Supplements

First we provide more detailed information about a closely-related economic model, dynamic discrete choice models, is deferred to Section A. Then, theoretical details including the proof and extra assumptions for our main results are provided in Section B. We provide extended experiment details in Section C.

A. Dynamic Discrete Choice Models

In this Section, we detail the formulation of DDCs (Rust, 1987) to support our argument in Section 3. The results here are mainly a review of existing work. DDCs assume that agents make decisions following a Markov decision process described by the tuple $\{\{\mathcal{S}, \mathcal{E}\}, \mathcal{A}, \mathrm{P}, \gamma, r\}$, where

- $\{\mathcal{S}, \mathcal{E}\}$ denotes the space of state variables.
- $\mathcal{A}$ is a set of J actions as the action space.
- r is the reward function.
- $\gamma \in [0, 1)$ is the discount factor.
- P denotes the transition distribution.

At time point t, agents observe state variable $\mathbf{S}_t \in \mathcal{S}$, and $\mathbf{e}_t = \{e_{t1}, e_{t2}, \cdots, e_{tJ}\} \in \mathcal{E}$. While $\mathbf{S}_t$ is observable and recorded in the dataset, $\mathbf{e}_t$ is not observable in the dataset, and is only known by the agent when making decisions at time point t.

The decision variable is defined as a $J \times 1$ vector, $\mathbf{A}_t = \{A_{t1}, A_{t2}, \cdots, A_{tJ}\}^\top \in \mathcal{A}$, satisfying

- $\sum_{j=1}^{J} A_{tj} = 1$,
- $A_{tj} \in \{0, 1\}$.

Thus, the decision is indeed a selection over J choices.

The control problem agents are solving is formulated by the following value function:

$$V^{DDC}(\mathbf{s}, \epsilon) = \max_{\{\mathbf{a}_t\}_{t=0}^{\infty}} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(\mathbf{S}_t, \mathbf{e}_t, \mathbf{a}_t) \mid \mathbf{s}, \mathbf{e}\right]. \tag{6}$$

The reward function of DDCs is defined by decomposing over different actions. For $\mathbf{s} \in \mathcal{S}$, $\mathbf{e} \in \mathcal{E}$, and $\mathbf{a} \in \mathcal{A}$, the reward function $r(\mathbf{s}, \epsilon, \mathbf{a})$ is defined by

$$r(\mathbf{s}, \mathbf{e}, \mathbf{a}) = r^*(\mathbf{s}, \mathbf{a}) + \mathbf{a}^\top \mathbf{e},$$

where r^* is the estimation target.

A.1. Assumptions

We first discuss the assumptions for DDCs:

Assumption 4. *The transition distribution from $\mathbf{S}_t$ to $\mathbf{S}_{t+1}$ is independent from $\mathbf{e}_t$*

$$\mathrm{P}(\mathbf{S}_{t+1} = \mathbf{s}' \mid \mathbf{S}_t = \mathbf{s}, \mathbf{e}_t = \mathbf{e}, \mathbf{A}_t = \mathbf{a}) = \mathrm{P}(\mathbf{S}_{t+1} = \mathbf{s}' \mid \mathbf{S}_t = \mathbf{s}, \mathbf{A}_t = \mathbf{a}).$$

Assumption 5. *The $\mathbf{e}_t$ are independent and identically distributed (IID) according to a Type-I extreme value distribution (aka the Gumbel distribution).*

Note that ϵ_t could also follow other parametric distributions. As suggested by (Arcidiacono and Ellickson, 2011), Assumption 5 is virtually standard for all dynamic discrete choice models. We use Type-I extreme value distribution as an example, other distributions follow a similar analysis.

A.2. Likelihood for DDCs

Next, we derive the likelihood of DDCs.

Definition 4 (Conditional Value Function)**.**

$$Q^{DDC}(\mathbf{s},\mathbf{a}) = r^*(\mathbf{A}_t,\mathbf{S}_t) + \gamma \int_{\mathcal{S}} \bar{V}(\mathbf{s}')f(\mathbf{S}_{t+1}=\mathbf{s}' \mid \mathbf{S}_t=\mathbf{s},\mathbf{A}_t=\mathbf{a})d\mathbf{s}', \tag{7}$$

where

$$\bar{V}(\mathbf{s}) = \mathbb{E}_\epsilon[V^{DDC}(\mathbf{s},\epsilon)]. \tag{8}$$

Lemma 3. *Let the assumptions in Section A.1 be satisfied. The likelihood of observation* $\mathbb{X} = \{\mathbf{s}_1,\mathbf{a}_1,\mathbf{s}_2,\mathbf{a}_2,\cdots,\mathbf{s}_T,\mathbf{a}_T\}$ *becomes:*

$$L^{DDC}(\mathbb{X}) = \prod_{t=1}^{T-1} \mathrm{P}(\mathbf{S}_{t+1}=\mathbf{s}_{t+1} \mid \mathbf{S}_t=\mathbf{s}_t,\mathbf{A}_t=\mathbf{a}_t) \prod_{t=1}^{T} \left[\exp\left\{Q^{DDC}(\mathbf{s}_t,\mathbf{a}_t)\right\}/Z^{DDC}(\mathbf{s}_t)\right], \tag{9}$$

where

$$Z^{DDC}(\mathbf{s}) = \sum_{\mathbf{a}\in\mathcal{A}} \exp\left\{v(\mathbf{s},\mathbf{a})\right\}.$$

Comparing (9) with the likelihood in Lemma 1, it is clear that DDC is a special case of the considered setting when the action space is discrete and $\alpha = 1$.

B. Extended Theoretical Results

In this section, we provide more information about our theoretical results including Lemma 2, Theorem 1, and Theorem 2.

B.1. Proof of Lemma 2

In this section, we prove Lemma 2. By Theorem 1, we know

$$r(\mathbf{s},\mathbf{a}) = Q(\mathbf{s},\mathbf{a}) - \gamma\mathbb{E}\left[-\alpha\log(\pi^*(\mathbf{s}',\mathbf{a}^A)) + Q(\mathbf{s}',\mathbf{a}^A) \mid \mathbf{s},\mathbf{a}\right]. \tag{10}$$

In other words, we can recover the true reward function if the Q input to R-ESTIMATOR is not shaped. Next, we take an input shaped by $\phi(\mathbf{s})$:

$$Q'(\mathbf{s},\mathbf{a}) := Q(\mathbf{s},\mathbf{a}) + \phi(\mathbf{s}).$$

The according reward estimator $r'(\mathbf{s},\mathbf{a})$ can be derived as

$$\begin{aligned} r'(\mathbf{s},\mathbf{a}) &= Q'(\mathbf{s},\mathbf{a}) - \gamma\mathbb{E}\left[-\alpha\log(\pi^*(\mathbf{s}',\mathbf{a}^A)) + Q'(\mathbf{s}',\mathbf{a}^A) \mid \mathbf{s},\mathbf{a}\right] \\ &= Q(\mathbf{s},\mathbf{a}) - \gamma\mathbb{E}\left[-\alpha\log(\pi^*(\mathbf{s}',\mathbf{a}^A)) + Q(\mathbf{s}',\mathbf{a}^A) \mid \mathbf{s},\mathbf{a}\right] + \phi(\mathbf{s}) - \gamma\mathbb{E}\left[\phi(\mathbf{s}') \mid \mathbf{s},\mathbf{a}\right]. \end{aligned}$$

Finally, by (10) and the definition of $\Phi(\mathbf{s},\mathbf{a})$, we have

$$r'(\mathbf{s},\mathbf{a}) = r(\mathbf{s},\mathbf{a}) + \Phi(\mathbf{s},\mathbf{a}).$$

B.2. Theoretical Results for Theorem 1

The details for the derivation of Theorem 1 is provided in this Section.

B.2.1. Auxiliary Lemmas for Theorem 1

To start with, we provide required lemmas.

Lemma 4. *Following the definitions and Lemma 1, we can derive:*

$$V(\mathbf{s}) = \mathbb{E}(Q(\mathbf{s}, \mathbf{A})) + \alpha\mathcal{H}(\pi^*(\mathbf{s}, \cdot)),$$

where the expectation is over the action variable $\mathbf{A}$ *following the optimal policy* (2)*.*

Proof. To start with, we insert the optimal policy to the definition of V:

$$\begin{aligned} V(\mathbf{s}) &= \max_{\pi} \sum_{t=0}^{\infty} \gamma^t \, \mathbb{E}[r(\mathbf{S}_t, \mathbf{A}_t) + \alpha\mathcal{H}(\pi(\mathbf{S}_t, \cdot)) \mid \mathbf{S}_0 = \mathbf{s}] \\ &= \sum_{t=0}^{\infty} \gamma^t \, \mathbb{E}[r(\mathbf{S}_t, \mathbf{A}_t) + \alpha\mathcal{H}(\pi^*(\mathbf{S}_t, \cdot)) \mid \mathbf{S}_0 = \mathbf{s}]. \end{aligned}$$

Next, we take the case $t = 0$ out of the summation, and derive

$$\begin{aligned} V(\mathbf{s}) &= \mathbb{E}[r(\mathbf{s}, \mathbf{A}_0) + \alpha\mathcal{H}(\pi^*(\mathbf{s}, \cdot)) \mid \mathbf{S}_0 = \mathbf{s}] + \sum_{t=1}^{\infty} \gamma^t \, \mathbb{E}[r(\mathbf{S}_t, \mathbf{A}_t) + \alpha\mathcal{H}(\pi^*(\mathbf{S}_t, \cdot)) \mid \mathbf{S}_0 = \mathbf{s}] \\ &= \alpha\mathcal{H}(\pi^*(\mathbf{s}, \cdot)) + \mathbb{E}\left[r(\mathbf{s}, \mathbf{A}) \mid \mathbf{S}_0 = \mathbf{s}\right] + \mathbb{E}\left\{ \sum_{t=1}^{\infty} \left[\gamma^t r(\mathbf{S}_t, \mathbf{A}_t) + \gamma^t \alpha\mathcal{H}(\pi^*(\mathbf{S}_t, \cdot))\right] \mid \mathbf{S}_0 = \mathbf{s} \right\}. \end{aligned}$$

