On the Estimation of Linear Softmax Parametrized Probability Distributions

KSC 2023 Oral Session #13

Murad Aghazada[1⋆], Mohammad Benabbasi[2⋆], Junghyun Lee[3], Se-Young Yun[3]

February 21, 2024

[1] School of Computing, KAIST
[2] Département d'informatique, Université de Sherbrooke
[3] Kim Jaechul Graduate School of AI, KAIST

KAIST AI
Kim Jaechul Graduate School

Optimization and
Statistical Inference LAB

Motivation

- Importance of **Softmax Parametrization:**
  - Conversion of raw scores to normalized probability distributions
  - Crucial role in designing effective algorithms due to probabilistic interpretations
- Application Domains:
  - Reinforcement Learning (RL) - Multinomial logistic MDP [Hwang and Oh, 2023]
  - Human decision-making [Reverdy and Leonard, 2016]
  - Deep neural networks [Seddik et al., 2021]
  - Statistical ranking theory - Bradley-Terry model [Bradley and Terry, 1952], Plackett-Luce model [Plackett, 1975]

Outline

1. Problem Settings

Linear Softmax Parametrization (LSP)

Given a state space $\mathcal{S}$ with $|\mathcal{S}| = S$, a discrete probability distribution p over $\mathcal{S}$ is said to have a **linear softmax parametrization** if there exists a $\boldsymbol{\theta}^\star \in \mathbb{R}^d$ such that

$$p^\star(s) \triangleq p_{\boldsymbol{\theta}^\star}(s) := \frac{\exp(\varphi(s)^\intercal \boldsymbol{\theta}^\star)}{\sum_{\tilde{s} \in \mathcal{S}} \exp\left(\varphi(\tilde{s})^\intercal \boldsymbol{\theta}^\star\right)}$$

- d is the dimension of latent space
- $\varphi : \mathcal{S} \to \mathbb{R}^d$ is feature mapping function, **fixed and known to the learner**

From a deep learning perspective, φ is the neural features outputted from the body.

Learner's Goal

- Given an offline dataset $\mathcal{D} = \{s_1, s_2, \cdots, s_N\}$ with $s_i \overset{i.i.d.}{\sim} p^\star$, the learner's goal is to obtain an accurate estimate of $p^\star$, which we now denote as $\widehat{p}$.
- The *total variation (TV) distance* is used to measure quality of estimation:

$$d_{TV}(p_1, p_2) := \frac{1}{2} \sum_{s \in \mathcal{S}} |p_1(s) - p_2(s)|, \quad p_1, p_2 \in \mathcal{P}(\mathcal{S}).$$

    - $\mathcal{P}(\mathcal{S})$ is the set of distributions whose support is $\mathcal{S}$.

2. Theoretical Analyses of LSP

Expressivity

Theorem 1 Let $\mathcal{P}(\mathcal{S})$ be the set of distributions whose support is $\mathcal{S}$ and $\mathcal{P}(\varphi)$ be those with an LSP. Denote $\mathbf{\Phi} = [\varphi(s_1) \cdots \varphi(s_n)]^\intercal \in \mathbb{R}^{S \times d}$. Then $\mathcal{P}(\varphi) = \mathcal{P}(\mathcal{S})$ if and only if $\mathbf{\Phi}$ has linearly independent columns or $\mathrm{col}(\mathbf{\Phi}) = \mathbb{R}^d$.

- **Theorem 1** states that with reasonable condition on $\mathbf{\Phi}$, the set of distributions with a LSP is **maximally expressive**!
- In other words, any "nonparametric" distribution estimation can be converted to a "parametric" distribution estimation by utilizing an appropriate feature matrix via **LSP**.

Nonidentifiability

Theorem 2 The following holds:

- $d_{TV}(p_{\boldsymbol{\theta}^\star}, p_{\widehat{\boldsymbol{\theta}}}) \leqslant \frac{1}{2}\|\boldsymbol{\theta}^\star - \widehat{\boldsymbol{\theta}}\|_2 \sum_{s\in\mathcal{S}} \max_{s'\in\mathcal{S}} \|\varphi(s) - \varphi(s')\|_2$.
- If $\mathbf{1}_S \in \mathrm{col}(\boldsymbol{\Phi})$ or $\mathrm{null}(\boldsymbol{\Phi}) \neq \{\mathbf{0}_S\}$, then the following holds:

$$\forall v > 0 \;\; \exists \widetilde{\boldsymbol{\theta}}^\star \in \mathbb{R}^d \text{ s.t. } p_{\widetilde{\boldsymbol{\theta}}^\star} = p_{\boldsymbol{\theta}^\star}, \text{ yet } \|\widetilde{\boldsymbol{\theta}}^\star - \boldsymbol{\theta}^\star\|_2 \geqslant v.$$

- Tight bound on $\|\boldsymbol{\theta}^\star - \widehat{\boldsymbol{\theta}}\|_2$ implies low TV distance, but *not* vice versa.
- **Nonidentifiability** persists due to softmax's translation invariance, regardless of whether we are in the overparametrized regime ($d \geqslant S$) or not.
  - Previous works make additional assumptions or use regularized M-estimator [Negahban et al., 2012] to resolve nonidentifiability

Our goal is to *estimate distribution*, not the *parameter*!

