pi(s, a) - \pi_{t+1}(a|s)Q^\pi(s, a) \right] - \pi_{t+1}(a|s)Q^\pi(s, a) \right] + \epsilon_2 + C\sqrt{\epsilon_1 + CV_{\max} + \gamma K_{\max} + \epsilon_3} \\
(\text{Lemma} [10]) \\
\leq E_{\nu} \left[ E_{\pi} \left[ \zeta(s, a)Q^\pi(s, a) - \pi_{t+1}(a|s)Q^\pi(s, a) \right] - \pi_{t+1}(a|s)Q^\pi(s, a) \right] + \epsilon_2 + 2C\sqrt{\epsilon_1 + 3CV_{\max} + \gamma K_{\max} + \epsilon_3} \\
(\text{Lemma} [9]) \\
= E_{\nu} \left[ \zeta(s, a)Q^\pi(s, a) - \pi_{t+1}(a|s)Q^\pi(s, a) \right] + \epsilon_2 + 2C\sqrt{\epsilon_1 + 3CV_{\max} + \gamma K_{\max} + \epsilon_3} \\
= E_{\nu} \left[ Q^\pi(s, a) - Q^\pi_{t+1}(s, a) \right] + \epsilon_2 + 2C\sqrt{\epsilon_1 + 3CV_{\max} + \gamma K_{\max} + \epsilon_3} \\
(\pi_t \in \Pi^{ll}_{SC}) \\
\leq \gamma E_{p(\nu \times \tilde{\pi})}[V^\pi - V^{\pi_{t+1}}] + \epsilon_2 + 2C\sqrt{\epsilon_1 + 3CV_{\max} + \gamma K_{\max} + \epsilon_3} \\
(92)
\]

The second to last step follows from $\pi_t \in \Pi^{ll}_{SC}$: for all $s, a$ such that $\tilde{\pi}(a|s) > 0$, either $\zeta(s, a) = 1$, or $a = a_{\text{abs}}$. The later two indicate that $Q^\pi_{t+1}(s, a) = Q^\pi(s, a) = 0$. So for all $s, a$ such that $\tilde{\pi}(a|s) > 0$, $Q^\pi(s, a) = \zeta(s, a)Q^\pi(s, a)$ and $Q^\pi_{t+1}(s, a) = \zeta(s, a)Q^\pi_{t+1}(s, a)$. Now we proved

\[
E_{\nu} [V^\tilde{\pi} - V^{\pi_{t+1}}] \leq \gamma E_{p(\nu \times \tilde{\pi})}[V^\tilde{\pi} - V^{\pi_{t}}] + \epsilon_2 + \epsilon_3 + 2C\sqrt{\epsilon_1 + 3CV_{\max} + \gamma K_{\max}} \\
(92)
\]
Then $v$ satisfies Assumption 3 and 4, Theorem 1. Given an MDP $\mathcal{M} = < S, \mathcal{A}, R, \gamma, P >$, a dataset $D = \{(s, a, r, s')\}$ with $n$ samples that is draw i.i.d. from $\mu \times R \times P$, and a finite $Q$-function classes $\mathcal{F}$ and a finite policy class $\Pi$ satisfying Assumption 5 and 6, $v_\pi$ from Algorithm 1 satisfies that with probability at least $1 - 3\delta$,

$$v_\pi - V_\mu \leq \frac{4C}{(1 - \gamma)^3} \left( \sqrt{\frac{419V_{\max}^2 \ln \frac{|\mathcal{F}| |\Pi|}{\delta}}{3n} + 2V_{\max}^2} \right) + \frac{6C\epsilon_\mu}{(1 - \gamma)^3} + \frac{2C\epsilon_\Pi + 3\gamma K - 1 V_{\max}}{(1 - \gamma)^2} + \frac{V_{\max} \epsilon_\xi}{1 - \gamma}$$

for any policy $\pi \in \Pi^{\text{opt}}_0$ and only take action over $\mathcal{A}$.

**Proof.** For any policy $\tilde{\pi}$ that only take action over $\mathcal{A}$, Lemma 5 tells that $v_{\tilde{\pi}} \leq v_{\tilde{\pi}}^{\Xi(\tilde{\pi})} + \frac{V_{\max} \epsilon_\xi}{1 - \gamma}$.

Since $\pi_t = \Xi(\tilde{\pi}_t)$ and $\tilde{\pi}_t$ only takes action in $\mathcal{A}$, by Lemma 6 and Lemma 8 $v_{\tilde{\pi}_t} - v_{\tilde{\pi}_t} \leq v_{\tilde{\pi}_t}^{\Xi(\tilde{\pi}_t)} - v_{\tilde{\pi}_t}^{\Xi(\tilde{\pi}_t)} + \frac{V_{\max} \epsilon_\xi}{1 - \gamma}$ and Theorem 2 completes the proof.

When there exist an optimal policy that is supported well by $\mu$. We can derive the following result about value gap between learned policy and optimal policy immediately from the main theorem about approximate policy iteration.
**Corollary 2.** If there exists an $\pi^*$ on $M$ such that $\Pr(\mu(s,a) \leq 2b|\pi^*) \leq \epsilon$, then under the assumptions of Theorem 2, $\hat{\pi}_t$ from Algorithm 1 satisfies that with probability at least $1 - 3\delta$,

$$v^\pi_M - v^\pi_M \leq \frac{4C}{(1 - \gamma)^3} \left( \sqrt{\frac{419V_{\max}^2 \ln \frac{|F||B|}{\delta}}{3n}} + 2\sqrt{\epsilon_\pi} \right) + 6CV_{\max}\epsilon_\mu \left(1 - \gamma\right)^3 + \frac{2C\epsilon_\pi + 3\gamma K^{-1}V_{\max}}{(1 - \gamma)^2} + \frac{V_{\max}(\epsilon + C\epsilon_\mu)}{1 - \gamma}$$

**Proof.** Given the condition of $\pi^*$,

$$\Pr\left(\hat{\mu}(s,a) \leq b|\pi^*\right) \leq \Pr\left(\mu(s,a) \leq 2b|\pi^*\right) + \Pr\left(|\mu(s,a) - \hat{\mu}(s,a)| \geq b|\pi^*\right)$$

$$\leq \epsilon + \frac{E_{\pi^*} \left||\mu(s,a) - \hat{\mu}(s,a)\right||}{b}$$

$$\leq \epsilon + \frac{U_{dTV}(\mu(s,a), \hat{\mu}(s,a))}{b}$$

$$\leq \epsilon + C\epsilon_\mu$$

Then $\pi^* \in \Pi_C^U$ with $\epsilon_\pi = \epsilon + C\epsilon_\mu$, and applying Theorem 2 finished the proof. 

