Optimal Learning for Mallows Block Model

Róbert Busa-Fekete BUSAFEKETE@VERIZONMEDIA.COM
Yahoo! Research

Dimitris Fotakis FOTAKIS@CS.NTUA.GR
National Technical University of Athens

Balázs Szörényi BALAZS.SZORENYI@VERIZONMEDIA.COM
Yahoo! Research

Manolis Zampetakis MZAMPET@MIT.EDU
Massachusetts Institute of Technology

Editors: Alina Beygelzimer and Daniel Hsu

[1]Abstract

The Mallows model, introduced in the seminal paper of Mallows (1957), is one of the most fundamental ranking distribution over the symmetric group S_m. To analyze more complex ranking data, several studies considered the Generalized Mallows model (Fligner and Verducci, 1986; Doignon et al., 2004; Marden, 1995). Despite the significant research interest of ranking distributions, the exact sample complexity of estimating the parameters of a Mallows and a Generalized Mallows Model is not well-understood.

The main result of the paper is a tight sample complexity bound for learning Mallows and Generalized Mallows Model. We approach the learning problem by analyzing a more general model which interpolates between the single parameter Mallows Model and the m parameter Mallows model. We call our model *Mallows Block Model* – referring to the Block Models that are popular models in theoretical statistics. Our sample complexity analysis gives tight bound for learning the Mallows Block Model for any number of blocks.

We also consider the problem of learning from one sample when the central ranking is known. As a corollary of our analysis, we obtain tight finite time bounds on the optimal rate at which the error of the spread parameter estimate goes to zero as the number of items goes to infinity. This is a strengthening of the asymptotic result of Mukherjee (2016).

Keywords: Ranking distributions, Mallows model, Generalized Mallows, Exponential family

1. Introduction

The Mallows model is one of the most fundamental ranking distribution since it was introduced in the seminal paper of Mallows (1957). The model has two parameters, the *central ranking* $\pi_0 \in S_m$ and the *spread parameter* $\phi \in [0, 1]$. Based on these, the probability of observing a ranking $\pi \in S_m$ is proportional to $\phi^{d(\pi,\pi_0)}$, where d is a ranking distance, such as the number of discordant pairs, a.k.a Kendall's tau distance.

To capture more complicated distributions over rankings, several studies considered the generalized Mallows model (Fligner and Verducci, 1986; Doignon et al., 2004; Marden, 1995), which assigns a different spread parameter $\phi_i \in [0, 1]$ to each alternative i. Now the probability of observ-

[1] Extended abstract. Full version is available on arXiv with the same title.

ing $\pi \in S_m$ decreases exponentially in a weighted sum over the discordant pairs, where the weights are determined by the spread parameters of discordant items. Statistical estimation of the distribution and the parameters of the Mallows model has been of interest in a wide range of scientific areas including theoretical statistics (Mukherjee, 2016), machine learning (Lu and Boutilier, 2011; Awasthi et al., 2014; Chen et al., 2009; Meila and Bao, 2010), social choice (Caragiannis et al., 2016), theoretical computer science (Liu and Moitra, 2018) and many more.

Despite this extensive literature, to the best of our knowledge, no optimal results are known on the sample complexity of learning the parameters of a Mallows or a generalized Mallows model. In this work, we fill this gap by proving: (1) an upper bound on the number of samples needed by some simple estimators to accurately estimate the parameters of the Mallows model, (2) an essentially matching lower bound on the sample complexity of any accurate estimator. Using our tight sample analysis, we are able to quantify in the finite sample regime some results that were only known in the asymptotic regime (e.g., Mukherjee (2016)).

Additionally, we introduce the *Mallows Block model*, which interpolates between the simple Mallows and the generalized Mallows models. The definition of the Mallows Block model is similar in spirit to the (fundamental in theoretical statistics) Stochastic Block model (Klopp et al., 2017), which admits similar statistical properties. Also, Berthet et al. (2016) recently introduced the Ising Block model, which is conceptually similar to the Stochastic Block Model. As we prove, the Mallows Block model combines two nice properties: (a) like the generalized Mallows model, it describes a wider range of distributions over rankings than the Mallows model; and (b) it allows accurate estimation of the spread parameters even from one sample, as it has been proved in (Mukherjee, 2016) for the Mallows model. We analyze the sample complexity of the Mallows Block model by proving essentially tight upper and lower bounds when the block structure is known.