3. Distribution Estimators

Nonparametric Estimator

The **nonparametric estimator** is defined as follows:

$$\hat{p}_{nonparam}(s) := \frac{\sum_{i=1}^{N} \mathbf{1}[s_i = s]}{N}$$

Some known facts [Han et al., 2015]:

- $d_{TV}(p, \hat{p}_{nonparam}) = \mathcal{O}\left(\sqrt{\frac{S}{N}}\right)$.
- This rate is *minimax optimal*, i.e., the best performing in the worst case.

Parametric Estimator

The **parametric estimator** is defined as $\hat{p}_{param}(s) := p_{\hat{\theta}}(s)$ where

$$\widehat{\boldsymbol{\theta}} := \operatorname*{arg\,max}_{\boldsymbol{\theta} \in \mathbb{R}^d} \sum_{n=1}^{N} \log p_{\boldsymbol{\theta}}(s_n) - \frac{\lambda}{2} \|\boldsymbol{\theta}\|^2$$
$$= \operatorname*{arg\,max}_{\boldsymbol{\theta} \in \mathbb{R}^d} \sum_{s \in \mathcal{S}} \hat{N}(s) \log p_{\boldsymbol{\theta}}(s) - \frac{\lambda}{2} \|\boldsymbol{\theta}\|^2.$$

- $\hat{N}(s) := \sum_{n=1}^{N} \mathbf{1}[s_n = s]$ is the empirical visitation frequency of s
- $\lambda \geq 0$ is a regularization parameter
- We use gradient descent (GD) for computing the optimization problem.

More on Estimators

- *Nonparametric* estimator does not assume any particular functional form for the distribution, while *parametric* does, namely, the LSP parametrization.
- For *parametric* estimator, due to **nonidentifiability** issues, there is no known error rate for the resulting estimated distribution.
- From existing works on **identifiable** cases [Negahban et al., 2012], one could make an educated guess that the error rate for *parametric* estimator would be approximately $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$, at least in terms of the sample size N.

4. Experiments
(Our Contributions)

Setting

- Teacher-student setting where φ and $\theta^\star$ is fixed
- Three estimators are used: nonparametric, parametric with $\lambda = 0$ ("unregularized"), and parametric with optimal λ ("regularized")
  - The regularization parameter λ is found via grid-search for each set of experiments.

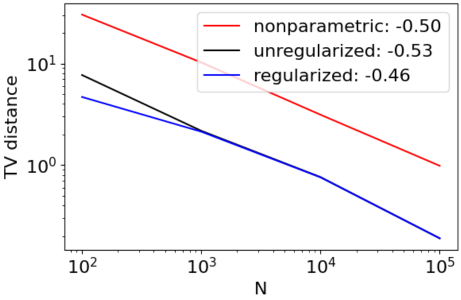
Setting

- Teacher-student setting where φ and $\theta^\star$ is fixed
- Three estimators are used: nonparametric, parametric with $\lambda = 0$ ("unregularized"), and parametric with optimal λ ("regularized")
  - The regularization parameter λ is found via grid-search for each set of experiments.

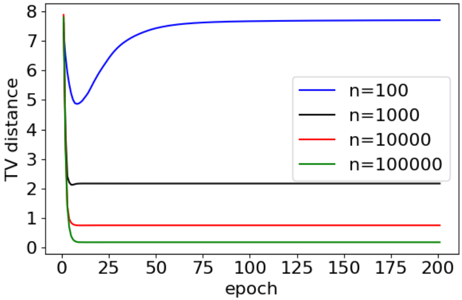
Research Questions:

1. What is the dependency of the error rate of the parametric estimator on N?
2. Which is the best estimator?
3. What is the dependency on d?
4. Is regularization effective?

Experiment #1. Vary N

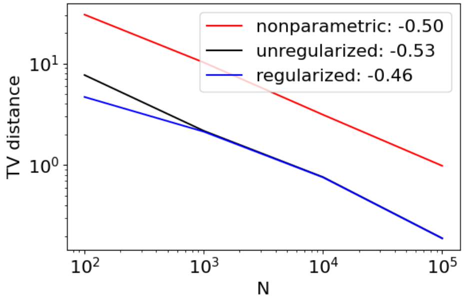


(a) TV vs. N (log-log plot)