**C.5 Safe Policy Improvement Result**

In many scenarios we aim to have a policy improvement that is guaranteed to be no worse than the data collection policy, which is called safe policy improvement. By abusing the notation a bit, let $\mu(a|s)$ be a policy that generate the data set. For our algorithm, the safe policy improvement will hold if $\mu \in \Pi_C^U$. To show $\mu \in \Pi_C^U$, we only need that $\Pr(\mu(s,a) \leq b|\mu) \leq \epsilon_\mu$. When the state-action space is finite, there must exist a minimum value for all non-zero $\mu(s,a)$'s. Let $\mu_{\min} = \min_{s,a \neq 0} \mu(s,a)$. Then we have that, if $b \leq \mu_{\min}$, $\Pr(\mu(s,a) \leq b|\mu) = 0$. Thus we have:

**Corollary 3.** With finite state action space and $b \leq \mu_{\min}$, under the assumptions as Theorem 2, $\hat{\pi}_t$ from Algorithm 1 satisfies that with probability at least $1 - 3\delta$,

$$v^\mu_M - v^\pi_M \leq \frac{52V_{\max} \sqrt{|S||A|} \left(\sqrt{\ln(2|S||A|/\delta)} + \sqrt{\ln(1 + nV_{\max})}\right) + 8}{\sqrt{n}(1 - \gamma)^3} + \frac{12V_{\max} |S||A| \ln(2|S||A|/\delta)}{nb(1 - \gamma)^3} + \frac{3\gamma K^{-1}V_{\max}}{(1 - \gamma)^2}$$

**Proof.** In finite state action space, the number of all deterministic policies is less than $|A|^{|S|}$. Thus we have a policy class with $\epsilon_{\Pi} = 0$ and $|\Pi| \leq |A|^{|S|}$. Since the $Q$ value is bounded in $[0, V_{\max}]$, we can construct a $\epsilon$ covering set $F$ of all value functions in $[0, V_{\max}]^{|S||A|}$ with $(V_{\max}^\epsilon + 1)^{|S||A|}$ functions. Then $\epsilon = \min_{f \in F} \|f - g\|_{\mu,2} \leq \max_{f \in F} \|f - g\|_{\infty} \leq \epsilon$.

We can also bound $\epsilon_\mu$ in finite state action space. For any fixed $s, a$, by Berstein’s inequality we have that with probability of $1 - \frac{\delta}{|S||A|}$,

$$|\hat{\mu}(s,a) - \mu(s,a)| = \left|\frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(s(i) = s, a(i) = a) - \mathbb{E}[\mathbb{I}(s(i) = s, a(i) = a)]\right|$$

$$\leq \frac{2V[\mathbb{I}(s(i) = s, a(i) = a)] \ln(2|S||A|/\delta)}{n} + \frac{4 \ln(2|S||A|/\delta)}{n}$$

$$= \frac{2\mu(s,a)(1 - \mu(s,a)) \ln(2|S||A|/\delta)}{n} + \frac{4 \ln(2|S||A|/\delta)}{n}$$
By taking summation of $|\hat{\mu}(s, a) - \mu(s, a)|$ and union bound over all $(s, a)$, we can bound the total variation bounds between $\hat{\mu}$ and $\mu$, with probability at least $1 - \delta$,

$$\|\hat{\mu} - \mu\|_{TV} = \frac{1}{2} \sum_{s,a} |\hat{\mu}(s, a) - \mu(s, a)|$$

(101)

$$\leq \frac{1}{2} \sum_{s,a} \left( \sqrt{\frac{2\mu(s,a)(1 - \mu(s,a)) \ln(2|S||A|/\delta)}{n}} + 4 \ln(2|S||A|/\delta) \right)$$

(102)

$$= \frac{2|S||A| \ln(2|S||A|/\delta)}{n} + \frac{1}{2} \sum_{s,a} \sqrt{\frac{2\mu(s,a)(1 - \mu(s,a)) \ln(2|S||A|/\delta)}{n}}$$

(103)

$$\leq \frac{2|S||A| \ln(2|S||A|/\delta)}{n} + \frac{1}{2} \left( \sum_{s,a} \frac{2\mu(s,a) \ln(2|S||A|/\delta)}{n} \sum_{s,a} \frac{(1 - \mu(s,a))}{\delta} \right)$$

(Cauchy-Schwartz’s inequality)

$$= \frac{2|S||A| \ln(2|S||A|/\delta)}{n} + \frac{1}{2} \sqrt{\frac{2 \ln(2|S||A|/\delta)}{n} (|S||A| - 1)}$$

(104)

$$\leq \frac{2|S||A| \ln(2|S||A|/\delta)}{n} + \sqrt{\frac{|S||A| \ln(2|S||A|/\delta)}{2n}}$$

(105)