Finally, by the definition of Q in Lemma 1, we can finish the proof:

$$V(\mathbf{s}) = \mathbb{E}[Q(\mathbf{s}, \mathbf{A})] + \alpha\mathcal{H}(\pi^*(\mathbf{s}, \cdot)).$$

□

Lemma 5. *The value function* (1) *satisfies the following equation:*

$$V(\mathbf{s}) = \alpha \log \int_{\mathbf{a} \in \mathcal{A}} \exp\left(\frac{Q(\mathbf{s}, \mathbf{a})}{\alpha}\right) d\mathbf{a}. \tag{11}$$

Proof. By Lemma 4, we can derive the relationship between Q and V as

$$V(\mathbf{s}) = \mathbb{E}[Q(\mathbf{s}, \mathbf{A})] + \alpha\mathcal{H}(\pi^*(\mathbf{s}, \cdot)),$$

where the expectation is over the action variable $\mathbf{A}$ following the optimal policy (2). Next, by the definitions of expectation and information entropy, we can derive

$$\begin{aligned} V(\mathbf{s}) &= \int_{\mathbf{a} \in \mathcal{A}} Q(\mathbf{s}, \mathbf{a}) \pi^*(\mathbf{s}, \mathbf{a}) d\mathbf{a} - \alpha \int_{\mathbf{a} \in \mathcal{A}} \log(\pi^*(\mathbf{s}, \mathbf{a})) \pi^*(\mathbf{s}, \mathbf{a}) d\mathbf{a} \\ &= \int_{\mathbf{a} \in \mathcal{A}} Q(\mathbf{s}, \mathbf{a}) \pi^*(\mathbf{s}, \mathbf{a}) d\mathbf{a} - \alpha \int_{\mathbf{a} \in \mathcal{A}} \frac{Q(\mathbf{s}, \mathbf{a})}{\alpha} \pi^*(\mathbf{s}, \mathbf{a}) d\mathbf{a} \\ &\quad + \alpha \int_{\mathbf{a} \in \mathcal{A}} \log\left[\int_{\mathbf{a}' \in \mathcal{A}} \exp\left(\frac{Q(\mathbf{s}, \mathbf{a}')}{\alpha}\right) d\mathbf{a}'\right] \pi^*(\mathbf{s}, \mathbf{a}) d\mathbf{a} \\ &= \alpha \log\left[\int_{\mathbf{a}' \in \mathcal{A}} \exp\left(\frac{Q(\mathbf{s}, \mathbf{a}')}{\alpha}\right) d\mathbf{a}'\right] \\ &= \alpha \log\left[\int_{\mathbf{a} \in \mathcal{A}} \exp\left(\frac{Q(\mathbf{s}, \mathbf{a})}{\alpha}\right) d\mathbf{a}\right]. \end{aligned}$$

□

As α approaches 0, Lemma 5 is consistent with the case with deterministic policies corresponding to $\alpha = 0$. We summarize the analysis in Remark 1

Remark 1. *As α approaches 0^+, (11) is consistent with the Bellman equation with deterministic polices:*

$$\lim_{\alpha \to 0^+} V(\mathbf{s}) = \max_a Q'(\mathbf{s}, \mathbf{a}),$$

where

$$Q'(\mathbf{s}, \mathbf{a}) = \lim_{\alpha \to 0^+} Q(\mathbf{s}, \mathbf{a}).$$

Proof. To start with, we review the definition of L^p-norm of functions:

Definition 5. *Given a measurable space $(\mathcal{A}, \mu)$ and a real number $p \in [1, \infty)$, the L^p-norm of a function $f : \mathcal{A} \mapsto \mathbb{R}$ is defined as*

$$\|f\|_p = \left(\int_{\mathcal{A}} |f|^p d\mu \right)^{1/p}$$

Therefore, (11) can be represented by the L^p-norm of the function of $\mathbf{a}$ $\exp(Q(\mathbf{s}, \cdot))$ with $p = 1/\alpha$:

$$\begin{aligned} V(\mathbf{s}) &= \alpha \log \int_{\mathbf{a} \in \mathcal{A}} \exp\left(\frac{Q(\mathbf{s}, \mathbf{a})}{\alpha} \right) d\mathbf{a} \\ &= \log \left| \int_{\mathbf{a} \in \mathcal{A}} \exp\left(\frac{Q(\mathbf{s}, \mathbf{a})}{\alpha} \right) d\mathbf{a} \right|^{\alpha} \\ &= \log \left| \int_{\mathbf{a} \in \mathcal{A}} \left[\exp\left(Q(\mathbf{s}, \mathbf{a}) \right) \right]^{1/\alpha} d\mathbf{a} \right|^{\alpha} \\ &= \log \|\exp(Q(\mathbf{s}, \cdot))\|_{1/\alpha}. \end{aligned}$$

Then, we take limit to the both sides of $V(\mathbf{s})$ and derive

$$\lim_{\alpha \to 0^+} V(\mathbf{s}) = \lim_{\alpha \to 0^+} \log \|\exp(Q(\mathbf{s}, \cdot))\|_{1/\alpha} = \log \|\exp(Q'(\mathbf{s}, \cdot))\|_{\infty} = \|Q'(\mathbf{s}, \cdot)\|_{\infty},$$

with $Q'(\mathbf{s}, \mathbf{a}) = \lim_{\alpha \to 0^+} Q(\mathbf{s}, \mathbf{a})$. Note that the second equation is true since both $\log$ and $\exp$ are monotonic functions.

□

Lemma 6. *Let $\pi^*(\mathbf{s}, \mathbf{a})$ be the optimal policy of agents, $Q(\mathbf{s}, \mathbf{a})$ the ground-truth Q-function, and $r(\mathbf{s}, \mathbf{a})$ the true reward. Under the formulation in Section 2, we have*

$$Q(\mathbf{s}, \mathbf{a}) = r(\mathbf{s}, \mathbf{a}) + \gamma \mathbb{E} \left[-\alpha \log(\pi^*(\mathbf{s}', \mathbf{a}^A)) + Q(\mathbf{s}', \mathbf{a}^A) \mid \mathbf{s}, \mathbf{a} \right].$$

Proof. According to Lemma 1, we have

$$\pi^*(\mathbf{s}, \mathbf{a}) = \frac{\exp\left(\frac{1}{\alpha} Q(\mathbf{s}, \mathbf{a}) \right)}{\int_{\mathbf{a}' \in \mathcal{A}} \exp\left(\frac{1}{\alpha} Q(\mathbf{s}, \mathbf{a}') \right) d\mathbf{a}'}.$$

By Lemma 5, we have

$$V(\mathbf{s}) = \alpha \log \int_{\mathbf{a} \in \mathcal{A}} \exp\left(\frac{Q(\mathbf{s}, \mathbf{a})}{\alpha} \right) d\mathbf{a}. \tag{12}$$

For the next step, we consider a specific action $\mathbf{a}^A$, and extract $\alpha \log \left[\exp\left(\frac{Q(\mathbf{s}, \mathbf{a}^A)}{\alpha} \right) \right]$ from (12):

$$\begin{aligned} V(\mathbf{s}) &= \alpha \log \left[\frac{\int_{\mathbf{a} \in \mathcal{A}} \exp\left(\frac{Q(\mathbf{s}, \mathbf{a})}{\alpha} \right) d\mathbf{a}}{\exp\left(\frac{Q(\mathbf{s}, \mathbf{a}^A)}{\alpha} \right)} \right] + \alpha \log \left[\exp\left(\frac{Q(\mathbf{s}, \mathbf{a}^A)}{\alpha} \right) \right] \\ &= \alpha \log \left(\frac{1}{\pi^*(\mathbf{s}, \mathbf{a}^A)} \right) + Q(\mathbf{s}, \mathbf{a}^A) \\ &= -\alpha \log \left(\pi^*(\mathbf{s}, \mathbf{a}^A) \right) + Q(\mathbf{s}, \mathbf{a}^A). \end{aligned} \tag{13}$$

According to Theorem 1 in Haarnoja et al. (2017), we have

$$Q(\mathbf{s},\mathbf{a}) = r(\mathbf{s},\mathbf{a}) + \gamma\mathbb{E}\left[V(\mathbf{s}') \mid \mathbf{s},\mathbf{a}\right] \tag{14}$$

Finally, by taking (13) into (14), we prove the result.

□

Lemma 7. *Let $\pi^*(\mathbf{s},\mathbf{a})$ be the optimal policy of agents, $Q(\mathbf{s},\mathbf{a})$ the ground-truth Q-function, and $r(\mathbf{s},\mathbf{a})$ the true reward. Define $\mathcal{T}$ as an operator on the set of continuous bounded functions $f : \mathcal{S} \longmapsto \mathbb{R}$:*

$$\mathcal{T}f(\mathbf{s}) := g(\mathbf{s}) + \gamma\mathbb{E}\big[-\alpha\log(\pi^*(\mathbf{s}',\mathbf{a}^A)) + f(\mathbf{s}') \mid \mathbf{s},\mathbf{a}^A\big]. \tag{15}$$

Under the formulation in Section 2 and Assumption 1, $Q^A(\mathbf{s}) := Q(\mathbf{s},\mathbf{a}^A)$ is the unique solution to

$$f(\mathbf{s}) = \mathcal{T}f(\mathbf{s}). \tag{16}$$

Proof. It is obvious that $\mathcal{T}$ satisfies the monotonicity and discounting condition. Therefore, we can conclude that $\mathcal{T}$ is a contraction, and (16) has the unique solution.

