A Data-Driven State Aggregation Approach for Dynamic Discrete-Choice Models (Supplementary Material)

Sinong Geng[1] **Houssam Nassif**[2] **Carlos A. Manzanares**[3]

[1]Computer Science Department, Princeton University, Princeton, NJ, USA
[2]Meta, Seattle, WA, USA
[3]Amazon, Seattle, WA, USA

1 DYNAMIC DISCRETE CHOICE MODELS IN THEIR ORIGINAL FORMULATION

In this section, we formulate dynamic discrete choice models (DDMs) using the original formulation [Rust, 1987], and discuss its connection with the IRL formulation in Section 2.1. Note that the setup in this section is an alternative to the IRL formulation which our main results are based on and just is provided for completeness and comparison. SamQ does not require the assumptions listed in this section.

1.1 MODEL

Agents choose actions according to a Markov decision process described by the tuple $\{\{\mathcal{S}, \mathcal{E}\}, \mathcal{A}, r, \gamma, P\}$, where

- $\{\mathcal{S}, \mathcal{E}\}$ denotes the space of state variables;
- $\mathcal{A}$ represents a set of n_a actions;
- r represents an agent utility function;
- $\gamma \in [0, 1)$ is a discount factor;
- P represents the transition distribution.

At time t, agents observe state S_t taking values in $\mathcal{S}$, and ϵ_t taking values in $\mathcal{E}$ to make decisions. While S_t is observable to researchers, ϵ_t is observable to agents but not to researchers. The action is defined as a $n_a \times 1$ indicator vector, A_t, satisfying

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

In other words, at each time point, agents make a distinct choice over n_a possible actions. Meanwhile, ϵ_t is also a $n_a \times 1$ representing the potential shock of taking a choice.

The agent's control problem has the following value function:

$$V(s, \epsilon) = \max_{\{a_t\}_{t=0}^{\infty}} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(S_t, \epsilon_t, A_t) \mid s, \epsilon\right], \tag{1}$$

where the expectation is taken over realizations of ϵ_t, as well as transitions of S_t and ϵ_t as dictated by P. The utility function $r(s_t, \epsilon_t, a_t)$ can be further decomposed into

$$r(s_t, \epsilon_t, a_t) = u(s_t, a_t) + a_t^\top \epsilon_t,$$

where u represents the deterministic part of the utility function. Agents, but not researchers, observe ϵ_t before making a choice in each time period.

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

1.2 ASSUMPTIONS AND DEFINITIONS

We study DDMs under the following common assumptions.

Assumption 1. *The transition from S_t to S_{t+1} is independent of ϵ_t*

$$\mathrm{P}(S_{t+1} \mid S_t, \epsilon_t, A_t) = \mathrm{P}(S_{t+1} \mid S_t, A_t).$$

Assumption 2. *The random shocks ϵ_t at each time point are independent and identically distributed (IID) according to a type-I extreme value distribution.*

Assumption 1 ensures that unobservable state variables do not influence state transitions. This assumption is common, since it drastically simplifies the task of identifying the impact of changes in observable versus unobservable state variables. In our setting, Assumption 2 is convenient but not necessary, and $\boldsymbol{\epsilon}_t$ could follow other parametric distributions. As pointed out by Arcidiacono and Ellickson [2011], Assumptions 1 and 2 are nearly standard for applications of dynamic discrete choice models. Such a formulation is proved to be equivalent to the IRL formulation in Section 2.1 by Geng et al. [2020], Fu et al. [2018], Ermon et al. [2015].

2 PROOF OF THEOREM 1

Proof. By definition of L and $\tilde{L}$, we can derive

$$\begin{aligned} L(\mathbb{D};\theta^*) - \tilde{L}(\mathbb{D};\theta^*) =& \frac{1}{T}\sum_{(s,a)\in\mathbb{D}} \Bigg[Q^{\theta^*}(s,a) - \tilde{Q}^{\theta^*}(\Pi(s),a) \\ &+ \log\Bigg(\sum_{a'\in\mathcal{A}} \exp(\tilde{Q}^{\theta}(\Pi(s),a'))\Bigg) - \log\Bigg(\sum_{a'\in\mathcal{A}} \exp(Q^{\theta}(s,a'))\Bigg)\Bigg] \\ \le& \frac{1}{T}\sum_{(s,a)\in\mathbb{D}} \Bigg[\Big|Q^{\theta^*}(s,a) - \tilde{Q}^{\theta^*}(\Pi(s),a)\Big| + \max_{a'\in\mathcal{A}}\Big|Q^{\theta^*}(s,a') - \tilde{Q}^{\theta^*}(\Pi(s),a')\Big|\Bigg] \\ \le& 2\max_{a'\in\mathcal{A}}\Big|Q^{\theta^*}(s,a') - \tilde{Q}^{\theta^*}(\Pi(s),a')\Big|, \end{aligned} \tag{2}$$