Now in a finite state action space we can construct the policy and $Q$ function sets with $|F| \leq (V^{\max} + 1)|S||A|$, $|\Pi| \leq |A||S|$, $\epsilon_{\Pi} = 0$, $\epsilon_{\pi} \leq \epsilon$, and bounded $\epsilon_{\mu}$. By plugging these terms into the result of Theorem II, we have the following bound:

$$v^\mu_M - v^\pi_M \leq 4C \frac{1}{(1 - \gamma)^3} \left( \sqrt{\frac{419V^2_{\max} |S| \ln |A| + |S||A| \ln(1 + V_{\max}/\epsilon) + \ln(1/\delta)}{3n}} \right) + 2\sqrt{\epsilon}$$

$$+ \frac{6CV_{\max}}{(1 - \gamma)^3} \left( \sqrt{\frac{|S||A| \ln(2|S||A|/\delta)}{2n}} + \frac{2|S||A| \ln(2|S||A|/\delta)}{n} \right) + \frac{3\gamma^{K-1}V_{\max}}{(1 - \gamma)^2}$$

(106)

for any chosen $\epsilon > 0$. So we can set that $\epsilon = 1/n$ to upper bound the the infimum of this upper bound.

$$v^\mu_M - v^\pi_M \leq 4C \frac{1}{(1 - \gamma)^3} \left( \sqrt{\frac{419V^2_{\max} |S| \ln |A| + |S||A| \ln(1 + nV_{\max}) + \ln(1/\delta)}{3n}} \right) + 2\sqrt{\frac{1}{n}}$$

$$+ \frac{6CV_{\max}}{(1 - \gamma)^3} \left( \sqrt{\frac{|S||A| \ln(2|S||A|/\delta)}{2n}} + \frac{2|S||A| \ln(2|S||A|/\delta)}{n} \right) + \frac{3\gamma^{K-1}V_{\max}}{(1 - \gamma)^2}$$

(107)

Notice that in discrete space we have that $U \leq 1$. By replacing $C$ with $1/b$ and simplify some terms, we have that:

$$v^\mu_M - v^\pi_M \leq \sqrt{\frac{6704V^2_{\max} |S| (\ln(|A|/\delta) + |A| \ln(1 + nV_{\max}))}{3nb^2(1 - \gamma)^6}} + \frac{8}{b\sqrt{b}(1 - \gamma)^3}$$

$$+ \frac{18V^2_{\max} |S||A| \ln(2|S||A|/\delta)}{nb^2(1 - \gamma)^6} + \frac{12V_{\max} |S||A| \ln(2|S||A|/\delta)}{nb(1 - \gamma)^3} + \frac{3\gamma^{K-1}V_{\max}}{(1 - \gamma)^2}$$

$$\leq \frac{52V_{\max} \sqrt{|S||A| (\ln(2|S||A|/\delta) + \ln(1 + nV_{\max})) + 8}}{\sqrt{nb(1 - \gamma)^3}}$$

$$+ \frac{12V_{\max} |S||A| \ln(2|S||A|/\delta)}{nb(1 - \gamma)^3} + \frac{3\gamma^{K-1}V_{\max}}{(1 - \gamma)^2}$$

(108)
Proof. Given a MDP $M = < S, A, R, P, \gamma, P >$, a dataset $D = \{ (s, a, r, s') \} \text{ with } n$ samples that is drawn i.i.d. from $\mu \times R \times P$, and a finite Q-function classes $\mathcal{F}$ satisfying Assumption 5, $\hat{\pi}_t$ from Algorithm 2 satisfies that with probability at least $1 - \delta$, $v^\pi_t - v^\pi_{\hat{\pi}_t} \leq \frac{2C}{(1 - \gamma)^2} \left( \sqrt{\frac{208V^\pi_{\max} \ln |\mathcal{F}|}{3n}} + 2\sqrt{\epsilon_F} + V^{\max}_\epsilon + \|Q^\hat{\pi} - T_{\hat{\pi}}Q^\hat{\pi}\|_2,\mu \right) + \frac{(2\gamma^t + \epsilon_c)V^{\max}}{1 - \gamma}$ for any policy $\pi \in \Pi^\text{all}_{SC}$.

We will first give a proof sketch before we start the proof formally. The proof follows a similar structural as the policy iteration case. To prove Theorem 4 we first prove a similar version of Theorem 5 later. Then we show an upper bound of $v^\pi_t - v^\pi_{\hat{\pi}_t}$, where $\pi \in \Pi^\text{all}_{SC}$, and $\pi_t$ is the output of Algorithm (Theorem 5) will be formally stated later). Then we are going to show that for any policy $\pi$ in the $\zeta$-constrained policy set, after a projection $\mathcal{E}$ it is in the strong $\zeta$-constrained policy set and $v^{\pi}_{\hat{\pi}_t} \leq v^{\pi_{\hat{\pi}_t} + V^{\max}_\epsilon / (1 - \gamma)}$. Then we can provide the upper bound for $v^\pi_{\hat{\pi}_t} - v^\pi_{\hat{\pi}_t}$ for any $\pi$ in $\zeta$-constrained policy set (Theorem 4).

The proof sketch of Theorem 5 goes as follows. One key step to prove this error bound is to convert the performance difference between any policy $\pi \in \Pi^\text{all}_{SC}$ and $\pi_t$ to a value function gap that is filtered by $\zeta$:

$$v^\pi_t - v^\pi_{\hat{\pi}_t} \leq \|\zeta (Q^\hat{\pi} - f_t)\|_{1, \nu_{\hat{\pi}_t} / (1 - \gamma)},$$

where $\nu_1$ is some admissible distribution over $S \times A$. The filter $\zeta$ allows the change of measure from $\nu_t$ to $\mu$ without constraining the density ratio between an arbitrary distribution $\nu$ and $\mu$. Instead for any $s, a$ where $\zeta$ is one, by definition $\mu$ is lower bounded and the density ratio is bounded by $C$ (details in Lemma 13).

