Computing Optimal Bayesian Decisions for Rank Aggregation via MCMC Sampling

David Hughes and **Kevin Hwang** and **Lirong Xia**
Rensselaer Polytechnic Institute, Troy, NY, USA
{hughed2,hwangk2}@rpi.edu, xial@cs.rpi.edu

Abstract

We propose two efficient and general MCMC algorithms to compute optimal Bayesian decisions for Mallows' model and Condorcet's model w.r.t. any loss function and prior. We show that the mixing time of our Markov chain for Mallows' model is polynomial in $\varphi^{-k_{max}}$, d_{max}, and the input size, where φ is the dispersion of the model, k_{max} measures agents' largest total bias in bipartitions of alternatives, and d_{max} is the maximum ratio between prior probabilities. We also show that in some cases the mixing time is at least $\Theta(\varphi^{-k_{max}/2})$. For Condorcet's model, our Markov chain is rapid mixing for moderate prior distributions. Efficiency of our algorithms are illustrated by experiments on real-world datasets.

1 INTRODUCTION

In many social choice (a.k.a. rank aggregation) problems we want to compute an objectively optimal joint decision based on agents' preferences. For example, the *Condorcet Jury Theorem* (Condorcet, 1785) studies how to select a "correct" leader in political elections when the votes are noisy perceptions of the ground truth. Principles and algorithms for social choice have also been used to aggregate rankings in meta-search engines (Dwork et al., 2001), recommender systems (Ghosh et al., 1999), crowdsourcing (Mao et al., 2013), semantic webs (Porello and Endriss, 2013), and peer grading for MOOC (Raman and Joachims, 2014).

In these social choice applications it is natural to take a Bayesian approach. Given a statistical model that describes agents' noisy perception of the ground truth, a prior distribution, and a loss function that evaluates the joint decision w.r.t. the ground truth, we compute an optimal joint decision that has the minimum Bayesian expected loss w.r.t. the posterior distribution. This was recently formalized as a *statistical decision-theoretic framework for social choice* by Azari Soufiani et al. (2014).

A major challenge in previous research, especially in the Bayesian approaches, is the high computational complexity of decision making. For example, the maximum likelihood estimator (MLE) of a popular ranking model called *Mallows' model* (Mallows, 1957) is NP-hard to compute (Bartholdi et al., 1989). Computing optimal Bayesian decisions for rank aggregation is a hard combinatorial optimization problem because the parameter space is often discrete and its size is often exponential. Most previous work focused on designing efficient case-by-case algorithms for computing MLEs and MAPs of popular ranking models. However, the following question is left unanswered:

Are there general and efficient algorithms that compute optimal Bayesian decisions for a wide range of rank aggregation problems?

Our contributions. We give positive answers to this question for two popular ranking models: *Mallows' model* (Mallows, 1957) and *Condorcet's model* (Condorcet, 1785; Young, 1988) by proposing two general Markov chain Monte Carlo (MCMC) algorithms. Our algorithms work for any prior distribution, any decision space, and any loss function. In both algorithms, we first generate multiple samples by a Markov chain whose stationary distribution is the posterior distribution, which are used to compute the optimal decision that minimizes the empirical loss. Our Markov chains for both models are Metropolis-Hastings samplers (Metropolis et al., 1953; Hastings, 1970). For Mallows' model, we apply a random transposition on adjacent pairs of alternatives in each step, and for Condorcet's model, we adopt an *independent sampler*, which samples a candidate parameter with probability that is proportional to its likelihood in each step.

We prove that our Markov chains have good theoretical guarantees on the *mixing time*, which measures the rate of convergence to the stationary distribution and is closely related to the overall running time of the algorithms (Theorem 1). For Mallows' model, we show that the mixing

time of our Markov chain is polynomial in $\varphi^{-k_{max}}$, d_{max}, and the input size, where φ is the dispersion of the model, k_{max} measures agents' largest total bias in bipartitions of alternatives (Theorem 3), and d_{max} is the maximum ratio between prior probabilities. We also show that in some cases the mixing time is at least $\Theta(\varphi^{-k_{max}/2})$ (Proposition 1). For Condorcet's model, our Markov chain is rapid mixing for moderate prior distributions—its mixing time is polynomial in the input size and $(\ln \frac{d_{max}}{d_{max}-1})^{-1}$ (Theorem 5). Therefore, we can efficiently generate samples to estimate the Bayesian expected loss with high accuracy, for Mallows' model when $\varphi^{-k_{max}}$ and d_{max} are not too large, and for Condorcet's model when d_{max} is not too large, for a wide range of social choice problems including those that are provably NP-hard to compute or even approximate (Theorem 2 and 4).

Computational and statistical efficiency of our algorithms are shown in preliminary experiments using real-world election data from Preflib (Mattei and Walsh, 2013) (`www.Preflib.org`). The key observation is that for Mallows' model, when the number of alternatives is at least 11 the brute-force search takes much more time than generating 10 million samples, based on which we can achieve a high precision for estimating the Bayesian loss and making the optimal Bayesian decision.

Related Work and Discussions. MCMC methods have been widely applied in Bayesian statistics (Smith and Roberts, 1993). However, obtaining non-trivial bounds on the mixing time is a "considerable challenge" (Jerrum and Sinclair, 1996). Our proofs involve novel applications of a wide range of analytical tools, which we believe to have independent interest. To the best of our knowledge, this is the first time that MCMC methods for Bayesian inference for ranking models have been analytically studied. The theoretical bounds on the mixing time of the proposed Markov chains are our main theoretical contribution.

Practically, the main advantage of our algorithms is their generality because they work for any decision space, any loss function, and any prior distribution. For many natural loss functions (e.g. those in Definition 4), no efficient algorithm was previously known. There have been much work on computing MLE and MAP for Mallows' model and Condorcet's model in (computational) social choice (Bartholdi et al., 1989; Hemaspaandra et al., 2005; Betzler et al., 2008), theory (Ailon et al., 2005; Kenyon-Mathieu and Schudy, 2007), and machine learning (Kuo et al., 2009; Long et al., 2010; Lu and Boutilier, 2011; Liu, 2011; Negahban et al., 2012; Azari Soufiani et al., 2012, 2013a,b; Raman and Joachims, 2015). Also see the experimental study by Ali and Meila (2012) for a comparison of 104 algorithms and combinations. Our algorithms are more general, as MLEs and MAPs are special combinations of decision space and loss function.

Specifically, we have not seen much work on MCMC algorithms for rank aggregation. The work that is closest to ours is by Raman and Joachims (2015), who proposed a different Markov chain for Mallows' model and tested it on MOOC grading data. However, their paper did not analyze the mixing time. Diaconis and Hanlon (1992) proposed a Metropolis-Hastings algorithm to generate data according to a Mallows-like model that is based on the *Cayley distance*, which is different from both models studied in this paper. Moreover, it is not clear whether the algorithm by Diaconis and Hanlon can be leveraged to generate samples from the posterior distribution.

2 PRELIMINARIES

Let $\mathcal{C}$ denote a set of m alternatives. Let $\mathcal{L}(\mathcal{C})$ denote the set of *linear orders* over $\mathcal{C}$, that is, the set of all transitive, antisymmetric, and total binary relations. Let $\mathcal{B}(\mathcal{C})$ denote the set of all possibly cyclic orders over $\mathcal{C}$, that is, all irreflexive, antisymmetric, and total binary relations over $\mathcal{C}$. Clearly $\mathcal{L}(\mathcal{C}) \subseteq \mathcal{B}(\mathcal{C})$. Each agent uses a (possibly cyclic) order over $\mathcal{C}$ to represent her preferences. Let R denote the *(preference) profile* containing preferences from n agents. Given R, we want to make a joint decision from a decision space $\mathcal{D}$, which can be different from $\mathcal{C}$.

For any profile R, its *weighted majority graph*, denoted by $\mathrm{WMG}(R)$, is a weighted directed graph whose vertices are $\mathcal{C}$, and there is an edge between each pair of alternatives (a, b) with weight $w_R(a, b) = \#\{V \in R : a \succ_V b\} - \#\{V \in R : b \succ_V a\}$. Clearly $w_R(a, b) + w_R(b, a) = 0$.