where the first inequality is due to the fact that the log sum exp function is Lipschitz continuous with constant 1. Then, we take f in Lemma 2 as $Q^{\theta^*}(s,a)$, and derive

$$\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\Big|Q^{\theta^*}(s,a) - \tilde{Q}^{\theta^*}(\Pi(s),a)\Big| \le \frac{2}{1-\gamma}\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\Big|Q^{\theta^*}(s,a) - Q^{\theta^*}(\Pi(s),a)\Big|. \tag{3}$$

By taking (3) to (2),

$$L(\mathbb{D};\theta^*) - \tilde{L}(\mathbb{D};\theta^*) \le \frac{4}{1-\gamma}\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\Big|Q^{\theta^*}(s,a) - Q^{\theta^*}(\Pi(s),a)\Big|.$$

Finally, by Lemma 1

$$\epsilon_{asy} \le \frac{4}{c_H(1-\gamma)}\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\Big|Q^{\theta^*}(s,a) - Q^{\theta^*}(\Pi(s),a)\Big| = \epsilon_Q,$$

which finishes the proof. □

Lemma 1. *Under Assumption 1 and Assumption 2,*

$$\Big\|\tilde{\theta} - \theta^*\Big\|^2 \le \frac{E[L(\mathbb{D};\theta^*) - \tilde{L}(\mathbb{D};\theta^*)]}{c_H}.$$

Proof. By the definition of $\tilde{\theta}$,

$$0 \le \mathbb{E}[\tilde{L}(\mathbb{D};\tilde{\theta}) - \tilde{L}(\mathbb{D};\theta^*)] \le \mathbb{E}[L(\mathbb{D};\theta^*) - \tilde{L}(\mathbb{D};\theta^*)]. \tag{4}$$

Further, by Taylor expansion, we have

$$\mathbb{E}[\tilde{L}(\mathbb{D};\tilde{\theta}) - \tilde{L}(\mathbb{D};\theta^*)] = (\tilde{\theta}-\theta^*)^\top \mathbb{E}\left[-\frac{\partial^2 \tilde{L}(\mathbb{D};\bar{\theta})}{\partial \theta^2}\right](\tilde{\theta}-\theta^*),$$

where $\bar{\theta} = k\theta^* + (1-k)\tilde{\theta}$ with some $k \in [0,1]$. Note that the first order term is zero, since $\tilde{\theta}$ maximizes $\mathbb{E}[\tilde{L}(\mathbb{D},\theta)]$. By Assumption 1, we finish the proof.

$$\mathbb{E}[\tilde{L}(\mathbb{D};\tilde{\theta}) - \tilde{L}(\mathbb{D};\theta^*)] = (\tilde{\theta}-\theta^*)^\top \mathbb{E}\left[-\frac{\partial^2 \tilde{L}(\mathbb{D};\bar{\theta})}{\partial \theta^2}\right](\tilde{\theta}-\theta^*) \geq C_H \left\|\tilde{\theta}-\theta^*\right\|^2.$$

□

Lemma 2. *For any projection function Π defined in Section 3.1 and its aggregated Q function $\tilde{Q}$, the following inequality is true:*

$$\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left|Q^{\theta^*}(s,a) - \tilde{Q}^{\theta^*}(\Pi(s),a)\right| \leq \frac{2}{1-\gamma}\min_f \max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left|Q^{\theta^*}(s,a) - f(\Pi(s),a)\right|,$$

where $f(s,a) : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ is any function.