The rest of the proof has a similar structure with the standard FQI analysis. In Lemma 15, we bound the norm $\|\zeta (Q^F - f_t)\|_{2, \nu}$ by $C \| (f_t - T_{\hat{\pi}}f_t)\|_{2, \mu} / (1 - \gamma)$ and one additional sub-optimality error $\|Q^\pi - T_{\hat{\pi}}Q^\hat{\pi}\|_{2, \mu}$. The additional sub-optimality error term comes from the fact that $\tilde{\pi}$ may not be an optimal policy since the optimal policy may not be a $\zeta$-constrained policy. The last step to finish the proof is to bound the expected Bellman residual by concentration inequality. Lemma 16 shows how to bound that following a similar approach as [12]. Then the main theorem is proved by combine all those steps. After that we prove when we can bound the value gap with respect to optimal value in Corollary 4.

Now we start the proof. We are going to condition on the high probability bounds in Assumption 2 holds when we proof the lemmas.

Lemma 13. For $\pi_t = \mathcal{E}(\hat{\pi}_t)$ in Algorithm 2 for any policy $\hat{\pi} \in \Pi^\text{all}_{SC}$ we have

$$v^\pi_t - v^\pi_{\hat{\pi}_t} \leq \sum_{h=0}^\infty \gamma^h \left( \|\zeta (Q^\hat{\pi} - f_t)\|_{1, \nu_{\hat{\pi}_t} \times \hat{\pi}} + \|\zeta (Q^\pi - f_t)\|_{1, \nu_{\hat{\pi}_t} \times \pi_t} \right).$$

Proof. Given a deterministic greedy policy $\tilde{\pi}_t$, $\pi_t = \mathcal{E}(\tilde{\pi}_t)$ is also a deterministic policy and $\pi_t(s)$ equals $\tilde{\pi}_t(s)$ unless $\zeta(s, \tilde{\pi}_t(s)) = 0$, where $\pi_t(s) = \alpha_{\text{abs}}$. Notice $\tilde{\pi}_t(s)$ is the maximizer of $\zeta(s, \cdot) f_t(s, \cdot)$. If $\zeta(s, \tilde{\pi}_t(s)) = 0$ then $\zeta(s, \alpha) f_t(s, \alpha) = 0$ for all $\alpha$. We have that $\pi_t(s)$ is also the
maximizer of $\zeta(s, \cdot) f_t(s, \cdot)$.

\[
v^\pi - v^\pi_t = \sum_{h=0}^{\infty} \gamma^h E_{s \sim \eta_h} [Q^\pi(s, \bar{\pi}) - Q^\pi(s, \pi_t)]
\]

(3) Lemma 6.1)

\[
\leq \sum_{h=0}^{\infty} \gamma^h E_{s \sim \eta_h} \left[ \zeta(s, \bar{\pi})Q^\pi(s, \bar{\pi}) - \zeta(s, \pi_t)Q^\pi(s, \pi_t) \right] 
\]

(108)

\[
\leq \sum_{h=0}^{\infty} \gamma^h \left[ \zeta(s, \bar{\pi})Q^\pi(s, \bar{\pi}) - \zeta(s, \bar{\pi})f_t(s, \bar{\pi}) + \zeta(s, \pi_t)f_t(s, \pi_t) - \zeta(s, \pi_t)Q^\pi(s, \pi_t) \right] 
\]

(109)

\[
\leq \sum_{h=0}^{\infty} \gamma^h \left( \| \zeta (Q^\pi - f_t) \|_{1, \eta_h^t \times \bar{\pi}} + \| \zeta (Q^\pi - f_t) \|_{1, \eta_h^t \times \pi_t} \right) 
\]

(110)

Equation (108) follows from the fact that for all $s, a$ such that $\bar{\pi}(a|s) > 0$, either $\zeta(s, a) = 1$, or $\alpha = \alpha_{abs}$, $\alpha = \alpha_{abs}$ indicates that $Q^\pi(s, a) = 0$. So for all $s, a$ such that $\bar{\pi}(a|s) > 0$, $Q^\pi(s, a) = \zeta(s, a)Q^\pi(s, a)$. The second part follows from that for any $s, a, Q^\pi(s, a) \geq \zeta(s, a)Q^\pi(s, a)$. Equation (109) follows from the fact that $\pi_t(s)$ is the maximizer of $\zeta(s, \cdot) f_t(s, \cdot)$.

Lemma 14. For any two function $f_1, f_2 : S' \times A' \to \mathbb{R}^+$, define $\pi_{f_1, f_2}(s) = \arg\max_{a \in A} |f_1(s, a) - f_2(s, a)|$. Then we have $\forall \nu : S' \to \Delta(A')$,

\[
\left\| \max_{a \in A} f_1 - \max_{a \in A} f_2 \right\|_{1, \nu_\pi} \leq \left\| f_1 - f_2 \right\|_{1, \nu} \times \pi_{f_1, f_2}.
\]

Proof.

\[
\left\| \max_{a \in A} f_1 - \max_{a \in A} f_2 \right\|_{1, \nu_\pi} = E_{s \sim \nu} \max_{a \in A} f_1(s, a) - \max_{a \in A} f_2(s, a)
\]

\[
\leq E_{s \sim \nu} \max_{a \in A} |f_1(s, a) - f_2(s, a)|
\]

\[
= E_{s \sim \nu, a \sim \pi_{f_1, f_2}} |f_1(s, a) - f_2(s, a)|
\]

\[
= \left\| f_1 - f_2 \right\|^2_{1, \nu_\pi \times \pi_{f_1, f_2}}.
\]

