A. Equivalence between Regularization and Constraint Satisfaction

A.1. Formulating Different Regularized Policy Learning Problems as Constrained Policy Learning

In this section, we provide connections between regularized policy learning and our constrained formulation (OPT). Although the main paper focuses on batch policy learning, here we are agnostic between online and batch learning settings.

Entropy regularized RL. The standard reinforcement learning objective, either in online or batch setting, is to find a policy π^*_{std} that minimizes the long-term cost (equivalent to maximizing the accumuted rewards): $\pi^*_{\text{std}} = \arg\min_\pi \sum_t \mathbb{E}_{(x_t,a_t)\sim\pi}[c(x_t,a_t)] = \arg\min_\pi \mathbb{E}_{(x,a)\sim\mu_\pi}[c(x,a)]$. Maximum entropy reinforcement learning (Haarnoja et al., 2017) augments the cost with an entropy term, such that the optimal policy maximizes its entropy at each visited state: $\pi^*_{\text{MaxEnt}} = \arg\min_\pi \mathbb{E}_{(x,a)\sim\mu_\pi}[c(x,a)] - \lambda\mathbb{H}(\pi(\cdot|x))$. As discuseed by (Haarnoja et al., 2017), the goal is for the agent to maximize the entropy of the entire trajectory, and not greedily maximizing entropy at the current time step (i.e., Boltzmann exploration). Maximum entropy policy learning was first proposed by (Ziebart et al., 2008; Ziebart, 2010) in the context of learning from expert demonstrations. Entropy regulazed RL/IL is equivalent to our problem (OPT) by simply set $C(\pi) = \mathbb{E}_{(x_t,a_t)\sim\pi}[c(x_t,a_t)]$ (standard RL objective), and $g(x,a) = \pi(a|x)\log\pi(a|x)$, thus $G(\pi) = -\mathbb{H}(\pi) \le \tau$

Smooth imitation learning (& Regularized imitation learning). This is a constrained imitation learning problem studied by (Le et al., 2016): learning to mimic smooth behavior in continuous space from human desmonstrations. The data collected from human demonstrations is considered to be fixed and given a priori, thus the imitation learning task is also a batch policy learning problem. The proposed approach from (Le et al., 2016) is to view policy learning as a function regularization problem: policy $\pi = (f,g)$ is a combination of functions f and h, where f belongs to some expressive function class F (e.g., decision trees, neural networks) and $h \in$ H with certifiable smoothness property (e.g., linear models). Policy learning is the solution to the functional regularization problem $\pi = \arg\min_{f,g} \mathbb{E}_{x\sim\mu_\pi} \|f(x) - \pi_E(x)\| + \lambda\|h(x) - \pi_E(x)\|$, where π_E is the expert policy. This constrained imitation learning setting is equivalent to our problem (OPT) as follows: $C(\pi) = C((f,h)) = \mathbb{E}_{x\sim\mu_\pi}\|f(x) - \pi_E(x)\|$ and $G(\pi) = G((f,h)) = \min_{h'\in\text{H}}\|h'(x) - \pi_E(x)\| \le \tau$

Regularizing RL with expert demonstrations / Learning from imperfect demonstrations. Efficient exploration in RL is a well-known challenge. Expert demonstrations provide a way around online exploration to reduce the sample complexity for learning. However, the label budget for expert demonstrations may be limited, resulting in a sparse coverage of the state space compared to what the online RL agent can explore (Hester et al., 2018). Additionally, expert demonstrations may be imperfect (Oh et al., 2018). Some recent work proposed to regularize standard RL objective with some deviation measure between the learning policy and (sparse) expert data (Hester et al., 2018; Oh et al., 2018; Henaff et al., 2019).

For clarity we focus on the regularized RL objective for Q-learning in (Hester et al., 2018), which is defined as $J(\pi) = J_{DQ}(Q) + \lambda_1 J_n(Q) + \lambda_2 J_E(Q) + \lambda_3 J_{L2}(Q)$, where $J_{DQ}(Q)$ is the standard deep Q-learning loss, $J_n(Q)$ is the n-step return loss, $J_E(Q)$ is the imitation learning loss defined as $J_E(Q) = \max_{a\in\text{A}}[Q(x,a) + \ell(a_E,a) - Q(x,a_E)]$, and $J_{L2}(Q)$ is an L2 regularization loss applied to the Q-network to prevent overfitting to a small expert dataset. The regularization parameters λ's are obtained by hyperparameter tuning. This approach provides a bridge between RL and IL, whose objective functions are fundamentally different (see AggreVate by (Ross & Bagnell, 2014) for an alternative approach).

We can cast this problem into (OPT) as: $C(\pi) = C_{DQ}(Q) + \lambda_3 C_{L2}(Q)$ (standard RL objective), and two constraints: $g_1(\pi) = \mathbb{E}_{x\sim\mu_\pi}[\max_{a\in\text{A}} Q(x,a) + \ell(a_E,a) - Q(x,a_E)]$, and $g_2(x,a) = \mathbb{E}_{x\sim\mu_\pi}[c_t + \gamma c_{t+1} + \ldots + \gamma^{n-1}c_{t+n-1} + \min'_a \gamma^n Q(x_{t+n},a') - Q(x_t,a)]$. Here g_1 captures the loss w.r.t. expert demonstrations and g_2 reflects the n-step return constraint.

More generally, one can define the imitation learning constraint as $G(\pi) = \mathbb{E}_{x\sim\mu_\pi}\ell(\pi(x),\pi_E(x))$ for an appropriate divergence definition between $\pi(x)$ and $\pi_E(x)$ (defined at states where expert demonstrations are available).

Conservative policy improvement. Many policy search algorithms perform small policy update steps, requiring the new policy π to stay within a neighborhood of the most recent policy iterate π_k to ensure learning stability (Levine & Abbeel, 2014; Schulman et al., 2015; Montgomery & Levine, 2016; Achiam et al., 2017). This simply corresponds to the definition of $G(\pi) = \texttt{distance}(\pi,\pi_k) \le \tau$, where `distance` is typically KL-divergence or total variation distance between the distribution induced by π and π_k. For KL-divergence, the single timestep cost $g(x,a) = -\pi(a|x)\log(\frac{\pi_k(a|x)}{\pi(a|x)})$

A.2. Equivalence of Regularization and Constraint Viewpoint - Proof of Proposition 2.1

`Regularization` $\Longrightarrow$ `Constraint`: Let $\lambda > 0$ and π^* be optimal policy in `Regularization`. Set $\tau = G(\pi^*)$. Suppose that π^* is not optimal in `Constraint`. Then $\exists \pi \in \Pi$ such that $G(\pi) \leq \tau$ and $C(\pi) < C(\pi^*)$. We then have

$$C(\pi) + \lambda^\top G(\pi) < C(\pi^*) + \lambda^\top \tau = C(\pi^*) + \lambda^\top G(\pi^*)$$

which contradicts the optimality of π^* for `Regularization` problem. Thus π^* is also the optimal solution of the `Constraint` problem.

`Constraint` $\Longrightarrow$ `Regularization`: Given τ and let π^* be the corresponding optimal solution of the `Constraint` problem. The Lagrangian of `Constraint` is given by $L(\pi, \lambda) = C(\pi) + \lambda^\top G(\pi), \lambda \geq 0$. We then have $\pi^* = \arg\min_{\pi \in \Pi} \max_{\lambda \geq 0} L(\pi, \lambda)$. Let

$$\lambda^* = \arg\max_{\lambda \geq 0} \min_{\pi \in \Pi} L(\pi, \lambda)$$

Slater's condition implies strong duality. By strong duality and the strong max-min property (Boyd & Vandenberghe, 2004), we can exchange the order of maximization and minimization. Thus π^* is the optimal solution of

$$\min_{\pi \in \Pi} \quad C(\pi) + (\lambda^*)^\top (G(\pi) - \tau)$$

Removing the constaint $(\lambda^*)^\top \tau$, we have that π^* is the optimal solution of the `Regularization` problem with $\lambda = \lambda^*$. And since $\pi^* \neq \arg\min_{\pi \in \Pi} C(\pi)$, we must have $\lambda^* \geq 0$.

B. Convergence Proofs

B.1. Convergence of Meta-algorithm - Proof of Proposition 3.1

Let us evaluate the empirical primal-dual gap of the Lagrangian after T iterations:

$$\max_{\lambda} L(\widehat{\pi}_T, \lambda) = \max_{\lambda} \frac{1}{T} \sum_t L(\pi_t, \lambda) \tag{1}$$

$$\leq \frac{1}{T} \sum_t L(\pi_t, \lambda_t) + \frac{o(T)}{T} \tag{2}$$

$$\leq \frac{1}{T} \sum_t L(\pi, \lambda_t) + \frac{o(T)}{T} \quad \forall \pi \in \Pi \tag{3}$$

$$= L(\pi, \widehat{\lambda}_T) + \frac{o(T)}{T} \quad \forall \pi \tag{4}$$

Equations (1) and (4) are due to the definition of $\widehat{\pi}_T$ and $\widehat{\lambda}_T$ and linearity of $L(\pi, \lambda)$ wrt λ and the distribution over policies in Π. Equation (2) is due to the no-regret property of `Online-algorithm`. Equation (3) is true since π_t is best response wrt λ_t. Since equation (4) holds for all π, we can conclude that for T sufficiently large such that $\frac{o(T)}{T} \leq \omega$, we have $\max_\lambda L(\widehat{\pi}_T, \lambda) \leq \min_\pi L(\pi, \widehat{\lambda}_T) + \omega$, which will terminate the algorithm.

Note that we always have $\max_\lambda L(\widehat{\pi}_T, \lambda) \geq L(\widehat{\pi}_T, \widehat{\lambda}_T) \geq \min_\pi L(\pi, \widehat{\lambda}_T)$. Algorithm 1's convergence rate depends on the regret bound of the `Online-algorithm` procedure. Multiple algorithms exist with regret scaling as $\Omega(\sqrt{T})$ (e.g., online gradient descent with regularizer, variants of online mirror descent). In that case, the algorithm will terminate after $O(\frac{1}{\omega^2})$ iterations.

B.2. Empirical Convergence Analysis of Main Algorithm - Proof of Theorem 4.1

By choosing normalized exponentiated gradient as the online learning subroutine, we have the following regret bound after T iterations of the main algorithm 2 (chapter 2 of (Shalev-Shwartz et al., 2012)) for any $\lambda \in \mathbb{R}_+^{m+1}, \|\lambda\|_1 = B$:

$$\frac{1}{T} \sum_{t=1}^{T} \widehat{L}(\pi_t, \lambda) \leq \frac{1}{T} \sum_{t=1}^{T} \widehat{L}(\pi_t, \lambda_t) + \frac{\frac{B \log(m+1)}{\eta} + \eta \overline{G}^2 BT}{T} \tag{5}$$

Denote $\omega_T = \frac{\frac{B \log(m+1)}{\eta} + \eta \overline{G}^2 BT}{T}$ to simplify notations. By the linearity of $\widehat{L}(\pi, \lambda)$ in both π and λ, we have for any λ that

$$\widehat{L}(\widehat{\pi}_T, \lambda) \overset{\text{linearity}}{=} \frac{1}{T} \sum_{t=1}^{T} \widehat{L}(\pi_t, \lambda) \overset{\text{eqn (5)}}{\leq} \frac{1}{T} \sum_{t=1}^{T} \widehat{L}(\pi_t, \lambda_t) + \omega_T \overset{\text{best response } \pi_t}{\leq} \frac{1}{T} \sum_{t=1}^{T} \widehat{L}(\widehat{\pi}_T, \lambda_t) + \omega_T \overset{\text{linearity}}{=} \widehat{L}(\widehat{\pi}_T, \widehat{\lambda}_T) + \omega_T$$

Since this is true for any λ, $\max_\lambda \widehat{L}(\widehat{\pi}_T, \lambda) \leq \widehat{L}(\widehat{\pi}_T, \widehat{\lambda}_T) + \omega_T$.

On the other hand, for any policy π, we also have

$$\widehat{L}(\pi, \widehat{\lambda}_T) \overset{\text{linearity}}{=} \frac{1}{T} \sum_{t=1}^{T} \widehat{L}(\pi, \lambda_t) \overset{\text{best response } \pi_t}{\geq} \frac{1}{T} \sum_{t=1}^{T} \widehat{L}(\pi_t, \lambda_t) \overset{\text{eqn (5)}}{\geq} \frac{1}{T} \sum_{t=1}^{T} \widehat{L}(\pi_t, \widehat{\lambda}_T) - \omega_T \overset{\text{linearity}}{=} \widehat{L}(\widehat{\pi}_T, \widehat{\lambda}_T) - \omega_T$$

Thus $\min_\pi \widehat{L}(\pi, \widehat{\lambda}_T) \geq \widehat{L}(\widehat{\pi}_T, \widehat{\lambda}_T) - \omega_T$, leading to

$$\max_{\lambda} \widehat{L}(\widehat{\pi}_T, \lambda) - \min_{\pi} \widehat{L}(\pi, \widehat{\lambda}_T) \leq \widehat{L}(\widehat{\pi}_T, \widehat{\lambda}_T) + \omega_T - (\widehat{L}(\widehat{\pi}_T, \widehat{\lambda}_T) - \omega_T) = 2\omega_T$$

After T iterations of the main algorithm 2, therefore, the empirical primal-dual gap is bounded by

$$\max_{\lambda} \widehat{L}(\widehat{\pi}_T, \lambda) - \min_{\pi} \widehat{L}(\pi, \widehat{\lambda}_T) \leq \frac{2\frac{B \log(m+1)}{\eta} + 2\eta \overline{G}^2 BT}{T}$$

In particular, if we want the gap to fall below a desired threshold ω, setting the online learning rate $\eta = \frac{\omega}{4\overline{G}^2 B}$ will ensure that the algorithm converges after $\frac{16 B^2 \overline{G}^2 \log(m+1)}{\omega^2}$ iterations.