For any $V, W \in \mathcal{B}(\mathcal{C})$, we let $\mathrm{KT}(V, W)$ denote the *Kendall-tau distance* between V and W, that is, the number of different pairwise comparisons in V and W. In this paper, we focus on the following two ranking models.

Definition 1 (Mallows' model with fixed dispersion)**.** *The parameter space is $\Theta_M = \mathcal{L}(\mathcal{C})$, the sample space $\mathcal{S}_M$ is composed of n i.i.d. generated data in $\mathcal{L}(\mathcal{C})$, and for any $W \in \mathcal{L}(\mathcal{C})$ and any profile R, we have $\Pr_M(R|W) = \prod_{V \in R} \left(\frac{1}{Z_M} \varphi^{KT(V,W)} \right)$, where Z_M is the normalization factor such that $Z_M = \sum_{U \in \mathcal{L}(\mathcal{C})} \varphi^{KT(U,W)}$.*

Definition 2 (Condorcet's model with fixed dispersion)**.** *The parameter space is $\Theta_C = \mathcal{B}(\mathcal{C})$, the sample space $\mathcal{S}_C$ is composed of n i.i.d. generated data in $\mathcal{B}(\mathcal{C})$, and for any $W \in \mathcal{B}(\mathcal{C})$ and any profile R, we have $\Pr_C(R|W) = \prod_{V \in R} \left(\frac{1}{Z_C} \varphi^{KT(V,W)} \right)$, where Z_C is the normalization factor such that $Z_C = \sum_{U \in \mathcal{B}(\mathcal{C})} \varphi^{KT(U,W)}$.*[1]

In the above two definitions, the normalization factors Z_M and Z_C are independent of the selection of W. We now recall the *statistical decision-theoretic framework for social*

[1]Our results also work for the variant where the sample space is $(\mathcal{L}(\mathcal{C}))^n$.

choice defined by Azari Soufiani et al. (2014) to formulate the computational problem.

Definition 3 ((Azari Soufiani et al., 2014))**.** *A statistical decision-theoretic framework for social choice (*SDT framework *for short) is a tuple $\mathcal{F} = (\mathcal{M}_{\mathcal{C}}, \mathcal{D}, L)$, where $\mathcal{C}$ is the set of alternatives, $\mathcal{M}_{\mathcal{C}} = (\Theta, \Pr, \mathcal{S})$ is a ranking model, $\mathcal{D}$ is the decision space, and $L : \Theta \times \mathcal{D} \to \mathbb{R}_{\geq 0}$ is a loss function that is easy to compute.*

In this paper, we focus on computing the *Bayesian estimators* f_B of a SDT framework that minimizes the expected Bayesian loss for any prior and profile. More precisely, given a SDT framework $\mathcal{F}$, a prior over Θ, and a profile R, $f_B(R) \in \min_{d \in \mathcal{D}} E_{\theta \sim \Pr(\cdot|R)} L(\theta, d)$. For example, the *maximum a posteriori (MAP)* is the Bayesian estimator for the SDT framework with $\mathcal{D} = \Theta$ and the 0-1 loss function.

Natural choices of the decision space for ranking models are: (1) $\mathcal{D} = \mathcal{C}$, and f_B is called a *resolute voting rule*. (2) $\mathcal{D} = 2^{\mathcal{C}} - \{\emptyset\}$, and f_B is called a *irresolute voting rule*. (3) $\mathcal{D} = \mathcal{L}(\mathcal{C})$, and f_B is called a *preference function* or *social welfare function*.

In this paper, we focus on case (1) $\mathcal{D} = \mathcal{C}$ and the exact Top-k loss functions defined below. The proposed algorithms and theorems on the mixing time work for any SDT framework.

Definition 4. *Let $\mathcal{D} = \mathcal{C}$. For any $k \leq m - 1$, the* exact Top-k *loss function $L_{ETop\text{-}k}$ is defined as: for any $W \in \mathcal{B}(\mathcal{C})$ and $d \in \mathcal{C}$, $L_{ETop\text{-}k}(W, d) = 0$ if there exits $A \subseteq \mathcal{C}$ such that $|A| = k$, $d \in A$, and for all $a \in A, b \in \mathcal{C} - A$ we have $a \succ_W b$; otherwise $L_{ETop\text{-}k}(W, d) = 1$.*

In words, the loss of an alternative d under the exact Top-k loss function is 0 if d is clearly ranked within top k positions in the ground truth; otherwise the loss is 1.[2]

A Markov chain over a state space Θ is characterized by a *transition matrix* P such that for any $V, W \in \Theta$, $P(V, W)$ is the probability for the next state to be W given that the current state is V. Therefore, for any $V \in \Theta$, we have $\sum_{W \in \Theta} P(V, W) = 1$. In this paper, the state space is the parameter space because we want to sample from the posterior distribution. The *stationary distribution* π of a Markov chain with transition matrix P is a probability distribution over Θ such that for any $V \in \Theta$ we have $\sum_{W \in \Theta} \pi(W) P(W, V) = \pi(V)$, that is, π (as a row vector) is a left eigenvector of P with eigenvalue 1.

Given a Markov chain with transition matrix $P(V, W)$ and a unique stationary distribution $\pi(\cdot)$, the *variation distance* at time t w.r.t. starting state V is defined to be

$$\Delta_V(t) = \max_{S \subseteq \Theta} |P^t(V, S) - \pi(S)|$$

where $P^t(V, S)$ is the probability for the Markov chain starting at V to end in a state in S after t steps.

The convergence rate of a Markov chain to the stationary distribution is measured by its *mixing time* $\tau_V(\epsilon)$, which is the number of steps that guarantee a variation distance below ϵ. Formally, we define

$$\tau_V(\epsilon) = \min\{t : \Delta_V(t') \leq \epsilon \text{ for all } t' \geq t\}$$

Let $\tau(\epsilon)$ denote the maximum mixing time for all starting states, that is, $\tau(\epsilon) = \max_V \tau_V(\epsilon)$.

Our algorithms first use a Markov chain sampler $\mathfrak{M}$ that runs a Markov chain for multiple steps to generate samples from the parameter space Θ, then compute the optimal decision w.r.t. these samples. Formally, we use Algorithm 1 to estimate the expected Bayesian loss of all decisions, then choose one with minimum expected loss.

Algorithm 1 CompBayesianLoss

1: **Input:** a profile R, a SDT framework $\mathcal{F} = (\mathcal{M}_{\mathcal{C}}, \mathcal{D}, L)$ with prior $\Pr(\cdot)$, a Markov chain sampler $\mathfrak{M}$ over Θ whose stationary distribution is $\Pr(\cdot|R)$, and a decision $d \in \mathcal{D}$.
2: Use $\mathfrak{M}$ to generate N independent samples, denoted by Q.
3: **return** $\sum_{W \in Q} L(W, d)/|Q|$.

It is well-known that the mixing time is closely related to the running time of approximate algorithms based on samples generated from the Markov chain. For SDT frameworks, this relation is formalized in Theorem 1. For completeness we include a short proof.

Theorem 1. *For any $d \in \mathcal{D}$ and any $\epsilon > 0, \delta > 0$, Algorithm 1 can compute the expected Bayesian loss of d with no more than ϵ additive error with probability at least $1 - \delta$ in $O(\frac{l_{max}^2}{\epsilon^2} \ln \delta^{-1} \tau(\frac{\epsilon}{2l_{max}})\eta)$ time, where l_{max} is the maximum loss in L, $\tau(\cdot)$ is the mixing time of the Markov chain used by $\mathfrak{M}$, and η is running time for one step in $\mathfrak{M}$.*

Proof: Let π denote the posterior distribution over Θ given R and let π^* denote the distribution by the Markov chain sampler. We first prove that using $O(\frac{l_{max}^2}{\epsilon^2} \ln \delta^{-1})$ independent samples, the output of Algorithm 1 is no more than $\epsilon/2$ away from $E_{V \sim \pi^*} L(V, d)$. We then choose a sampler $\mathfrak{M}$ with mixing time $\frac{\epsilon}{2l_{max}}$ and show that $|E_{V \sim \pi^*} L(V, d) - E_{V \sim \pi} L(V, d)| \leq \frac{\epsilon}{2}$.