Lemma 15. For the data distribution $\mu$ and any admissible distribution $\nu$ over $S' \times A'$, $f, f' : S \times A \to \mathbb{R}^+$ and any $\bar{\pi} \in \Pi_{\Sigma}^{\mathbb{R}^+}$, we have

\[
\left\| \zeta \left( f - Q^\pi \right) \right\|_{1, \nu} \leq C \left( \left\| f - T_\zeta f' \right\|_{2, \mu} + \left\| T_\zeta Q^\pi - Q^\pi \right\|_{2, \mu} + V_{\max} e_\mu \right)
\]

\[
+ \gamma \left\| \zeta \left( f' - Q^\pi \right) \right\|_{2, \nu_\pi \times \pi_{f', f' \sim Q^\pi}}.
\]
The change of norms from $\| \cdot \|_{1, \nu}$ follows from that $\zeta(s, a) \neq 0$ iff $\tilde{\mu}(s, a) \geq b$ and thus $\nu(s, a) \leq \tilde{\mu}(s, a) U/b = C \tilde{\mu}(s, a)$. The last step follows from Lemma 14 $\| \zeta \|_{1, \nu} \leq \gamma \| \max_{a \in A} \zeta f' - \max_{a \in A} \zeta Q^\pi \|_{1, P(\nu)}$ follows from:

$$
\| \zeta \|_{1, \nu} = \mathbb{E}_{(s, a) \sim P(s, a)} \zeta(s, a) \left| \zeta f'(s, a) - \zeta Q^\pi(s, a) \right|
$$

Now we are going to use Berstein’s inequality to bound $\| f_{t+1} - T_{\zeta} f_{t} \|_{2, \mu}$, which mostly follows from [11]’s proof for the vanilla value iteration.

**Lemma 16.** With Assumption 5 holds, let $g_f^* = \arg \min_{g \in F} \| g - T_{\zeta} f \|_{2, \mu}$, then $\| g_f^* - T_{\zeta} f \|_{2, \mu} \leq \epsilon_F$. The dataset $D$ is generated i.i.d. from $M$ as follows: $(s, a) \sim \mu$, $r \sim P(s, a)$. Define $L_{\mu}(f; f') = \mathbb{E} [L_D(f; f')]$. We have that $\forall f \in F$, with probability at least $1 - \delta$, $L_{\mu}(T_{\zeta} D f; f) - L_{\mu}(g_f^*; f) \leq \frac{208 V_{\max}^2 \ln |F|}{3n} + \epsilon_F$ where $T_{\zeta} D f = \arg \min_{g \in F} L_D(g, f)$.

**Proof.** This proof is similar with the proof of Lemma 7 and we adapt it to operator $T_{\zeta}$. The only change is the definition of $V_f(\cdot)$ and $X(\cdot, \cdot, \cdot)$. The definition of $L_D$ and $L_{\mu}$ would not change between $M$ and $M'$, and the right hand side is also the same constant for $M$ and $M'$. So the result we prove here does not change from $M$ to $M'$.

For the simplicity of notations, let $V_f(s) = \max_{a \in A} \zeta(s, a) f(s, a)$. Fix $f, g \in F$, and define $X(g, f, g_f^*) := (g(s, a) - r - \gamma V_f(s'))^2 - (g_f^*(s, a) - r - \gamma V_f(s'))^2$.
Plugging each \((s, a, r, s') \in D\) into \(X(g, f, g_f)\), we get i.i.d. variables \(X_1(g, f, g_f), X_2(g, f, g_f), \ldots, X_n(g, f, g_f)\). It is easy to see that
\[
\frac{1}{n} \sum_{i=1}^{n} X_i(g, f, g_f) = \mathcal{L}_D(g; f) - \mathcal{L}_D(g_f; f).
\]

By the definition of \(\mathcal{L}_\mu\), it is also easy to see that
\[
\mathcal{L}_\mu(g; f) = \|g - T_c f\|_2^2 + \mathbb{E}_{s,a \sim \mu} \left[ \mathbb{V}_{r,s'} \left( r + \gamma \max_{a' \in A} \zeta(s', a') f(s', a') \right) \right].
\]
Notice that the second part does not depend on \(g\). Then
\[
\mathcal{L}_\mu(g; f) - \mathcal{L}_\mu(T_c f; f) = \|g - T_c f\|_2^2,
\]
and
\[
\mathbb{E}[X(g, f, g_f)] \leq \mathbb{E}[X(g, f, g_f)^2]
= \mathbb{E}_\mu \left[ \left( (g(s, a) - r - \gamma V_f(s'))^2 - (g_f(s, a) - r - \gamma V_f(s'))^2 \right)^2 \right]
= \mathbb{E}_\mu \left[ \left( (g(s, a) - g_f(s, a))^2 (g(s, a) + g_f(s, a) - 2r - 2\gamma V_f(s')) \right)^2 \right]
\leq 4V^2 \max_s \mathbb{E}_\mu \left[ (g(s, a) - g_f^*(s, a))^2 \right]
= 4V^2 \max_s \|g - g_f^*\|_2^2\]
(122)
\[
\leq 8V^2 \max_s (\mathbb{E}[X(g, f, g_f)] + 2\epsilon_f).
\]

Step (*) holds because
\[
\|g - g_f^*\|_2^2
\leq 2 \left( \|g - T_c f\|_2^2 + \|T_c f - g_f^*\|_2^2 \right)
= 2 \left[ (\mathcal{L}_\mu(g; f) - \mathcal{L}_\mu(T_c f; f)) - (\mathcal{L}_\mu(g_f^*; f) - \mathcal{L}_\mu(T_c f; f)) + 2\|T_c f - g_f^*\|_2^2 \right]
= 2 \left( \mathbb{E}[X(g, f, g_f)] + 2\|T_c f - g_f^*\|_2^2 \right)
\leq 2(\mathbb{E}[X(g, f, g_f)] + 2\epsilon_f)
\]

Next, we apply (one-sided) Bernstein’s inequality and union bound over all \(f \in F\) and \(g \in F\). With probability at least \(1 - \delta\), we have
\[
\mathbb{E}[X(g, f, g_f)] - \frac{1}{n} \sum_{i=1}^{n} X_i(g, f, g_f) \leq \sqrt{\frac{2\mathbb{V}[X(g, f, g_f)] \ln \frac{\|F\|_2^2}{\delta}}{n}} + \frac{4V^2 \max \ln \frac{\|F\|_2^2}{\delta}}{3n}
\]
\[
= \sqrt{\frac{32V^2 \max \mathbb{E}[X(g, f, g_f)] + 2\epsilon_f \ln \frac{\|F\|_2^2}{\delta}}{n}} + \frac{8V^2 \max \ln \frac{\|F\|_2^2}{\delta}}{3n}
\]