Proof. The proof follows Theorem 3 of Tsitsiklis and Van Roy [1996]. □

3 PROOF OF THEOREM 2

3.1 TECHNICAL LEMMAS FOR THEOREM 2

Lemma 3. *Given $\theta \in \Theta$, for any $\delta \in (0,1)$, we provide the following probabilistic bound for the estimated aggregated likelihood $\hat{L}$*

$$\mathrm{P}\Bigg(\left|\hat{L}(\mathbb{D};\theta) - \mathbb{E}[\tilde{L}(\mathbb{D};\theta)]\right| \leq \frac{2(R_{max}+1)}{1-\gamma}\sqrt{\frac{\log(\frac{4}{\delta})}{2N}} + \frac{R_{max}+1}{1-\gamma}\sqrt{\frac{\log(\frac{8|\tilde{\mathcal{S}}||\mathcal{A}|}{\delta})}{2N}}\frac{2}{C_{uni} - \sqrt{\frac{\log(\frac{4|\tilde{\mathcal{S}}||\mathcal{A}|}{\delta})}{2N}}}\Bigg) \geq 1-\delta,$$

where the expectation is over the sample $\mathbb{D}$.

Proof. By inserting $\tilde{L}(\mathbb{D};\theta)$, we have

$$\left|\hat{L}(\mathbb{D};\theta) - \mathbb{E}[\tilde{L}(\mathbb{D};\theta)]\right| \leq \left|\hat{L}(\mathbb{D};\theta) - \tilde{L}(\mathbb{D};\theta)\right| + \left|\tilde{L}(\mathbb{D};\theta) - \mathbb{E}[\tilde{L}(\mathbb{D};\theta)]\right|. \tag{5}$$

First term on the RHS of (5) To start with, we consider $\left|\hat{L}(\mathbb{D};\hat{\theta}) - \tilde{L}(\mathbb{D};\hat{\theta})\right|$. To this end, we aim to bound $\max_{(\tilde{s},a)\in\tilde{\mathcal{S}}\times\mathcal{A}}\left|\tilde{Q}^\theta(\tilde{s},a) - \hat{Q}^\theta(\tilde{s},a)\right|$. We insert $\hat{\mathcal{T}}(\tilde{Q}^\theta(\tilde{s},a))$:

$$\tilde{Q}^\theta(\tilde{s},a) - \hat{Q}^\theta(\tilde{s},a) = \tilde{\mathcal{T}}(\tilde{Q}^\theta(\tilde{s},a)) - \hat{\mathcal{T}}(\tilde{Q}^\theta(\tilde{s},a)) + \hat{\mathcal{T}}(\tilde{Q}^\theta(\tilde{s},a)) - \hat{\mathcal{T}}(\hat{Q}^\theta(\tilde{s},a)).$$

Since $\hat{\mathcal{T}}$ is a contraction with γ, we further derive

$$\left|\tilde{Q}^\theta(\tilde{s},a) - \hat{Q}^\theta(\tilde{s},a)\right| \leq \frac{\left|\tilde{\mathcal{T}}(\tilde{Q}^\theta(\tilde{s},a)) - \hat{\mathcal{T}}(\tilde{Q}^\theta(\tilde{s},a))\right|}{1-\gamma}. \tag{6}$$

By the definition of $\tilde{\mathcal{T}}$ and $\hat{\mathcal{T}}$, it can be seen that $\hat{\mathcal{T}}(\tilde{Q}^\theta(\tilde{s},a))$ is a sample average estimation to $\tilde{\mathcal{T}}(\tilde{Q}^\theta(\tilde{s},a))$. Therefore, we aim to bound the difference between the two by concentration inequalities. Specifically, by assumption 6 and Hoeffding's inequality, we have

$$\mathrm{P}\Big(\sum_{i=1,2,\cdots,N} \mathbb{1}_{\{\Pi(s_i)=\tilde{s},a_i=a\}} \geq NC_{uni} - \sqrt{-\frac{1}{2}N\log(\frac{\delta}{2})}\Big) \geq 1-\frac{\delta}{2}. \tag{7}$$

Further, conditional on the event $\Big\{\sum_{i=1,2,\cdots,N} \mathbb{1}_{\{\Pi(s_i)=\tilde{s},a_i=a\}} \geq NC_{uni} - \sqrt{-N\log(\frac{\delta}{2})}\Big\}$, by Hoeffding's inequality and Assumption 7, for any $(\tilde{s},a) \in \tilde{\mathcal{S}} \times \mathcal{A}$

$$\mathrm{P}\Bigg(\Big|\tilde{\mathcal{T}}(\tilde{Q}^\theta(\tilde{s},a)) - \hat{\mathcal{T}}(\tilde{Q}^\theta(\tilde{s},a))\Big| \leq \frac{R_{max}+1}{1-\gamma}\sqrt{\frac{\log(\frac{4}{\delta})}{2N}}\frac{1}{C_{uni}-\sqrt{\frac{\log(\frac{2}{\delta})}{2N}}}\Bigg) \geq 1-\frac{\delta}{2}. \tag{8}$$