On the other hand, by taking $\mathbf{a} = \mathbf{a}^A$ to Lemma 6, we have

$$Q^A(\mathbf{s}) = \mathcal{T}Q^A(\mathbf{s}). \tag{17}$$

Thus, $Q^A(\mathbf{s})$ is the unique solution. □

B.2.2. Proof of Theorem 1

Proof. Since the transition probabilities are given, and $\hat{\pi}(\mathbf{s},\mathbf{a})$ is accurate, we have $\hat{\mathbb{E}} = \mathbb{E}$, $\hat{\pi} = \pi^*$, and thus $\hat{\mathcal{T}} = \mathcal{T}$. According to Lemma 7, Q^A-Estimator (denoted as $\hat{Q}^A(\mathbf{s})$) uniquely recovers the true $Q^A(\mathbf{s})$:

$$\hat{Q}^A(\mathbf{s}) = Q^A(\mathbf{s}).$$

By Lemma 1, we have

$$\pi^*(\mathbf{s},\mathbf{a}) = \frac{\exp\left(\frac{Q(\mathbf{s},\mathbf{a})}{\alpha}\right)}{\int_{\mathbf{a}'\in\mathcal{A}}\exp\left(\frac{Q(\mathbf{s},\mathbf{a}')}{\alpha}\right)d\mathbf{a}'}.$$

Then, we have

$$\begin{aligned}
&\log(\hat{\pi}(\mathbf{s},\mathbf{a})) - \log(\hat{\pi}(\mathbf{s},\mathbf{a}^A))\\
&= \log\left[\exp\left(\frac{Q(\mathbf{s},\mathbf{a})}{\alpha}\right)\right] - \log\left[\exp\left(\frac{Q^A(\mathbf{s})}{\alpha}\right)\right]\\
&\quad + \log\left(\int_{\mathcal{A}}\exp(\frac{Q(\mathbf{s},\mathbf{a})}{\alpha})d\mathbf{a}\right) - \log\left(\int_{\mathcal{A}}\exp(\frac{Q(\mathbf{s},\mathbf{a})}{\alpha})d\mathbf{a}\right)\\
&= \frac{1}{\alpha}(Q(\mathbf{s},\mathbf{a}) - Q^A(\mathbf{s})),
\end{aligned}$$

suggesting

$$\alpha\log(\hat{\pi}(\mathbf{s},\mathbf{a})) - \alpha\log(\hat{\pi}(\mathbf{s},\mathbf{a}^A)) + \hat{Q}^A(\mathbf{s}) = Q(\mathbf{s},\mathbf{a}).$$

Therefore, Q-Estimator recovers the true Q-function:

$$\hat{Q}(\mathbf{s},\mathbf{a}) = Q(\mathbf{s},\mathbf{a}). \tag{18}$$

Finally, by taking (18) into R-Estimator, we derive

$$\begin{aligned}
\hat{r}(\mathbf{s},\mathbf{a}) &= \hat{Q}(\mathbf{s},\mathbf{a}) - \gamma\hat{\mathbb{E}}\left[-\alpha\log(\hat{\pi}(\mathbf{s}',\mathbf{a}^A)) + \hat{Q}(\mathbf{s}',\mathbf{a}^A) \mid \mathbf{s},\mathbf{a}\right]\\
&= Q(\mathbf{s},\mathbf{a}) - \gamma\mathbb{E}\left[-\alpha\log(\pi^*(\mathbf{s}',\mathbf{a}^A)) + Q(\mathbf{s}',\mathbf{a}^A) \mid \mathbf{s},\mathbf{a}\right],
\end{aligned}$$

where $\hat{\pi} = \pi^*$ according to the setting of Theorem 1. Finally, by Lemma 6, we have

$$\hat{r}(\mathbf{s}, \mathbf{a}) = Q(\mathbf{s}, \mathbf{a}) - \gamma \mathbb{E}\left[-\alpha \log(\pi^*(\mathbf{s}', \mathbf{a}^A)) + Q(\mathbf{s}', \mathbf{a}^A) \mid \mathbf{s}, \mathbf{a}\right] = r(\mathbf{s}, \mathbf{a}).$$

□

B.3. Theoretical Results for Theorem 2

In this Section, we provide all the assumptions required by Theorem 2, and prove Theorem 2.

B.3.1. Extra Definitions and Assumptions for Theorem 2

We first list the notions required for the proof of Theorem 2. To start with we define $\bar{r}$ and $\bar{Q}$ with $\hat{\pi}$ and the exact expectation $\mathbb{E}$.

Definition 6. *We use Q_k to denote the result of Algorithm 2 after the k^{th} iteration. Accordingly, $Q_k^A(\mathbf{s}) = Q_k(\mathbf{s}, \mathbf{a}^A)$. Define $\bar{Q}^A(\mathbf{s})$ as the solution to a fixed point to the operator $\bar{\mathcal{T}}$:*

$$\bar{\mathcal{T}}\bar{Q}^A(\mathbf{s}) = \bar{Q}^A(\mathbf{s}),$$

with

$$\bar{\mathcal{T}}\bar{Q}^A(\mathbf{s}) = g(\mathbf{s}) + \gamma \mathbb{E}\big[-\alpha \log(\hat{\pi}(\mathbf{s}', \mathbf{a}^A)) + \bar{Q}^A(\mathbf{s}') \mid \mathbf{s}, \mathbf{a}^A\big],$$

with an exact expectation and the estimated policy.

Definition 7. *Denote the estimated reward function by Algorithm 1 as $\hat{r}(\mathbf{s}, \mathbf{a})$. Define*

$$\bar{r}(\mathbf{s}, \mathbf{a}) := \hat{Q}(\mathbf{s}, \mathbf{a}) - \gamma \mathbb{E}\Big[-\alpha \log(\hat{\pi}(\mathbf{s}', \mathbf{a}^A)) + \hat{Q}(\mathbf{s}', \mathbf{a}^A) \mid \mathbf{s}, \mathbf{a}\Big],$$

with an exact expectation and the estimated policy.

Definition 8 (Definition 5.1 in Arora et al. (2019))**.** *A distribution is (λ, P, n)-non-degenerate if for n IID samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ with $\mathbf{x}_i \in \mathcal{S} \times \mathcal{A}$, $\lambda_{min}(\mathbf{H}^\infty) \geq \lambda \geq 0$ with probability at least $1 - P$. λ_{min} denotes the smallest eigen value. $\mathbf{H}^\infty$ is defined in Theorem 2.*

Definition 9. *$\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ follows a $(\lambda, \frac{P}{3}, n)$-non-degenerate distribution $\mathcal{D}$, and $\{(\mathbf{x}_i, y_i^k)\}_{i=1}^n$ follows a $(\lambda, \frac{P}{3}, n)$-non-degenerate distribution $\mathcal{D}^k$. We use subscripts like $D_{\mathbf{s}}$ to denote the marginal distribution of the state variables according to $\mathcal{D}$.*

It can be noticed that 9 is trying to quantify the quality of sample generation process. In practice, it is very hard to exactly derive λ.

Definition 10. *Define*

$$\rho_k(\mathbf{s}) := g(\mathbf{s}) + \gamma \mathbb{E}\left[-\alpha \log(\hat{\pi}(\mathbf{s}', \mathbf{a}^A)) + Q_{k-1}^A(\mathbf{s}') \mid \mathbf{s}, \mathbf{a}^A\right] - Q_k^A(\mathbf{s})$$

and

$$\epsilon_{Max} := \max_{k \in [N]} \left[\int_{\mathcal{S}} |\rho_k(\mathbf{s})|^2 d\mathcal{D}_{\mathbf{s}}(\mathbf{s})\right]^{1/2}.$$

Definition 11. *We use m to denote the number of neurons in the hidden layer of neural networks in Algorithm 1 and Algorithm 2.*

To quantify the generalization errors of neural networks, we require the following assumptions.

Assumption 6. *The numbers of training iterations of neural networks in Algorithm 1 and Algorithm 2 equal to $\Omega\left(\frac{1}{\eta\lambda} \log \frac{n}{P}\right)$, where $\eta > 0$ is the learning rate. m in Definition 11 satisfies $m \geq c^{-2}\mathrm{poly}(n, \lambda^{-1}, \frac{3}{P})$, where $c = O\left(\frac{\lambda P}{n}\right)$.*

Assumption 7. *The neural networks in Algorithm 1 and Algorithm 2 are two-layer ReLU networks trained by the randomly initialized gradient descent.*

Further, for the MDP in Algorithm 2, we pose the following assumptions commonly used for the FQI method (Munos and Szepesvári, 2008; Yang et al., 2020).

Assumption 8. *The true reward function can be bounded by* $\max_{\mathbf{s},\mathbf{a}}|r(\mathbf{s},\mathbf{a})| \leq R$*.*

Assumption 9. *Define the operator* $\bar{\mathcal{T}}$ *on the set of continuous bounded functions* $f : \mathcal{S} \longmapsto \mathbb{R}$

$$\bar{\mathcal{T}} f(\mathbf{s}) := g(\mathbf{s}) + \gamma \mathbb{E}\big[-\alpha \log(\hat{\pi}(\mathbf{s}', \mathbf{a}^A)) + f(\mathbf{s}') \mid \mathbf{s}, \mathbf{a}^A\big].$$

Define the concentration coefficient as

$$\kappa(m) = \left[\mathbb{E}_{\mathcal{D}_\mathbf{s}}\left|\frac{d\bar{\mathcal{T}}^m \mathcal{D}_\mathbf{s}}{d\mathcal{D}_\mathbf{s}}\right|^2\right]^{1/2}.$$

We assume that $(1-\gamma)\sum_{m\geq 1}\gamma^{m-1}\kappa(m) \leq \psi$*.*

B.3.2. Auxiliary Lemmas for Theorem 2

Lemma 8 (Hoeffding's Inequality)**.** *Let* $X_1, X_2, \cdots, X_n$ *be* n *IID random variables drawn from distribution* $\mathcal{D}$*, with* $0 \leq X_i \leq a$*,* $\forall i \in \{1, 2, \cdots, n\}$*. Let* $\bar{X} := \frac{1}{n}\sum_{i=1}^n X_i$*. Then, for any* $t > 0$*,*

$$P(\big|\bar{X} - \mathbb{E}[\bar{X}]\big| \geq t) \leq 2\exp\left(-\frac{2nt^2}{a^2}\right).$$

To start with, we consider the errors induced by Algorithm 1 by studying the difference between $\hat{r}(\mathbf{s},\mathbf{a})$ and $\bar{r}(\mathbf{s},\mathbf{a})$.