Let $X^1, \ldots, X^N$ denote N i.i.d. random variables distributed as $L(V, d)$, where V is generated from π^*. Let $Y^N = (\sum_{i=1}^N X^i)/N$. Because for any $\theta \in \Theta$ and $d \in \mathcal{D}$, $0 \leq L(\theta, d) \leq l_{max}$, we have $\text{Var}(X^1) \leq l_{max}^2$ and $\text{Var}(Y^N) \leq l_{max}^2/N$. Also it is easy to check that $E_{V \sim \pi^*} L(V, d) = E(X^1) = E(Y^N)$. Therefore, by Chebyshev's inequality we have: $\Pr(|Y^N - E(Y^N)| \geq$

[2] We note that for some $W \in \mathcal{B}(\mathcal{C})$ no alternative is clearly ranked within top k. For any $W \in \mathcal{L}(\mathcal{C})$, there are always k alternatives clearly ranked in top k.

$\frac{\epsilon}{2}) \leq \frac{4\text{Var}(Y^N)}{\epsilon^2} \leq \frac{4l_{max}^2}{\epsilon^2 N}$. When $N \geq \frac{16}{3}\frac{l_{max}^2}{\epsilon^2}$ we have $\Pr(|Y^N - E(Y^N)| \geq \frac{\epsilon}{2}) \leq \frac{3}{4}$. This can be leveraged to an algorithm that outputs an estimation to $E(X^1)$ with no more than $\frac{\epsilon}{2}$ additive error with probability at least $1-\delta$, using $O(\frac{l_{max}^2}{\epsilon^2}\ln\delta^{-1})$ calls to the sampler $\mathfrak{M}$. For any $\mathfrak{M}$ with mixing time $\tau(\frac{\epsilon}{2l_{max}})$, we have $|E_{V\sim\pi^*}L(V,d) - E_{V\sim\pi}L(V,d)| \leq \sum_{V\in\Theta} L(V,d)|\pi(V) - \pi^*(V)| \leq l_{max}\frac{\epsilon}{2l_{max}} = \frac{\epsilon}{2}$.

Therefore, the total running time of Algorithm 1 is $O(\frac{l_{max}^2}{\epsilon^2}\ln\delta^{-1}\tau(\frac{\epsilon}{2l_{max}})\eta)$. □

In the remainder of this paper we focus on the Markov chains for Mallows' model and Condorcet's model. For both models, we will design *Metropolis-Hastings sampling algorithms* (Metropolis et al., 1953; Hastings, 1970), which work as follows. For each state V we first generate a candidate W for the next state from a proposal distribution $p_V(\cdot)$. Then, with probability $\min\{1, \frac{\pi(W)p_W(V)}{\pi(V)p_V(W)}\}$ the next state is W; otherwise the next state remains at V.

3 MARKOV CHAIN FOR MALLOWS' MODEL

To motivate the study, we first prove that the expected Bayesian loss is hard to approximate for Mallows' model w.r.t. the exact Top-1 loss function and uniform prior. In the BAYESIANLOSS problem, we are given a SDT framework, a prior, and a decision $d \in \mathcal{D}$. We are asked to compute the expected Bayesian loss of d.[3]

Theorem 2. *If* BAYESIANLOSS *for Mallows' model w.r.t. the exact Top-*1 *loss function and the uniform prior has a polynomial-time approximation algorithm with constant approximation ratio, then* $\mathsf{P} = \mathsf{NP} = \mathsf{P}_{||}^{\mathsf{NP}}$.

$\mathsf{P}_{||}^{\mathsf{NP}}$ is the class of problems that can be computed by a P oracle machine using polynomial number of parallel access to NP oracles. $\mathsf{P}_{||}^{\mathsf{NP}}$ contains NP and co-NP.

Proof: The hardness is proved by a reduction from the KEMENYWINNER problem, which is NP-hard (Bartholdi et al., 1989) and $\mathsf{P}_{||}^{\mathsf{NP}}$-complete (Hemaspaandra et al., 2005). Given a profile R, for any alternative c, the *Kemeny score* of c is the smallest Kendall-tau distance between the profile and any linear order where c is ranked at the top. An alternative with the minimum Kemeny score is called a *Kemeny winner*. In a KEMENYWINNER problem, we are given a profile R and an alternative c, and we are asked if c is a Kemeny winner.

Given a KEMENYWINNER instance (R, c), we construct a BAYESIANLOSS instance where there is a new alternative d and the profile R' satisfies that $\text{WMG}(R')$ equals to $\text{WMG}(R)$ plus the edges $d \to a$ for all $a \neq c$, whose weights are 1 (if weights on edges in $\text{WMG}(R)$ are odd) or 2 (if weights on edges in $\text{WMG}(R)$ are even). Given any constant $\alpha > 1$, we let $\varphi = \frac{\alpha^2}{2(m!)^2}$. We note that φ can be represented using polynomial number of bits. This instance can be constructed in polynomial time using McGarvey's trick (McGarvey, 1953).

Clearly d is a Kemeny winner in R'. If the KEMENYWINNER is a "yes" instance, then c is also a Kemeny winner in R' (where d is ranked in the second place in the winning ranking). The Bayesian expected loss of d is at least $\frac{1}{m!}$ because in at least one ranking d is not at the top.

If the KEMENYWINNER is a "no" instance, then d is the unique Kemeny winner. Let V denote a ranking where d is ranked in the top and $\text{KT}(V, R')$ equals to the Kemeny score of d. It is easy to check that for any ranking W where d is not ranked in the top, $\text{KT}(V, R') \leq \text{KT}(W, R') - 1$, which means that $\frac{\Pr(W|R')}{\Pr(V|R')} \leq \varphi$. Therefore, the posterior probability that d is not ranked in the top, which equals to the expected Bayesian loss of d, is at most $\frac{(m!-1)\varphi}{1+(m!-1)\varphi}$. When $\varphi = \frac{\alpha^2}{2(m!)^2}$, we have $\frac{(m!-1)\varphi}{1+(m!-1)\varphi} < \frac{\alpha^2}{m!}$.

It follows that if there exists a polynomial-time α-approximation algorithm for BAYESIANLOSS, then we can use this algorithm to solve KEMENYWINNER in polynomial time: if the output of the algorithm is no more than $\frac{\alpha}{m!}$ then the KEMENYWINNER instance is a "no" instance; otherwise the KEMENYWINNER instance is a "yes" instance. This means that $\mathsf{P} = \mathsf{NP} = \mathsf{P}_{||}^{\mathsf{NP}}$. □

We now present the Markov chain $\mathfrak{M}_M$ for Mallows' model in Algorithm 2, which runs $\mathfrak{M}_M$ for N steps. In each step, we first apply a random transposition of adjacent alternatives in the current state V (changing $d \succ c$ to $c \succ d$) to obtain a candidate ranking W, then with probability $\frac{1}{2}\min\{\frac{\Pr_M(W)}{\Pr_M(V)}\varphi^{R[d\succ c]-R[c\succ d]}, 1\}$ we let $V = W$, otherwise the next state stays at V, where $R[c \succ d]$ is the number of times $c \succ d$ in R. The $\frac{1}{2}$ factor is a popular trick to prove bounds on the mixing time in Lemma 2.

Algorithm 2 Markov chain $\mathfrak{M}_M$ for Mallows' model.

1: **Inputs:** a profile R, a prior $\Pr_M$, an initial ranking V, and the number of iterations N.
2: **for** t =1 **to** N **do**
3: Switch a pair of adjacent alternatives uniformly at random (from $d \succ c$ to $c \succ d$). Let W denote the new ranking.
4: With probability $\frac{1}{2}\min\{\frac{\Pr_M(W)}{\Pr_M(V)}\varphi^{R[d\succ c]-R[c\succ d]}, 1\}$ let $V = W$.
5: **end for**
6: **return** V.

For any profile R, let k_{max} denote the max cut of the undirected $\text{WMG}(R)$. That is, $k_{max} = \max_{A\subseteq\mathcal{C}}\sum_{a\in A, b\in\mathcal{C}-A}|w_R(a,b)|$. Let d_{max} denote the

[3] This problem is known to be NP-hard and $\mathsf{P}_{||}^{\mathsf{NP}}$-hard to compute exactly (Procaccia et al., 2012; Azari Soufiani et al., 2014).

maximum ratio between prior probabilities. That is, $d_{max} = \max_{V,W\in\Theta} \frac{\Pr_M(V)}{\Pr_M(W)}$.