Since \(T_{\zeta,D} f\) minimizes \(\mathcal{L}_D(\cdot; f)\), it also minimizes \(\frac{1}{n} \sum_{i=1}^{n} X_i(\cdot, f, g_f)\). This is because the two objectives only differ by a constant \(\mathcal{L}_D(g_f^*; f)\). Hence,
\[
\frac{1}{n} \sum_{i=1}^{n} X_i(T_{\zeta,D} f, f, g_f) \leq \frac{1}{n} \sum_{i=1}^{n} X_i(g_f^*, f, g_f) = 0.
\]

Then,
\[
\mathbb{E}[X(T_{\zeta,D} f, f, g_f)] \leq \sqrt{\frac{32V^2 \max \mathbb{E}[X(T_{\zeta,D} f, f, g_f)] + 2\epsilon_f \ln \frac{\|F\|_2^2}{\delta}}{n}} + \frac{8V^2 \max \ln \frac{\|F\|_2^2}{\delta}}{3n}.
\]
Solving for the quadratic formula,
\[
\mathbb{E}[X(T_{\tilde{\pi},D}f,f,g^*_f)] \leq \left( \frac{112V_{\max}^2 \ln \frac{[f]}{\delta}}{3n} + \frac{128V_{\max}^2 \ln \frac{[f]}{\delta}}{n} + \frac{64V_{\max}^2 \ln \frac{[f]}{\delta}}{n} + \frac{56V_{\max}^2 \ln \frac{[f]}{\delta}}{3n} \right) + \epsilon_f.
\]
Noticing that \(\mathbb{E}[X(T_{\tilde{\pi},D}f,f,g^*_f)] = \mathcal{L}_\mu(T_{\tilde{\pi},D}f;f) - \mathcal{L}_\mu(g^*_f;f)\), we complete the proof. \(\square\)

Now we could prove the main theorem about fitted Q iteration.

**Theorem 5.** Given a MDP \(M = \langle S,A,R,P,\gamma,p \rangle\), a dataset \(D = \{(s,a,r,s')\}\) with \(n\) samples that is draw i.i.d. from \(\mu \times R \times P\), and a finite Q-function classes \(\mathcal{F}\) satisfying Assumption 5 \(\pi_t = \Xi(\tilde{\pi}_t)\) from Algorithm 2 satisfies that with probability at least \(1 - 2\delta\), \(v^\pi - v^\pi_t \leq \frac{2C}{(1 - \gamma)^2} \left( \sqrt{\frac{208V_{\max}^2 \ln \frac{[f]}{\delta}}{3n}} + 2\sqrt{\epsilon_f} + V_{\max}\epsilon_\mu + \left\| Q^\pi - T_{\tilde{\pi}}Q^\pi \right\|_{1,\mu} \right) + \frac{2\gamma^4V_{\max}}{1 - \gamma}
\]
for any policy \(\tilde{\pi} \in \Pi_{SC}^{\mathcal{H}}\).

**Proof.** Firstly, we can let \(f = f_t\) and \(f' = f_{t-1}\) in Lemma 15. This gives us that
\[
\left\| f_t - Q^\pi \right\|_{1,\mu} \leq C \left( \left\| f_t - T_{\tilde{\pi}}f_{t-1}\right\|_{2,\mu} + \left\| Q^\pi - T_{\tilde{\pi}}Q^\pi \right\|_{1,\mu} + 2V_{\max}\epsilon_\mu \right) + \gamma\left\| f_{t-1} - Q^\pi \right\|_{1,P(\nu) \times \pi_{f_{t-1}},Q^\pi}
\]
Note that we can apply the same analysis on \(P(\nu) \times \pi_{f_{t-1},Q^\pi}\) and expand the inequality \(t\) times. It then suffices to upper bound \(\left\| f_t - T_{\tilde{\pi}}f_{t-1}\right\|_{2,\mu}\).

\[
\begin{align*}
\left\| f_t - T_{\tilde{\pi}}f_{t-1}\right\|_{2,\mu} &= \mathcal{L}_\mu(f_t;f_{t-1}) - \mathcal{L}_\mu(T_{\tilde{\pi}}f_{t-1};f_{t-1}) \
&= \mathcal{L}_\mu(f_t;f_{t-1}) - \mathcal{L}_\mu(g_{f_{t-1}}^*;f_{t-1}) + \mathcal{L}_\mu(g_{f_{t-1}}^*;f_{t-1}) - \mathcal{L}_\mu(T_{\tilde{\pi}}f_{t-1};f_{t-1}) \
&\leq \epsilon_4 + \left\| g_{f_{t-1}}^* - T_{\tilde{\pi}}f_{t-1}\right\|_{2,\mu} \
&\leq \epsilon_4 + \epsilon_f.
\end{align*}
\]
(Definition of \(g_{Q_{\tilde{\pi}}^{k-1}}\) and Assumption 5)

The inequality holds with probability at least \(1 - \delta\) and \(\epsilon_4 = \frac{208V_{\max}^2 \ln \frac{[f]}{\delta}}{3n} + \epsilon_f\). Noticing that \(\epsilon_4\) and \(\epsilon_f\) do not depend on \(t\), and the inequality holds simultaneously for different \(t\), we have that
\[
\left\| f_t - Q^\pi \right\|_{1,\mu} \leq \frac{1}{1 - \gamma} C \left( \sqrt{(\epsilon_4 + \epsilon_f)} + V_{\max}\epsilon_\mu + \left\| Q^\pi - T_{\tilde{\pi}}Q^\pi \right\|_{1,\mu} \right) + \gamma^t V_{\max}.
\]
Applying this to Lemma 13, we have that
\[
v^\bar{\pi} - v^{\pi_t} \leq \frac{2}{1 - \gamma} \left( \frac{1 - \gamma^t}{1 - \gamma} C \left( \sqrt{\epsilon_4 + \epsilon_F} + V_{\text{max}} \epsilon_\mu + \| Q^\bar{\pi} - T_{\zeta} Q^\bar{\pi} \|_{1, \mu} + \gamma^t V_{\text{max}} \right) \right)
\]
\[
\leq \frac{2C}{(1 - \gamma)^2} \left( \sqrt{\epsilon_4 + \epsilon_F} + V_{\text{max}} \epsilon_\mu + \| Q^\bar{\pi} - T_{\zeta} Q^\bar{\pi} \|_{1, \mu} \right) + \frac{2\gamma^t V_{\text{max}}}{1 - \gamma}.
\]