Lemma 9. *With Definition 7 and 9, under Assumption 3, 7, and 6, we have*

$$\mathbb{E}_{(\mathbf{s},\mathbf{a})\sim\mathcal{D}_{\mathbf{s},\mathbf{a}}}[|\hat{r}(\mathbf{s},\mathbf{a}) - \bar{r}(\mathbf{s},\mathbf{a})|] \leq \sqrt{\frac{2\mathbf{y}^T(\mathbf{H}^\infty)^{-1}\mathbf{y}}{n}} + O\left(\sqrt{\frac{\log(\frac{n}{\lambda P})}{n}}\right),$$

with probability at least $1-P$*.*

Proof. Note that the difference between $\hat{r}$ and $\bar{r}$ is no more than the estimation error of the expectation in $\hat{r}$, which is conducted by neural networks. Therefore, we apply Theorem 5.1 in Arora et al. (2019). □

Next, we consider the error of Algorithm 2. First, we bound the difference between $\hat{Q}^A(\mathbf{s})$ and $\bar{Q}^A(\mathbf{s})$.

Lemma 10. *With Definitions 6 and 10, under Assumptions 8 and 9, we have*

$$\mathbb{E}_{(\mathbf{s},\mathbf{a})\sim\mathcal{D}_{\mathbf{s},\mathbf{a}}}\left[\left|\hat{Q}^A(\mathbf{s}) - \bar{Q}^A(\mathbf{s})\right|\right] \leq \sum_{k=0}^{N-1}\gamma^{N-k-1}\kappa(N-k-1)\epsilon_{Max} + \gamma^N\frac{2R}{1-\gamma},$$

with probability at least $1-P$*.*

Proof. By Lemma C.2 in Yang et al. (2020), we can derive

$$\hat{Q}^A(\mathbf{s}) - \bar{Q}^A(\mathbf{s}) = \sum_{k=0}^{N-1}\left(\gamma^{N-k-1}\bar{\mathcal{T}}^{N-k-1}\rho_{k+1}(\mathbf{s})\right) + \gamma^N\bar{\mathcal{T}}^N(\bar{Q}^A(\mathbf{s}) - Q_0^A(\mathbf{s})),$$

where Q_0^A denotes the initialized estimator for Algorithm 2. Then, we apply the expectation to the absolute value of both sides of the equation above:

$$\begin{aligned}\mathbb{E}_{\mathbf{s}\sim\mathcal{D}_\mathbf{s}}\left[\left|\hat{Q}^A(\mathbf{s}) - \bar{Q}^A(\mathbf{s})\right|\right] &\leq \sum_{k=0}^{N-1}\left(\gamma^{N-k-1}\mathbb{E}_{\mathbf{s}\sim\mathcal{D}_\mathbf{s}}\left[\left|\bar{\mathcal{T}}^{N-k-1}\rho_{k+1}(\mathbf{s})\right|\right]\right) \\ &\quad + \gamma^N\mathbb{E}_{\mathbf{s}\sim\mathcal{D}_\mathbf{s}}\left[\left|\bar{\mathcal{T}}^N(\bar{Q}^A(\mathbf{s}) - Q_0^A(\mathbf{s}))\right|\right] \\ &\leq \sum_{k=0}^{N-1}\gamma^{N-k-1}\mathbb{E}_{\mathbf{s}\sim\mathcal{D}_\mathbf{s}}\left[\left|\bar{\mathcal{T}}^{N-i-1}\rho_{k+1}(\mathbf{s})\right|\right] + \gamma^N\frac{2R}{1-\gamma}.\end{aligned}$$

By Cauchy-Schwarz inequality, we have

$$\begin{aligned}\mathbb{E}_{(\mathbf{s},\mathbf{a})\sim\mathcal{D}_{\mathbf{s},\mathbf{a}}}\left[\left|\bar{\mathcal{T}}^{N-k-1}\rho_{k+1}\right|\right] &\leq \left[\int_{\mathcal{S}}\left|\frac{d(\bar{\mathcal{T}}^{N-k-1}\mathcal{D}_{\mathbf{s}})}{d\mathcal{D}_{\mathbf{s}}}(\mathbf{s})\right|^2 d\mathcal{D}_{\mathbf{s}}(\mathbf{s})\right]^{1/2}\left[\int_{\mathcal{S}}\left|\rho_{k+1}(\mathbf{a})\right|^2 d\mathcal{D}_{\mathbf{s}}(\mathbf{s})\right]^{1/2}\\ &\leq \kappa(N-k-1)\left[\int_{\mathcal{S}}\left|\rho_{k+1}(\mathbf{a})\right|^2 d\mathcal{D}_{\mathbf{s}}(\mathbf{s})\right]^{1/2},\end{aligned}$$

where the last inequality is according to Assumption 9. Therefore,

$$\begin{aligned}\mathbb{E}_{\mathbf{s}\sim\mathcal{D}_{\mathbf{s}}}\left[\left|\hat{Q}^A(\mathbf{s})-\bar{Q}^A(\mathbf{s})\right|\right] &\leq \sum_{i=0}^{N-1}\gamma^{N-k-1}\mathbb{E}_{\mathbf{s}\sim\mathcal{D}_{\mathbf{s}}}\left[\left|P^{N-k-1}\rho_{k+1}(\mathbf{s})\right|\right]+\gamma^N\frac{2R}{1-\gamma}\\ &\leq \sum_{i=0}^{N-1}\gamma^{N-k-1}\kappa(N-k-1)\left[\int_{\mathcal{S}}\left|\rho_{k+1}(\mathbf{s})\right|^2 d\mathcal{D}_{\mathbf{s}}(\mathbf{s})\right]^{1/2}+\gamma^N\frac{2R}{1-\gamma}\\ &\leq \sum_{k=0}^{N-1}\gamma^{N-k-1}\kappa(N-k-1)\epsilon_{Max}+\gamma^N\frac{2R}{1-\gamma},\end{aligned}$$

where the last inequality is according to Definition 10. □

We study the $\rho_k(\mathbf{s})$ in ϵ_{Max}:

$$\rho_k(\mathbf{s}) = g(\mathbf{s})+\gamma\mathbb{E}\left[-\alpha\log(\hat{\pi}(\mathbf{s}',\mathbf{a}^A))+Q^A_{k-1}(\mathbf{s}')\mid\mathbf{s},\mathbf{a}^A\right]-Q^A_k.$$

According to the the fifth line of 2, Q^A_k is a deep function trained using the samples $g(\mathbf{s})+\gamma-\alpha\log(\hat{\pi}(\mathbf{s}',\mathbf{a}^A))+Q^A_{k-1}(\mathbf{s}')$, where $\mathbf{s}'$ is the next-step state variable. In other words, the Q^A_k is trained to estimate $g(\mathbf{s})+\gamma\mathbb{E}\left[-\alpha\log(\hat{\pi}(\mathbf{s}',\mathbf{a}^A))+Q^A_{k-1}(\mathbf{s}')\mid\mathbf{s},\mathbf{a}^A\right]$, the first part of $\rho_k(\mathbf{s})$. Therefore, it can be seen that ϵ_{Max} is really caused by the neural networks used in Algorithm 2. We use the results in Arora et al. (2019) again to bound ϵ_{Max}.

Lemma 11. *With Definition 9, under Assumptions 3, 7, and 6, we have*

$$\epsilon_{Max}\leq\max_{k\in[N]}\sqrt{\frac{2\mathbf{y}_k^T(\mathbf{H}^\infty)^{-1}\mathbf{y}_k}{n}}+O\left(\sqrt{\frac{\log(\frac{n}{\lambda P})}{n}}\right).$$

with probability at least $1-(N+1)P$.

Proof. We consider

$$\begin{aligned}&\left[\int_{\mathcal{S}}\left|\rho_k(\mathbf{s})\right|^2 dD_{\mathbf{s}}(\mathbf{s})\right]^{1/2}\\ &=\left[\int_{\mathcal{S}}\left|Q^A_k(\mathbf{s})-g(\mathbf{s})-\gamma\mathbb{E}\left[-\alpha\log(\hat{\pi}(\mathbf{s}',\mathbf{a}^A))+Q^A_{k-1}(\mathbf{s}')\mid\mathbf{s},\mathbf{a}^A\right]\right|^2 dD_{\mathbf{s}}(\mathbf{s})\right]^{1/2}\\ &\leq\left[\int_{\mathcal{S}\times\mathcal{Y}}\left|y-g(\mathbf{s})-\gamma\mathbb{E}\left[-\alpha\log(\hat{\pi}(\mathbf{s}',\mathbf{a}^A))+Q^A_{k-1}(\mathbf{s}')\mid\mathbf{s},\mathbf{a}^A\right]\right|^2 dD_{\mathbf{s},y}(\mathbf{s},y)\right]^{1/2}\\ &\quad+\left[\int_{\mathcal{S}\times\mathcal{Y}}\left|Q^A_k(\mathbf{s})-y\right|^2 dD_{\mathbf{s},y}(\mathbf{s},y)\right]^{1/2}\end{aligned}\tag{19}$$

It should be noticed that the second part on the right hand side is the generation error of the deep estimation. According to Theorem 5.1 in Arora et al. (2019), with probability at least $1-P$, we have

$$\left[\int_{\mathcal{S}\times\mathcal{Y}}\left|Q^A_k(\mathbf{s})-y\right|^2 dD_{\mathbf{s},y}(\mathbf{s},y)\right]^{1/2}\leq\sqrt{\frac{2\mathbf{y}_k^T(\mathbf{H}^\infty)^{-1}\mathbf{y}_k}{n}}+O\left(\sqrt{\frac{\log(\frac{n}{\lambda P})}{n}}\right).$$