1.1. Results and Techniques

In this work, we fully determine the sample complexity of learning Mallows and Generalized Mallows distributions, in a unified way, via the definition of the Mallows Block model. In a nutshell, we show how to estimate the parameters of these distributions in a (sample and time) efficient way, and how this implies efficient density estimation in KL-divergence and in total variation distance. Our approach is general and relies on some properties of the exponential family which properties might be leveraged to prove the exact learning rates for other complicated exponential families, such as the Ising model.

Learning in KL-divergence. Our learning algorithm for the spread parameters essentially finds the maximum likelihood solution, but in a provably computationally efficient way. The sample complexity analysis of the consistency of our estimator is based on some known and some novel results about exponential families. We show that the KL-divergence of two distributions in an exponential family is equal to the square difference of their parameters multiplied by the variance of a corresponding distribution in the exponential family. This lead us to a new strong concentration inequality for distributions in an exponential family which allows us to get a systematic way of proving upper bounds on the number of samples required to learn an exponential family in KL-divergence. Thus, we depart from the (only known) upper bounds on density estimation in total variation distance. We apply our technique to the Mallows Block model and get tight upper bounds of $O\left(\frac{d}{\varepsilon^2} + \log(m)\right)$ samples, where d is the (known) number of blocks in the Mallows Block model. We sketch the statement of this result below.

Informal Theorem 1 *Given $n = \tilde{\Omega}\left(\frac{d}{\varepsilon^2} + \log(m)\right)$ samples from a Mallows d-Block distribution $\mathcal{P}$, we can learn a distribution $\hat{\mathcal{P}}$ such that $\mathrm{D}_{\mathrm{KL}}\left(\mathcal{P} || \hat{\mathcal{P}}\right) \leq \varepsilon^2$ and hence $\mathrm{d}_{\mathrm{TV}}\left(\mathcal{P}, \hat{\mathcal{P}}\right) \leq \varepsilon$.*

Parameter Estimation. Extending a result of Caragiannis et al. (2016), we show that a logarithmic number of samples is both sufficient and necessary to estimate the central ranking of a generalized Mallows distribution. Then, using our results on exponential families, we show that estimating the spread parameter ϕ of a Mallows distribution boils down to obtaining a lower bound on the KL-divergence between two Mallows distributions with the same central ranking and parameters $|\phi - \phi'| = \Theta(\varepsilon)$. With such a lower bound on the KL-divergence, we can apply the concentration inequality for exponential family which we shall present in our paper, and show that once we learn the central ranking, with additional $O\left(\frac{d}{m^\star \varepsilon^2}\right)$ i.i.d. samples, we can estimate the parameter vector $\boldsymbol{\phi}$ of the underlying Mallows Block model within ℓ_2 error at most ε. Here, d denotes the number of blocks of the Mallows Block model and $m^\star$ is the minimum size of any block. We put everything together in the following informal theorem.

Informal Theorem 2 *Given $n = \tilde{\Omega}\left(\frac{d}{m^\star \varepsilon^2} + \log(m)\right)$ samples from a Mallows d-Block distribution $\mathcal{P}$ with parameters $\pi^\star$ and $\boldsymbol{\phi}^\star$, we can estimate $\hat{\pi}$ and $\hat{\boldsymbol{\phi}}$ so that $\hat{\pi} = \pi^\star$ and $\left\|\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}^\star\right\|_2 \leq \varepsilon$.*

A key observation in the proof of Theorem 2 is that the sufficient statistics for a generalized Mallows model with known central ranking are provided by an m-variate distribution where the i-th coordinate is an independent *truncated geometric distribution*. Truncated geometric distributions interpolate between Bernoulli and geometric distributions. The sufficient statistics of the Mallows Block model correspond to sums of truncated geometric distributions, which interpolate between Binomial and Negative Binomial distributions. We hence believe that the study of sums of truncated geometric distribution may be of independent interest. We should also highlight that in our approach, only the lower bound on the variance depends on Kendall's tau distance. Once we have such a bound for other exponential families, we can immediately apply our technique, e.g., to Mallows models with Spearman's Footrule and Spearman's Rank Correlation, as in (Mukherjee, 2016).