Now we are going to use the fact that there is an no-value-loss projection from the \( \zeta \)-constrained policy set to the strong \( \zeta \)-constrained policy set to prove an error bound w.r.t any \( \bar{\pi} \in \Pi_{\text{all}}^{\zeta} \).

**Theorem 2.** Given a MDP \( M = \langle S, A, R, P, \gamma, p \rangle \), a dataset \( D = \{ (s, a, r, s') \} \) with \( n \) samples that is drawn i.i.d. from \( \mu \times R \times P \), and a finite Q-function classes \( F \) satisfying Assumption 5 \( \bar{\pi}_t \) from Algorithm 2 satisfies that with probability at least \( 1 - 2\delta \),
\[
v^\bar{\pi} - v^{\pi_t} \leq \frac{2C}{(1 - \gamma)^2} \left( \sqrt{\frac{208V_{\text{max}} \ln \frac{|F|}{\delta}}{3n}} + 2\sqrt{\epsilon_F} + V_{\text{max}} \epsilon_\mu + \| Q^\bar{\pi} - T_{\zeta} Q^\bar{\pi} \|_{2, \mu} \right) + \frac{(2\gamma^t + \epsilon_\zeta) V_{\text{max}}}{1 - \gamma}
\]
for any policy \( \bar{\pi} \in \Pi_{\text{all}}^{\zeta} \).

**Proof.** The difference between this theorem and Theorem 5 is that \( \bar{\pi} \) is in \( \Pi_{\text{all}}^{\zeta} \) which is significantly larger than \( \Pi_{\text{SC}}^{\zeta} \).

This prove mimics the proof of Theorem 1. For any policy \( \bar{\pi} \in \Pi_{\text{all}}^{\zeta} \), Lemma 3 tells that \( v^\bar{\pi}_M \leq v^{\Xi(\bar{\pi})}_M + V_{\text{max}} \epsilon_\zeta \). Since \( \pi_t = \Xi(\bar{\pi}_t) \), \( v^\bar{\pi}_{M'} \geq v^{\pi_t}_M \). Then \( v^\bar{\pi}_{M'} - v^{\pi_t}_M \leq v^{\Xi(\bar{\pi})}_M - v^{\pi_t}_M + V_{\text{max}} \epsilon_\zeta \) and Theorem 5 completes the proof.

**Remark:** The first term in the theorem comes from that the best policy in the \( \zeta \)-constrained policy set is not optimal. Note that the \( \zeta \)-constrained policy set does not requires any realizability to do with our function approximation but merely about the density ratio of a policy. When there is an optimal policy of \( M \) such in \( \Pi_{\text{all}}^{\zeta} \), we have the same type of bound as standard approximate value iteration analysis.

**Corollary 4.** If there exists an \( \pi^* \) on \( M \) such that \( \Pr(\mu(s, a) \leq 2b|\pi^*) \leq \epsilon \), then under the condition as Theorem 4 \( \bar{\pi}_t \) from Algorithm 2 satisfies that with probability at least \( 1 - 2\delta \),
\[
v^\bar{\pi} - v^{\pi_t} \leq \frac{2C}{(1 - \gamma)^2} \left( \sqrt{\frac{208V_{\text{max}} \ln \frac{|F|}{\delta}}{3n}} + 2\sqrt{\epsilon_F} + V_{\text{max}} \epsilon_\mu + \| Q^\bar{\pi} - T_{\zeta} Q^\bar{\pi} \|_{2, \mu} \right) + \frac{V_{\text{max}}(2\gamma^t + \epsilon + CU \epsilon_\mu)}{1 - \gamma}
\]

**Proof.** The proof of \( \pi^* \in \Pi_{\text{C}}^{\zeta} \) is same as the proof in Corollary 1. Then proof is finished by applying Theorem 4.

**E** Details of CartPole Experiment

**E.1** Full results of Discretized CartPole-v0

In section 5.1 we compare AVI, BCQL 2, SPIBB 4. Behavior cloning and our algorithm PQL in CartPole-V0 with discretized state space. The data is generated by a \( \epsilon \)-greedy policy (\( \epsilon \) from 0.1 to 0.9) and we report the resulting policies from different algorithm with the best hyper-parameter in each \( \epsilon \). In this section we show the learning curve for each \( \epsilon \) and each hyper-parameter value. We run the BCQ algorithm with the threshold of \( \hat{p}(a|s) \) in \{0, 0.05, 0.1, 0.2\}, and we run the SPIBB
algorithm with the threshold of $\hat{\mu}(s, a)$ in \{0.01, 0.005, 0.001, 0.0005, 0.0001\} and PQI with the threshold of $\hat{\mu}(s, a)$ in a smaller set \{0.005, 0.001, 0.0005\}. Figure 1 shows for most of the $\epsilon$ and threshold our algorithm tie with the best baseline (SPIBB), and the best threshold of our algorithm outperform all baseline algorithms in 8 out of 9 cases.

In Figure 1, we observe the trend that smaller $\epsilon$ will prefer a smaller $b$. This is verified by more results in the next section, and we discuss the reasons for this phenomenon there.