We bound the first part of the RHS of (19) by Hoeffding's Inequality (Lemma 8). This provides an error with smaller order, and thus can be ignored. Therefore, by the union bound, we can conclude that, with probability at least $1-(N+1)P$,

$$\epsilon_{Max} \leq \max_{k\in[N]} \sqrt{\frac{2\mathbf{y}_k^T(\mathbf{H}^\infty)^{-1}\mathbf{y}_k}{n}} + O\left(\sqrt{\frac{\log(\frac{n}{\lambda P})}{n}}\right).$$

□

Lemma 12. *Under Assumption 2, the estimation error of $\hat{Q}$ can be bounded by:*

$$\mathbb{E}_{(\mathbf{s},\mathbf{a})\sim\mathcal{D}_{\mathbf{s},\mathbf{a}}}\left[\left|\hat{Q}(\mathbf{s},\mathbf{a})-Q(\mathbf{s},\mathbf{a})\right|\right] \leq \mathbb{E}_{\mathbf{s}\sim\mathcal{D}_{\mathbf{s}}}\left[\left|\hat{Q}^A(\mathbf{s})-Q^A(\mathbf{s})\right|\right] + 2\alpha\epsilon_\pi.$$

Proof. By the definition of π^* in (2), we can derive

$$\log(\pi^*(\mathbf{s},\mathbf{a})) - \log(\pi^*(\mathbf{s},\mathbf{a}^A)) = \frac{Q(\mathbf{s},\mathbf{a})-Q(\mathbf{s},\mathbf{a}^A)}{\alpha}.$$

Therefore,

$$Q(\mathbf{s},\mathbf{a}) = \alpha\log(\pi^*(\mathbf{s},\mathbf{a})) - \alpha\log(\pi^*(\mathbf{s},\mathbf{a}^A)) + Q^A(\mathbf{s}).$$

Similarly, according to Algorithm 2,

$$\hat{Q}(\mathbf{s},\mathbf{a}) = \alpha\log(\hat{\pi}(\mathbf{s},\mathbf{a})) - \alpha\log(\hat{\pi}(\mathbf{s},\mathbf{a}^A)) + \hat{Q}^A(\mathbf{s}).$$

Therefore, by Assumption 2,

$$\mathbb{E}_{(\mathbf{s},\mathbf{a})\sim\mathcal{D}_{\mathbf{s},\mathbf{a}}}\left[\left|\hat{Q}(\mathbf{s},\mathbf{a})-Q(\mathbf{s},\mathbf{a})\right|\right] \leq \mathbb{E}_{\mathbf{s}\sim\mathcal{D}_{\mathbf{s}}}\left[\left|\hat{Q}^A(\mathbf{s})-Q^A(\mathbf{s})\right|\right] + 2\alpha\epsilon_\pi.$$

□

Lemma 13. *With Definitions 6, under Assumption 2, we have*

$$\mathbb{E}_{\mathbf{s}\sim\mathcal{D}_{\mathbf{s}}}\left[\left|\bar{Q}^A(\mathbf{s})-Q^A(\mathbf{s})\right|\right] \leq \frac{\gamma\alpha\epsilon_\pi}{1-\gamma}.$$

Proof. According to Lemma 7, we have

$$Q^A(\mathbf{s}) := g(\mathbf{s}) + \gamma\mathbb{E}\big[-\alpha\log(\pi^*(\mathbf{s}',\mathbf{a}^A)) + Q^A(\mathbf{s}') \mid \mathbf{s},\mathbf{a}^A\big].$$

On the other hand, by Definition 6

$$\bar{Q}^A(\mathbf{s}) = g(\mathbf{s}) + \gamma\mathbb{E}\big[-\alpha\log(\hat{\pi}(\mathbf{s}',\mathbf{a}^A)) + \bar{Q}^A(\mathbf{s}') \mid \mathbf{s},\mathbf{a}^A\big].$$

Therefore, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{s}\sim\mathcal{D}_{\mathbf{s}}}\left[\left|\bar{Q}^A(\mathbf{s})-Q^A(\mathbf{s})\right|\right]\\
=\,&\gamma\mathbb{E}_{\mathbf{s}\sim\mathcal{D}_{\mathbf{s}}}\left\{\left|\mathbb{E}\left[\alpha\log(\pi^*(\mathbf{s}',\mathbf{a}^A)) - \alpha\log(\hat{\pi}(\mathbf{s}',\mathbf{a}^A)) + \bar{Q}^A(\mathbf{s}') - Q^A(\mathbf{s}') \mid \mathbf{s},\mathbf{a}^A\right]\right|\right\}\\
\leq\,&\gamma\mathbb{E}_{\mathbf{s}\sim\mathcal{D}_{\mathbf{s}}}\left[\left|Q^A(\mathbf{s})-\bar{Q}^A(\mathbf{s})\right|\right] + \gamma\alpha\epsilon_\pi,
\end{aligned}$$

which means

$$\mathbb{E}_{\mathbf{s}\sim\mathcal{D}_{\mathbf{s}}}\left[\left|Q^A(\mathbf{s})-\bar{Q}^A(\mathbf{s})\right|\right] \leq \frac{\gamma\alpha\epsilon_\pi}{1-\gamma}.$$

□

Lemma 14. *With Definition 6 and 7, under Assumption 2, we have*

$$\begin{aligned}
&\mathbb{E}_{(\mathbf{s},\mathbf{a})\sim\mathcal{D}_{\mathbf{s},\mathbf{a}}}[|r(\mathbf{s},\mathbf{a})-\hat{r}(\mathbf{s},\mathbf{a})|]\\
&\leq (1+\gamma)\mathbb{E}_{\mathbf{s}\sim\mathcal{D}_{\mathbf{s}}}\left[\left|Q^A(\mathbf{s})-\hat{Q}^A(\mathbf{s})\right|\right] + \mathbb{E}_{(\mathbf{s},\mathbf{a})\sim\mathcal{D}_{\mathbf{s},\mathbf{a}}}[|\bar{r}(\mathbf{s},\mathbf{a})-\hat{r}(\mathbf{s},\mathbf{a})|]\\
&+ (\gamma+2)\alpha\epsilon_\pi.
\end{aligned}$$

Proof. To start with, by the tower principle, the triangle inequality, Lemma 6 and Definition 6, we can derive

$$\begin{aligned}\mathbb{E}_{(\mathbf{s},\mathbf{a})\sim\mathcal{D}_{\mathbf{s},\mathbf{a}}}\left[|r(\mathbf{s},\mathbf{a})-\hat{r}(\mathbf{s},\mathbf{a})|\right] \leq & \mathbb{E}_{(\mathbf{s},\mathbf{a})\sim\mathcal{D}_{\mathbf{s},\mathbf{a}}}\left[\left|Q(\mathbf{s},\mathbf{a})-\hat{Q}(\mathbf{s},\mathbf{a})\right|\right] \\ & + \gamma\alpha\mathbb{E}_{\mathbf{s}\sim\mathcal{D}_{\mathbf{s}}}\left[\left|\log(\pi^*(\mathbf{s},\mathbf{a}^A)) - \log(\hat{\pi}(\mathbf{s},\mathbf{a}^A))\right|\right] \\ & + \gamma\mathbb{E}_{(\mathbf{s},\mathbf{a})\sim\mathcal{D}_{\mathbf{s},\mathbf{a}}}\left[\left|Q^A(\mathbf{s})-\hat{Q}^A(\mathbf{s})\right|\right] \\ \leq & \mathbb{E}_{(\mathbf{s},\mathbf{a})\sim\mathcal{D}_{\mathbf{s},\mathbf{a}}}\left[\left|Q(\mathbf{s},\mathbf{a})-\hat{Q}(\mathbf{s},\mathbf{a})\right|\right] \\ & + \mathbb{E}_{(\mathbf{s},\mathbf{a})\sim\mathcal{D}_{\mathbf{s},\mathbf{a}}}\left[|\bar{r}(\mathbf{s},\mathbf{a})-\hat{r}(\mathbf{s},\mathbf{a})|\right] + \gamma\alpha\epsilon_\pi \\ & + \gamma\mathbb{E}_{\mathbf{s}\sim\mathcal{D}_{\mathbf{s}}}\left[\left|Q^A(\mathbf{s})-\hat{Q}^A(\mathbf{s})\right|\right].\end{aligned}$$

Next, we use the results of Lemma 12, and derive

$$\begin{aligned}\mathbb{E}_{(\mathbf{s},\mathbf{a})\sim\mathcal{D}_{\mathbf{s},\mathbf{a}}}\left[|r(\mathbf{s},\mathbf{a})-\hat{r}(\mathbf{s},\mathbf{a})|\right] \leq (1+\gamma)\mathbb{E}_{\mathbf{s}\sim\mathcal{D}_{\mathbf{s}}}\left[\left|Q^A(\mathbf{s})-\hat{Q}^A(\mathbf{s})\right|\right] \\ + \mathbb{E}_{(\mathbf{s},\mathbf{a})\sim\mathcal{D}_{\mathbf{s},\mathbf{a}}}\left[|\bar{r}(\mathbf{s},\mathbf{a})-\hat{r}(\mathbf{s},\mathbf{a})|\right] + (\gamma+2)\alpha\epsilon_\pi,\end{aligned}$$

which proves the lemma. □

B.3.3. Proof of Theorem 2

Now, we prove Theorem 2. To start with, we study the error of Algorithm 2. Combining Lemma 10 and Lemma 11, we can derive that with the probability of at least $1-(N+1)P$:

$$\begin{aligned}&\mathbb{E}_{(\mathbf{s},\mathbf{a})\sim\mathcal{D}_{\mathbf{s},\mathbf{a}}}\left[\left|\hat{Q}^A(\mathbf{s})-\bar{Q}^A(\mathbf{s})\right|\right] \\ &\leq \sum_{k=0}^{N-1}\kappa(N-k-1)\gamma^{N-k-1}\left[\max_{k\in[N]}\sqrt{\frac{2\mathbf{y}_k^T(\mathbf{H}^\infty)^{-1}\mathbf{y}_k}{n}} + O\left(\sqrt{\frac{\log(\frac{n}{\lambda P})}{n}}\right)\right] + \gamma^N\frac{2R}{1-\gamma} \\ &\leq \frac{\phi}{1-\gamma}\max_{k\in[N]}\sqrt{\frac{2\mathbf{y}_k^T(\mathbf{H}^\infty)^{-1}\mathbf{y}_k}{n}} + \gamma^N\frac{2R}{1-\gamma} + O\left(\sqrt{\frac{\log(\frac{n}{\lambda P})}{n}}\right),\end{aligned} \tag{20}$$

where the last inequality is by Assumption 9. Further, by using Lemma 12 and Lemma 13, we can derive

$$\begin{aligned}&\mathbb{E}_{(\mathbf{s},\mathbf{a})\sim\mathcal{D}_{\mathbf{s},\mathbf{a}}}\left[\left|\hat{Q}(\mathbf{s},\mathbf{a})-Q(\mathbf{s},\mathbf{a})\right|\right] \\ &\leq \mathbb{E}_{\mathbf{s}\sim\mathcal{D}_{\mathbf{s}}}\left[\left|\hat{Q}^A(\mathbf{s})-Q^A(\mathbf{s})\right|\right] + 2\alpha\epsilon_\pi \\ &\leq \mathbb{E}_{\mathbf{s}\sim\mathcal{D}_{\mathbf{s}}}\left[\left|\hat{Q}^A(\mathbf{s})-\bar{Q}^A(\mathbf{s})\right|\right] + \mathbb{E}_{\mathbf{s}\sim\mathcal{D}_{\mathbf{s}}}\left[\left|\bar{Q}^A(\mathbf{s})-Q^A(\mathbf{s})\right|\right] + 2\alpha\epsilon_\pi \\ &\leq \frac{\phi}{1-\gamma}\max_{k\in[N]}\sqrt{\frac{2\mathbf{y}_k^T(\mathbf{H}^\infty)^{-1}\mathbf{y}_k}{n}} + \gamma^N\frac{2R}{1-\gamma} + O\left(\sqrt{\frac{\log(\frac{n}{\lambda P})}{n}}\right) + \frac{\gamma\alpha\epsilon_\pi}{1-\gamma}\end{aligned}$$

As a result, the error of Q-function estimation is bounded. Next, we proceed to the reward estimation. According to Lemma 9 and Lemma 14, we have

$$
\begin{aligned}
&\mathbb{E}_{(\mathbf{s},\mathbf{a})\sim\mathcal{D}_{\mathbf{s},\mathbf{a}}}[|\hat{r}(\mathbf{s},\mathbf{a}) - r(\mathbf{s},\mathbf{a})|] \\
&\leq (1+\gamma)\mathbb{E}_{(\mathbf{s},\mathbf{a})\sim\mathcal{D}_{\mathbf{s},\mathbf{a}}}\left[\left|Q(\mathbf{s},\mathbf{a}) - \hat{Q}(\mathbf{s},\mathbf{a})\right|\right] + \mathbb{E}_{(\mathbf{s},\mathbf{a})\sim\mathcal{D}_{\mathbf{s},\mathbf{a}}}[|\bar{r}(\mathbf{s},\mathbf{a}) - \hat{r}(\mathbf{s},\mathbf{a})|] + (2+\gamma)\alpha\epsilon_\pi \\
&\leq (1+\gamma)\left\{\frac{\psi}{1-\gamma}\max_{k\in[N]}\sqrt{\frac{2\mathbf{y}_k^T(\mathbf{H}^\infty)^{-1}\mathbf{y}_k}{n}} + \gamma^N\frac{2R}{1-\gamma} + O\left(\sqrt{\frac{\log(\frac{n}{\lambda P})}{n}}\right) + \frac{\gamma\alpha\epsilon_\pi}{1-\gamma}\right\} \\
&+ \sqrt{\frac{2\mathbf{y}^T(\mathbf{H}^\infty)^{-1}\mathbf{y}}{n}} + (2+\gamma)\alpha\epsilon_\pi + O\left(\sqrt{\frac{\log(\frac{n}{\lambda P})}{n}}\right) \\
&\leq \frac{(1+\gamma)\psi}{1-\gamma}\max_{k\in[N]}\sqrt{\frac{2\mathbf{y}_k^T(\mathbf{H}^\infty)^{-1}\mathbf{y}_k}{n}} + \gamma^N\frac{2R(1+\gamma)}{1-\gamma} + \frac{2\alpha\epsilon_\pi}{1-\gamma} \\
&+ \sqrt{\frac{2\mathbf{y}^T(\mathbf{H}^\infty)^{-1}\mathbf{y}}{n}} + O\left(\sqrt{\frac{\log(\frac{n}{\lambda P})}{n}}\right),
\end{aligned}
$$

with probability $1-(N+2)P$. Note that the proved results is slightly different from the Theorem 2. The two are consistent if we select $C = \max(\psi, R, 1/\lambda)$, and treat $P(N+2)$ as a constant for the tail probability.

C. Extended Experiments

We now provide extended experiments.

C.1. Details of Synthetic Experiments

In this section, we provide more details about the synthetic experiments implemented in Section 6.1.

C.1.1. Data Generation Process

We consider an MDP defined by the tuple $\{\mathcal{S}, \mathcal{A}, \mathrm{P}, \gamma, r\}$.

- $\mathcal{S} = [-p, p]^p$, with $p \in \{5, 10, 20, 40\}$, denotes the state space.
- $\mathcal{A} = [5]$ is the action space.
- The reward function is defined as
$$r(\mathbf{s},\mathbf{a}) = \frac{\mathbf{a}\tanh(\{\mathbf{s}^\top/p, \mathbf{a}/4\}\cdot\boldsymbol{\omega})}{4\mathbf{1}^\top\boldsymbol{\omega}},$$
where $\boldsymbol{\omega}$ is a $(p+1)\times 1$ vector.
- $\gamma = 0.9$ is the discount factor.
- $\alpha = 1$.
- The transition is defined as
$$\mathrm{P}(\mathbf{s}' \mid \mathbf{s},\mathbf{a}) = \begin{cases} 1 & s' = \mathbf{s} + \mathbf{a}/5 - 0.5 \in \mathcal{S} \\ \frac{1}{|\mathcal{S}|} & \text{otherwise} \end{cases}.$$

Next, we solve the MDP with a deep energy-based policy with $\gamma = 0.9$ and $\alpha = 1$, using the soft Q-learning method in Haarnoja et al. (2017). By conducting the learned policy for 50000 steps, we obtain the demonstration dataset, on which we compare PQR, MaxEnt-IRL, AIRL, and SPL-GD. We assume that γ and α are known as required by existing methods. Specifically, the hyperparameters used for the soft Q-learning are reported in Table 1.

C.1.2. Competing Methods

In this section, we provide more details regarding how we implement existing methods.

p	Optimization Method	Learning Rate	Final Loss for Q-Function	Number of Epochs
5	Adagrad	10^{-4}	4×10^{-3}	100
10	Adagrad	10^{-4}	6×10^{-3}	150
20	Adagrad	10^{-4}	9×10^{-3}	175
40	Adagrad	10^{-4}	1.2×10^{-2}	175

Table 1: Hyperparameters for soft Q-learning.

MaxEnt-IRL According to our analysis, without an appropriate identification procedure, the estimated Q-function may be shaped by any function of $\mathbf{s}$. Therefore, we implement MaxEnt-IRL with a grounding procedure by Algorithm 4.

Algorithm 4 MaxEnt-IRL with Grounding Procedure

```
Input: $\mathbb{X} = \{\mathbf{s}_0, \mathbf{a}_0, \mathbf{s}_1, \mathbf{a}_1, \cdots, \mathbf{s}_T, \mathbf{a}_T\}$, and $Q(0, 0)$
Output: $\hat{Q}(\mathbf{s}, \mathbf{a})$.
$\hat{Q}(\mathbf{s}, \mathbf{a}) \leftarrow$MAXENT-IRL($\mathbb{X}$)
$\hat{Q}(\mathbf{s}, \mathbf{a}) \leftarrow \hat{Q}(\mathbf{s}, \mathbf{a}) - \hat{Q}(0, 0) + Q(0, 0)$.
return $\hat{Q}(\mathbf{s}, \mathbf{a})$.
```

In other words, we allow MaxEnt-IRL to access the ground truth Q value $Q(0, 0)$. Empirically, this procedure significantly improves the performance of MaxEnt-IRL.

SPL-GD The original SPL-GD can only deal with discrete and finite state variables. To make the SPL-GD method available for our setting, we provide access to the true value function and Q-function. The procedure is summarized in Algorithm 5.

Algorithm 5 SPL-GD for Continuous States

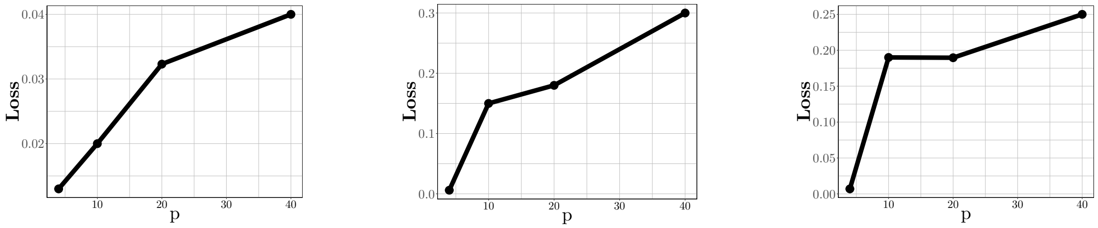
```
Input: Dataset: $\mathbb{X} = \{\mathbf{s}_0, \mathbf{a}_0, \mathbf{s}_1, \mathbf{a}_1, \cdots, \mathbf{s}_T, \mathbf{a}_T\}$, $Q(\mathbf{s}, \mathbf{a})$ and $V(\mathbf{s})$.
Output: $\hat{r}(\mathbf{s}, \mathbf{a})$.
for $t \in [T]$ do
    $y_t \leftarrow Q(\mathbf{s}_t, \mathbf{a}_t) - \gamma V(\mathbf{s}_{t+1})$
end for
Train a linear regression with $\{\mathbf{s}_t\}_{t=0}^{T-1}$ and $\{\mathbf{a}_t\}_{t=0}^{T-1}$ for $\hat{r}(\mathbf{s}, \mathbf{a})$.
return $\hat{r}(\mathbf{s}, \mathbf{a})$.
```

C.1.3. PQR Deep Networks Hyperparameters

In this section, we provide hyperparameters for training the deep neural networks in PQR. There are three types of deep neural networks in PQR: networks for policy estimation, Q estimation, and reward estimation. The parameters are provided in Table 2. The training errors of the neural networks are provided in Figure 7

Types of Networks	Optimization Method	Learning Rate	Number of Steps
Policy Estimation	Adagrad	5×10^{-4}	600
Q Estimation	Adagrad	1×10^{-3}	1000
Reward Estimation	Adagrad	1×10^{-3}	1000

Table 2: Hyperparameters for PQR.