C. End-to-end Generalization Analysis of Main Algorithm

In this section, we prove the following full statement of theorem 4.4 of the main paper. Note that to lessen notation, we define $\overline{V} = \overline{C} + B\overline{G}$ to be the bound of value functions under considerations in algorithm 2.

Theorem C.1 (Generalization bound of algorithm 2)**.** *Let π^* be the optimal policy to problem OPT. Let K be the number of iterations of FQE and FQI. Let $\widehat{\pi}$ be the policy returned by our main algorithm 2, with termination threshold ω. For any $\epsilon > 0, \delta \in (0,1)$, when $n \geq \frac{24 \cdot 214 \cdot \overline{V}^4}{\epsilon^2}\big(\log\frac{K(m+1)}{\delta} + \mathtt{dim}_{\mathrm{F}}\log\frac{320\overline{V}^2}{\epsilon^2} + \log(14e(\mathtt{dim}_{\mathrm{F}}+1))\big)$, we have with probability at least $1-\delta$:*

$$C(\widehat{\pi}) \leq C(\pi^*) + \omega + \frac{(4+B)\gamma}{(1-\gamma)^3}\big(\sqrt{C_\rho}\epsilon + 2\gamma^{K/2}\overline{V}\big)$$

and

$$G(\widehat{\pi}) \leq \tau + 2\frac{\overline{V}+\omega}{B} + \frac{\gamma^{1/2}}{(1-\gamma)^{3/2}}\big(\sqrt{C_\rho}\epsilon + \frac{2\gamma^{K/2}\overline{V}}{(1-\gamma)^{1/2}}\big)$$

Let $\widehat{\pi} = \frac{1}{T}\sum_t \pi_t$ be the returned policy T iterations, with corresponding dual variable $\widehat{\lambda} = \frac{1}{T}\sum_t \lambda_t$.

By the stopping condition, the empirical duality gap is less than some threshold ω, i.e., $\max_{\lambda \in \mathbb{R}_+^{m+1}, \|\lambda\|_1 = B} \widehat{L}(\widehat{\pi}, \lambda) - \min_{\pi \in \Pi} \widehat{L}(\pi, \widehat{\lambda}) \leq \omega$ where $\widehat{L}(\pi,\lambda) = \widehat{C}(\pi) + \lambda^\top(\widehat{G}(\pi) - \tau)$. We first show that the returned policy approximately satisfies the constraints. The proof of theorem C.1 will make use of the following empirical constraint satisfaction bound:

Lemma C.2 (Empirical constraint satisfactions)**.** *Assume that the constraints $\widehat{G}(\pi) \leq \tau$ are feasible. Then the returned policy $\widehat{\pi}$ approximately satisfies all constraints*

$$\max_{i=1:m+1}(\widehat{g}_i(\widehat{\pi}) - \tau_i) \leq 2\frac{\overline{C}+\omega}{B}$$

Proof. We consider $\max_{i=1:m+1}(\widehat{g}_i(\widehat{\pi}) - \tau_i) > 0$ (otherwise the lemma statement is trivially true). The termination condition implies that $\widehat{L}(\widehat{\pi}, \widehat{\lambda}) - \max_{\lambda \in \mathbb{R}_+^{m+1}, \|\lambda\|_1 = B} \widehat{L}(\widehat{\pi}, \lambda) \geq -\omega$

$$\implies \widehat{\lambda}^\top(\widehat{G}(\widehat{\pi}) - \widehat{\tau}) \geq \max_{\lambda \in \mathbb{R}_+^{m+1}, \|\lambda\|_1 = B} \lambda^\top(\widehat{G}(\widehat{\pi}) - \widehat{\tau}) - \omega \tag{6}$$

Relaxing the RHS of equation (6) by setting $\lambda[j] = B$ for $j = \arg\max_{i=1:m+1}[\widehat{g}_i(\widehat{\pi}) - \tau_i]$ and $\lambda[i] = 0 \ \ \forall i \neq j$ yields:

$$B \max_{i=1:m+1}[\widehat{g}_i(\widehat{\pi}) - \tau_i] - \omega \leq \widehat{\lambda}^\top(\widehat{G}(\widehat{\pi}) - \tau) \tag{7}$$

Given π such that $\widehat{G}(\pi) \leq \tau$, also by the termination condition:

$$\widehat{L}(\widehat{\pi}, \widehat{\lambda}) - \widehat{L}(\pi, \widehat{\lambda}) \leq \max_{\lambda \in \mathbb{R}_+^{m+1}, \|\lambda\|_1 = B} \widehat{L}(\widehat{\pi}, \lambda) - \min_{\pi \in \Pi} \widehat{L}(\pi, \widehat{\lambda}) \leq \omega$$

Thus implies

$$\widehat{L}(\widehat{\pi}, \widehat{\lambda}) \leq \widehat{L}(\pi, \widehat{\lambda}) + \omega = \widehat{C}(\pi) + \widehat{\lambda}^\top(\widehat{G}(\pi) - \tau) \leq \widehat{C}(\pi) + \omega \tag{8}$$

combining what we have from equation (8) and (7):

$$B \max_{i=1:m+1}[\widehat{g}_i(\widehat{\pi}) - \widehat{\tau}_i] - \omega \leq \widehat{\lambda}^\top(\widehat{G}(\widehat{\pi}) - \widehat{\tau}) = \widehat{L}(\widehat{\pi}, \widehat{\lambda}) - \widehat{C}(\widehat{\pi}) \leq \widehat{C}(\pi) + \omega - \widehat{C}(\widehat{\pi})$$

Rearranging and bounding $\widehat{C}(\pi) \leq \overline{C}$ and $\widehat{C}(\widehat{\pi}) \leq -\overline{C}$ finishes the proof of the lemma. □

We now return to the proof of theorem C.1, our task is to lift empirical error to generalization bound for main objective and constraints.

Denote by ϵ_{FQE} the (generalization) error introduced by the Fitted Q Evaluation procedure (algorithm 3) and ϵ_{FQI} the (generalization) error introduced by the Fitted Q Iteration procedure (algorithm 4). For now we keep ϵ_{FQE} and ϵ_{FQI} unspecified (to be specified shortly). That is, for each $t = 1, 2, \ldots, T$, we have with probability at least $1-\delta$:

$$C(\pi_t) + \lambda_t^\top(G(\pi_t) - \tau) \leq C(\pi^*) + \lambda_t^\top(G(\pi^*) - \tau) + \epsilon_{FQI}$$

Since π^* satisfies the constraints, i.e., $G(\pi^*) - \tau \leq 0$ componentwise, and $\lambda_t \geq 0$, we also have with probability $1-\delta$

$$L(\pi_t, \lambda_t) = C(\pi_t) + \lambda_t^\top (G(\pi_t) - \tau) \leq C(\pi^*) + \epsilon_{FQI} \tag{9}$$

Similarly, with probability $1-\delta$, all of the following inequalities are true

$$\widehat{C}(\pi_t) + \epsilon_{FQE} \geq C(\pi_t) \geq \widehat{C}(\pi_t) - \epsilon_{FQE} \tag{10}$$

$$\widehat{G}(\pi_t) + \epsilon_{FQE}\mathbf{1} \geq G(\pi_t) \geq \widehat{G}(\pi_t) - \epsilon_{FQE}\mathbf{1} \text{ (row wise for all } m \text{ constraints)} \tag{11}$$

Thus with probability at least $1-\delta$

$$\begin{aligned} L(\pi_t, \lambda_t) = C(\pi_t) + \lambda_t^\top (G(\pi_t) - \tau) &\geq \widehat{C}(\pi_t) + \lambda_t^\top (\widehat{G}(\pi_t) - \tau) - \epsilon_{FQE}(1 + \lambda_t^\top \mathbf{1}) \\ &\geq \widehat{C}(\pi_t) + \lambda_t^\top (\widehat{G}(\pi_t) - \tau) - \epsilon_{FQE}(1 + B) \\ &= \widehat{L}(\pi_t, \lambda_t) - \epsilon_{FQE}(1 + B) \end{aligned} \tag{12}$$

Recall that the execution of mixture policy $\widehat{\pi}$ is done by uniformly sampling one policy π_t from $\{\pi_1, \ldots, \pi_T\}$, and rolling-out with π_t. Thus from equations (9) and (12), we have $\mathbb{E}_{t\sim U[1:T]} \widehat{L}(\pi_t, \lambda_t) \leq C(\pi^*) + \epsilon_{FQI} + (1+B)\epsilon_{FQE}$ w.p. $1-\delta$. In other words, with probability $1-\delta$:

$$\frac{1}{T}\sum_{t=1}^{T} \widehat{L}(\pi_t, \lambda_t) \leq C(\pi^*) + \epsilon_{FQI} + (1+B)\epsilon_{FQE}$$

Due to the no-regret property of our online algorithm (EG in this case):

$$\frac{1}{T}\sum_{t=1}^{T} \widehat{L}(\pi_t, \lambda_t) \geq \max_{\lambda} \widehat{L}(\widehat{\pi}, \lambda) - \omega = \widehat{C}(\widehat{\pi}) + \max_{\lambda} \lambda^\top (\widehat{G}(\widehat{\pi}) - \tau) - \omega$$

If $\widehat{G}(\widehat{\pi}) - \tau \leq 0$ componentwise, choose $\lambda[i] = 0, i = 1, 2, \ldots, m$ and $\lambda[m+1] = B$. Otherwise, we can choose $\lambda[j] = B$ for $j = \underset{i=1:m+1}{\arg\max} [\widehat{g}_i(\widehat{\pi}) - \tau[i]]$ and $\lambda[i] = 0 \ \ \forall i \neq j$. We can see that $\max_{\lambda \in \mathbb{R}_+^{m+1}, \|\lambda\|_1 = B} \lambda^\top (\widehat{G}(\widehat{\pi}) - \tau) \geq 0$. Therefore:

$$\widehat{C}(\widehat{\pi}) - \omega \leq C(\pi^*) + \epsilon_{FQI} + (1+B)\epsilon_{FQE} \text{ with probability at least } 1-\delta$$

Combined with the first term from equation (10):

$$C(\widehat{\pi}) - \epsilon_{FQE} - \omega \leq C(\pi^*) + \epsilon_{FQI} + (1+B)\epsilon_{FQE}$$

or

$$C(\widehat{\pi}) \leq C(\pi^*) + \omega + \epsilon_{FQI} + (2+B)\epsilon_{FQE} \tag{13}$$

We now bring in the generalization error results from our standalone analysis of FQI (appendix F) and FQE (appendix E) into equation (13).

Specifically, when $n \geq \frac{24 \cdot 214 \cdot \overline{V}^4}{\epsilon^2}\left(\log \frac{K(m+1)}{\delta} + \texttt{dim}_\text{F} \log \frac{320\overline{V}^2}{\epsilon^2} + \log(14e(\texttt{dim}_\text{F} + 1))\right)$, when FQI and FQE are run with K iterations, we have the guarantee that for any $\epsilon > 0$, with probability at least $1-\delta$

$$\begin{aligned} C(\widehat{\pi}) &\leq C(\pi^*) + \omega + \underbrace{\frac{2\gamma}{(1-\gamma)^3}\left(\sqrt{C_\mu}\epsilon + 2\gamma^{K/2}\overline{V}\right)}_{\text{FQI generalization error}} + \underbrace{\frac{\gamma^{1/2}(2+B)}{(1-\gamma)^{3/2}}\left(\sqrt{C_\mu}\epsilon + \frac{\gamma^{K/2}}{(1-\gamma)^{1/2}}2\overline{V}\right)}_{(2+B)\times \text{ FQE generalization error}} \\ &\leq C(\pi^*) + \omega + \frac{(4+B)\gamma}{(1-\gamma)^3}\left(\sqrt{C_\mu}\epsilon + 2\gamma^{K/2}\overline{V}\right) \end{aligned} \tag{14}$$

From lemma C.2, $\widehat{G}(\widehat{\pi}) \leq \tau + 2\frac{\overline{C}+\omega}{B} \leq \tau + 2\frac{\overline{V}+\omega}{B}$. From equation (11), for each t=1,2,…,T, we have $\widehat{G}(\pi_t) \geq G(\pi_t) - \epsilon_{FQE}\mathbf{1}$ with probability $1-\delta$. Thus

$$\mathbf{P}\left(\widehat{G}(\widehat{\pi}) \geq G(\widehat{\pi}) - \epsilon_{FQE}\mathbf{1}\right) = \sum_{t=1}^{T} \mathbf{P}(\widehat{G}(\pi_t) \geq G(\pi_t) - \epsilon_{FQE}\mathbf{1} | \widehat{\pi} = \pi_t)\mathbf{P}(\widehat{\pi} = \pi_t) \geq T(1-\delta)\frac{1}{T} = 1-\delta$$

Therefore, we have the following generalization guarantee for the approximate satisfaction of all constraints:

$$G(\widehat{\pi}) \leq \tau + 2\frac{\overline{V}+\omega}{B} + \frac{\gamma^{1/2}}{(1-\gamma)^{3/2}}\left(\sqrt{C_\mu}\epsilon + \frac{\gamma^{K/2}}{(1-\gamma)^{1/2}}2\overline{V}\right) \tag{15}$$

Inequalities (14) and (15) complete the proof of theorem C.1 (and theorem 4.4 of the main paper)

D. Preliminaries to Analysis of Fitted Q Evaluation (FQE) and Fitted Q Iteration (FQI)

In this section, we set-up necessary notations and definitions for the theoretical analysis of FQE and FQI. To simplify the presentation, we will focus exclusively on weighted ℓ_2 norm for error analysis.

With the definitions and assumptions presented in this section, we will present the sample complexity guarantee of Fitted-Q-Evaluation (FQE) in appendix E. The proof for FQI will follow similarly in appendix F.

While it is possible to adapt proofs from related algorithms (Munos & Szepesvári, 2008; Antos et al., 2008b) to analyze FQE and FQI, in the next two sections we show improved convergence rate from $O(n^{-4})$ to $O(n^{-2})$, where n is the number of samples in data set D.