Learning from one sample. Arguably, the most interesting corollary of our tight analysis is that a single sample from a Mallows d-Block model with known central ranking is enough to estimate $\boldsymbol{\phi}$ within error $O\left(\sqrt{d/m^\star}\right)$, where again $m^\star$ is the minimum size of any block in the Mallows Block model. This result provides the exact rate of an asymptotic result by Mukherjee (2016).

Informal Theorem 3 *Given a single sample from a Mallows d-Block distribution $\mathcal{P}$ with known central ranking $\pi^\star$ and spread parameters $\boldsymbol{\phi}^\star$, we can estimate $\hat{\boldsymbol{\phi}}$ so that $\left\|\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}^\star\right\|_2 \leq \tilde{O}\left(\sqrt{\frac{d}{m^\star}}\right)$.*

Lower Bounds. On the lower bound side, we use Fano's inequality and show that $\Omega(\log(m))$ samples are necessary even for learning a simple Mallows distribution in total variation distance. Then, we show that $\Omega\left(\frac{d}{\varepsilon^2}\right)$ samples are necessary for learning a Mallows d-Block distribution in total variation distance.

Informal Theorem 4 *Any distribution estimation $\hat{\mathcal{P}}$ that is based only on $o\left(\frac{d}{\varepsilon^2} + \log(m)\right)$ samples from a Mallows d-Block distribution $\mathcal{P}$ satisfies $\mathrm{d}_{\mathrm{TV}}\left(\mathcal{P}, \hat{\mathcal{P}}\right) \geq \varepsilon$.*

Interestingly, our lower bound uses a general way to compute the total variation distance of two distributions that belong to the same exponential family. This theorem states that the total variation of two distributions in the same exponential family is equal to the distance between their parameters times the absolute deviation of a corresponding distribution in the family. This should be compared with the expression of the KL-divergence between two distributions in the same exponential family. Then our lower bound boils down to showing that for some range of parameters, the absolute deviation is within a constant from the standard deviation. With this proven, we get that the total variation distance is within a constant factor from the square root of the KL-divergence, and Fano's inequality can be applied.

References

Pranjal Awasthi, Avrim Blum, Or Sheffet, and Aravindan Vijayaraghavan. Learning mixtures of ranking models. In *Advances in Neural Information Processing Systems*, pages 2609–2617, 2014.

Quentin Berthet, Philippe Rigollet, and Piyush Srivastava. Exact recovery in the ising blockmodel. *arXiv preprint arXiv:1612.03880*, 2016.

Ioannis Caragiannis, Ariel D Procaccia, and Nisarg Shah. When do noisy votes reveal the truth? *ACM Transactions on Economics and Computation (TEAC)*, 4(3):15, 2016.

Harr Chen, S. R. K. Branavan, Regina Barzilay, and David R. Karger. Content modeling using latent permutations. *J. Artif. Intell. Res.*, 36:129–163, 2009.

Jean-Paul Doignon, Aleksandar Pekeč, and Michel Regenwetter. The repeated insertion model for rankings: Missing link between two subset choice models. *Psychometrika*, 69(1):33–54, 2004.

Michael A Fligner and Joseph S Verducci. Distance based ranking models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 359–369, 1986.

Olga Klopp, Alexandre B Tsybakov, Nicolas Verzelen, et al. Oracle inequalities for network models and sparse graphon estimation. *The Annals of Statistics*, 45(1):316–354, 2017.

Allen Liu and Ankur Moitra. Efficiently learning mixtures of Mallows models. In *FOCS*, pages 627–638. IEEE Computer Society, 2018.

Tyler Lu and Craig Boutilier. Learning Mallows models with pairwise preferences. In *Proceedings of the 28th International Conference on Machine Learning,*, pages 145–152, 2011.

C. Mallows. Non-null ranking models. *Biometrika*, 44(1):114–130, 1957.

John I. Marden. *Analyzing and Modeling Rank Data*. Chapman & Hall, 1995.

Marina Meila and Le Bao. An exponential model for infinite rankings. *Journal of Machine Learning Research*, 11:3481–3518, 2010.

Sumit Mukherjee. Estimation in exponential families on permutations. *The Annals of Statistics*, 44 (2):853–875, 2016.