(a) Neural networks for policy.

(b) Neural networks for Q.

(c) Neural Networks for reward.

Figure 7: Final loss on the training datasets of neural networks for different tasks for MDPs with $p = \{5, 10, 20, 40\}$.

C.2. Extended Synthetic Experiments

C.2.1. Robustness Analysis

It should be noticed that the performance of the proposed PQR method relies on the existence of anchor actions in Assumption 1. While as we have mentioned this assumption is reasonable in many real-world scenarios (Hotz and Miller, 1993; Bajari et al., 2010; Manzanares et al., 2015), it contradicts the setting of SPL-GD and AIRL, where the reward function is assumed to only depend on the state variables. Therefore, in this section, we study the performance of the proposed method when the underlying reward is indeed a state-only function.

Specifically, we replace the reward function in Section C.1 with the following function:

$$r(\mathbf{s}) = \tanh(\{\mathbf{s}^\top/p, \mathbf{a}/4\} \cdot \boldsymbol{\omega}).$$

Then, we use the same configurations as in Section 6.1, and compare the performances of PQR, SPL-GD, and AIRL in terms of reward recovery for an MDP with a 10-dimension state variable. When implement the PQR method, we randomly choose one action as the anchor action. The results are summarized in Figure 8.

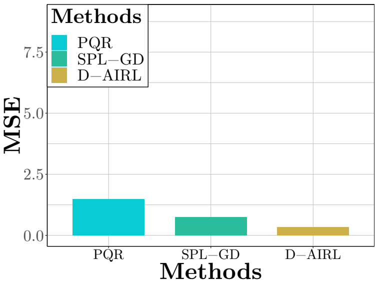


Figure 8: MSE for reward recovery

Note that D-AIRL performs the best if the reward function is indeed indecent with the action variable. The performance of the proposed method may not be guaranteed if anchor-action assumption is not satisfied. This is not surprising since the PQR method wrongly takes one action as the anchor action for the reward estimation under this setting.

C.2.2. Sensitivity Analysis

Theoretically, it is hard to identify the true α and γ. However, in some applications, there are some well-justified and widely used values for γ and α, like taking γ as the risk-free return rate in financial applications (Merton, 1973). Therefore, we treat α and γ as two hyperparameters, and provide a sensitivity analysis for PQR.

To analyze γ, we set $p = 10$ and $\alpha = 1$, and solve an MDP to generate three datasets with $\gamma = [0, 0.5, 0.9]$. We apply PQR with $\gamma = 0.9$ to the 3 datasets and check the reward estimation accuracy. We compare with the MaxEnt-IRL method, which is equivalent to taking $\gamma = 0$. Figure 9 plots the results. Recall from Section 2 that, when true $\gamma = 0$, the reward function equals to the Q-function. Therefore, MaxEnt-IRL gives a very accurate estimation, outperforming PQR with its mis-specified $\gamma = 0.9$. As true γ increases, PQR performance improves, while MaxEnt-IRL performance deteriorates. At true $\gamma = 0.9$, PQR achieves low error, while the error of MaxEnt-IRL blows up. It is important to note that the PQR error does not blow, as the identification procedure Algorithm 2 constrains the reward estimation to the right scale.

To analyze α, we set $p = 10$ and $\gamma = 0.9$, and solve an MDP to generate three datasets with $\alpha = [0.1, 0.5, 1]$. We apply PQR with $\alpha = 1$ and report results in Figure 9. As true α approaches 1, PQR performance improves. When true $\alpha = 0.1$, PQR error increases significantly. This indicates that PQR may be relatively more sensitive to α. Meanwhile, most IRL methods operate under $\alpha = 1$, with no ability to tune it. We provide a strategy to choose α in Section C.2.3.

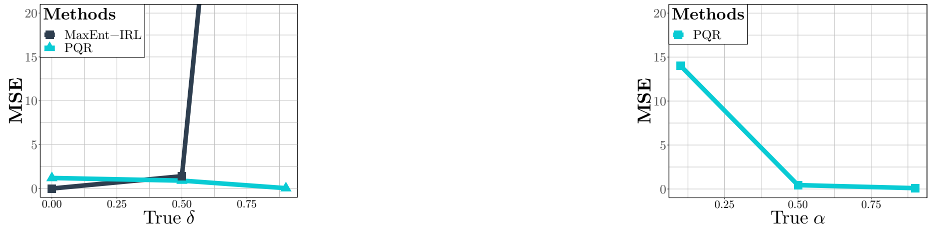


Figure 9: Sensitivity analysis for γ and α

C.2.3. α Selection

In this section, we propose a strategy to select α, under the assumption that we approximately know the scale of the true reward function, such as the average reward, maximal reward, or the minimal reward achieved in the dataset. To start with, we study the effect of α.

Theorem 3. *Given the demonstration dataset $\mathbb{X} = \{\mathbf{s}_0, \mathbf{a}_0, \mathbf{s}_1, \mathbf{a}_1, \cdots, \mathbf{s}_T, \mathbf{a}_T\}$, we use $\hat{r}_{\alpha_1}$ and $\hat{r}_{\alpha_2}$ to denote two estimated reward functions by PQR, using $\alpha_1, \alpha_2 \in (0, \infty)$ respectively. We assume that $\hat{\pi}$ is an accurate estimation to π^*, and the expectation in PQR estimators are exactly calculated. Therefore,*

$$\frac{\hat{r}_{\alpha_1}(\mathbf{s}, \mathbf{a}) - \frac{g(\mathbf{s})}{1-\gamma} + \gamma \mathbb{E}\left[\frac{g(\mathbf{s}')}{1-\gamma} \mid \mathbf{s}, \mathbf{a}\right]}{\hat{r}_{\alpha_2}(\mathbf{s}, \mathbf{a}) - \frac{g(\mathbf{s})}{1-\gamma} + \gamma \mathbb{E}\left[\frac{g(\mathbf{s}')}{1-\gamma} \mid \mathbf{s}, \mathbf{a}\right]} = \frac{\alpha_1}{\alpha_2},$$

where the expectation is on $\mathbf{s}'$ over one-step transition.

Proof. First, it should be noticed that α does not affect the estimation to π^*.

Then, according to Q^A-Estimator, we have

$$\hat{Q}^A_{\alpha_1}(\mathbf{s}) = \hat{\mathcal{T}}_{\alpha_1} \hat{Q}^A_{\alpha_1}(\mathbf{s})$$
$$\hat{Q}^A_{\alpha_2}(\mathbf{s}) = \hat{\mathcal{T}}_{\alpha_2} \hat{Q}^A_{\alpha_2}(\mathbf{s}),$$

where

$$\hat{\mathcal{T}}_{\alpha_1} f(\mathbf{s}) := g(\mathbf{s}) + \gamma \hat{\mathbb{E}}_{\mathbf{s}'}\left[-\alpha_1 \log(\hat{\pi}(\mathbf{s}', \mathbf{a}^A)) + f(\mathbf{s}') \mid \mathbf{s}, \mathbf{a}^A\right]$$
$$\hat{\mathcal{T}}_{\alpha_2} f(\mathbf{s}) := g(\mathbf{s}) + \gamma \hat{\mathbb{E}}_{\mathbf{s}'}\left[-\alpha_2 \log(\hat{\pi}(\mathbf{s}', \mathbf{a}^A)) + f(\mathbf{s}') \mid \mathbf{s}, \mathbf{a}^A\right].$$

Therefore,

$$\hat{Q}^A_{\alpha_1}(\mathbf{s}) - \frac{g(\mathbf{s})}{1-\gamma} := \gamma \hat{\mathbb{E}}_{\mathbf{s}'}\left[-\alpha_1 \log(\hat{\pi}(\mathbf{s}', \mathbf{a}^A)) + \hat{Q}^A_{\alpha_1}(\mathbf{s}) - \frac{g(\mathbf{s})}{1-\gamma} \mid \mathbf{s}, \mathbf{a}^A\right]$$
$$\hat{Q}^A_{\alpha_2}(\mathbf{s}) - \frac{g(\mathbf{s})}{1-\gamma} := \gamma \hat{\mathbb{E}}_{\mathbf{s}'}\left[-\alpha_2 \log(\hat{\pi}(\mathbf{s}', \mathbf{a}^A)) + \hat{Q}^A_{\alpha_2}(\mathbf{s}) - \frac{g(\mathbf{s})}{1-\gamma} \mid \mathbf{s}, \mathbf{a}^A\right],$$

which suggests that

$$\frac{\hat{Q}^A_{\alpha_1}(\mathbf{s}) - \frac{g(\mathbf{s})}{1-\gamma}}{\hat{Q}^A_{\alpha_2}(\mathbf{s}) - \frac{g(\mathbf{s})}{1-\gamma}} = \frac{\alpha_1}{\alpha_2}.$$