Theorem 3. *The mixing time of the Markov chain in Algorithm 2 for any starting state is* $O(m^4 d_{max}^3 \varphi^{-k_{max}}(nm^3 \log m \ln \varphi^{-1} + \ln \epsilon^{-1}))$.

Proof: It is easy to check that the Markov chain $\mathfrak{M}_M$ in Algorithm 2 is finite, ergodic, reversible and the transition matrix $P_M(V,W)$ is diagonally dominant. Therefore, all eigenvalues of $P_M(V,W)$ are real and positive, and the largest one is 1. The following lemma shows that the mixing time is closely related to the *spectral gap* $1-\lambda_{max}$, where λ_{max} is the second largest eigenvalue of $P_M(V,W)$. It is easy to verify that the stationary distribution π_M equals to the posterior distribution $\Pr_M(\cdot|R)$.

Lemma 1 (e.g. (Sinclair, 1992))**.**
(i) $\tau_V(\epsilon) \le (1-\lambda_{max})^{-1}(\ln \pi(V)^{-1} + \ln \epsilon^{-1})$*;*

(ii) $\max_{V\in\Theta} \tau_V(\epsilon) \ge \frac{1}{2}(1-\lambda_{max})^{-1} \ln(2\epsilon)^{-1}$.

It is often hard to directly obtain lower bounds on the spectral gap. We will take the *canonical path* approach, whose idea is the following. Any reversible Markov chain can be visualized as an undirected graph G where the vertices are Θ and the weight on the edge between V and W is $Q(V,W) = \pi(V)P(V,W)$. For each pair of states $V, W \in \Theta$, we fix a directed path (canonical path) from V to W in G, denoted by γ_{VW}. Let $|\gamma_{VW}|$ denote the length of γ_{VW}. Let Γ denote the set of all canonical paths defined above, one for each pair (V,W). The maximum *loading* of a single edge in Γ provides a lower bound on the spectral gap, thus it can be used to upper-bound the mixing time. Formally, this was proved by Sinclair (1992) in the following Lemma.

Lemma 2 ((Sinclair, 1992; Jerrum and Sinclair, 1996))**.** *Let $\mathfrak{M}$ be a finite, reversible, and ergodic Markov chain with loop probabilities $P(V,V) \ge \frac{1}{2}$ for all states V. Let Γ be a set of canonical paths with maximum edge loading*

$$\rho = \max_e \frac{1}{Q(e)} \sum_{\gamma_{VW}\ni e} \pi(V)\pi(W)|\gamma_{VW}|$$

Then the mixing time satisfies $\tau_V(\epsilon) \le \rho(\ln \pi(V)^{-1} + \ln \epsilon^{-1})$ for all $V \in \Theta$.

To apply Lemma 2, we consider the following canonical paths Γ_M for $\mathfrak{M}_M$. We note that the graph G for canonical paths, whose vertices are rankings over alternatives, is different from the weighted majority graph, whose vertices are the alternatives.

Definition 5. *In the canonical paths Γ_M, for any pair of different rankings $V, W \in \mathcal{L}(\mathcal{C})$, γ_{VW} contains the rankings obtained in $m-1$ stages of adjacent transpositions, where in stage k, the alternative ranked at the k-th position in W is moved up to the k-th position in V.*

W.l.o.g. let $W = [a_1 \succ \cdots \succ a_m]$. In stage 1, we apply the minimum number of adjacent transpositions on V to move a_1 to the top position. This process passes through no more than $m-1$ rankings, and let V_1 denote the ranking at the end of the process, where a_1 is ranked at the top position and the other part of V_1 is the same as in V. In stage 2, if a_2 is not already ranked at the second position of V_1, then we apply the minimum number of adjacent transpositions on V_1 to move a_2 to the second position. The process continues until we reach W. For example, when $m=4$, $V = [a_4 \succ a_1 \succ a_3 \succ a_2]$, and $W = [a_1 \succ a_2 \succ a_3 \succ a_4]$, we have $\gamma_{VW} = V \to [a_1 \succ a_4 \succ a_3 \succ a_2] \to [a_1 \succ a_4 \succ a_2 \succ a_3] \to [a_1 \succ a_2 \succ a_4 \succ a_3] \to W$.

It is not hard to see that $|\gamma_{VW}| \le m^2$. For any edge $e = E \to E'$ in a canonical path, we have $Q_M(E,E') = \pi_M(E)P_M(E,E') = \pi_M(E')P_M(E',E) \ge \frac{1}{2(m-1)} \min\{\pi_M(E), \pi_M(E')\}$. Therefore, we have

$$\rho \le \max_{e=E\to E'} \left(\frac{2m^2(m-1)}{\min\{\pi_M(E), \pi_M(E')\}} \sum_{\gamma_{VW}\ni e} \pi_M(V)\pi_M(W)\right) \tag{1}$$

Lemma 3. *Given the canonical paths Γ_M defined in Definition 5 and any edge $e = E \to E'$, we have:*

(i) $\sum_{V,W:\gamma_{VW}\ni e} \pi_M(V)\pi_M(W)/\pi_M(E) \le m\varphi^{-k_{max}}$*, and*

(ii) $\sum_{V,W:\gamma_{VW}\ni e} \pi_M(V)\pi_M(W)/\pi_M(E') \le m\varphi^{-k_{max}}$.

Proof: Let $E = [T \succ d \succ c \succ B]$ and $E' = [T \succ c \succ d \succ B]$. For any $0 \le k \le |T|$, we let T_k denote the top k ordering of T and let T_k^* denote the remaining ordering. That is, for any $k \le |T|$ we have $T = [T_k \succ T_k^*]$. It is easy to check that $e \in \gamma_{VW}$ if and only if there exists $0 \le k \le |T| \le m-2$ such that the following conditions hold.

(1) $T_k^* \succ d \succ B$ and $d \succ c$ hold in V. Let $\mathcal{V}_k$ denote the set of all such V's.

(2) The top $k+1$ alternatives in W are ranked as $[T_k \succ c]$. Let $\mathcal{W}_k$ denote the set of all such W's.

We first prove the inequality for E. Let A_k denote the alternatives in T_k, let A_k^* denote the alternatives in T_k^* plus d, let S denote the alternatives in B, and let $\bar{A}_k = \mathcal{C} - A_k$. For each pairwise comparison $a \succ_E b$ in E, either we have (1) $a \succ_V b$ for all $V \in \mathcal{V}_K$ or (2) $a \succ_W b$ for all $W \in \mathcal{V}_K$. This relationship is shown in Table 1.

For example, "W" at (A_k, A_k) in Table 1 means that for all $W \in \mathcal{W}_k$, $(a,b) \in A_k \times A_k$, $a \succ_W b$ if and only if $a \succ_E b$. (c,c) is marked N/A because the pairwise comparison between c and c is not well defined.

For any $J \subseteq \mathcal{L}(\mathcal{C}) \times \mathcal{L}(\mathcal{C})$ and $V, W \in \mathcal{L}(\mathcal{C})$, we let $D_J(V,W)$ denote the number of different pairwise comparisons between V and W for all paris $\{a,b\}$ such that

	A_k	A_k^*	c	S
A_k	W	W	W	W
A_k^*	W	V	V	V
c	W	V	N/A	W
S	W	V	W	V

Table 1: Pairwise comparisons in E that are the same as in $V \in \mathcal{V}_k$ or $W \in \mathcal{W}_k$.