Combining (7) and (8), for a given $(\tilde{s},a) \in \tilde{\mathcal{S}} \times \mathcal{A}$, for any $\delta \in (0,1)$

$$\mathrm{P}\Bigg(\Big|\tilde{\mathcal{T}}(\tilde{Q}^\theta(\tilde{s},a)) - \hat{\mathcal{T}}(\tilde{Q}^\theta(\tilde{s},a))\Big| \leq \frac{R_{max}+1}{1-\gamma}\sqrt{\frac{\log(\frac{4}{\delta})}{2N}}\frac{1}{C_{uni}-\sqrt{\frac{\log(\frac{2}{\delta})}{2N}}}\Bigg) \geq 1-\delta.$$

Next, by union bound again, we can extend the results to any $(\tilde{s},a) \in \tilde{\mathcal{S}} \times \mathcal{A}$

$$\mathrm{P}\Bigg(\max_{\tilde{s}\in\tilde{\mathcal{S}},a\in\mathcal{A}}\Big|\tilde{\mathcal{T}}(\tilde{Q}^\theta(\tilde{s},a)) - \hat{\mathcal{T}}(\tilde{Q}^\theta(\tilde{s},a))\Big| \leq \frac{R_{max}+1}{1-\gamma}\sqrt{\frac{\log(\frac{4|\tilde{\mathcal{S}}||\mathcal{A}|}{\delta})}{2N}}\frac{1}{C_{uni}-\sqrt{\frac{\log(\frac{2|\tilde{\mathcal{S}}||\mathcal{A}|}{\delta})}{2N}}}\Bigg) \geq 1-\delta. \tag{9}$$

Combined with (6), we derive:

$$\mathrm{P}\Bigg(\max_{(\tilde{s},a)\in\tilde{\mathcal{S}}\times\mathcal{A}}\Big|\tilde{Q}^\theta(\tilde{s},a) - \hat{Q}^\theta(\tilde{s},a)\Big| \leq \frac{R_{max}+1}{(1-\gamma)^2}\sqrt{\frac{\log(\frac{4|\tilde{\mathcal{S}}||\mathcal{A}|}{\delta})}{2N}}\frac{1}{C_{uni}-\sqrt{\frac{\log(\frac{2|\tilde{\mathcal{S}}||\mathcal{A}|}{\delta})}{2N}}}\Bigg) \geq 1-\delta.$$

By the definition of $\tilde{L}$ in (6) and (2), we have

$$\mathrm{P}\Bigg(\Big|\tilde{L}(\mathbb{D};\theta) - \hat{L}(\mathbb{D};\theta)\Big| \leq \frac{R_{max}+1}{(1-\gamma)^2}\sqrt{\frac{\log(\frac{4|\tilde{\mathcal{S}}||\mathcal{A}|}{\delta})}{2N}}\frac{2}{C_{uni}-\sqrt{\frac{\log(\frac{2|\tilde{\mathcal{S}}||\mathcal{A}|}{\delta})}{2N}}}\Bigg) \geq 1-\delta.$$

Second term on the RHS of (5) Now, we consider $\Big|\tilde{L}(\mathbb{D};\theta) - \mathbb{E}[\tilde{L}(\mathbb{D};\theta)]\Big|$. By (2) and Assumption 7, $\tilde{L}(\mathbb{D};\hat{\theta})$ is bounded by $\frac{2(R_{max}+1)}{1-\gamma}$. Thus, by Hoeffding's inequality, for any $\delta \in (0,1)$

$$\mathrm{P}\Bigg(\Big|\mathbb{E}[\tilde{L}(\mathbb{D};\hat{\theta})] - \tilde{L}(\mathbb{D};\hat{\theta})\Big| \leq \frac{2(R_{max}+1)}{1-\gamma}\sqrt{\frac{\log(\frac{2}{\delta})}{2N}}\Bigg) \geq 1-\delta.$$