To be consistent with the notations in the main paper, we use the convention $C(\pi)$ as the value function that denotes long-term accumulated cost, instead of using $V(\pi)$ denoting long-term rewards in the traditional RL literature. Our notation for Q function is similar to the RL literature - the only difference is that the optimal policy minimizes $Q(x,a)$ instead of maximizing. We denote the bound on the value function as $\overline{C}$ (alternatively if the single timestep cost is bounded by $\bar{c}$, then $\overline{C} = \frac{\bar{c}}{1-\gamma}$). For simplicity, the standalone analysis of FQE and FQI concerns only with the cost objective c. Dealing with cost $c + \lambda^\top g$ offers no extra difficulty - in that case we simply augment the bound of the value function to $\overline{V} = \overline{C} + B\overline{G}$.

D.1. Bellman operators

The *Bellman optimality operator* $\mathbb{T} : \mathcal{B}(\mathrm{X} \times \mathrm{A}; \overline{C}) \mapsto \mathcal{B}(\mathrm{X} \times \mathrm{A}; \overline{C})$ as

$$(\mathbb{T}Q)(x,a) = c(x,a) + \gamma \int_{\mathrm{X}} \min_{a' \in \mathrm{A}} Q(x',a') p(dx'|x,a) \tag{16}$$

The optimal value functions are defined as usual by $C^*(x) = \sup_\pi C^\pi(x)$ and $Q^*(x,a) = \sup_\pi Q^\pi(x,a) \ \ \forall x \in \mathrm{X}, a \in \mathrm{A}$.

For a given policy π, the *Bellman evaluation operator* $\mathbb{T}^\pi : \mathcal{B}(\mathrm{X} \times \mathrm{A}; \overline{C}) \mapsto \mathcal{B}(\mathrm{X} \times \mathrm{A}; \overline{C})$ as

$$(\mathbb{T}^\pi Q)(x,a) = c(x,a) + \gamma \int_{\mathrm{X}} Q(x', \pi(x')) p(dx'|x,a) \tag{17}$$

It is well known that $\mathbb{T}^\pi Q^\pi = Q^\pi$, a fixed point of the $\mathbb{T}^\pi$ operator.

D.2. Data distribution and weighted ℓ_2 norm

Denote the state-action data generating distribution as μ, induced by some data-generating (behavior) policy π_{D}, that is, $(x_i, a_i) \sim \mu$ for $(x_i, a_i, x_i', c_i) \in \mathrm{D}$.

Note that data set D is formed by multiple trajectories generated by π_{D}. For each (x_i, a_i), we have $x_i' \sim p(\cdot|x_i, a_i)$ and $c_i = c(x_i, a_i)$. For any (measurable) function $f : \mathrm{X} \times \mathrm{A} \mapsto \mathbb{R}$, define the μ-weighted ℓ_2 norm of f as $\|f\|_\mu^2 = \int_{\mathrm{X}\times\mathrm{A}} f(x,a)^2 \mu(dx, da) = \int_{\mathrm{X}\times\mathrm{A}} f(x,a)^2 \mu_x(dx) \pi_{\mathrm{D}}(a|dx)$. Similarly for any other state-action distribution ρ, $\|f\|_\rho^2 = \int_{\mathrm{X}\times\mathrm{A}} f(x,a)^2 \rho(dx, da)$

D.3. Inherent Bellman error

FQE and FQI depend on a chosen function class F to approximate $Q(x,a)$. To express how well the Bellman operator $\mathbb{T}g$ can be approximated by a function in the policy class F, when $\mathbb{T}g \notin \mathrm{F}$, a notion of distance, known as inherent Bellman error was first proposed by (Munos, 2003) and used in the analysis of related ADP algorithms (Munos & Szepesvári, 2008; Munos, 2007; Antos et al., 2008a;b; Lazaric et al., 2010; 2012; Lazaric & Restelli, 2011; Maillard et al., 2010).

Definition D.1 (Inherent Bellman Error)**.** Given a function class F and a chosen distribution ρ, the *inherent Bellman error* of F is defined as

$$d_{\mathrm{F}} = d(\mathrm{F}, \mathbb{T}\mathrm{F}) = \sup_{h \in \mathrm{F}} \inf_{f \in \mathrm{F}} \|f - \mathbb{T}h\|_\rho$$

where $\|\cdot\|_\rho$ is the $\rho-$weighted ℓ_2 norm and $\mathbb{T}$ is the Bellman optimality operator defined in (16)

To analyze FQE, we will form a similar definition for the Bellman evaluation operator

Definition D.2 (Inherent Bellman Evaluation Error)**.** Given a function class F and a policy π, the *inherent Bellman*

evaluation error of F is defined as

$$d_{\mathrm{F}}^{\pi} = d(\mathrm{F}, \mathbb{T}^{\pi}\mathrm{F}) = \sup_{h \in \mathrm{F}} \inf_{f \in \mathrm{F}} \|f - \mathbb{T}^{\pi} h\|_{\rho_\pi}$$

where $\|\cdot\|_{\rho_\pi}$ is the ℓ_2 norm weighted by ρ_π. ρ_π is defined as the state-action distribution induced by policy π, and $\mathbb{T}^{\pi}$ is the Bellman operator defined in (17)

D.4. Concentrability coefficients

Let P^{π} denote the operator acting on $f : \mathrm{X} \times \mathrm{A} \mapsto \mathbb{R}$ such that $(P^{\pi} f)(x, a) = \int_{\mathrm{X}} f(x', \pi(x')) p(x'|x, a) dx'$. Acting on f (e.g., approximates Q), P^{π} captures the transition dynamics of taking action a and following π thereafters.

The following definition and assumption are standard in the analysis of related approximate dynamic programming algorithms (Lazaric et al., 2012; Munos & Szepesvári, 2008; Antos et al., 2008a). As approximate value iteration and policy iteration algorithms perform policy update, the new policy at each round will induce a different stationary state-action distribution. One way to quantify the distribution shift is the notion of concentrability coefficient of future state-action distribution, a variant of the notion introduced by (Munos, 2003).

Definition D.3 (Concentrability coefficient of state-action distribution)**.** Given data generating distribution $\mu \sim \pi_{\mathrm{D}}$, initial state distribution χ. For $m \geq 0$, and an arbitrary sequence of stationary policies $\{\pi_m\}_{m \geq 1}$ let

$$\beta_{\mu}(m) = \sup_{\pi_1, \ldots, \pi_m} \left\| \frac{d(\chi P^{\pi_1} P^{\pi_2} \ldots P^{\pi_m})}{d\mu} \right\|_{\infty}$$

($\beta_{\mu}(m) = \infty$ if the future state distribution $\chi P^{\pi_1} P^{\pi_2} \ldots P^{\pi_m}$ is not absolutely continuous w.r.t. μ, i.e, $\chi P^{\pi_1} P^{\pi_2} \ldots P^{\pi_m}(x, a) > 0$ for some $\mu(x, a) = 0$)

Assumption 3. $\beta_{\mu} = (1 - \gamma)^2 \sum_{m \geq 1} m \gamma^{m-1} \beta_{\mu}(m) < \infty$

Combination Lock Example. An example of an MDP that violates Assumption 3 is the "combination lock" example proposed by (Koenig & Simmons, 1996). In this finite MDP, we have N states $\mathrm{X} = \{1, 2, \ldots, N\}$, and 2 actions: going L or R. The initial state is $x_0 = 1$. In any state x, action L takes agent back to initial state x_0, and action R advances the agent to the next state $x + 1$ in a chain fashion. Suppose that the reward is 0 everywhere except for the very last state N. One can see that for an MDP such that any behavior policy π_D that has a bounded from below probability of taking action L from any state x, i.e., $\pi_D(L|x) \geq \nu > 0$, then it takes an exponential number of trajectories to learn or evaluate a policy that always takes action R. In this setting, we can see that the concentration coefficient β_{μ} can be designed to be arbitrarily large.

D.5. Complexity measure of function class F

Definition D.4 (Random L_1 Norm Covers)**.** Let $\epsilon > 0$, let F be a set of functions $\mathrm{X} \mapsto \mathbb{R}$, let $x_1^n = (x_1, \ldots, x_n)$ be n fixed points in X. Then a collection of functions $\mathrm{F}_{\epsilon} = \{f_1, \ldots, f_N\}$ is an ϵ-cover of F on x_1^n if

$$\forall f \in \mathrm{F}, \exists f' \in \mathrm{F}_{\epsilon} : |\frac{1}{n} \sum_{i=1}^{n} f(x_i) - \frac{1}{n} \sum_{i=1}^{n} f'(x_i)| \leq \epsilon$$

The empirical covering number, denote by $\mathcal{N}_1(\epsilon, \mathrm{F}, x_1^n)$, is the size of the smallest ϵ-cover on x_1^n. Take $\mathcal{N}_1(\epsilon, \mathrm{F}, x_1^n) = \infty$ if no finite ϵ-cover exists.

Definition D.5 (Pseudo-Dimension)**.** A real-valued function class F has pseudo-dimension $\mathtt{dim}_{\mathrm{F}}$ defined as the VC dimension of the function class induced by the sub-level set of functions of F. In other words, define function class $\mathrm{H} = \{(x, y) \mapsto \mathrm{sign}(f(x) - y : f \in \mathrm{F}\}$, then

$$\mathtt{dim}_{\mathrm{F}} = \mathtt{VC\text{-}dimension}(\mathrm{H})$$

E. Generalization Analysis of Fitted Q Evaluation

In this section we prove the following statement for Fitted Q Evaluation (FQE).

Theorem E.1 (Guarantee for FQE - General Case (theorem 4.2 in main paper)). *Under Assumption 3, for $\epsilon > 0$ & $\delta \in (0,1)$, after K iterations of Fitted Q Evaluation (Algorithm 3), for $n = O\big(\frac{\overline{C}^4}{\epsilon^2}(\log\frac{K}{\delta} + \texttt{dim}_\text{F}\log\frac{\overline{C}^2}{\epsilon^2} + \log\texttt{dim}_\text{F})\big)$, we have with probability $1-\delta$:*

$$\big|C(\pi) - \widehat{C}(\pi)\big| \leq \frac{\gamma^{1/2}}{(1-\gamma)^{3/2}}\big(\sqrt{\beta_\mu}\,(2d_\text{F}^\pi + \epsilon) + \frac{2\gamma^{K/2}\overline{C}}{(1-\gamma)^{1/2}}\big).$$

Theorem E.2 (Guarantee for FQE - Bellman Realizable Case). *Under Assumptions 3-4, for any $\epsilon > 0, \delta \in (0,1)$, after K iterations of Fitted Q Evaluation (Algorithm 3), when $n \geq \frac{24\cdot 214\cdot\overline{C}^4}{\epsilon^2}\big(\log\frac{K}{\delta} + \texttt{dim}_\text{F}\log\frac{320\overline{C}^2}{\epsilon^2} + \log(14e(\texttt{dim}_\text{F}+1))\big)$, we have with probability $1-\delta$:*

$$\big|C(\pi) - \widehat{C}(\pi)\big| \leq \frac{\gamma^{1/2}}{(1-\gamma)^{3/2}}\big(\sqrt{\beta_\mu}\epsilon + \frac{2\gamma^{K/2}\overline{C}}{(1-\gamma)^{1/2}}\big)$$

We first focus on theorem E.2, analyzing FQE assuming a sufficiently rich function class F so that the Bellman evaluation update $\mathbb{T}^\pi$ is closed wrt F (thus inherent Bellman evaluation error is 0). We call this the *Bellman evaluation realizability assumption*. This assumption simplifies the presentation of our bounds and also simplifies the final error analysis of Algo. 2.

After analyzing FQE under this Bellman realizable setting, we will turn to error bound for general, non-realizable setting in theorem E.1 (also theorem 4.2 in the main paper). The main difference in the non-realizable setting is the appearance of an extra term d_F^π our final bound.

E.1. Error bound for single iteration - Bellman realizable case

Assumption 4 (Bellman evaluation realizability). *We consider function classes* F *sufficiently rich so that* $\forall f, \mathbb{T}^\pi f \in \text{F}$.

We begin with the following result bounding the error for a single iteration of FQE, under "training" distribution $\mu \sim \pi_D$

Proposition E.3 (Error bound for single iteration). *Let the functions in* F *also be bounded by $\overline{C}$, and let $\texttt{dim}_\text{F}$ denote the pseudo-dimension of the function class* F. *We have with probability at least $1-\delta$:*

$$\|Q_k - \mathbb{T}^\pi Q_{k-1}\|_\mu < \epsilon$$

when $n \geq \frac{24\cdot 214\cdot\overline{C}^4}{\epsilon^2}\left(\log\frac{1}{\delta} + \texttt{dim}_\text{F}\log\frac{320\overline{C}^2}{\epsilon^2} + \log(14e(\texttt{dim}_\text{F}+1))\right)$

Remark E.4. *Note from proposition E.3 that the dependence of sample complexity n here on ϵ is $\widetilde{O}(\frac{1}{\epsilon^2})$, which is better than previously known analysis for Fitted Value Iteration (Munos & Szepesvári, 2008) and FittedPolicyQ (continuous version of Fitted Q Iteration (Antos et al., 2008a)) dependence of $\widetilde{O}(\frac{1}{\epsilon^4})$. The finite sample analysis of LSTD (Lazaric et al., 2010) showed an $\widetilde{O}(\frac{1}{\epsilon^2})$ dependence using linear function approximation. Here we prove similar convergence rate for general non-linear (bounded) function approximators.*

Proof of Proposition E.3. Recall the training target in round k is $y_i = c_i + \gamma Q_{k-1}(x_i', \pi(x_i'))$ for $i = 1, 2, \ldots, n$, and $Q_k \in \text{F}$ is the solution to the following regression problem:

$$Q_k = \arg\min_{f\in\text{F}} \frac{1}{n}\sum_{i=1}^n (f(x_i,a_i) - y_i)^2$$

Consider random variables $(x,a) \sim \mu$ and $y = c(x,a) + \gamma Q_{k-1}(x', \pi(x'))$ where $x' \sim p(\cdot|x,a)$. By this definition, $\mathbb{T}^\pi Q_{k-1}$ is the *regression function* that minimizes square loss $\min_{h:\mathbb{R}^{X\times A}\mapsto\mathbb{R}} \mathbb{E}|h(x,a) - y|^2$ out of all functions h (not necessarily in F). This is due to $(\mathbb{T}^\pi Q_{k-1})(\tilde{x},\tilde{a}) = \mathbb{E}\left[y|x=\tilde{x}, a=\tilde{a}\right]$ by definition of the Bellman operator. Consider Q_{k-1} fixed and we now want to relate the learned function Q_k over finite set of n samples with the regression function over the whole data distribution via uniform deviation bound. We use the following lemma:

Lemma E.5 ((Györfi et al., 2006), theorem 11.4. Original version (Lee et al., 1996), theorem 3). *Consider random vector (X,Y) and n i.i.d samples (X_i, Y_i). Let $m(x)$ be the (optimal) regression function under square loss $m(x) = \mathbb{E}[Y|X=x]$. Assume $|Y| \leq B$ a.s. and $B \leq 1$. Let* F *be a set of function $f : \mathbb{R}^d \mapsto \mathbb{R}$ and let $|f(x)| \leq B$. Then for each $n \geq 1$*

$$\mathbf{P}\Bigg\{\exists f \in \text{F} : \mathbb{E}|f(X) - Y|^2 - \mathbb{E}|m(X) - Y|^2 - \frac{1}{n}\sum_{i=1}^n \big(|f(X_i) - Y_i|^2 - |m(X_i) - Y_i|^2\big) \geq$$

$$\epsilon \cdot \left(\alpha + \beta + \mathbb{E}|f(X) - Y|^2 - \mathbb{E}|m(X) - Y|^2\right) \Bigg\}$$
$$\leq 14 \sup_{x_1^n} \mathcal{N}_1\left(\frac{\beta\epsilon}{20B}, \mathrm{F}, x_1^n\right) \exp\left(-\frac{\epsilon^2(1-\epsilon)\alpha n}{214(1+\epsilon)B^4}\right)$$

where $\alpha, \beta > 0$ *and* $0 < \epsilon < 1/2$

To apply this lemma, first note that since $\mathbb{T}^\pi Q_{k-1}$ is the optimal regression function[6], we have

$$\mathbb{E}_\mu\left[(Q_k(x,a) - y)^2\right] = \mathbb{E}_\mu\left[\left(Q_k(x,a) - \mathbb{T}^\pi Q_{k-1}(x,a) + \mathbb{T}^\pi Q_{k-1}(x,a) - y\right)^2\right]$$
$$= \mathbb{E}_\mu\left[\left(Q_k(x,a) - \mathbb{T}^\pi Q_{k-1}(x,a)\right)^2\right] + \mathbb{E}_\mu[\left(\mathbb{T}^\pi Q_{k-1}(x,a) - y\right)^2]$$

thus

$$\|Q_k - \mathbb{T}^\pi Q_{k-1}\|_\mu^2 = \mathbb{E}\left[\left(Q_k(x,a) - \mathbb{T}^\pi Q_{k-1}(x,a)\right)^2\right] = \mathbb{E}\left[(Q_k(x,a) - y)^2\right] - \mathbb{E}\left[(\mathbb{T}^\pi Q_{k-1}(x,a) - y)^2\right]$$

where by definition

$$\mathbb{E}\left[\left(Q_k(x,a) - \mathbb{T}^\pi Q_{k-1}(x,a)\right)^2\right] = \int \left(Q_k(x,a) - \mathbb{T}^\pi Q_{k-1}(x,a)\right)^2 \mu(dx, da)$$
$$= \int (Q_k(x,a) - \mathbb{T}^\pi(x,a))^2 \mu_x(dx)\pi_{\mathrm{D}}(a|dx)$$

Next, given a fixed data set $\widetilde{D}_k \sim \mu$

$$\mathbf{P}\left\{\|Q_k - \mathbb{T}^\pi Q_{k-1}\|_\mu^2 > \epsilon\right\} = \mathbf{P}\Bigg\{\mathbb{E}\left[(Q_k(x,a) - y)^2\right] - \mathbb{E}\left[(\mathbb{T}^\pi Q_{k-1}(x,a) - y)^2\right] > \epsilon\Bigg\}$$

$$\leq \mathbf{P}\Bigg\{\mathbb{E}\left[(Q_k(x,a) - y)^2\right] - \mathbb{E}\left[(\mathbb{T}^\pi Q_{k-1}(x,a) - y)^2\right]$$
$$- 2 \cdot \left(\frac{1}{n}\sum_{i=1}^n (Q_k(x_i,a_i) - y_i)^2 - \frac{1}{n}\sum_{i=1}^n (\mathbb{T}^\pi Q_{k-1}(x_i,a_i) - y_i)^2\right) > \epsilon\Bigg\} \tag{18}$$

$$= \mathbf{P}\Bigg\{\mathbb{E}\left[(Q_k(x,a) - y)^2\right] - \mathbb{E}\left[(\mathbb{T}^\pi Q_{k-1}(x,a) - y)^2\right]$$
$$- \frac{1}{n}\sum_{i=1}^n \left[(Q_k(x_i,a_i) - y_i)^2 - (\mathbb{T}^\pi Q_{k-1}(x_i,a_i) - y_i)^2\right]$$
$$> \frac{1}{2}(\epsilon + \mathbb{E}\left[(Q_k(x,a) - y)^2\right] - \mathbb{E}\left[(\mathbb{T}^\pi Q_{k-1}(x,a) - y)^2\right])\Bigg\} \tag{19}$$

$$\leq \mathbf{P}\Bigg\{\exists f \in \mathrm{F} : \mathbb{E}\left[(f(x,a) - y)^2\right] - \mathbb{E}\left[(\mathbb{T}^\pi Q_{k-1}(x,a) - y)^2\right]$$
$$- \frac{1}{n}\sum_{i=1}^n \left[(f(x_i,a_i) - y_i)^2 - (\mathbb{T}^\pi Q_{k-1}(x_i,a_i) - y_i)^2\right]$$
$$\geq \frac{1}{2}(\frac{\epsilon}{2} + \frac{\epsilon}{2} + \mathbb{E}\left[(f(x,a) - y)^2\right] - \mathbb{E}\left[(\mathbb{T}^\pi Q_{k-1}(x,a) - y)^2\right])\Bigg\}$$

$$\leq 14 \sup_{x_1^n} \mathcal{N}_1\left(\frac{\epsilon}{80\overline{C}}, \mathrm{F}, x_1^n\right) \cdot \exp\left(-\frac{n\epsilon}{24 \cdot 214\overline{C}^4}\right) \tag{20}$$

Equation (18) uses the definition of $Q_k = \arg\min_{f \in \mathrm{F}} \frac{1}{n}\sum_{i=1}^n (f(x_i,a_i) - y_i)^2$ and the fact that $\mathbb{T}^\pi Q_{k-1} \in \mathrm{F}$, thus making the extra term a positive addition. Equation (19) is due to rearranging the terms. Equation (20) is an application of lemma E.5. We can further bound the empirical covering number by invoking the following lemma due to Haussler (Haussler, 1995):

Lemma E.6 ((Haussler, 1995), Corollary 3)**.** *For any set* X*, any points* $x^{1:n} \in \mathrm{X}^n$*, any class* F *of functions on* X *taking*

[6] It is easy to see that if $m(x) = \mathbb{E}[y|x]$ is the regression function then for any function $f(x)$, we have $\mathbb{E}\left[(f(x) - m(x))(m(x) - y) = 0\right]$

values in $[0, \overline{C}]$ *with pseudo-dimension* $\dim_{\mathrm{F}} < \infty$*, and any* $\epsilon > 0$

$$\mathcal{N}_1(\epsilon, \mathrm{F}, x_1^n) \leq e(\dim_{\mathrm{F}} + 1)\left(\frac{2e\overline{C}}{\epsilon}\right)^{\dim_{\mathrm{F}}}$$

Applying lemma E.6 to equation (20), we have the inequality

$$\mathbf{P}\left\{ \|Q_k - \mathbb{T}^\pi Q_{k-1}\|_\mu^2 > \epsilon \right\} \leq 14 \cdot e \cdot (\dim_{\mathrm{F}} + 1)\left(\frac{320\overline{C}^2}{\epsilon}\right)^{\dim_{\mathrm{F}}} \cdot \exp\left(-\frac{n\epsilon}{24 \cdot 214\overline{C}^4}\right) \tag{21}$$

We thus have that when $n \geq \frac{24 \cdot 214 \cdot \overline{C}^4}{\epsilon^2}\left(\log\frac{1}{\delta} + \dim_{\mathrm{F}} \log\frac{320\overline{C}^2}{\epsilon^2} + \log(14e(\dim_{\mathrm{F}} + 1))\right)$:

$$\|Q_k - \mathbb{T}^\pi Q_{k-1}\|_\rho < \epsilon$$

with probability at least $1 - \delta$. Notice that the dependence of sample complexity n here on ϵ is $\widetilde{O}(\frac{1}{\epsilon^2})$, which is better than previously known analyses for other approximate dynamic programming algorithms such as Fitted Value Iteration (Munos & Szepesvári, 2008), FittedPolicyQ (Antos et al., 2008b;a) with dependence of $O(\frac{1}{\epsilon^4})$.

E.2. Error bound for single iteration - Bellman non-realizable case

We now give similar error bound for the general case, where Assumption 4 does not hold. Consider the decomposition

$$\begin{aligned}
\|Q_k - \mathbb{T}^\pi Q_{k-1}\|_\mu^2 &= \mathbb{E}\left[(Q_k(x,a) - y)^2\right] - \mathbb{E}\left[(\mathbb{T}^\pi Q_{k-1}(x,a) - y)^2\right] \\
&= \Bigg\{ \mathbb{E}\left[(Q_k(x,a) - y)^2\right] - \mathbb{E}\left[(\mathbb{T}^\pi Q_{k-1}(x,a) - y)^2\right] \\
&\qquad - 2 \cdot \left(\frac{1}{n}\sum_{i=1}^n (Q_k(x_i,a_i) - y_i)^2 - \frac{1}{n}\sum_{i=1}^n (\mathbb{T}^\pi Q_{k-1}(x_i,a_i) - y_i)^2 \right) \Bigg\} \\
&\qquad + \left\{ 2 \cdot \left(\frac{1}{n}\sum_{i=1}^n (Q_k(x_i,a_i) - y_i)^2 - \frac{1}{n}\sum_{i=1}^n (\mathbb{T}^\pi Q_{k-1}(x_i,a_i) - y_i)^2 \right) \right\} \\
&= \texttt{component_1} + \texttt{component_2}
\end{aligned}$$

Splitting the probability of error into two separate bounds. We saw from the previous section (equation (21)) that

$$\mathbf{P}(\texttt{component\_1} > \epsilon/2) \leq 14 \cdot e \cdot (\dim_{\mathrm{F}} + 1)\left(\frac{640\overline{C}^2}{\epsilon}\right)^{\dim_{\mathrm{F}}} \cdot \exp\left(-\frac{n\epsilon}{48 \cdot 214\overline{C}^4}\right) \tag{22}$$

We no longer have $\texttt{component\_2} \leq 0$ since $\mathbb{T}^\pi Q_{k-1} \notin \mathrm{F}$. Let $f^* = \inf_{f \in \mathrm{F}} \|f - \mathbb{T}^\pi Q_{k-1}\|_\mu^2$. Since $Q_k = \arg\min_{f \in \mathrm{F}} \frac{1}{n}\sum_{i=1}^n (f(x_i,a_i) - y_i)^2$, we can upper-bound $\texttt{component\_2}$ by

$$\texttt{component\_2} \leq 2 \cdot \left(\frac{1}{n}\sum_{i=1}^n (f^*(x_i,a_i) - y_i)^2 - \frac{1}{n}\sum_{i=1}^n (\mathbb{T}^\pi Q_{k-1}(x_i,a_i) - y_i)^2 \right)$$

We can treat f^* as a fixed function, unlike random function Q_k, and use standard concentration inequalities to bound the empirical average from the expectation. Let random variable $z = ((x,a), y)$, $z_i = ((x_i,a_i), y_i), i = 1, \ldots, n$ and let

$$h(z) = (f^*(x,a) - y)^2 - (\mathbb{T}^\pi Q_{k-1}(x,a) - y)^2$$

We have $|h(z)| \leq 4\overline{C}^2$. We will derive a bound for

$$\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^n h(z_i) - \mathbb{E}h(z) > \frac{\epsilon}{4} + \mathbb{E}h(z) \right)$$

using Bernstein inequality(Mohri et al., 2012). First, using the relationship $h(z) = (f^*(x,a) + \mathbb{T}^\pi Q_{k-1}(x,a) - 2y)(f^*(x,a) - \mathbb{T}^\pi Q_{k-1}(x,a))$, the variance of $h(z)$ can be bounded by a constant factor of $\mathbb{E}h(z)$, since

$$\begin{aligned}
\mathbf{Var}(h(z)) &\leq \mathbb{E}h(z)^2 \leq 16\overline{C}^2 \mathbb{E}\left[(f^*(x,a) - \mathbb{T}^\pi Q_{k-1}(x,a))^2\right] \\
&= 16\overline{C}^2\left(\mathbb{E}\left[(f^*(x,a) - y)^2\right] - \mathbb{E}\left[(\mathbb{T}^\pi Q_{k-1}(x,a) - y)^2\right]\right) \qquad (23) \\
&= 16\overline{C}^2 \mathbb{E}h(z) \qquad (24)
\end{aligned}$$