$(a,b) \in J$ or $(b,a) \in J$. Formally, $D_J(V,W) = \#\{\{a,b\} : [(a,b) \in J \text{ or } (b,a) \in J] \text{ and } [[a \succ_V b \text{ and } b \succ_W a] \text{ or } [a \succ_W b \text{ and } b \succ_V a]]\}$. In other words, $D_J(V,W)$ is the Kendall-tau distance between the restriction of V on J and the restriction of W on J. We note that any unordered pair of alternatives $\{a,b\}$ is counted only once in J. In particular, for any $a \in \mathcal{C}$ and $A \subset \mathcal{C}$, $D_{\{a\}\times\{a\}}(V,W) = 0$ and $D_{(\{a\}\cup A)\times(\{a\}\cup A)}(V,W) = D_{A\times(\{a\}\cup A)}(V,W) = D_{(\{a\}\cup A)\times A}(V,W)$. We have

$$\begin{aligned}
&\sum_{V,W:\gamma_{VW}\ni e} \pi_M(V)\pi_M(W)/\pi_M(E) \\
&= \sum_{0\le k\le |T|} \sum_{V\in\mathcal{V}_k, W\in\mathcal{W}_k} \pi_M(V)\pi_M(W)/\pi_M(E) \\
&\le \sum_{0\le k\le |T|} \sum_{V\in\mathcal{V}_k, W\in\mathcal{W}_k} d_{max}^3 \frac{\Pr_M(R|V)\Pr_M(R|W)}{\Pr_M(R|E)\sum_{U\in\mathcal{L}(\mathcal{C})}\Pr_M(R|U)} \qquad (2) \\
&= \sum_{0\le k\le |T|} \sum_{V\in\mathcal{V}_k, W\in\mathcal{W}_k} \frac{d_{max}^3 f_1(V,E) f_2(W,E)}{\sum_{U\in\mathcal{L}(\mathcal{C})}\varphi^{\mathrm{KT}(U,R)}} \qquad (3)
\end{aligned}$$

where $f_1(V,E) = \varphi^{D_{A_k\times\mathcal{C}}(V,R)+D_{\{c\}\times S}(V,R)}$ and $f_2(W,E) = \varphi^{D_{A_k^*\times\bar{A}_k}(W,R)+D_{S\times S}(W,R)}$. (2) is due to the following lemma.

Lemma 4. *For any ranking model, any $V \in \Theta$, and any profile R, we have $\frac{\Pr(V|R)}{d_{max}} \le \frac{\Pr(R|V)}{\sum_{U\in\Theta}\Pr(R|U)} \le d_{max}\Pr(V|R)$.*

For (3), according to Table 1, we have (for all grids with "V" in Table 1)

$$\begin{aligned}
&D_{A_k^*\times\bar{A}_k}(E,R) + D_{S\times S}(E,R) \\
&= D_{A_k^*\times\bar{A}_k}(V,R) + D_{S\times S}(V,R)
\end{aligned}$$

and (for all grids with "W" in Table 1)

$$\begin{aligned}
&D_{A_k\times\mathcal{C}}(E,R) + D_{\{c\}\times S}(E,R) \\
&= D_{A_k\times\mathcal{C}}(W,R) + D_{\{c\}\times S}(W,R)
\end{aligned}$$

We note that for any $U \in \mathcal{L}(\mathcal{C})$, $\Pr(R|U) \propto \varphi^{\mathrm{KT}(U,R)}$ and

$$\begin{aligned}
\mathrm{KT}(U,R) = &D_{A_k\times\mathcal{C}}(U,R) + D_{A_k^*\times\bar{A}_k}(U,R) \\
&+ D_{\{c\}\times S}(U,R) + D_{S\times S}(U,R)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathrm{KT}(E,R) = &D_{A_k^*\times\bar{A}_k}(V,R) + D_{S\times S}(V,R) \\
&+ D_{A_k\times\mathcal{C}}(W,R) + D_{\{c\}\times S}(W,R)
\end{aligned} \qquad (4)$$

(3) follows after substituting (4) into (2).

We next prove that $\sum_{V\in\mathcal{V}_k, W\in\mathcal{W}_k} \frac{f_1(V,E)f_2(W,E)}{\sum_{U\in\mathcal{L}(\mathcal{C})}\varphi^{\mathrm{KT}(U,R)}} \le \varphi^{-k_{max}}$. To do so, we define a function $g : \mathcal{V}_k \times \mathcal{W}_k \to \mathcal{L}(\mathcal{C})$ as follows: for any $(V,W) \in \mathcal{V}_k \times \mathcal{W}_k$, $g(V,W)$ is obtained from V by applying a permutation over $A_k^* \cup S$ so that the preferences of $g(V,W)$ over $A_k^* \cup S$ become the preferences of W over $A_k^* \cup S$, while the other pairwise comparisons stay the same as in V. This means that the positions of $A_k \cup \{c\}$ in $g(V,W)$ are the same as in V. It is not hard to verify that for all pairs $(V_1,W_1) \ne (V_2,W_2)$, we have $g(V_1,W_1) \ne g(V_2,W_2)$, which means that $\{g(V,W) : V \in \mathcal{V}_k, W \in \mathcal{W}_k\} \subseteq \mathcal{L}(\mathcal{C})$.

Claim 1. *For any $(V,W) \in \mathcal{V}_k \times \mathcal{W}_k$, $f_1(V,E)f_2(W,E) \le \varphi^{KT(g(V,W),R)-k_{max}}$.*

Proof: By the definition of $g(V,W)$ we have $D_{A_k\times(A_k\cup\{c\})}(V,R) = D_{A_k\times(A_k\cup\{c\})}(g(V,W),R)$ and $D_{A_k^*\times(A_k^*\cup S)}(W,R) = D_{A_k^*\times(A_k^*\cup S)}(g(V,W),R)$. Therefore,

$$\begin{aligned}
&f_1(V,E)f_2(W,E)/\varphi^{g(V,W)} \\
=&\frac{\varphi^{D_{A_k\times(A_k^*\cup S)}(V,R)+D_{\{c\}\times S}(V,R)+D_{\{c\}\times A_k^*}(W,R)}}{\varphi^{D_{(A_k\cup\{c\})\times(A_k^*\cup S)}(g(V,W),R)}} \\
\le&\varphi^{-\sum_{a\in A_k\cup\{c\}, b\in A_k^*\cup S}|w_R(a,b)|} \le \varphi^{-k_{max}}
\end{aligned}$$

□

By Claim 1 we have

$$\begin{aligned}
&\sum_{0\le k\le |T|} \sum_{V\in\mathcal{V}_k, W\in\mathcal{W}_k} \frac{f_1(V,E)f_2(W,E)}{\sum_{U\in\mathcal{L}(\mathcal{C})}\varphi^{\mathrm{KT}(U,P)}} \\
\le &\sum_{0\le k\le |T|} \frac{\sum_{V\in\mathcal{V}_k, W\in\mathcal{W}_k} f_1(V,E)f_2(W,E)}{\sum_{V\in\mathcal{V}_k, W\in\mathcal{W}_k}\varphi^{\mathrm{KT}(g(V,W),R)}} \\
\le &\sum_{0\le k\le |T|} \max_{V\in\mathcal{V}_k, W\in\mathcal{W}_k} f_1(V,E)f_2(W,E)/\varphi^{\mathrm{KT}(g(V,W),R)} \\
\le &m\varphi^{-k_{max}}
\end{aligned}$$

This proves the inequality for E. The inequality for E' is proved similarly by letting A_k^* denote the alternatives in T_k^* and letting S denote the alternatives in B plus d. □

Combining Lemma 3 and inequality (1), we have $\rho \le 2m^4 d_{max}^3 \varphi^{-k_{max}}$. We also note that for any state V, $\pi(V) \ge \varphi^{nm^2}/m!$, which means that $\ln\pi(V)^{-1}$ is $O(nm^3 \log m \ln\varphi^{-1})$. The theorem follows after applying Lemma 2. □

Remarks: The upper bound proved in Theorem 3 is polynomial in $m, n, d_{max}, \ln \epsilon^{-1}$, and is exponential in k_{max} (with base φ^{-1}). Therefore, the algorithm is efficient if d_{max} and $\varphi^{-k_{max}}$ are small, that is, either φ is close to 1 or k_{max} is small. The next proposition shows that the mixing time of $\mathfrak{M}_M$ is sometimes $\Omega(m\varphi^{-k_{max}/2})$.

Proposition 1. *There exists a constant α so that for any $m \geq 3$, there exists a profile R and the mixing time of $\mathfrak{M}_M$ is at least $\alpha m \varphi^{-k_{max}/2} \ln(2\epsilon)^{-1}$.*

Proof: For any $m \geq 3$ and any even number l we can construct a profile R^l with polynomial many votes using McGarvey's trick (McGarvey, 1953) such that the WMG(R^l) contains only three edges: $a_1 \to a_2$, $a_2 \to a_3$, $a_3 \to a_1$, and the weight on all three edges is l.

