On the Estimation of Linear Softmax Parametrized Markov Chains

[KCC 2024 Oral Session]

Kunwoo Na[1,2,3], Junghyun Lee[3], Seyoung Yun[3]

[1]Department of Chemistry Education, SNU [2]Department of Mathematical Sciences, SNU
[2]Kim Jaechul Graduate School of AI, KAIST

June 27, 2024

Outline

Introduction

Theoretical Analyses
- Expressivity results
- Non-identifiability results

Experiments

Outline

Introduction

Theoretical Analyses
- Expressivity results
- Non-identifiability results

Experiments

- Softmax parametrization is highly ubiquitous when one wishes to estimate discrete probability distribution.
- Due to its simplicity, it is employed in a wide range of applications such as multinomial logistic Markov Decision Processes [HO23], deep learning [S^+21], and human decision-making [RL15].
- In this work, we compare three distinct choices of softmax-type parametrization of a **transition probability distribution**.

Problem Setup

- Given a finite set $\mathcal{S}$ with $|\mathcal{S}| = N$, let $P \in \Delta(\mathcal{S})$ where $\Delta(\mathcal{S})$ denote the set of all transition probability distributions over $\mathcal{S}$.
- For example, if $\mathcal{S} = \{s_1, \cdots, s_N\}$, $P(\cdot \mid s_i)$ is the probability distribution over $\mathcal{S}$ given that the current state is s_i.
- The transition probability distribution P can be canonically identified as an element of $\mathrm{Mat}_{N\times N}(\mathbb{R})$.
- We analyze three softmax-type parametrizations of P that exploit $\mathrm{softmax} : \mathbb{R}^N \to \mathbb{R}$ to generate probability distributions $P(\cdot \mid s)$ for each $s \in \mathcal{S}$.

Softmax-type parametrizations

In this work, we provide theoretical and empirical analyses on three popular ways to estimate P, which are summarized below.

1. $p(s' \mid s) = \mathrm{softmax}(\{\varphi(s)^\intercal \theta_\star(s')\}_{s'})$, where $\varphi : \mathcal{S} \to \mathbb{R}^d$ is *known* and $\theta_\star : \mathcal{S} \to \mathbb{R}^d$ is unknown.
2. $p(s' \mid s) = \mathrm{softmax}(\{\varphi(s, s')^\intercal \theta_\star\}_{s'})$, where $\varphi : \mathcal{S} \times \mathcal{S} \to \mathbb{R}^d$ is *known* and $\theta_\star \in \mathbb{R}^d$ is unknown.
3. $p(s' \mid s) = \mathrm{softmax}(\{\varphi(s, s')^\intercal \theta_\star\}_{s'})$, where $\varphi : \mathcal{S} \times \mathcal{S} \to \mathbb{R}^d$ is unknown and $\theta_\star \in \mathbb{R}^d$ is also unknown.

In any case, we are given a trajectory $(X_1, \cdots, X_T)$ of length T, where

$$X_1 \sim \mu, \ X_{t+1} \sim p(\cdot \mid X_t), \quad t = 1, 2, \cdots, T-1.$$

Here, μ is some unknown probability distribution over $\mathcal{S}$.

We consider the performance of two MLE's:

- **Non-parametric model**:

$$\widehat{p}_{\text{nonparam}}(s' \mid s) = \frac{\#[s \to s']}{\#[s]}, \quad \forall s, s' \in \mathcal{S}.$$

  This is known to be minimax over ergodic Markov chains [WK21].

- **Parametric model**:

$$\widehat{p} := p_{\widehat{\theta}_T}, \quad \widehat{\theta}_T = \operatorname*{argmax}_{\theta \in \mathbb{R}^d} \sum_{t=1}^{T} \Big\{ \log p_\theta(X_{t+1} \mid X_t) \Big\}.$$

  One reasonable expectation is that when d is small, the latter MLE may be able to break the barrier of the minimax rate.

Outline

Introduction

Theoretical Analyses
Expressivity results
Non-identifiability results

Experiments

We say that the parametrization scheme is **fully expressive** if every Markov chain can be expressed as that scheme.