Equation (23) stems from $\mathbb{T}^\pi Q_{k-1}$ being the optimal regression function. Now we can apply equation (24) and Bernstein inequality to obtain

$$\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^{n}h(z_i)-\mathbb{E}h(z)>\frac{\epsilon}{4}+\mathbb{E}h(z)\right)\leq\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^{n}h(z_i)-\mathbb{E}h(z)>\frac{\epsilon}{4}+\frac{\mathbf{Var}(h(z))}{16\overline{C}^2}\right)\leq\dots$$
$$\leq\exp\left(-\frac{n\left(\frac{\epsilon}{4}+\frac{\mathbf{Var}}{16\overline{C}^2}\right)^2}{2\mathbf{Var}+2\frac{4\overline{C}^2}{3}\left(\frac{\epsilon}{4}+\frac{\mathbf{Var}}{16\overline{C}^2}\right)}\right)$$
$$\leq\exp\left(-\frac{n\left(\frac{\epsilon}{4}+\frac{\mathbf{Var}}{16\overline{C}^2}\right)^2}{\left(32\overline{C}^2+\frac{8\overline{C}^2}{3}\right)\left(\frac{\epsilon}{4}+\frac{\mathbf{Var}}{16\overline{C}^2}\right)}\right)=\exp\left(-\frac{n\left(\frac{\epsilon}{4}+\frac{\mathbf{Var}}{16\overline{C}^2}\right)}{32\overline{C}^2+\frac{8\overline{C}^2}{3}}\right)\leq\exp\left(-\frac{1}{128+\frac{32}{3}}\cdot\frac{n\epsilon}{\overline{C}^2}\right)$$

Thus

$$\mathbf{P}\left(2\cdot\left[\frac{1}{n}\sum_{i=1}^{n}h(z_i)-2\mathbb{E}h(z)\right]>\frac{\epsilon}{2}\right)\leq\exp\left(-\frac{3}{416}\cdot\frac{n\epsilon}{\overline{C}^2}\right)\tag{25}$$

Now we have

$$\texttt{component\_2}\leq 2\cdot\frac{1}{n}\sum_{i=1}^{n}h(z_i)=2\cdot\left[\frac{1}{n}\sum_{i=1}^{n}h(z_i)-2\mathbb{E}h(z)\right]+4\mathbb{E}h(z)$$

Using again the fact that $\mathbb{T}^\pi Q_{k-1}$ is the optimal regression function

$$\mathbb{E}h(z)=\mathbb{E}_D\left[(f^*(x,a)-y)^2\right]-\mathbb{E}_D\left[(\mathbb{T}^\pi Q_{k-1}(x,a)-y)^2\right]=\mathbb{E}_D\left[(f^*(x,a)-\mathbb{T}^\pi Q_{k-1}(x,a))^2\right]$$
$$=\inf_{f\in\mathrm{F}}\|f-\mathbb{T}^\pi Q_{k-1}\|_\mu^2\tag{26}$$

Combining equations (22), (25) and (26), we can conclude that

$$\mathbf{P}\left\{\|Q_k-\mathbb{T}^\pi Q_{k-1}\|_\mu^2-4\inf_{f\in\mathrm{F}}\|f-\mathbb{T}^\pi Q_{k-1}\|_\mu^2>\epsilon\right\}\leq 14\cdot e\cdot(\texttt{dim}_\mathrm{F}+1)\left(\frac{640\overline{C}^2}{\epsilon}\right)^{\texttt{dim}_\mathrm{F}}\cdot\exp\left(-\frac{n\epsilon}{48\cdot 214\overline{C}^4}\right)$$
$$+\exp\left(-\frac{3}{416}\cdot\frac{n\epsilon}{\overline{C}^2}\right)$$

thus implying

$$\mathbf{P}\left\{\|Q_k-\mathbb{T}^\pi Q_{k-1}\|_\mu-2\inf_{f\in\mathrm{F}}\|f-\mathbb{T}^\pi Q_{k-1}\|_\mu>\epsilon\right\}\leq 14\cdot e\cdot(\texttt{dim}_\mathrm{F}+1)\left(\frac{640\overline{C}^2}{\epsilon^2}\right)^{\texttt{dim}_\mathrm{F}}\cdot\exp\left(-\frac{n\epsilon^2}{48\cdot 214\overline{C}^4}\right)$$
$$+\exp\left(-\frac{3}{416}\cdot\frac{n\epsilon^2}{\overline{C}^2}\right)\tag{27}$$

We now can further upper-bound the term $2\inf_{f\in\mathrm{F}}\|f-\mathbb{T}^\pi Q_{k-1}\|_\mu\leq 2\sup_{f'\in\mathrm{F}}\inf_{f\in\mathrm{F}}\|f-\mathbb{T}^\pi f'\|_\mu=2d_\mathrm{F}^\pi$ (the worst-case *inherent Bellman evaluation error*), leading to the final bound for the Bellman non-realizable case.

One may wish to further remove the inherent Bellman evaluation error from our error bound. However, counter-examples exist where the inherent Bellman error cannot generally be estimated using function approximation (see section 11.6 of (Sutton & Barto, 2018)). Fortunately, inherent Bellman error can be driven to be small by choosing rich function class F (low bias), at the expense of more samples requirement (higher variance, through higher pseudo-dimension $\texttt{dim}_\mathrm{F}$).

While the bound in (27) looks more complicated than the Bellman realizable case in equation 21, note that the convergence rate will still be $O(\frac{1}{n^2})$.

E.3. Bounding the error across iterations

Previous sub-sections E.2 and E.2 have analyzed the error of FQE for a single iteration in Bellman realizable and non-realizable case. We now analyze how errors from different iterations flow through the FQE algorithm. The proof borrows the idea from lemma 3 and 4 of (Munos & Szepesvári, 2008) for fitted value iteration (for value function V instead of Q), with appropriate modifications for our off-policy evaluation context.

Recall that C^π, Q^π denote the true value function and action-value function, respectively, under the evaluation policy π.

And $C_K = \mathbb{E}[Q_K(x, \pi(x))]$ denote the value function associated with the returned function Q_K from algorithm 3. Our goal is to bound the difference $C^\pi - C_K$ between the true value function and the estimated value of the returned function Q_K.

Let the unknown state-action distribution induced by the evaluation policy π be ρ. We first bound the loss $\|Q^\pi - Q_K\|_\rho$ under the "test-time "distribution ρ of (x, a), which differs from the state-action μ induced by data-generating policy π_{D}. We will then lift the loss bound from Q_K to C_K.

Step 1: Upper-bound the value estimation error

Let $\epsilon_{k-1} = Q_k - \mathbb{T}^\pi Q_{k-1} \in \mathrm{X} \times \mathrm{A}, \bar{\mathcal{C}}$. We have for every k that

$$\begin{aligned} Q^\pi - Q_k &= \mathbb{T}^\pi Q^\pi - \mathbb{T}^\pi Q_{k-1} + \epsilon_{k-1} \quad (Q^\pi \text{ is fixed point of } T^\pi) \\ &= \gamma P^\pi (Q^\pi - Q_{k-1}) + \epsilon_{k-1} \end{aligned}$$

Thus by simple recursion

$$\begin{aligned} Q^\pi - Q_K &= \sum_{k=0}^{K-1} \gamma^{K-k-1} (P^\pi)^{K-k-1} \epsilon_k + \gamma^K (P^\pi)^K (Q^\pi - Q_0) \\ &= \frac{1-\gamma^{K+1}}{1-\gamma} \left[\sum_{k=0}^{K-1} \frac{(1-\gamma)\gamma^{K-k-1}}{1-\gamma^{K+1}} (P^\pi)^{K-k-1} \epsilon_k + \frac{(1-\gamma)\gamma^K}{1-\gamma^{K+1}} (P^\pi)^K (Q^\pi - Q_0) \right] \\ &= \frac{1-\gamma^{K+1}}{1-\gamma} \left[\sum_{k=0}^{K-1} \alpha_k A_k \epsilon_k + \alpha_K A_K (Q^\pi - Q_0) \right] \end{aligned} \tag{28}$$

where for simplicity of notations, we denote

$$\begin{aligned} \alpha_k &= \frac{(1-\gamma)\gamma^{K-k-1}}{1-\gamma^{K+1}} \text{ for } k < K, \alpha_K = \frac{(1-\gamma)\gamma^K}{1-\gamma^{K+1}} \\ A_k &= (P^\pi)^{K-k-1}, A_K = (P^\pi)^K \end{aligned}$$

Note that A_k's are probability kernels and α_k's are deliberately chosen such that $\sum_k \alpha_k = 1$.

We can apply point-wise absolute value on both sides of (28) with $|f|$ being the short-hand notation for $|f(x, a)|$ and inequality holds point-wise. By triangle inequalities:

$$|Q^\pi - Q_K| \leq \frac{1-\gamma^{K+1}}{1-\gamma} \left[\sum_{k=0}^{K-1} \alpha_k A_k |\epsilon_k| + \alpha_K A_K |Q^\pi - Q_0| \right] \tag{29}$$

Step 2: Bounding $\|Q^\pi - Q_K\|_\rho$ for any unknown distribution ρ. To handle distribution shift from μ to ρ, we decompose the loss as follows:

$$\begin{aligned} \|Q^\pi - Q_K\|_\rho^2 &= \int \rho(dx, da) \left(Q^\pi(x, a) - Q_K(x, a)\right)^2 \\ &\leq \left[\frac{1-\gamma^{K+1}}{1-\gamma}\right]^2 \int \rho(dx, da) \left[\left(\sum_{k=0}^{K-1} \alpha_k A_k |\epsilon_k| + \alpha_K A_K |Q^\pi - Q_0| \right)(x, a) \right]^2 \text{ (from(29))} \\ &\leq \left[\frac{1-\gamma^{K+1}}{1-\gamma}\right]^2 \int \rho(dx, da) \left[\sum_{k=0}^{K-1} \alpha_k (A_k \epsilon_k)^2 + \alpha_K (A_K (Q^* - Q_0))^2 \right](x, a) \text{ (Jensen)} \\ &\leq \left[\frac{1-\gamma^{K+1}}{1-\gamma}\right]^2 \int \rho(dx, da) \left[\sum_{k=0}^{K-1} \alpha_k A_k \epsilon_k^2 + \alpha_K A_K (Q^* - Q_0)^2 \right](x, a) \text{ (Jensen)} \end{aligned}$$

Using assumption 3 (assumption 1 of the main paper), we can bound each term ρA_k as

$$\rho A_k = \rho (P^\pi)^{K-k-1} \leq \mu \beta_\mu (K-k-1) \text{ (definition D.3)}$$

Thus

$$\|Q^\pi - Q_K\|_\rho^2 \leq \left[\frac{1-\gamma^{K+1}}{1-\gamma}\right]^2 \left[\frac{1}{1-\gamma^{K+1}} \sum_{k=0}^{K-1} (1-\gamma)\gamma^{K-k-1} \beta_\mu (K-k-1) \|\epsilon_k\|_\mu^2 + \alpha_K (2\overline{C})^2 \right]$$

Assumption 3 (stronger than necessary for proof of FQE) can be used to upper-bound the first order concentration coefficient:

$$(1-\gamma)\sum_{m\geq 0}\gamma^m\beta_\mu(m) \leq \frac{\gamma}{1-\gamma}\left[(1-\gamma)^2\sum_{m\geq 1}m\gamma^{m-1}\beta_\mu(m)\right] = \frac{\gamma}{1-\gamma}\beta_\mu$$

This gives the upper-bound for $\|Q^\pi - Q_K\|_\rho^2$ as

$$\begin{aligned}\|Q^\pi - Q_K\|_\rho^2 &\leq \left[\frac{1-\gamma^{K+1}}{1-\gamma}\right]^2\left[\frac{\gamma}{(1-\gamma)(1-\gamma^{K+1})}\beta_\mu\max_k\|\epsilon_k\|_\mu^2 + \frac{(1-\gamma)\gamma^K}{1-\gamma^{K+1}}(2\overline{C})^2\right]\\ &\leq \frac{1-\gamma^{K+1}}{(1-\gamma)^2}\left[\frac{\gamma}{1-\gamma}\beta_\mu\max_k\|\epsilon_k\|_\mu^2 + (1-\gamma)\gamma^K(2\overline{C})^2\right]\\ &\leq \frac{\gamma}{(1-\gamma)^3}\beta_\mu\max_k\|\epsilon_k\|_\mu^2 + \frac{\gamma^K}{1-\gamma}(2\overline{C})^2\end{aligned}$$

Using $a^2+b^2 \leq (a+b)^2$ for nonnegative a, b, we conclude that

$$\|Q^\pi - Q_K\|_\rho \leq \frac{\gamma^{1/2}}{(1-\gamma)^{3/2}}\left(\sqrt{\beta_\mu}\max_k\|\epsilon_k\|_\mu + \frac{\gamma^{K/2}}{(1-\gamma)^{1/2}}2\overline{C}\right) \tag{30}$$

Step 3: Turning error bound from Q to $|C^\pi - C_K|$ Now we can choose ρ to be the state-action distribution by the evaluation policy π. The error bound on the value function C follows simply by integrating inequality (30) over state-action pairs induced by π. The final error across iterations can be related to individual iteration error by

$$|C^\pi - C_K| \leq \frac{\gamma^{1/2}}{(1-\gamma)^{3/2}}\left(\sqrt{\beta_\mu}\max_k\|\epsilon_k\|_\mu + \frac{\gamma^{K/2}}{(1-\gamma)^{1/2}}2\overline{C}\right) \tag{31}$$

E.4. Finite-sample guarantees for Fitted Q Evaluation

Combining results from (21), (27) and (31), we have the final guarantees for FQE under both realizable and general cases.

Realizable Case - Proof of theorem E.2. From (21), when $n \geq \frac{24\cdot 214\cdot\overline{C}^4}{\epsilon^2}\left(\log\frac{K}{\delta} + \mathtt{dim}_\mathrm{F}\log\frac{320\overline{C}^2}{\epsilon^2} + \log(14e(\mathtt{dim}_\mathrm{F}+1))\right)$, we have $\|\epsilon_k\|_\mu < \epsilon$ with probability at least $1-\delta/K$ for any $0 \leq k < K$. Thus we conclude that for any $\epsilon > 0, 0 < \delta < 1$, after K iterations of Fitted Q Evaluation, the value estimate returned by Q_K satisfies:

$$|C^\pi - C_K| \leq \frac{\gamma^{1/2}}{(1-\gamma)^{3/2}}\left(\sqrt{\beta_\mu}\epsilon + \frac{\gamma^{K/2}}{(1-\gamma)^{1/2}}2\overline{C}\right)$$

holds with probability $1-\delta$ when $n \geq \frac{24\cdot 214\cdot\overline{C}^4}{\epsilon^2}\left(\log\frac{K}{\delta} + \mathtt{dim}_\mathrm{F}\log\frac{320\overline{C}^2}{\epsilon^2} + \log(14e(\mathtt{dim}_\mathrm{F}+1))\right)$. This concludes the proof of theorem E.2.