E.2 Ablation study of threshold $b$

A key aspect of our algorithm is to filter the state space by a threshold on the estimated probability $\hat{\mu}(s, a)$. This prevents the algorithm from updating using low-confidence state, action pairs when bootstrapping values. Then the choice of threshold $b$ is a key trade-off in our algorithm: if $b$ is too small it can not remove the low-confident state, action pairs effectively; if $b$ is too large it might remove too many state, action pairs and prevent learning from more data. In order to demonstrate the effect of $b$ and how should we choose $b$ in different settings, we show the performance of PQI in a larger range of $b$ and several $\epsilon$ values.

In figure 2, we show the trend that smaller $b$ works better for larger $\epsilon$ and larger $b$ works better for smaller $\epsilon$ in general. This can be explained in the following way: with a larger $\epsilon$ the data distribution is more exploratory and hence the probabilities on individual state, action pairs are smaller. So a the same threshold that performs well with low exploration now censors a much larger part of the state, action space, necessitating a smaller threshold as $\epsilon$ is increased. In general, we find that having the largest threshold which still retains a significant fraction of the state, action space is a good heuristic for setting the $b$ parameter.
F Details of D4RL Experiment

In this section we introduce some missing details about the PQL algorithm and the experimental details in D4RL tasks. Our code is available at https://github.com/yaoliuc/PQL.

PQL algorithm is implemented based on the architecture of Batch-Constrained deep Q-learning (BCQ) \(^2\) algorithm. More specifically, we use the similar Clipped Double Q-Learning (CDQ) update rule for the \(Q\) learning part, and employ a similar variational auto-encoder to fit the conditional action distribution in the batch. We use an additional variational auto-encoder to fit the marginalized state distribution of the batch. To implement an actual \(Q\) learning algorithm instead of an actor-critic algorithm, we did not sample from the actor in the Bellman backup but sample a larger batch from the fitted conditional action distribution. Algorithm \(^4\) shows the pseudo-code of PQL to provide more details. We highlight the difference with the BCQ algorithm in red.

Algorithm 4 Pessimistic Q-learning (PQL)

**Input:** Batch \(D\), ELBO threshold \(b\), maximum perturbation \(\Phi\), target update rate \(\tau\), mini-batch size \(N\), max number of iteration \(T\). Number of actions \(k\).

Initialize two Q network \(Q_{\theta_1}\) and \(Q_{\theta_2}\), policy (perturbation) model: \(\xi_\phi\). (\(\xi_\phi \in [-\Phi, \Phi]\)), action VAE \(G_{\omega_1}^a\) and state VAE \(G_{\omega_2}^s\).

Pretrain \(G_{\omega_1}^a\): \(\omega_2 \leftarrow \arg \min_{\omega_2} \text{ELBO}(B; G_{\omega_2}^s)\).

for \(t = 1\) to \(T\) do

Sample a minibatch \(B\) with \(N\) samples from \(D\).

\(\omega_1 \leftarrow \arg \min_{\omega_2} \text{ELBO}(B; G_{\omega_2}^s)\).

Sample \(k\) actions \(a_i^j\) from \(G_{\omega_1}^a(s')\) for each \(s'\).

Compute the target \(y\) for each \((s, a, r, s')\) pair:

\[
y = r + \gamma \mathbb{I}(\text{ELBO}(s'; G_{\omega_2}^a) \geq b) \left[ \max_{a'} \left( 0.75 \times \min_{j=1,2} Q_{\theta_j} + 0.25 \times \max_{j=1,2} Q_{\theta_j} \right) \right]
\]

\(\theta \leftarrow \arg \min_{\theta} \sum(y - Q_{\theta}(s, a))^2\)

Sample \(k\) actions \(a_i\) from \(G_{\omega_1}^a(s)\) for each \(s\).

\(\phi \leftarrow \arg \max_{\phi} \sum \max_{a_i} Q_{\theta_1}(s, a_i + \xi_\phi(s, a_i))\)

Update target network: \(\theta' = (1 - \tau)\theta + \tau \phi, \phi' = (1 - \tau)\phi + \tau \phi\)

end for

When evaluate the resulting policy: select action \(a = \arg \max_{a_i} Q_{\theta_1}(s, a_i + \xi_\phi(s, a_i))\) where \(a_i\) are \(k\) actions sampled from \(G_{\omega_1}^a(s)\) given \(s\).

In practice, the indicator function \(\mathbb{I}(\text{ELBO}(s'; G_{\omega_2}^a) \geq b)\) is implemented by sigmoid(100(\(\text{ELBO}(s'; G_{\omega_2}^a) - b)\)) to provide a slightly more smooth target. The evidence lower bound (ELBO) in VAE is:

\[
\text{ELBO}(s; G_{\omega_2}^a) = \sum(s - \hat{s})^2 + D_{\text{KL}}(N(\mu, \sigma)||N(0, 1)) \tag{123}
\]

where \(\mu\) and \(\sigma\) is sampled from the encoder of VAE with input \(s\) and \(\hat{s}\) is sampled from the decoder with the hidden state generated from \(N(\mu, \sigma)\). \(\text{ELBO}(B; G_{\omega_2}^a)\) is the averaged ELBO on the minibatch \(B\). So does \(G_{\omega_1}^a\). Note that this ELBO objective make the implicit assumption that the decoder’s distribution is a Gaussian distribution with mean equals to the output of decoder network.
So when we generate the sample \( a' \) for computing \( y \), we add a Gaussian noise to recover a sample from the full posterior distribution.

For most of the hyper-parameters in Algorithm 4, we use the same value with the BCQ algorithm. We run all algorithms with \( T = 5 \times 10^5 \) gradient steps as other reported results in D4RL tasks, and the minibatch size \( N = 100 \) at each step. The number of sampled action when running the policy is \( k = 100 \). Target network update rate is 0.005. The threshold \( b \) of ELBO is selected as 2-percentile of the \( ELBO(s) \) in the whole dataset after pretrain the VAE.

References