It is easy to check that $k_{max} = 2l$. We prove the lower bound on the mixing time by applying Lemma 2(ii) and the *conductance* approach.

Definition 6 (Sinclair and Jerrum (1989))**.** *The* conductance *of a Markov chain $\mathfrak{M}$ is defined as*

$$\Phi(\mathfrak{M}) = \min_{S \subset \Theta : \pi(S) \leq 1/2} \frac{Q(S, \bar{S})}{\pi(S)},$$

where $Q(S, \bar{S}) = \sum_{V \in S, W \in \bar{S}} Q(V, W)$.

The spectral gap is related to the conductance in the following lemma proved by Sinclair and Jerrum (1989).

Lemma 5 ((Sinclair and Jerrum, 1989))**.** *For any reversible Markov chain whose conductance is Φ, the second eigenvalue λ_1 satisfies $1 - 2\Phi \leq \lambda_1 \leq 1 - \frac{\Phi^2}{2}$.*

We recall that all eigenvalues of $\mathfrak{M}_M$ are non-negative. Therefore, the spectral gap of $\mathfrak{M}_M$ is $1 - \lambda_1$, which is at most 2Φ. The following claim gives an upper bound on the conductance for all R^l.

Claim 2. *There exists $\beta > 0$ so that for all even number l, $\Phi(\mathfrak{M}_M) \leq \beta \frac{1}{m} \varphi^{k_{max}/2}$ for all R^l.*

Proof: We let $S \subseteq \Theta$ denote the set of rankings where $a_1 \succ a_2 \succ a_3$. For any $V \in S$ and $W \in \bar{S}$, if $P(V, W) > 0$ then either $a_2 \succ_W a_1 \succ_W a_3$ or $a_1 \succ_W a_3 \succ_W a_2$, which means that $Q(V, \bar{S})/\pi(V) \leq \frac{1}{m-1} \varphi^{k_{max}/2}$. Therefore, $\frac{Q(S,\bar{S})}{\pi(S)} = \frac{\sum_{V \in S} Q(V, \bar{S})}{\sum_{V \in S} \pi(V)} \leq \frac{1}{m-1} \varphi^{k_{max}/2}$. It is easy to check that $1/6 \leq \pi(S) \leq 1/3$, which means that there exists $\beta > 0$ so that $\Phi(\mathfrak{M}_M) \leq \frac{Q(S,\bar{S})}{\pi(S)} \leq \beta \frac{1}{m} \varphi^{k_{max}/2}$. □

Combining Lemma 5 and Claim 2 we have $1 - \lambda_{max} = 1 - \lambda_1 \leq \beta \frac{1}{m} \varphi^{k_{max}/2}$. It follows from Lemma 1(ii) that $\max_V \tau_V(\epsilon) \geq \frac{\beta}{2} m \varphi^{-k_{max}/2} \ln(2\epsilon)^{-1}$. □

4 MARKOV CHAIN FOR CONDORCET'S MODEL

For Condorcet's model it has been shown by Young (1988) that for any $W \in \mathcal{B}(\mathcal{C})$ and any profile R, $\Pr_C(R|W) \propto \prod_{a \succ_W b} (\frac{\varphi}{1-\varphi})^{R[a \succ b]}$, where we recall that $R[a \succ b]$ is the number of times $a \succ b$ in R. This leads to the following observation.

Proposition 2. *Let R denote a profile of binary relations. $V \in \mathcal{B}(\mathcal{C})$ maximizes the likelihood if and only if for any pair of alternatives (a, b), we have $(R[a \succ b] > R[b \succ a]) \Rightarrow (a \succ_V b)$.*

An immediate corollary is that for any profile R, computing the MLE is in P. While computing the expected Bayesian loss w.r.t. the exact Top-1 loss function is in P (Young, 1988; Elkind and Shah, 2014; Azari Soufiani et al., 2014), for some natural loss functions computing the minimum expected Bayesian loss is NP-hard. Formally, in a MINBAYESIANLOSS problem, we are given a SDT framework, a prior, a decision $d \in \mathcal{D}$, and a number l. We are asked whether there exists a decision whose expected Bayesian loss is no more than l.

Theorem 4. MINBAYESIANLOSS *can be computed in polynomial time for Condorcet's model w.r.t. $L_{ETop\text{-}k}$ and the uniform prior for any fixed k. It is* NP*-hard to compute* MINBAYESIANLOSS *for Condorcet's model w.r.t. $L_{ETop\text{-}\frac{m}{2}}$ and the uniform prior for even m.*

Proof: For any fixed k, the Bayesian loss of any decision can be computed by enumerating all combinations of alternatives ranked at top k positions in the ground truth, the probability of which can be computed by Claim 3 below.

We prove the NP-hardness of MINBAYESIANLOSS for Condorcet's model w.r.t. $L_{\text{ETop-}\frac{m}{2}}$ by a reduction from an NP-hard problem called ONEWAYBISECTION (Feige and Yahalom, 2003). In a ONEWAYBISECTION instance, we are given an oriented graph $G = (\mathcal{V}, \mathcal{E})$ with m vertices, where m is even, and we asked whether there exists a partition of $\mathcal{V} = \mathcal{S} \cup \mathcal{T}$ so that $|\mathcal{S}| = |\mathcal{T}| = \frac{m}{2}$ and there is no edge from $\mathcal{T}$ to $\mathcal{S}$. For any ONEWAYBISECTION instance, we construct a MINBAYESIANLOSS instance as follows.

The alternatives are the vertices. The preferences R are obtained by McGarvey's trick (McGarvey, 1953) so that the positive edges in WMG(R) are the same as in G, and all positive weights are 2. $\varphi = 2^{-m^2}$. $l = 1 - 2^{-m^2/4}$.

For any $A \subseteq \mathcal{C}$, we let $\Pr(A \succ \bar{A}|R)$ denote the posterior probability that all alternatives in A are preferred to all alternatives in $\bar{A}$. That is, $h(A) = \sum_{W : A \succ_W \bar{A}} \Pr(W|R)$. The next claim follows after calculations in (Elkind and Shah, 2014; Azari Soufiani et al., 2014).

Claim 3. $\Pr(A \succ \bar{A}|R) = \prod_{a \in A, b \in \bar{A}} F(a, b)$, *where*

$$F(a,b) = \begin{cases} \frac{1}{1+\varphi^2} & \text{if } w_R(a,b) = 2 \\ \frac{\varphi^2}{1+\varphi^2} & \text{if } w_R(a,b) = -2 \\ 1/2 & \text{if } w_R(a,b) = 0 \end{cases}$$

Therefore, if the ONEWAYBISECTION instance has a solution A, then the expected Bayesian loss for any alternative in A is at most $1 - \Pr(A \succ \bar{A}|R) \leq 1 - 2^{-m^2/4}$,

which means that the MINBAYESIANLOSS instance is a "yes" instance. If the ONEWAYBISECTION instance does not have a solution, then for any alternative $a \in \mathcal{C}$, the expected Bayesian loss is at least $1 - \binom{m}{m/2}\frac{\varphi^2}{1+\varphi^2} > 1 - 2^{m \log m}\frac{\varphi^2}{1+\varphi^2} > 1 - 2^{-m^2/4}$, which means that the MINBAYESIANLOSS instance is a "no" instance. □

We now present the Markov chain $\mathfrak{M}_C$ for Condorcet's model in Algorithm 3, which runs $\mathfrak{M}_C$ for N steps. $\mathfrak{M}_C$ is an *independent sampler* and starts at an arbitrary state that maximizes the likelihood. In each step, a candidate next state is drawn independent of the current state, which means that $p_X(\cdot)$ is the same for all X. We let $p(\cdot)$ denote this proposal distribution.

In $\mathfrak{M}_C$, $p(\cdot)$ is the posterior probability *assuming that the prior is uniform*. In other words, for any $W \in \mathcal{B}(\mathcal{C})$, $p(W)$ is proportional to $\Pr(R|W)$. A binary relation in $\mathcal{B}(\mathcal{C})$ can be efficiently generated from $p(\cdot)$ by generating pairwise comparisons between alternatives independently, such that for each pair of alternatives (a, b), $\frac{\Pr(a \succ_W b)}{\Pr(b \succ_W a)} = \frac{\varphi^{R[b\succ a]}}{\varphi^{R[a\succ b]}} = \varphi^{R[b\succ a]-R[a\succ b]}$.