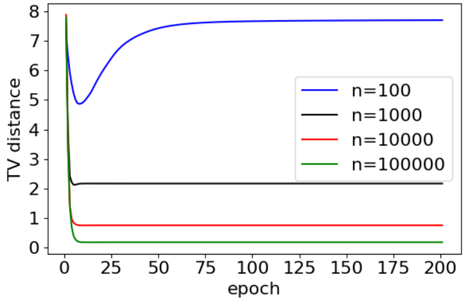


(b) TV vs. epoch (unregularized)

Experiment #1. Vary N



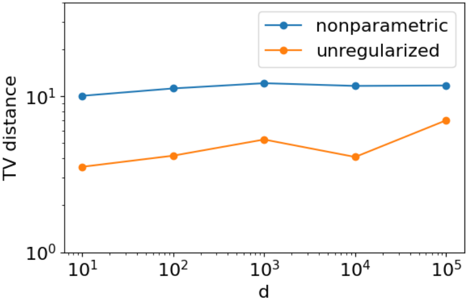
(a) TV vs. N (log-log plot)



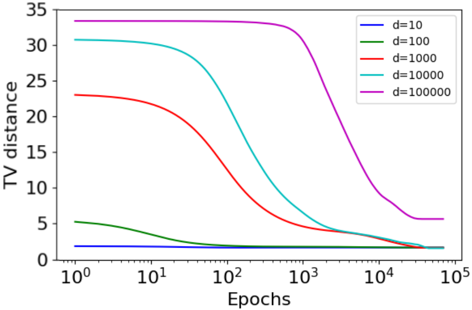
(b) TV vs. epoch (unregularized)

- All estimators have a slope of (approximately) -0.5 in the log-log plot
  $\rightarrow$ all estimators have the error rate of the form $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$ w.r.t. N.
- Still, parametric estimator attains smaller error than nonparametric
  $\rightarrow$ smaller multiplicative/additive constant in d or S?
- Sign of overfitting for small N, which is somewhat resolved via regularization

Experiment #2. Vary d

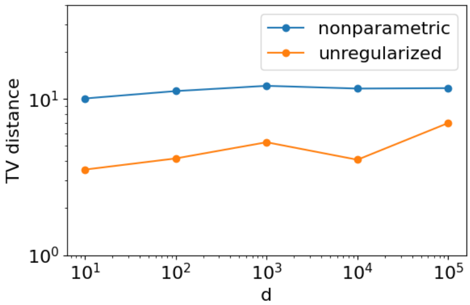


(c) TV vs. d (log-log plot)

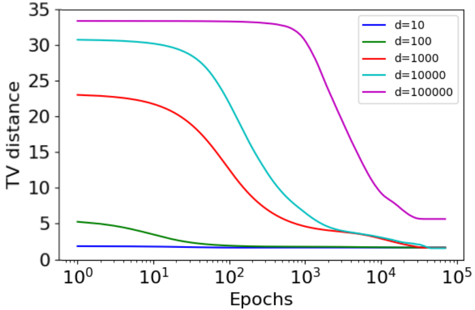


(d) TV vs. epoch (unregularized)

Experiment #2. Vary d



(c) TV vs. d (log-log plot)



(d) TV vs. epoch (unregularized)

- The error rate seems to be unaffected by varying d
  $\rightarrow$ the error rate doesn't depend on d?
- Still, parametric outperforms nonparametric across considered d's.
- High d results in harder optimization (e.g., longer plateau)

4. Concluding Remarks

Concluding Remarks

Our contributions: We introduce LSP and formally prove the expressivity and nonidentifiability guarantees of LSP. Experiments show that the parametric estimator leveraging LSP, although it has the same $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$ decay rate in N, results in a lower error rate than the nonparametric estimator.

Open problems:

- Rigorous statistical guarantees (e.g., minimax optimality) of the parametric estimator
- Linear softmax parametrization $\rightarrow$ Nonlinear softmax parametrization

References

Bradley, R. A. and Terry, M. E. (1952).
Rank Analysis of Incomplete Block Designs: The Method of Paired Comparisons.
Biometrika, 39(3-4):324–345.

Han, Y., Jiao, J., and Weissman, T. (2015).
Minimax Estimation of Discrete Distributions Under ℓ_1 Loss.
IEEE Transactions on Information Theory, 61(11):6343–6354.

Hwang, T. and Oh, M. (2023).
Model-Based Reinforcement Learning with Multinomial Logistic Function Approximation.
In *AAAI*.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012).
A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers.
Statistical Science, 27(4):538 – 557.

Plackett, R. L. (1975).
The Analysis of Permutations.
Journal of the Royal Statistical Society Series C: Applied Statistics, 24(2):193–202.

Reverdy, P. and Leonard, N. E. (2016).
Parameter Estimation in Softmax Decision-Making Models With Linear Objective Functions.
IEEE Transactions on Automation Science and Engineering, 13(1):54–67.

Seddik, E. A. et al. (2021).
The Unexpected Deterministic and Universal Behavior of Large Softmax Classifiers.
In *AISTATS*.