Non-realizable Case - Proof of theorem E.1 and theorem 4.2 of main paper. Similarly, from (27) we have

$$\begin{aligned}\mathbf{P}\left\{\|Q_k - \mathbb{T}^\pi Q_{k-1}\|_\mu - 2\inf_{f\in\mathrm{F}}\|f - \mathbb{T}^\pi Q_{k-1}\|_\mu > \epsilon\right\} \leq 14\cdot e\cdot(\mathtt{dim}_\mathrm{F}+1)\left(\frac{640\overline{C}^2}{\epsilon^2}\right)^{\mathtt{dim}_\mathrm{F}}\cdot\exp\left(-\frac{n\epsilon^2}{48\cdot 214\overline{C}^4}\right)\\ + \exp\left(-\frac{3}{416}\cdot\frac{n\epsilon^2}{\overline{C}^2}\right)\end{aligned}$$

Since $\inf_{f\in\mathrm{F}}\|f - \mathbb{T}^\pi Q_{k-1}\|_\mu \leq \sup_{h\in\mathrm{F}}\inf_{f\in\mathrm{F}}\|f - \mathbb{T}^\pi h\|_\mu = d_\mathrm{F}^\pi$ (the *inherent Bellman evaluation error*), similar arguments to the realizable case lead to the conclusion that for any $\epsilon > 0, 0 < \delta < 1$, after K iterations of FQE:

$$|C^\pi - C_K| \leq \frac{\gamma^{1/2}}{(1-\gamma)^{3/2}}\left(\sqrt{\beta_\mu}(2d_\mathrm{F}^\pi + \epsilon) + \frac{\gamma^{K/2}}{(1-\gamma)^{1/2}}2\overline{C}\right)$$

holds with probability $1-\delta$ when $n = O\big(\frac{\overline{C}^4}{\epsilon^2}(\log\frac{K}{\delta} + \mathtt{dim}_\mathrm{F}\log\frac{\overline{C}^2}{\epsilon^2} + \log\mathtt{dim}_\mathrm{F})\big)$, thus finishes the proof of theorem E.1.

Note that in both cases, the $\widetilde{O}(\frac{1}{\epsilon^2})$ dependency of n is significant improvement over previous finite-sample analysis of related approximate dynamic programming algorithms (Munos & Szepesvári, 2008; Antos et al., 2008b;a). This dependency matches that of previous analysis using linear function approximators from (Lazaric et al., 2012; 2010) for LSTD and LSPI algorithms. Here our analysis, using similar assumptions, is applicable for general non-linear, bounded function classes. , which is an improvement over convergence rate of $O(\frac{1}{n^4})$ in related approximate dynamic programming algorithms (Antos et al., 2008a;b; Munos & Szepesvári, 2008).

F. Finite-Sample Analysis of Fitted Q Iteration (FQI)

F.1. Algorithm and Discussion

Algorithm 4 Fitted Q Iteration with Function Approximation: $\texttt{FQI}(c)$ (Ernst et al., 2005)

Input: Collected data set $\mathrm{D} = \{x_i, a_i, x'_i, c_i\}_{i=1}^n$. Function class F
1: Initialize $Q_0 \in \mathrm{F}$ randomly
2: **for** $k = 1, 2, \ldots, K$ **do**
3: Compute target $y_i = c_i + \gamma \min_a Q_{k-1}(x'_i, a) \;\; \forall i$
4: Build training set $\widetilde{\mathrm{D}}_k = \{(x_i, a_i), y_i\}_{i=1}^n$
5: Solve a supervised learning problem:
$Q_k = \arg\min_{f \in \mathrm{F}} \frac{1}{n} \sum_{i=1}^n (f(x_i, a_i) - y_i)^2$
6: **end for**
Output: $\pi_K(\cdot) = \arg\min_a Q_K(\cdot, a)$ (greedy policy with respect to the returned function Q_K)

The analysis of FQI (algorithm 4) follows analogously from the analysis of FQE from the previous section (Appendix E). For brevity, we skip certain detailed derivations, especially those that are largely identical to FQE's analysis.

To the best of our knowledge, a finite-sample analysis of FQI with general non-linear function approximation has not been published (Continuous FQI from (Antos et al., 2008a) is in fact a Fitted Policy Iteration algorithm and is different from algo 4). In principle, one can adapt existing analysis of fitted value iteration (Munos & Szepesvári, 2008) and FittedPolicyQ (Antos et al., 2008b;a) to show that under similar assumptions, among policies greedy w.r.t. functions in F, FQI will find $\epsilon-$ optimal policy using $n = \widetilde{O}(\frac{1}{\epsilon^4})$ samples. We derive an improved analysis of FQI with general non-linear function approximations, with better sample complexity of $n = \widetilde{O}(\frac{1}{\epsilon^2})$. We note that the appendix of (Lazaric & Restelli, 2011) contains an analysis of LinearFQI showing similar rate to ours, albeit with linear function approximators.

In this section, we prove the following statement:

Theorem F.1 (Guarantee for FQI - General Case (theorem 4.3 in main paper))**.** *Under Assumption 3, for any $\epsilon > 0, \delta \in (0, 1)$, after K iterations of Fitted Q Iteration (algorithm 4), for $n = O\big(\frac{\overline{C}^4}{\epsilon^2}(\log \frac{K}{\delta} + \texttt{dim}_\mathrm{F} \log \frac{\overline{C}^2}{\epsilon^2} + \log \texttt{dim}_\mathrm{F})\big)$, we have with probability $1 - \delta$:*

$$C^* - C(\pi_K) \leq \frac{2\gamma}{(1-\gamma)^3}\left(\sqrt{\beta_\mu}\,(2d_\mathrm{F} + \epsilon) + 2\gamma^{K/2}\overline{C}\right)$$

where π_K is the policy greedy with respect to the returned function Q_K, and C^* is the value of the optimal policy.

The key steps to the proof follow similar scheme to the proof of FQE. We first bound the error for each iteration, and then analyze how the errors flow through the algorithm.

F.2. Single iteration error bound $\|Q_k - \mathbb{T}Q_{k-1}\|_\mu$

Here μ is the state-action distribution induced by the data-generating policy π_D.

We begin with the decomposition:

$$\|Q_k - \mathbb{T}Q_{k-1}\|_\mu^2 = \mathbb{E}\left[(Q_k(x,a) - y)^2\right] - \mathbb{E}\left[(\mathbb{T}Q_{k-1}(x,a) - y)^2\right]$$

$$= \left\{\mathbb{E}\left[(Q_k(x,a) - y)^2\right] - \mathbb{E}\left[(\mathbb{T}Q_{k-1}(x,a) - y)^2\right] - 2 \cdot \left(\frac{1}{n}\sum_{i=1}^n (Q_k(x_i,a_i) - y_i)^2 - \frac{1}{n}\sum_{i=1}^n (\mathbb{T}Q_{k-1}(x_i,a_i) - y_i)^2\right)\right\}$$

$$+ \left\{2 \cdot \left(\frac{1}{n}\sum_{i=1}^n (Q_k(x_i,a_i) - y_i)^2 - \frac{1}{n}\sum_{i=1}^n (\mathbb{T}Q_{k-1}(x_i,a_i) - y_i)^2\right)\right\}$$

$$= \texttt{component\_1} + \texttt{component\_2}$$

For $\mathbb{T}$ the Bellman (optimality) operator (equation 16), $\mathbb{T}Q_{k-1}$ is the *regression function* that minimizes square loss $\min_{h:\mathbb{R}^{X\times A}\mapsto\mathbb{R}} \mathbb{E}|h(x,a) - y|^2$, with the random variables $(x,a) \sim \mu$ and $y = c(x,a) + \gamma \min_{a'} Q_{k-1}(x', a')$ where $x' \sim p(x'|x,a)$. Invoking lemma E.5 and following the steps similar to equations (18),(19),(20) and (21) from appendix E, we

can bound the first component as

$$\mathbf{P}(\texttt{component\_1} > \epsilon/2) \leq 14 \cdot e \cdot (\texttt{dim}_{\mathrm{F}} + 1) \left(\frac{640\overline{C}^2}{\epsilon}\right)^{\texttt{dim}_{\mathrm{F}}} \cdot \exp\left(-\frac{n\epsilon}{48 \cdot 214\overline{C}^4}\right) \tag{32}$$

Let $f^* = \arg\inf_{f \in \mathrm{F}} \|f - \mathbb{T}Q_{k-1}\|_\mu^2$. Since $Q_k = \arg\min_{f \in \mathrm{F}} \frac{1}{n}\sum_{i=1}^n (f(x_i, a_i) - y_i)^2$, we can upper-bound `component_2` by

$$\texttt{component\_2} \leq 2 \cdot \left(\frac{1}{n}\sum_{i=1}^n (f^*(x_i, a_i) - y_i)^2 - \frac{1}{n}\sum_{i=1}^n (\mathbb{T}Q_{k-1}(x_i, a_i) - y_i)^2\right)$$

Let random variable $z = ((x, a), y)$, $z_i = ((x_i, a_i), y_i), i = 1, \ldots, n$ and let

$$h(z) = (f^*(x, a) - y)^2 - (\mathbb{T}Q_{k-1}(x, a) - y)^2$$

We have $|h(z)| \leq 4\overline{C}^2$. We can derive a bound for $\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^n h(z_i) - \mathbb{E}h(z) > \frac{\epsilon}{4} + \mathbb{E}h(z)\right)$ using Bernstein inequality, similar to equations (23) and (24) from appendix E to obtain:

$$\mathbf{P}\left(2 \cdot \left[\frac{1}{n}\sum_{i=1}^n h(z_i) - 2\mathbb{E}h(z)\right] > \frac{\epsilon}{2}\right) \leq \exp\left(-\frac{3}{416} \cdot \frac{n\epsilon}{\overline{C}^2}\right) \tag{33}$$

Now we have

$$\texttt{component\_2} \leq 2 \cdot \frac{1}{n}\sum_{i=1}^n h(z_i) = 2 \cdot \left[\frac{1}{n}\sum_{i=1}^n h(z_i) - 2\mathbb{E}h(z)\right] + 4\mathbb{E}h(z)$$

Since

$$\begin{aligned}\mathbb{E}h(z) &= \mathbb{E}_{\widetilde{D}_k}\left[(f^*(x, a) - y)^2\right] - \mathbb{E}_{\widetilde{D}_k}\left[(\mathbb{T}Q_{k-1}(x, a) - y)^2\right] = \mathbb{E}_{\widetilde{D}_k}\left[(f^*(x, a) - \mathbb{T}Q_{k-1}(x, a))^2\right] \\ &= \inf_{f \in \mathrm{F}} \|f - \mathbb{T}Q_{k-1}\|_\mu^2\end{aligned} \tag{34}$$

Combining equations (32), (33) and (34), we obtain that

$$\begin{aligned}\mathbf{P}\left\{\|Q_k - \mathbb{T}Q_{k-1}\|_\mu^2 - 4\inf_{f \in \mathrm{F}} \|f - \mathbb{T}Q_{k-1}\|_\mu^2 > \epsilon\right\} \leq 14 \cdot e \cdot (\texttt{dim}_{\mathrm{F}} + 1) \left(\frac{640\overline{C}^2}{\epsilon}\right)^{\texttt{dim}_{\mathrm{F}}} \cdot \exp\left(-\frac{n\epsilon}{48 \cdot 214\overline{C}^4}\right) \\ + \exp\left(-\frac{3}{416} \cdot \frac{n\epsilon}{\overline{C}^2}\right)\end{aligned} \tag{35}$$

F.3. Propagation of error bound for $\|Q^* - Q^{\pi_K}\|_\rho$

The analysis of error propagation for FQI is more involved than that of FQE, but the proof largely follows the error propagation analysis in lemma 3 and 4 of (Munos & Szepesvári, 2008) in the fitted value iteration context (for V function). We include the Q function's (slighly more complicated) derivation here for completeness.