Therefore, by union bound, (5) can be bounded by

$$\begin{aligned}\mathrm{P}\Bigg(\Big|\hat{L}(\mathbb{D};\theta) - \mathbb{E}[\tilde{L}(\mathbb{D};\theta)]\Big| \leq & \frac{2(R_{max}+1)}{1-\gamma}\sqrt{\frac{\log(\frac{4}{\delta})}{2N}} \\ & + \frac{R_{max}+1}{(1-\gamma)^2}\sqrt{\frac{\log(\frac{8|\tilde{\mathcal{S}}||\mathcal{A}|}{\delta})}{2N}}\frac{2}{C_{uni}-\sqrt{\frac{\log(\frac{4|\tilde{\mathcal{S}}||\mathcal{A}|}{\delta})}{2N}}}\Bigg) \geq 1-\delta.\end{aligned}$$

□

Lemma 4. *Let* $\tilde{\theta}^{\hat{\Pi}} := \arg\max_{\theta\in\Theta} \mathbb{E}[\tilde{L}(\mathbb{D};\theta,\hat{\Pi})]$. *Then,*

$$\left\|\theta^* - \tilde{\theta}^{\hat{\Pi}}\right\| \leq \frac{4}{C_H(1-\gamma)}\left(\frac{R_{\max}+1}{1-\gamma}\frac{4}{n_s^{\frac{1}{n_a}}-1} + 2\epsilon_Q + \epsilon_c\right).$$

Proof. A Euclidean ball of radius R in $\mathbb{R}^{n_a}$ can be covered by $\left(\frac{4R+\delta}{\delta}\right)^{n_a}$ balls of radius δ (see Lemma 2.5 of Van de Geer and van de Geer [2000]). Therefore, with n_s states after aggregation, by Assumption 4,

$$\hat{\epsilon}(\Pi^*) \leq \frac{R_{\max}+1}{1-\gamma}\frac{4}{n_s^{\frac{1}{n_a}}-1}.$$

Further by Assumption 4 and Assumption 5,

$$\epsilon(\hat{\Pi}) \leq \hat{\epsilon}(\Pi^*) + 2\epsilon_Q + \epsilon_c \leq \frac{R_{\max}+1}{1-\gamma}\frac{4}{n_s^{\frac{1}{n_a}}-1} + 2\epsilon_Q + \epsilon_c.$$

Therefore, by Theorem 1

$$\left\|\theta^* - \tilde{\theta}^{\hat{\Pi}}\right\| \leq \frac{4}{C_H(1-\gamma)}\left(\frac{R_{\max}+1}{1-\gamma}\frac{4}{n_s^{\frac{1}{n_a}}-1} + 2\epsilon_Q + \epsilon_c\right).$$

□

3.2 PROOF

We first aim to bound $\mathbb{E}[\tilde{L}(\mathbb{D};\tilde{\theta}^{\hat{\Pi}}) - \tilde{L}(\mathbb{D};\hat{\theta})]$, where the expectation is over $\mathbb{D}$ only instead of $\hat{\theta}$. To this end, we insert $\hat{L}(\mathbb{D};\tilde{\theta}^{\hat{\Pi}})$ and $\hat{L}(\mathbb{D};\hat{\theta})$:

$$\begin{aligned}\mathbb{E}[\tilde{L}(\mathbb{D};\tilde{\theta}^{\hat{\Pi}}) - \tilde{L}(\mathbb{D};\hat{\theta})] \leq & \mathbb{E}[\tilde{L}(\mathbb{D};\tilde{\theta}^{\hat{\Pi}}) - \hat{L}(\mathbb{D};\tilde{\theta}^{\hat{\Pi}})] + \hat{L}(\mathbb{D};\tilde{\theta}^{\hat{\Pi}}) - \hat{L}(\mathbb{D};\hat{\theta}) + \hat{L}(\mathbb{D};\hat{\theta}) - \mathbb{E}[\tilde{L}(\mathbb{D};\hat{\theta})] \\ \leq & \left|\mathbb{E}[\tilde{L}(\mathbb{D};\tilde{\theta}^{\hat{\Pi}}) - \hat{L}(\mathbb{D};\tilde{\theta}^{\hat{\Pi}})]\right| + \left|\hat{L}(\mathbb{D};\hat{\theta}) - \mathbb{E}[\tilde{L}(\mathbb{D};\hat{\theta})]\right|.\end{aligned}$$