Theorem 2.1 (Informal Version)

The **parametrizations #1, #2, #3** are fully expressive if there are no irrelevent or redundant features.

The formal statement can be organized as

Param #1 is fully expressive when:	**Param #2, #3** are fully expressive when:
the linear equation $L_\Phi x = y$ has solution for $\forall y \in \mathbb{R}^d$	the linear equation $L_\Psi x = y$ has solution for $\forall y \in \mathbb{R}^d$

where we define

$$\Phi = \begin{bmatrix} \varphi(s_1)^\intercal \\ \vdots \\ \varphi(s_N)^\intercal \end{bmatrix} \in \mathrm{Mat}_{N,d}(\mathbb{R}), \quad \Psi = \begin{bmatrix} \Phi(s_1) \\ \hline \vdots \\ \hline \Phi(s_N) \end{bmatrix} \in \mathrm{Mat}_{N^2,d}(\mathbb{R})$$

for parametrization #1 and {#2, #3}, respectively. Here, for parametrizations #2 and #3, $\Phi(s)$ is defined by

$$\Phi(s) = [\varphi(s, s_1)^\intercal \cdots \varphi(s, s_N)^\intercal]^\intercal \in \mathrm{Mat}_{N,d}(\mathbb{R}).$$

- An accurate estimate of $\theta_\star$ yields an accurate estimate of $p_{\theta_\star}$.
- An inaccurate estimate $\widehat{\theta}$ of $\theta_\star$ might still yield a good estimate of $p_{\theta_\star}$, due to the translation invariance of softmax. [**Non-identifiability**]

Theorem 2.2 (Accurate $\theta \Rightarrow$ Accurate p_θ; Parametrization #1)

Assume that the true transition probability distribution has representation $p_{\theta_\star}(s'|s) = \mathrm{softmax}\{(\varphi(s)^\intercal \theta_\star(s')\}_{s'})$, and consider the parametrization $p_\theta(s'|s) = \mathrm{softmax}(\{\varphi(s)^\intercal \theta(s')\}_{s'})$. Then, one has that

$$\|p_\theta - p_{\theta_\star}\|_{\infty,1} := \max_{s \in \mathcal{S}} d_{\mathrm{TV}}\left(p_\theta(\cdot|s), p_{\theta_\star}(\cdot|s)\right) \lesssim \frac{N}{2}\|\theta - \theta_\star\|_{\infty,2}.$$

Theorem 2.3 (Accurate $\theta \Rightarrow$ Accurate p_θ; Parametrization #2)

Consider the parametrization $p_\theta(s'|s) = \mathrm{softmax}(\{\varphi(s,s')^\intercal \theta\}_{s'})$, where $\varphi : \mathcal{S} \times \mathcal{S} \to \mathbb{R}^d$ is known. Assume that the true transition probability distribution has representation $p_{\theta_\star}(s'|s) = \mathrm{softmax}(\varphi(s,s')^\intercal \theta_\star)$. Then, one has that

$$\|p_\theta - p_{\theta_\star}\|_{\infty,1} \lesssim \frac{1}{2}\|\theta - \theta_\star\|_2.$$

Proposition 1 (Non-identifiability, Parametrization #1)
If $\mathbf{1}_{\mathbb{R}^d} \in \operatorname{Ran} L_\Phi$, then for any $\varepsilon > 0$, there exists a $\tilde{\theta}_\star : \mathcal{S} \to \mathbb{R}^d$ such that $p_{\theta_\star} = p_{\tilde{\theta}_\star}$, yet $\|\theta_\star - \tilde{\theta}_\star\|_{\infty,2} \geqslant \varepsilon$.

Proposition 2 (Non-identiability, Parametrization #2 & #3)
If $\mathbf{1} = \mathbf{1}_{\mathbb{R}^d} \in \operatorname{Ran} L_\Psi$, then for any given $\varepsilon > 0$, there exists some $\tilde{\theta}_\star$ such that $p_{\theta_\star} = p_{\tilde{\theta}_\star}$, yet $\|\theta_\star - \tilde{\theta}_\star\|_2 \geqslant \varepsilon$.