Recall that π_K is greedy wrt the learned function Q_K returned by FQI. We aim to bound the difference $C^* - C^{\pi_K}$ between the optimal value function and that π_K. For a (to-be-specified) distribution ρ of state-action pairs (different from the data distribution μ), we bound the generalization loss $\|Q^* - Q^{\pi_K}\|_\rho$

Step 1: Upper-bound the propagation error (value). Let $\epsilon_{k-1} = Q_k - \mathbb{T}Q_{k-1}$. We have that

$$\begin{aligned}Q^* - Q_k = \mathbb{T}^{\pi^*}Q^* - \mathbb{T}^{\pi^*}Q_{k-1} + \mathbb{T}^{\pi^*}Q_{k-1} - \mathbb{T}Q_{k-1} + \epsilon_{k-1} &\leq \mathbb{T}^{\pi^*}Q^* - \mathbb{T}^{\pi^*}Q_{k-1} + \epsilon_{k-1} \text{ (b/c } \mathbb{T}Q_{k-1} \geq \mathbb{T}^{\pi^*}Q_{k-1}) \\ &= \gamma P^{\pi^*}(Q^* - Q_{k-1}) + \epsilon_{k-1}\end{aligned}$$

Thus by recursion $Q^* - Q_K \leq \sum_{k=0}^{K-1} \gamma^{K-k-1}(P^{\pi^*})^{K-k-1}\epsilon_k + \gamma^K(P^{\pi^*})^K(Q^* - Q_0)$

Step 2: Lower-bound the propagation error (value). Similarly

$$\begin{aligned}Q^* - Q_k = \mathbb{T}Q^* - \mathbb{T}^{\pi_{k-1}}Q^* + \mathbb{T}^{\pi_{k-1}}Q^* - \mathbb{T}Q_{k-1} + \epsilon_{k-1} &\geq \mathbb{T}^{\pi_{k-1}}Q^* - \mathbb{T}Q_{k-1} + \epsilon_{k-1} \text{ (as } \mathbb{T}Q^* \geq \mathbb{T}^{\pi_{k-1}}Q^*) \\ &\geq \mathbb{T}^{\pi_{k-1}}Q^* - \mathbb{T}^{\pi_{k-1}}Q_{k-1} + \epsilon_{k-1} \text{ (b/c } \pi_{k-1} \text{ greedy wrt } Q_{k-1}) \\ &= \gamma P^{\pi_{k-1}}(Q^* - Q_{k-1}) + \epsilon_{k-1}\end{aligned}$$

And by recursion $Q^* - Q_K \geq \sum_{k=0}^{K-1} \gamma^{K-k-1}(P^{\pi_{K-1}}P^{\pi_{K-2}}\ldots P^{\pi_{k+1}})\epsilon_k + \gamma^K(P^{\pi_{K-1}}P^{\pi_{K-2}}\ldots P^{\pi_0})(Q^* - Q_0)$

Step 3: Upper-bound the propagation error (policy). Beginning with a decomposition of value wrt to policy π_K

$$\begin{aligned}
Q^* - Q^{\pi_K} &= \mathbb{T}^{\pi^*}Q^* - \mathbb{T}^{\pi^*}Q_K + \mathbb{T}^{\pi^*}Q_K - \mathbb{T}^{\pi_K}Q_K + \mathbb{T}^{\pi_K}Q_K - \mathbb{T}^{\pi_K}Q^{\pi_K} \\
&\leq (\mathbb{T}^{\pi^*}Q^* - \mathbb{T}^{\pi^*}Q_K) + (\mathbb{T}^{\pi_K}Q_K - \mathbb{T}^{\pi_K}Q^{\pi_K}) \qquad (\text{ since } \mathbb{T}^{\pi^*}Q_K \leq \mathbb{T}Q_K = \mathbb{T}^{\pi_K}Q_K) \\
&= \gamma P^{\pi^*}(Q^* - Q_K) + \gamma P^{\pi_K}(Q_K - Q^{\pi_K}) \\
&= \gamma P^{\pi^*}(Q^* - Q_K) + \gamma P^{\pi_K}(Q_K - Q^* + Q^* - Q^{\pi_K})
\end{aligned}$$

Thus leading to $(I - \gamma P^{\pi_K})(Q^* - Q^{\pi_K}) \leq \gamma(P^{\pi^*} - P^{\pi_K})(Q^* - Q_K)$ The operator $(I - \gamma P^{\pi_K})$ is invertible and $(I - \gamma P^{\pi_K})^{-1} = \sum_{m\geq 0}\gamma^m (P^{\pi_K})^m$ is monotonic. Thus

$$\begin{aligned}
Q^* - Q^{\pi_K} &\leq \gamma(I - \gamma P^{\pi_K})^{-1}(P^{\pi^*} - P^{\pi_K})(Q^* - Q_K) \\
&= \gamma(I - \gamma P^{\pi_K})^{-1}P^{\pi^*}(Q^* - Q_K) - \gamma(I - \gamma P^{\pi_K})^{-1}P^{\pi_K}(Q^* - Q_K) \qquad (36)
\end{aligned}$$

Applying inequalities from Step 1 and Step 2 to the RHS of (36), we have

$$\begin{aligned}
Q^* - Q^{\pi_K} \leq (I - \gamma P^{\pi_K})^{-1}\Bigg[&\sum_{k=0}^{K-1}\gamma^{K-k}\left((P^{\pi^*})^{K-k} - P^{\pi_K}P^{\pi_{K-1}}\dots P^{\pi_{k+1}}\right)\epsilon_k \\
&+ \gamma^{K+1}\left((P^{\pi^*})^{K+1} - (P^{\pi_K}P^{\pi_{K-1}}\dots P^{\pi_0})\right)(Q^* - Q_0)\Bigg] \qquad (37)
\end{aligned}$$

Next we apply point-wise absolute value on RHS of (37), with $|\epsilon_k|$ being the short-hand notation for $|\epsilon_k(x,a)|$ point-wise. Using triangle inequalities and rewriting (37) in a more compact form ((Munos & Szepesvári, 2008)):

$$Q^* - Q^{\pi_K} \leq \frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2}\left[\sum_{k=0}^{K-1}\alpha_k A_k|\epsilon_k| + \alpha_K A_K|Q^* - Q_0|\right]$$

where $\alpha_k = \frac{(1-\gamma)\gamma^{K-k-1}}{1-\gamma^{K+1}}$ for $k < K, \alpha_K = \frac{(1-\gamma)\gamma^K}{1-\gamma^{K+1}}$ and

$$\begin{aligned}
A_k &= \frac{1-\gamma}{2}(I - \gamma P^{\pi_K})^{-1}\left[(P^{\pi^*})^{K-k} + P^{\pi_K}P^{\pi_{K-1}}\dots P^{\pi_{k+1}}\right] \text{ for } k < K \\
A_K &= \frac{1-\gamma}{2}(I - \gamma P^{\pi_K})^{-1}\left[(P^{\pi^*})^{K+1} + P^{\pi_K}P^{\pi_{K-1}}\dots P^{\pi_0}\right]
\end{aligned}$$

Note that A_k's are probability kernels that combine the P^{π_i} terms and α_k's are chosen such that $\sum_k \alpha_k = 1$.

Step 4: Bounding $\|Q^* - Q^{\pi_K}\|_\rho^2$ for any test distribution ρ.

This step handles distribution shift from μ to ρ (similar to Step 2 from sub-section E.3 of appendix E)

$$\|Q^* - Q^{\pi_K}\|_\rho^2 \leq \left[\frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2}\right]^2 \int \rho(dx,da)\left[\sum_{k=0}^{K-1}\alpha_k A_k \epsilon_k^2 + \alpha_K A_K (Q^* - Q_0)^2\right](x,a) \text{ (twice Jensen)}$$

Using assumption 3 (assumption 1 in the main paper), each term ρA_k is bounded as

$$\begin{aligned}
\rho A_k &= \frac{1-\gamma}{2}\rho(I - \gamma P^{\pi_K})^{-1}\left[(P^{\pi^*})^{K-k} + P^{\pi_K}P^{\pi_{K-1}}\dots P^{\pi_{k+1}}\right] \\
&= \frac{1-\gamma}{2}\sum_{m\geq 0}\gamma^m \rho(P^{\pi_K})^m\left[(P^{\pi^*})^{K-k} + P^{\pi_K}P^{\pi_{K-1}}\dots P^{\pi_{k+1}}\right] \leq (1-\gamma)\sum_{m\geq 0}\gamma^m\beta_\mu(m+K-k)\mu \qquad (\text{def D.3})
\end{aligned}$$

Thus

$$\begin{aligned}
\|Q^* - Q^{\pi_K}\|_\rho^2 &\leq \left[\frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2}\right]^2\left[\frac{1}{1-\gamma^{K+1}}\sum_{k=0}^{K-1}(1-\gamma)^2\sum_{m\geq 0}\gamma^{m+K-k-1}\beta_\mu(m+K-k)\|\epsilon_k\|_\mu^2 + \alpha_K(2\overline{C})^2\right] \\
&\leq \left[\frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2}\right]^2\left[\frac{1}{1-\gamma^{K+1}}\beta_\mu\max_k\|\epsilon_k\|_\mu^2 + \frac{(1-\gamma)\gamma^K}{1-\gamma^{K+1}}(2\overline{C})^2\right] \qquad (\text{assumption 3}) \\
&\leq \left[\frac{2\gamma(1-\gamma^{K+1})}{(1-\gamma)^2}\right]^2\left[\frac{1}{1-\gamma^{K+1}}\beta_\mu\max_k\|\epsilon_k\|_\mu^2 + \frac{\gamma^K}{1-\gamma^{K+1}}(2\overline{C})^2\right] \\
&\leq \left[\frac{2\gamma}{(1-\gamma)^2}\right]^2\left[\beta_\mu\max_k\|\epsilon_k\|_\mu^2 + \gamma^K(2\overline{C})^2\right]
\end{aligned}$$

Using $a^2 + b^2 \leq (a+b)^2$ for nonnegative a, b, we thus conclude that

$$\|Q^* - Q^{\pi_K}\|_\rho \leq \frac{2\gamma}{(1-\gamma)^2}\left(\sqrt{\beta_\mu}\max_k \|\epsilon_k\|_\mu + 2\gamma^{K/2}\overline{C}\right) \tag{38}$$

Step 5: Bounding $C^* - C^{\pi_K}$ Using the performance difference lemma (lemma 6.1 of (Kakade & Langford, 2002), which states that $C^* - C^{\pi_K} = -\frac{1}{1-\gamma}\mathbb{E}_{\substack{x\sim d_{\pi_K}\\ a\sim\pi_K}} A^*[x,a]$. We can upper-bound the performance difference of value function as

$$\begin{aligned}
C^* - C^{\pi_K} &= \frac{1}{1-\gamma}\mathbb{E}_{\substack{x\sim d_{\pi_K}\\ a\sim\pi_K}}[C^*(x) - Q^*(x,a)] = \frac{1}{1-\gamma}\mathbb{E}_{x\sim d_{\pi_K}}[C^*(x) - Q^*(x,\pi_K(x))] \\
&\leq \frac{1}{1-\gamma}\mathbb{E}_{x\sim d_{\pi_K}}[Q^*(x,\pi^*(x)) - Q_K(x,\pi^*(x)) + Q_K(x,\pi_K(x) - Q^*(x,\pi_K(x))] \text{ (greedy)} \\
&\leq \frac{1}{1-\gamma}\mathbb{E}_{x\sim d_{\pi_K}}|Q^*(x,\pi^*(x)) - Q_K(x,\pi^*(x))| + |Q_K(x,\pi_K(x) - Q^*(x,\pi_K(x))| \\
&\leq \frac{1}{1-\gamma}\left(\|Q^* - Q^{\pi_K}\|_{d_{\pi_K}\times\pi^*} + \|Q^* - Q^{\pi_K}\|_{d_{\pi_K}\times\pi_K}\right) \text{ (upper-bound 1-norm by 2-norm)} \\
&\leq \frac{2\gamma}{(1-\gamma)^3}\left(\sqrt{\beta_\mu}\max_k \|\epsilon_k\|_\mu + 2\gamma^{K/2}\overline{C}\right)
\end{aligned} \tag{39}$$

Note that inequality (39) follows from (38) by specifying $\rho = \chi P^{\pi_K} P^{\pi^*}$ and $\rho = \chi P^{\pi_K} P^{\pi_K}$, respectively ($\chi$ is the initial state distribution).

F.4. Finite-sample guarantees for Fitted Q Iteration

From (35) we have:

$$\begin{aligned}
\mathbf{P}\left\{\|Q_k - \mathbb{T}Q_{k-1}\|_\mu - 2\inf_{f\in\mathrm{F}}\|f - \mathbb{T}Q_{k-1}\|_\mu > \epsilon\right\} \leq 14\cdot e\cdot(\mathtt{dim}_\mathrm{F}+1)\left(\frac{640\overline{C}^2}{\epsilon^2}\right)^{\mathtt{dim}_\mathrm{F}}\cdot\exp\left(-\frac{n\epsilon^2}{48\cdot 214\overline{C}^4}\right) \\
+\exp\left(-\frac{3}{416}\cdot\frac{n\epsilon^2}{\overline{C}^2}\right)
\end{aligned}$$

Note that $\inf_{f\in\mathrm{F}}\|f - \mathbb{T}Q_{k-1}\|_\mu \leq \sup_{h\in\mathrm{F}}\inf_{f\in\mathrm{F}}\|f - \mathbb{T}h\|_\mu = d_\mathrm{F}$ (the *inherent Bellman error* from equation 16). Combining with equation (39), we have the conclusion that for any $\epsilon > 0, 0 < \delta < 1$, after K iterations of Fitted Q Iteration, and for π_K the greedy policy wrt Q_K:

$$C^* - C_K^\pi \leq \frac{2\gamma}{(1-\gamma)^3}\left(\sqrt{\beta_\mu}(2d_\mathrm{F}+\epsilon) + 2\gamma^{K/2}\overline{C}\right)$$

holds with probability $1-\delta$ when $n = O\big(\frac{\overline{C}^4}{\epsilon^2}(\log\frac{K}{\delta} + \mathtt{dim}_\mathrm{F}\log\frac{\overline{C}^2}{\epsilon^2} + \log\mathtt{dim}_\mathrm{F})\big)$.

Note that compared to the Fitted Value Iteration analysis of (Munos & Szepesvári, 2008), our error includes an extra factor 2 for d_F.

F.5. Statement for the Bellman-realizable Case

To facilitate the end-to-end generalization analysis of theorem 4.4 in the main paper, we include a version of FQI analysis under Bellman-realizable assumption in this section. The theorem is a consequence of previous analysis in this section.

Assumption 5 (Bellman evaluation realizability)**.** *We consider function classes* F *sufficiently rich so that* $\forall f, \mathbb{T}f \in \mathrm{F}$.

Theorem F.2 (Guarantee for FQI - Bellman-realizable Case)**.** *Under Assumption 3 and 5, for any* $\epsilon > 0, \delta \in (0,1)$*, after* K *iterations of Fitted Q Iteration, for* $n \geq \frac{24\cdot 214\cdot\overline{C}^4}{\epsilon^2}\big(\log\frac{K}{\delta} + \mathtt{dim}_\mathrm{F}\log\frac{320\overline{C}^2}{\epsilon^2} + \log(14e(\mathtt{dim}_\mathrm{F}+1))\big)$*, we have with probability* $1-\delta$*:*

$$C^* - C(\pi_K) \leq \frac{2\gamma}{(1-\gamma)^3}\big(\sqrt{\beta_\mu}\epsilon + 2\gamma^{K/2}\overline{C}\big)$$

where π_K *is the policy greedy with respect to the returned function* Q_K*, and* C^* *is the value of the optimal policy.*

G. Additional Instantiation of Meta-Algorithm (algorithm 1)

We provide an additional instantiation of the meta-algorithm described in the main paper, with Online Gradient Descent (OGD) (Zinkevich, 2003) and Least-Squares Policy Iteration (LSPI) (Lagoudakis & Parr, 2003) as subroutines. Using LSPI requires a feature map ϕ such that any state-action pair can be represented by k features. The value function is linear in parameters represented by ϕ. Policy representation is simplified to a weight vector $w \in \mathbb{R}^k$.