Algorithm 3 Markov chain $\mathfrak{M}_C$ for Condorcet's model.

1: **Inputs:** a profile R, a prior $\Pr_C$ over $\mathcal{L}(\mathcal{C})$, and the number of iterations N.
2: Let $V \in \mathcal{B}(\mathcal{C})$ denote a binary relation with the maximum likelihood computed by Proposition 2.
3: **for** t =1 **to** N **do**
4: Generate $W \in \mathcal{B}(\mathcal{C})$ where all pairwise comparisons are generated independently such that for any (a, b), $\frac{\Pr(a\succ b)}{\Pr(b\succ a)} = \varphi^{R[b\succ a]-R[a\succ b]}$.
5: With probability $\min\{\frac{\Pr_C(W)}{\Pr_C(V)}, 1\}$ let $V = W$.
6: **end for**
7: **return** V.

Theorem 5. *The mixing time of $\mathfrak{M}_C$ in Algorithm 3 is $O((\ln \frac{d_{max}}{d_{max}-1})^{-1}(\ln d_{max} + m \ln m + \ln \epsilon^{-1}))$.*

Proof: We apply a result by Liu (1996) to prove the upper bound on the variation distance.

Lemma 6 ((Liu, 1996))**.** *For any independent sampler starting at V, we have*

$$\Delta_V(t) \leq \frac{(1-\min_W\{p(W)/\pi(W)\})^t}{2\sqrt{\pi(V)}}$$

For any W, we have $p(W) = \frac{\Pr(R|W)}{\sum_{U\in\mathcal{B}(C)}\Pr(R|U)}$ and $\pi(W) = \Pr(W|R) = \frac{\Pr(R|W)\cdot\Pr(W)}{\sum_{U\in\mathcal{B}(C)}\Pr(R|U)\cdot\Pr(U)}$. Therefore $\frac{p(W)}{\pi(W)} = \frac{1}{\Pr(W)} \cdot \frac{\sum_{U\in\mathcal{B}(C)}\Pr(R|U)\cdot\Pr(U)}{\sum_{U\in\mathcal{B}(C)}\Pr(R|U)} \geq \min_U\{\frac{\Pr(U)}{\Pr(Y)}\} \geq \frac{1}{d_{max}}$. By Lemma 6 we have $\Delta_V(t) \leq \frac{1}{2\sqrt{\pi(V)}} \cdot (1 - \frac{1}{d_{max}})^t$. Therefore, $\tau_V(\epsilon)$ is $O((\ln \frac{d_{max}}{d_{max}-1})^{-1}(\ln \pi(V)^{-1} + \ln \epsilon^{-1}))$. When V maximizes the likelihood, we have $\pi(V) \geq \frac{1}{d_{max}m!}$. Applying Stirling's formula we have $\ln \pi(V)^{-1}$ is $O(\ln d_{max} + m \ln m)$, which proves the theorem. □

5 EXPERIMENTS

Most theoretical results in this paper are based on worst-case analysis. In this section we present some preliminary experimental results to illustrate the efficiency of our algorithms for real-world ranking data.

Dataset: We use the weighted majority graph dataset from Preflib (Mattei and Walsh, 2013) (`www.Preflib.org`). Most of these datasets are collected from political elections. For each WMG, we normalize the weights to $[-1, 1]$ by dividing all weights by the heaviest one. This is without loss of generality because we can set φ appropriately.

All experiments were run on a laptop with Intel i7-4600U processor, 8GB memory, and 256 GB SSD hard drive, running Windows 8.1 (64bit) and Python 2.7.9 (32bit).

5.1 MALLOWS' MODEL

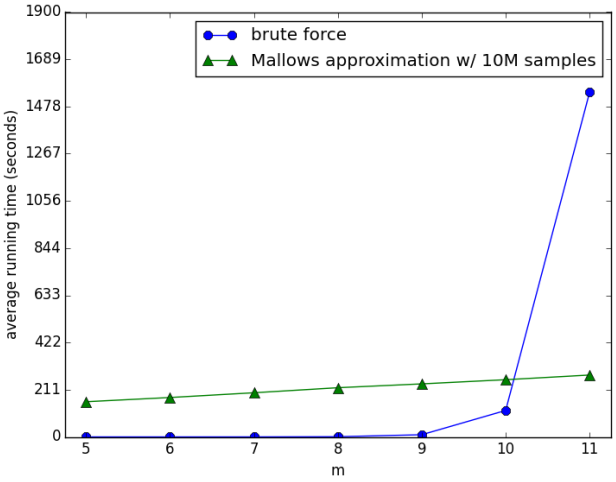
We tested Algorithm 1 with Algorithm 2 for the decision problem with Mallows' model, $\varphi = 0.9$, $\mathcal{D} = \mathcal{C}$, uniform prior, and the exact Top-1 loss function for $m = 5$ through 11. We use brute-force enumeration to compute the optimal decision and discard the first $1/8$ samples in our MCMC algorithm as burn-in. The average running time and convergence of Bayesian loss computed for all datasets with the same number of alternatives are shown in Figure 1.[4]

In Figure 1 (a) we observe that the running time of brute-force search grows exponentially in m (because the number of parameters is $m!$) while the running time for our MCMC algorithm grows linearly in m. Figure 1 (b) shows the reduction of the total difference between the estimated Bayesian loss via MCMC and the ground truth computed by brute-force search w.r.t. the number of samples. We note that this is *not* the variance distance. The average is taken over all datasets with the same number of alternatives. We observe that the total difference can be effectively reduced by increasing the number of samples. Moreover, when we use 10 million samples, the optimal Bayesian decision is correct in 109 out of 115 datasets ($\approx 95\%$) as shown in Table 2.

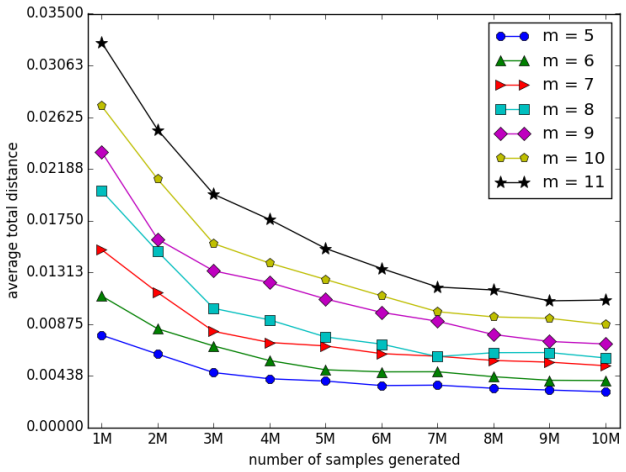
$m =$	5	6	7	8	9	10	11	Total
correct	18	13	18	11	17	20	12	109
incorrect	0	0	2	1	3	0	0	6

Table 2: The number of correct and incorrect Bayesian decisions for Mallows' model.

[4]We have also tested the efficiency of the algorithm with larger N in Algorithm 2, and observe that the efficiency is in general lower than the efficiency when all samples are used.



(a)Average running time.



(b)Average total difference.

Figure 1: The average running time and average total difference of our algorithm for Mallows' model.

5.2 CONDORCET'S MODEL

We tested Algorithm 1 with Algorithm 3 for the decision problem with Condorcet's model, $\varphi = 0.9$, $\mathcal{D} = \mathcal{C}$, uniform prior, and the exact Top-$\lfloor \frac{m}{2} \rfloor$ loss function for $m = 5$ through 11. Given a binary relation $W \in \mathcal{B}(\mathcal{C})$ and an alternative d, the exact Top-k loss can be computed by the following polynomial-time algorithm. We say an alternative a *dominates* another alternative b in W, if there is a directed path from a to b in W. For each alternative c, we first compute the set of all alternatives that dominate c, denoted by D_c. By definition we have $c \in D_c$. Then, the loss of d is 0 if and only if there exists an alternative c such that (1) $|D_c| = k$ and (2) $d \in D_c$; otherwise the loss of d is 1.