By Lemma 3 and the union bound,

$$\begin{aligned}\mathrm{P}\Bigg(\max_{\theta\in\Theta}\left|\hat{L}(\mathbb{D};\theta) - \mathbb{E}[\tilde{L}(\mathbb{D};\theta)]\right| \leq & \frac{2(R_{max}+1)}{1-\gamma}\sqrt{\frac{\log(\frac{4|\Theta|}{\delta})}{2N}} \\ & + \frac{R_{max}+1}{(1-\gamma)^2}\sqrt{\frac{\log(\frac{8|\tilde{\mathcal{S}}||\mathcal{A}||\Theta|}{\delta})}{2N}}\frac{2}{C_{uni} - \sqrt{\frac{\log(\frac{4|\tilde{\mathcal{S}}||\mathcal{A}||\Theta|}{\delta})}{2N}}}\Bigg) \geq 1-\delta.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathrm{P}\Bigg(\mathbb{E}[\tilde{L}(\mathbb{D};\tilde{\theta}^{\hat{\Pi}}) - \tilde{L}(\mathbb{D};\hat{\theta})] \leq & \frac{4(R_{max}+1)}{1-\gamma}\sqrt{\frac{\log(\frac{4|\Theta|}{\delta})}{2N}} \\ & + \frac{R_{max}+1}{(1-\gamma)^2}\sqrt{\frac{\log(\frac{8|\tilde{\mathcal{S}}||\mathcal{A}||\Theta|}{\delta})}{2N}}\frac{4}{C_{uni} - \sqrt{\frac{\log(\frac{4|\tilde{\mathcal{S}}||\mathcal{A}||\Theta|}{\delta})}{2N}}}\Bigg) \geq 1-\delta.\end{aligned}$$

By Assumption 1 and a similar analysis as Lemma 1,

$$\begin{aligned}\mathrm{P}\Bigg(\left|\hat{\theta} - \tilde{\theta}^{\hat{\Pi}}\right| \leq & \frac{4(R_{max}+1)}{(1-\gamma)C_H}\sqrt{\frac{\log(\frac{4|\Theta|}{\delta})}{2N}} \\ & + \frac{R_{max}+1}{(1-\gamma)^2 C_H}\sqrt{\frac{\log(\frac{8|\tilde{\mathcal{S}}||\mathcal{A}||\Theta|}{\delta})}{2N}}\frac{4}{C_{uni} - \sqrt{\frac{\log(\frac{4|\tilde{\mathcal{S}}||\mathcal{A}||\Theta|}{\delta})}{2N}}}\Bigg) \geq 1-\delta.\end{aligned}$$

Combined with Lemma 4,

$$\mathrm{P}\Bigg(\Big|\hat{\theta}-\theta^*\Big| \leq \frac{4}{C_H(1-\gamma)}\Bigg(\frac{R_{\max}+1}{1-\gamma}\frac{4}{n_s^{\frac{1}{n_a}}-1}+2\epsilon_Q+\epsilon_c\Bigg)+\frac{4(R_{max}+1)}{(1-\gamma)C_H}\sqrt{\frac{\log(\frac{4|\Theta|}{\delta})}{2N}}$$

$$+\frac{R_{max}+1}{(1-\gamma)^2C_H}\sqrt{\frac{\log(\frac{8n_sn_a|\Theta|}{\delta})}{2N}}\frac{4}{C_{uni}-\sqrt{\frac{\log(\frac{4n_sn_a|\Theta|}{\delta})}{2N}}}\Bigg) \geq 1-\delta.$$

References

Peter Arcidiacono and Paul B Ellickson. Practical methods for estimation of dynamic discrete choice models. *Annual Review of Economics*, 3(1):363–394, 2011.

Stefano Ermon, Yexiang Xue, Russell Toth, Bistra Dilkina, Richard Bernstein, Theodoros Damoulas, Patrick Clark, Steve DeGloria, Andrew Mude, Christopher Barrett, and Carla P. Gomes. Learning large-scale dynamic discrete choice models of spatio-temporal preferences with application to migratory pastoralism in east africa. In *AAAI Conference on Artificial Intelligence*, pages 644–650, 2015.

Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adverserial inverse reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2018.

Sinong Geng, Houssam Nassif, Carlos Manzanares, Max Reppen, and Ronnie Sircar. Deep PQR: Solving inverse reinforcement learning using anchor actions. In *International Conference on Machine Learning*, pages 3431–3441, 2020.

John Rust. Optimal replacement of gmc bus engines: An empirical model of harold zurcher. *Econometrica: Journal of the Econometric Society*, pages 999–1033, 1987.

John N Tsitsiklis and Benjamin Van Roy. Feature-based methods for large scale dynamic programming. *Machine Learning*, 22(1):59–94, 1996.

Sara A Van de Geer and Sara van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.