Similar to our main algorithm 2, OGD updates require bounded parameters λ. We thus introduce hyper-parameter B as the bound of λ in ℓ_2 norm. The gradient update is projected to the ℓ_2 ball when the norm of λ exceeds B (line 15 of algo 5).

Algorithm 5 Batch Learning under Constraints using Online Gradient Descent and Least-Squares Policy Iteration

Input: Dataset $\mathrm{D} = \{x_i, a_i, x'_i, c_i, g_i\}_{i=1}^n \sim \pi_{\mathrm{D}}$. Online algorithm parameters: ℓ_2 norm bound B, learning rate η
Input: Number of basis function k. Basis function ϕ (feature map for state-action pairs)
1: Initialize $\lambda_1 = (0, \dots, 0) \in \mathbb{R}^m$
2: **for** each round t **do**
3: Learn $w_t \leftarrow \texttt{LSPI}(c + \lambda_t^\top g)$ *// LSPI with cost $c + \lambda_t^\top g$*
4: Evaluate $\widehat{C}(w_t) \leftarrow \texttt{LSTDQ}(w_t, c)$ *// Algo 7 with π_t, cost c*
5: Evaluate $\widehat{G}(w_t) \leftarrow \texttt{LSTDQ}(w_t, g)$ *// Algo 7 with π_t,* cost g
6: $\widehat{w}_t \leftarrow \frac{1}{t}\sum_{t'=1}^t w_{t'}$
7: $\widehat{C}(\widehat{w}_t) \leftarrow \frac{1}{t}\sum_{t'=1}^t \widehat{C}(w_{t'}), \widehat{G}(\widehat{w}_t) \leftarrow \frac{1}{t}\sum_{t'=1}^t \widehat{G}(w_{t'})$
8: $\widehat{\lambda}_t \leftarrow \frac{1}{t}\sum_{t'=1}^t \lambda_{t'}$
9: Learn $\widetilde{w} \leftarrow \texttt{LSPI}(c + \widehat{\lambda}_t^\top g)$ *// LSPI with cost $c + \widehat{\lambda}_t^\top g$*
10: Evaluate $\widehat{C}(\widetilde{w}) \leftarrow \texttt{LSTDQ}(\widetilde{w}, c), \widehat{G}(\widetilde{w}) \leftarrow \texttt{LSTDQ}(\widetilde{w}, g)$
11: $\widehat{\mathrm{L}}_{\max} = \max_{\lambda, \|\lambda\|_2 \le B} \left(\widehat{C}(\widehat{w}_t) + \lambda^\top (\widehat{G}(\widehat{w}_t) - \tau)\right)$
12: $\widehat{\mathrm{L}}_{\min} = \widehat{C}(\widetilde{w}) + \widehat{\lambda}_t^\top (\widehat{G}(\widetilde{w}) - \tau)$
13: **if** $\widehat{\mathrm{L}}_{\max} - \widehat{\mathrm{L}}_{\min} \le \omega$ **then**
14: Return $\widehat{\pi}_t$ greedy w.r.t $\widehat{w}_t$ $\left(\text{i.e., } \widehat{\pi}_t(x) = \arg\min_{a \in \mathrm{A}} \widehat{w}_t^\top \phi(x, a) \ \ \forall x\right)$
15: **end if**
16: $\lambda_{t+1} = \mathcal{P}(\lambda_t - \eta(\widehat{G}(\pi_t) - \tau))$ where projection $\mathcal{P}(\lambda) = B\frac{\lambda}{\max\{B, \|\lambda\|_2\}}$
17: **end for**

Algorithm 6 Least-Squares Policy Iteration: $\texttt{LSPI}(c)$ (Lagoudakis & Parr, 2003)

Input: Stopping criterion ϵ
1: Initialize $w' \leftarrow w_0$
2: **repeat**
3: $w \leftarrow w'$
4: $w' \leftarrow \texttt{LSTDQ}(w, c)$
5: **until** $\|w - w'\| \le \epsilon$
Output: Policy weight w $\left(\text{i.e., } \pi(x) = \arg\min_{a \in \mathrm{A}} w^\top \phi(x, a) \ \ \forall x\right)$

Algorithm 7 $\texttt{LSTDQ}(w, c)$ (Lagoudakis & Parr, 2003)

1: Initialize $\widetilde{\mathbf{A}} \leftarrow \mathbf{0}$ *// $k \times k$ matrix*
2: Initialize $\widetilde{b} \leftarrow \mathbf{0}$ *// $k \times 1$ vector*
3: **for** each $(x, a, x', c) \in \mathrm{D}$ **do**
4: $a' = \arg\min_{\widetilde{a} \in \mathrm{A}} w^\top \phi(x', \widetilde{a})$
5: $\widetilde{\mathbf{A}} \leftarrow \widetilde{\mathbf{A}} + \phi(x, a)\big(\phi(x, a) - \gamma\phi(x', a')\big)^\top$
6: $\widetilde{b} \leftarrow \widetilde{b} + \phi(x, a)c$
7: **end for**
8: $\widetilde{w} \leftarrow \widetilde{\mathbf{A}}^{-1}\widetilde{b}$
Output: $\widetilde{w}$

H. Additional Experimental Details

H.1. Environment Descriptions and Procedures

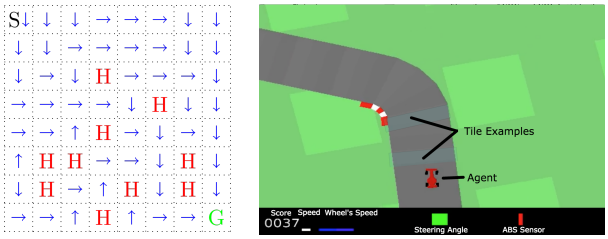

Figure 3. Depicting the *FrozenLake* and *CarRacing* environments.

Frozen Lake. The environment is a 8x8 grid as seen in Figure 3 (left), based on OpenAi's FrozenLake-v0. In each episode, the agent starts from S and traverse to goal G. While traversing the grid, the agent must avoid the pre-determined holes denoted by H. If the agent steps off of the grid, the agent returns to the same grid location. The episode terminates when the agent reaches the goal or falls into a hole. The arrows in Figure 3 (left) is an example policy returned by our algorithm, showing an optimal route.

Denote X_{holes} as the set of all holes in the grid and $\mathrm{X}_{goal} = \{x_{goal}\}$ is a singleton set representing the goal in the grid. The contrained batch policy learning problem is:

$$\begin{aligned} \min_{\pi \in \Pi} \quad & C(\pi) = \mathbb{E}[\mathbb{I}(x' \notin \mathrm{X}_{goals})] = \mathrm{P}(x' \notin \{x_{goal}\}) \\ \text{s.t.} \quad & G(\pi) = \mathbb{E}[\mathbb{I}(x' \in \mathrm{X}_{holes})] = \mathrm{P}(x' \in \mathrm{X}_{holes}) \leq \tau \end{aligned} \tag{40}$$

We collect 5000 trajectories by selecting an action randomly with probability .95 and an action from a DDQN-trained model with probability .05.Furthermore we set $B = 30$ and $\eta = 50$, the hyperparameters of our Exponentiated Gradient subroutine. We set the threshold for the constraint $\tau = .1$.

Car Racing. The environment is a racetrack as seen in Figure 3 (right), modified from OpenAi's CarRacing-v0. In each state, given by the raw pixels, the agent has 12 actions: $a \in A = \{(i, j, k) | i \in \{-1, 0, 1\}, j \in \{0, 1\}, k \in \{0, .2\}\}$. The action tuple (i, j, k) cooresponds to steering angle, amount of gas applied and amount of brake applied, respectively. In each episode, the agent starts at the same point on the track and must traverse over 95% of the track, given by a discretization of 281 tiles. The agent recieves a reward of $+\frac{1000}{281}$ for each unique tile over which the agent drives. The agent receives a penalty of $-.1$ per-time step. Our collected dataset takes the form: $\mathrm{D} = \{(x_{t-6}, x_{t-3}, x_t), a_t, (x_{t-3}, x_t, x_{t+3}), c_t, g_{0,t}, g_{1,t}\}$ where x_i denotes the image at timestep i and a_t is applied 3 times between x_t and x_{t+3}. This frame-stacking option is common practice in online RL for Atari and video games.In our collected dataset D, the maximum horizon is 469 time steps.

The first constraint concerns accumulated number of brakes, a proxy for smooth driving or acceleration. The second constraint concerns how far the agent travels away from the center of the track, given by the Euclidean distance between the agent and the closest point on the center of the track. Let N_t be the number of tiles that is collected by the agent in time t. The constrained batch policy learning problem is:

$$\begin{aligned} \min_{\pi \in \Pi} \quad & \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t (-\frac{1000}{281} N_t + .1)] \\ \text{s.t.} \quad & G_0(\pi) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t \mathbb{I}(a_t \in \mathrm{A}_{braking})] \leq \tau_0 \\ & G_1(\pi) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t d(u_t, v_t)] \leq \tau_1 \end{aligned} \tag{41}$$

We instatiate our subroutines, FQE and FQI, with multi-layered CNNs. Furthermore we set $B = 10$ and $\eta = .01$, the hyperparameters of our Exponentiated Gradient subroutine. We set the threshold for the constraint to be about 75% of the value exhibited by online RL agent trained by DDQN (Van Hasselt et al., 2016).

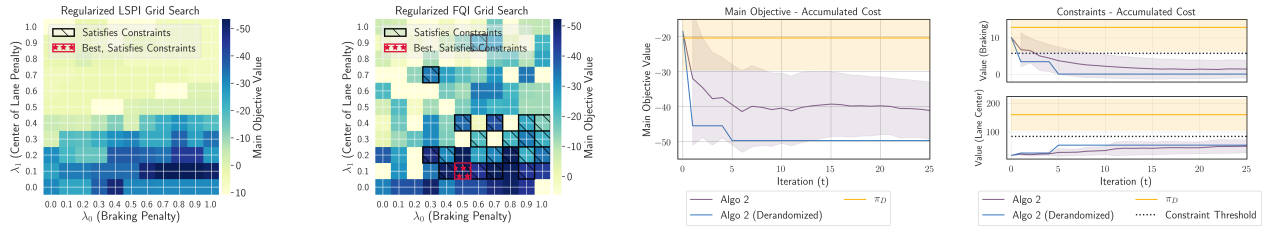

Figure 4. (First and Second figures) Result of 2-D grid-search for one-shot, regularized policy learning for *LSPI* (left) and *FQI* (right). (Third and Fourth figures) value range of individual policies in our mixtured policy and data generating policy π_D for main objective (left) and cost constraint (right)

H.2. Additional Discussion for the Car Racing Experiment

Regularized policy learning and grid-search. We perform grid search over a range of regularization parameters λ for both Least-Squares Policy Iteration - LSPI ((Lagoudakis & Parr, 2003)) and Fitted Q Iteration - FQI ((Ernst et al., 2005)). The results, seen from the the first and second plot of Figure 4, show that one-shot regularized learning has difficulty learning a policy that satisfies both constraints. We augment LSPI with non-linear feature mapping from one of our best performing FQI model (using CNNs representation). While both regularized LSPI and regularized FQI can achieve low main objective cost, the constraint cost values tend to be sensitive with the λ step. Overall for the whole grid search, about 10% of regularized policies satisfy both constraints, while none of the regularized LSPI policy satisfies both constraints.

Mixture policy and de-randomization. As our algorithm returned a mixture policy, it is natural to analyze the performance of individual policies in the mixture. The third and fourth plot from Figure 4 show the range of performance of individual policy in our mixture (purple band). We compare individual policy return with the stochastic behavior of the data generation policy. Note that our policies satisfy constraints almost always, while the individual policy returned in the mixture also tends to outperform π_D with respect to the main objective cost.

Off-policy evaluation standalone comparison. Typically, inverse propensity scoring based methods call for stochastic behavior and evaluation policies (Precup et al., 2000; Swaminathan & Joachims, 2015). However in this domain, the evaluation policy and environment are both deterministic, with long horizon (the max horizon is D is 469). Consequently Per-Decision Importance Sampling typically evaluates the policy as 0. In general, off-policy policy evaluation in long-horizon domains is known to be challenging (Liu et al., 2018; Guo et al., 2017). We augment PDIS by approximating the evaluation policy with a stochastic policy, using a softmin temperature parameter. However, PDIS still largely shows significant errors. For Doubly Robust and Weighted Doubly Robust methods, we train a model of the environment as follows:

- a 32-dimensional representation of state input is learned using variational autoencoder. Dimensionality reduction is necessary to aid accuracy, as original state dimension is $96 \times 96 \times 3$
- an LSTM is used to learn the transition dynamics $P(z(x')|z(x), a)$, where $z(x)$ is the low-dimensional representation learned from previous step. Technically, using a recurrent neural networks is an augmentation to the dynamical modeling, as true MDPs typically do not require long-term memory
- the model is trained separately on a different dataset, collected indendently from the dataset D used for evaluation

The architecture of our dynamics model is inspired by recent work in model-based online policy learning (Ha & Schmidhuber, 2018). However, despite our best effort, learning the dynamics model accurately proves highly challenging, as the horizon and dimensionality of this domain are much larger than popular benchmarks in the OPE literature (Jiang & Li, 2016; Thomas & Brunskill, 2016; Farajtabar et al., 2018). The dynamics model has difficulty predicting the future state several time steps away. Thus we find that the long-horizon, model-based estimation component of DR and WDR in this high-dimensional setting is not sufficiently accurate. For future work, a thorough benchmarking of off-policy evaluation methods in high-dimensional domains would be a valuable contribution.