Therefore, by Q-ESTIMATOR, we can conclude that

$$\frac{\hat{Q}_{\alpha_1}(\mathbf{s},\mathbf{a}) - \frac{g(\mathbf{s})}{1-\gamma}}{\hat{Q}_{\alpha_2}(\mathbf{s},\mathbf{a}) - \frac{g(\mathbf{s})}{1-\gamma}} = \frac{\alpha_1}{\alpha_2}. \tag{21}$$

By taking (21) into R-ESTIMATOR, we have

$$\frac{\hat{r}_{\alpha_1}(\mathbf{s},\mathbf{a}) - \frac{g(\mathbf{s})}{1-\gamma} + \gamma\mathbb{E}\left[\frac{g(\mathbf{s}')}{1-\gamma} \mid \mathbf{s},\mathbf{a}\right]}{\hat{r}_{\alpha_2}(\mathbf{s},\mathbf{a}) - \frac{g(\mathbf{s})}{1-\gamma} + \gamma\mathbb{E}\left[\frac{g(\mathbf{s}')}{1-\gamma} \mid \mathbf{s},\mathbf{a}\right]} = \frac{\alpha_1}{\alpha_2}.$$

□

As a special case, when $g(\mathbf{s}) = 0$, we have

$$\frac{\hat{r}_{\alpha_1}(\mathbf{s},\mathbf{a})}{\hat{r}_{\alpha_2}(\mathbf{s},\mathbf{a})} = \frac{\alpha_1}{\alpha_2}.$$

Therefore, the α used in PQR determines the scale of the estimated reward function. If we know the scale of the true reward function, we can tune α so that the estimated reward function has the right scale. When $g(\mathbf{s}) = 0$, the procedure to select α is summarized in Algorithm 6:

Algorithm 6 α Selection

```
Input: Dataset: $\mathbb{X} = \{\mathbf{s}_0, \mathbf{a}_0, \mathbf{s}_1, \mathbf{a}_1, \cdots, \mathbf{s}_T, \mathbf{a}_T\}$.
Input: $R$, the average reward achieved on $\mathbb{X}$.
Input: $\gamma$ and $N$.
$\hat{r}(\mathbf{s},\mathbf{a}) \leftarrow \mathrm{PQR}(\mathbb{X}, \alpha = 1, \gamma, N)$
$\hat{\alpha} = \frac{R\cdot(T+1)}{\sum_{(\mathbf{s},\mathbf{a})\in\mathbb{X}} \hat{r}(\mathbf{s},\mathbf{a})}$
return $\hat{\alpha}$
```

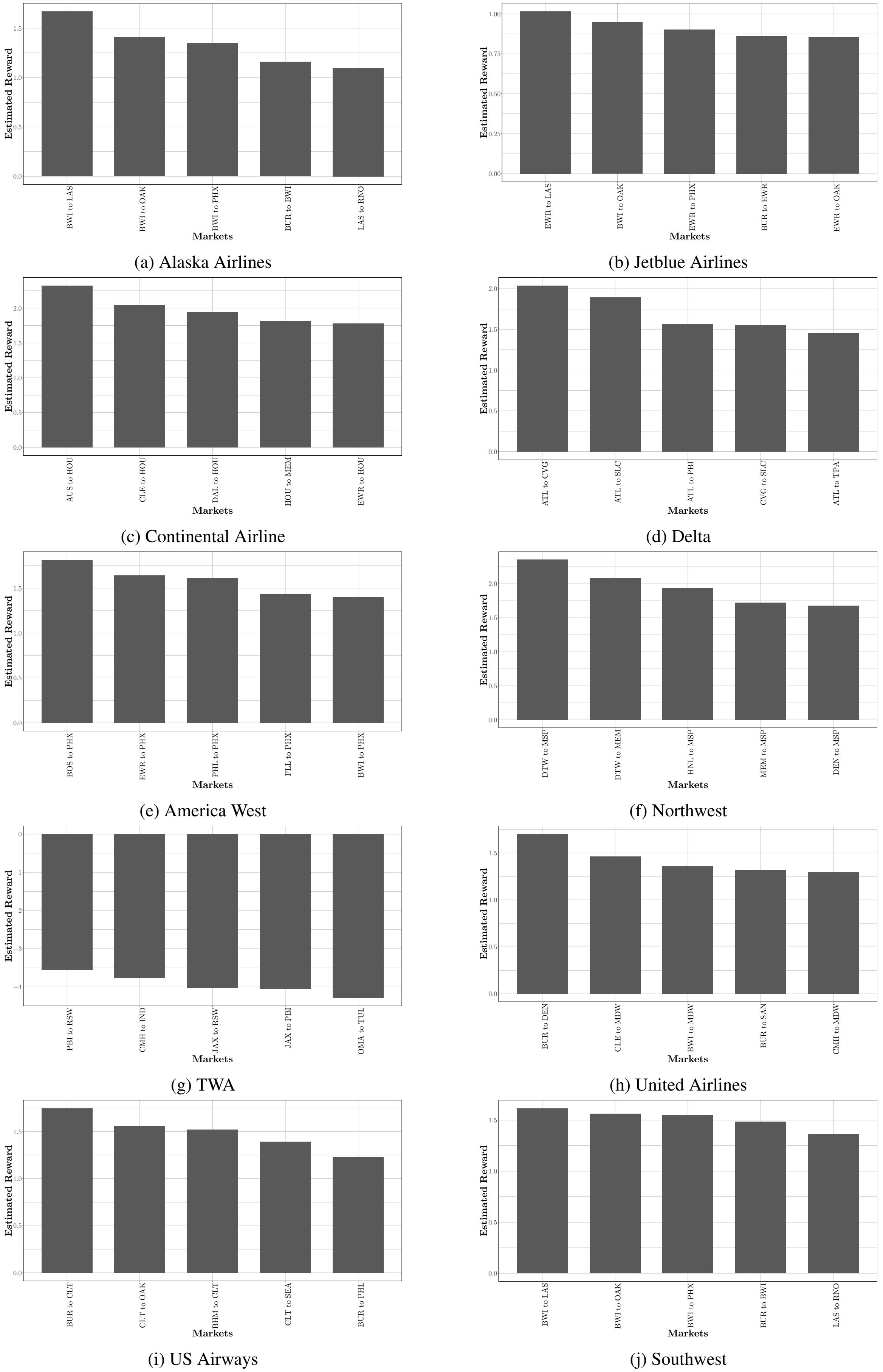
C.3. Details of Real-World Experiments

In this section, we provide detailed configurations of the PQR method when applied to the airline market entry analysis. To start with, we follow the convention in Benkard (2004); Berry and Jia (2010) and let $\alpha = 1$. As mentioned in Section 6.2, we only focus on the top 60 CSAs, summarized in Table 3, and 11 airline companies summarized in Table 4.

We take $\delta = 0.95$. Specifically, the hyperparameters used are provided in Table 5. We train the methods using the data from the start of 2013 to the end of 2014, and calculate the likelihood using the data of January 2015.

The state variables include the distance between CSAs, the populations of the origin and destination CSAs, the log of the current passenger density for the market, whether the carrier is already flying for the market, the number of nonstop competitors for the market, the share of flights operated out of the origin CSA and the destination CSA, whether the carrier operates any flights out of the origin or destination, and the number of flights operated by the carrier in the current period.

The estimated reward functions of all the 11 airline companies for the top 5 markets are provided in Figure 10.



(a) Alaska Airlines

(b) Jetblue Airlines

(c) Continental Airline

(d) Delta

(e) America West

(f) Northwest

(g) TWA

(h) United Airlines

(i) US Airways

(j) Southwest

Figure 10: Estimated reward functions for airline companies for top 5 markets.

CSAs	Airport Codes	CSAs	Airport Codes
Albuquerque	ABQ	Orlando	MCO
Albany	ALB	Chicago	ORD, MDW
Anchorage	ANC	Memphis	MEM
Atlanta	ATL	Milwaukee	MKE
Austin	AUS	Minneapolis-St.Paul	MSP
Hartford	BDL	NewOrleans	MSY
Birmingham	BHM	SanFrancisco	SFO, SJC, OAK
Nashville	BNA	Kahului	OGG
Boise	BOI	OklahomaCity	OKC
Boston	BOS, MHT, PVD	Omaha	OMA
Buffalo	BUF	Norfolk	ORF
LosAngeles	LAX, ONT, SNA, BUR	PalmBeach	PBI
WashingtonDC	IAD, DCA, BWI	Portland	PDX
Cleveland	CLE	Philadelphia	PHL
Charlotte	CLT	Phoenix	PHX
Columbus	CMH	Pittsburgh	PIT
Cincinnati	CVG	Raleigh-Durham	RDU
Dallas	DFW, DAL	Reno	RNO
Denver	DEN	Southwest Florida	RSW
Detroit	DTW	SanDiego	SAN
ElPaso	ELP	SanAntonio	SAT
New York	JFK, LGA, EWR	Louisville	SDF
Miami	MIA, FLL	Seattle	SEA
Spokane	GEG	SanJuan	SJU
Honolulu	HNL	SaltLakeCity	SLC
Houston	IAH, HOU	Sacramento	SMF
Indianapolis	IND	St.Louis	STL
Jacksonville	JAX	Tampa	TPA
LasVegas	LAS	Tulsa	TUL
KansasCity	MCI	Tuscon	TUS

Table 3: Top 60 CSAs with the most itineraries in 2002.

Airline Companies	Airline Company Codes
American Airlines	AA
Alaska Airlines	AS
Jetblue Airlines	B6
Continental Airlines	CO
Delta	DL
America West	HP
Northwest	NW
TWA	TW
United Airlines	UA
US Airways	US
Southwest	WN

Table 4: List of airline companies and their codes.

Types of Networks	Optimization Method	Learning Rate	Number of Steps
Policy Estimation	Adagrad	1.5×10^{-4}	1000
Q Estimation	Adagrad	0.5×10^{-4}	500
Reward Estimation	Adagrad	1×10^{-3}	1000

Table 5: Hyperparameters for PQR for the airline market entry analysis.