By Theorem 4, when k is small there exists a polynomial-time algorithm to compute the Bayesian losses. This makes experiments on Preflib data hard because the algorithm can efficiently compute the optimal Bayesian losses for reasonably large m (for example $m = 20$), and there are not enough datasets with $m > 20$. Therefore, we only show the reduction in average total difference in Figure 2. Extensive experimental studies on real-world datasets are left for future work.

In Figure 2 we observe that (i) for the same number of samples the total difference for Condorcet's model is smaller than the total difference for Mallows' model, and (ii) with the same number of samples, larger m corresponds to smaller total difference. (The total difference for $m = 11$ is too small to be seen clearly in Figure 2.) This may be due to two reasons. First, $\mathfrak{M}_C$ has a better theoretical guarantee (that it is fast mixing) than $\mathfrak{M}_M$. Second, for Condorcet's model the probability for the loss of any decision (alternative) to be 0 w.r.t. the exact Top-$\lfloor \frac{m}{2} \rfloor$ loss function is small, which means that for most generated samples the loss of all alternatives is 1.

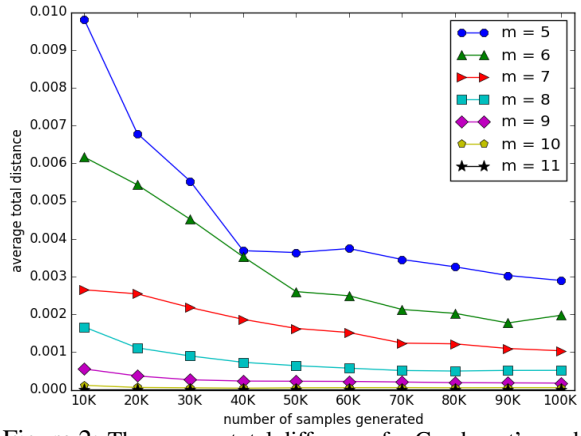


Figure 2: The average total difference for Condorcet's model.

6 SUMMARY AND FUTURE WORK

We have proposed and analyzed two MCMC algorithms for making optimal Bayesian decisions for two popular ranking models w.r.t. any prior and loss function. There are many open questions and future directions. Can we improve the analysis to show that the Markov chain for Mallows' model is rapid mixing or prove that the Bayesian decision problems are hard to approximate by efficient randomized algorithms? Can we design and analyze other Markov chain samplers? How to further improve the performance of the Markov chain sampler in practice? How does the Markov chain approach compare to other popular statistical and machine learning techniques, for example importance sampling?

ACKNOWLEDGMENTS

We thank Zhibing Zhao and anonymous reviewers of UAI-15 for helpful comments. This work is supported in part by NSF CAREER under award number IIS-1453542.

References

Nir Ailon, Moses Charikar, and Alantha Newman. Aggregating inconsistent information: Ranking and clustering. In *Proc. STOC*, pages 684–693, 2005.

Alnur Ali and Marina Meila. Experiments with Kemeny ranking: What works when? *Mathematical Social Sciences*, 64(1):28–40, 2012.

Hossein Azari Soufiani, David C. Parkes, and Lirong Xia. Random utility theory for social choice. In *Proc. NIPS*, pages 126–134, 2012.

Hossein Azari Soufiani, William Chen, David C. Parkes, and Lirong Xia. Generalized method-of-moments for rank aggregation. In *Proc. NIPS*, 2013a.

Hossein Azari Soufiani, Hansheng Diao, Zhenyu Lai, and David C Parkes. Generalized random utility models with multiple types. In *Proc. NIPS*, 2013b.

Hossein Azari Soufiani, David C. Parkes, and Lirong Xia. Statistical decision theory approaches to social choice. In *Proc. NIPS*, 2014.

John Bartholdi, III, Craig Tovey, and Michael Trick. Voting schemes for which it can be difficult to tell who won the election. *Social Choice and Welfare*, 6:157–165, 1989.

Nadja Betzler, Michael R. Fellows, Jiong Guo, Rolf Niedermeier, and Frances A. Rosamond. Fixed-Parameter Algorithms for Kemeny Scores. In *Algorithmic Aspects in Information and Management*, volume 5034 of *Lecture Notes in Computer Science*, pages 60–71, 2008.

Marquis de Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Paris: L'Imprimerie Royale, 1785.

Persi Diaconis and Phil Hanlon. Eigenanalysis for some examples of the Metropolis algorithm. *Contemporary Mathematics*, 138: 99–117, 1992.

Cynthia Dwork, Ravi Kumar, Moni Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Proc. WWW*, pages 613–622, 2001.

Edith Elkind and Nisarg Shah. How to Pick the Best Alternative Given Noisy Cyclic Preferences? In *Proc. UAI*, 2014.

Uriel Feige and Orly Yahalom. On the Complexity of Finding Balanced Oneway Cuts. *Information Processing Letters*, 87 (1):1–5, 2003.

Sumit Ghosh, Manisha Mundhe, Karina Hernandez, and Sandip Sen. Voting for movies: the anatomy of a recommender system. In *Proc. AGENTS*, pages 434–435, 1999.

W. Keith Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57(1):97–109, 1970.

Edith Hemaspaandra, Holger Spakowski, and Jörg Vogel. The complexity of Kemeny elections. *Theoretical Computer Science*, 349(3):382–391, December 2005.

Mark Jerrum and Alistair Sinclair. The Markov chain Monte Carlo method: an approach to approximate counting and integration. In Dorit S. Hochbaum, editor, *Approximation algorithms for NP-hard problems*, pages 482–519. PWS Publishing Company, 1996.

Claire Kenyon-Mathieu and Warren Schudy. How to Rank with Few Errors: A PTAS for Weighted Feedback Arc Set on Tournaments. In *Proc. STOC*, pages 95–103, 2007.

Jen-Wei Kuo, Pu-Jen Cheng, and Hsin-Min Wang. Learning to Rank from Bayesian Decision Inference. In *Proc. CIKM*, pages 827–836, 2009.

Jun S. Liu. Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and Computing*, 6(2):113–119, 1996.

Tie-Yan Liu. *Learning to Rank for Information Retrieval*. Springer, 2011.

Bo Long, Olivier Chapelle, Ya Zhang, Yi Chang, Zhaohui Zheng, and Belle Tseng. Active Learning for Ranking Through Expected Loss Optimization. In *Proc. SIGIR*, pages 267–274, 2010.

Tyler Lu and Craig Boutilier. Learning mallows models with pairwise preferences. In *Proc. ICML*, pages 145–152, 2011.

Colin L. Mallows. Non-null ranking model. *Biometrika*, 44(1/2): 114–130, 1957.

Andrew Mao, Ariel D. Procaccia, and Yiling Chen. Better human computation through principled voting. In *Proc. AAAI*, 2013.

Nicholas Mattei and Toby Walsh. PrefLib: A Library of Preference Data. In *Proc. ADT*, 2013.

David C. McGarvey. A theorem on the construction of voting paradoxes. *Econometrica*, 21(4):608–610, 1953.

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, 21(6):1087–1092, 1953.

Sahand Negahban, Sewoong Oh, and Devavrat Shah. Iterative ranking from pair-wise comparisons. In *Proc. NIPS*, pages 2483–2491, 2012.

Daniele Porello and Ulle Endriss. Ontology Merging as Social Choice: Judgment Aggregation under the Open World Assumption. *Journal of Logic and Computation*, 2013.

Ariel D. Procaccia, Sashank J. Reddi, and Nisarg Shah. A maximum likelihood approach for selecting sets of alternatives. In *Proc. UAI*, 2012.

Karthik Raman and Thorsten Joachims. Methods for Ordinal Peer Grading. In *Proc. SIGKDD*, pages 1037–1046, 2014.

Karthik Raman and Thorsten Joachims. Bayesian Ordinal Peer Grading. In *Proc. L@S*, 2015.

Alistair Sinclair. Improved Bounds for Mixing Rates of Markov Chains and Multicommodity Flow. *Combinatorics, Probability and Computing*, 1(4):351–370, 1992.

Alistair Sinclair and Mark Jerrum. Approximate counting, uniform generation and rapidly mixing Markov chains. *Information and Computation*, 82(1):93–133, 1989.

A. F. M. Smith and G. O. Roberts. Bayesian Computation Via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society. Series B*, 55 (1):3–23, 1993.

H. Peyton Young. Condorcet's theory of voting. *American Political Science Review*, 82:1231–1244, 1988.