Outline

Introduction

Theoretical Analyses

Expressivity results

Non-identifiability results

Experiments

Setup

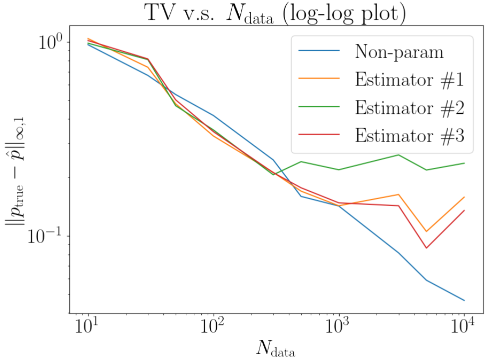
- We consider a Markov chain $\mathbb{M} = (\mathcal{S}, \mu, P)$ with $N = 10$ states and (randomly generated) fixed μ and P.
- We consider the non-parametric estimator and three distinct parametric estimators.
- For each parametric estimators, we perform the maximum likelihood estimator w.r.t. θ: precisely speaking,

$$\underset{\theta \in \mathbb{R}^d}{\text{maximize}} \quad \sum_{t=1}^{T} \log p_\theta(X_{t+1} \mid X_t)$$

  via gradient ascent on θ with the learning rate of 0.003.

Experiment #1

We vary the number of data points N_{data} over a set of values: $N_{\text{data}} \in \{10, 30, 100, 300, 1000, 3000, 10000, 30000\}$, and observe the decay rate of the metric $\|p_\theta - P\|_{\infty,1}$.



- For $N_{\text{data}} \leqslant 10^3$, we observe the slope of -1/2 on the log-log plot, indicating a decay rate of $\mathcal{O}(N_{\text{data}}^{-1/2})$.
- As N_{data} increases, the absolute value of the slope for parametric estimators decreases, indicating improved performance compared to the non-parametric estimator.

Experiment #2

In this experiment, we observe the decay rate of the discrepancy metric $\|p_\theta - P\|_{\infty,1}$ over the number of epochs.

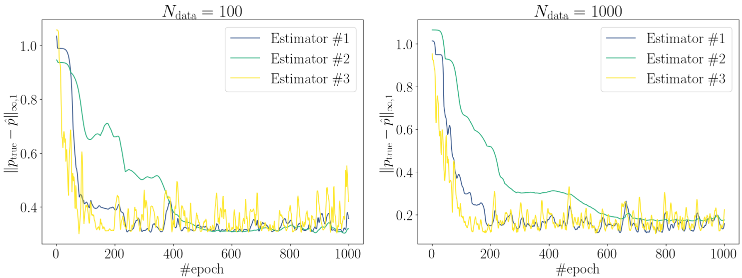


Figure: (Left) training curve for $N_{\text{data}} = 100$, (Right) training curve for $N_{\text{data}} = 1000$

- In both figures, the first and third estimators demonstrate superior performance compared to the second estimator, while the second estimator exhibits greater robustness.

Future work

- Theoretically exploring the decay rate of this metric with respect to N_{data}?
- Understand the observed decay in the absolute slope of the first figure as N_{data} increases?

Bibliography I

[HO23] Taehyun Hwang and Min-hwan Oh. Model-Based Reinforcement Learning with Multinomial Logistic Function Approximation. In *AAAI*, 2023.

[RL15] Paul Reverdy and Naomi Ehrich Leonard. Parameter estimation in softmax decision-making models with linear objective functions. *IEEE Transactions on Automation Science and Engineering*, 13(1):54–67, 2015.

[S+21] M. Seddik et al. The Unexpected Deterministic and Universal Behavior of Large Softmax Classifiers. In *AISTATS*, 2021.

[WK21] Geoffrey Wolfer and Aryeh Kontorovich. Statistical estimation of ergodic Markov chain kernel over discrete state space. *Bernoulli*, 27(1):532 – 553, 2021.