Speeding up the FMMR perfect sampling algorithm: A case study revisited

Robert P. Dobrow

Mathematics and Computer Science Department

Carleton College

rdobrow@carleton.edu and
http://www.mathcs.carleton.edu/faculty/bdobrow/

and

James Allen Fill[1]

Department of Mathematical Sciences

The Johns Hopkins University

jimfill@jhu.edu and http://www.mts.jhu.edu/~fill/

ABSTRACT

In a previous paper by the second author, two Markov chain Monte Carlo perfect sampling algorithms—one called coupling from the past (CFTP) and the other (FMMR) based on rejection sampling—are compared using as a case study the move-to-front (MTF) self-organizing list chain. Here we revisit that case study and, in particular, exploit the dependence of FMMR on the user-chosen initial state. We give a stochastic monotonicity result for the running time of FMMR applied to MTF and thus identify the initial state that gives the stochastically smallest running time; by contrast, the initial state used in the previous study gives the stochastically *largest* running time. By changing from worst choice to best choice of initial state we achieve remarkable speedup of FMMR for MTF; for example, we reduce the running time (as measured in Markov chain steps) from exponential in the length n of the list nearly down to n when the items in the list are requested according to a geometric distribution. For this same example, the running time for CFTP grows exponentially in n.

AMS 2000 *subject classifications.* Primary 60J10, 68U20; secondary 60G40, 68P05, 68P10, 65C05, 65C10, 65C40.

Key words and phrases. Perfect simulation, exact sampling, rejection sampling, Markov chain Monte Carlo, FMMR algorithm, Fill's algorithm, move-to-front rule, coupling from the past, Propp–Wilson algorithm, running time, monotonicity, separation, strong stationary time, partially ordered set.

Date. Revised June 4, 2003.

[1]Research for this author supported by NSF grants DMS–9803780 and DMS–0104167, and for both authors by The Johns Hopkins University's Acheson J. Duncan Fund for the Advancement of Research in Statistics.

1 Introduction and summary

Perfect sampling has had a substantial impact on the world of Markov Chain Monte Carlo (MCMC). In MCMC, one is interested in obtaining a sample from a distribution π from which it is computationally difficult (or even infeasible) to simulate directly. One constructs a Markov chain whose stationary distribution is π and after running the chain "a long time" takes an outcome from the chain as an (approximate) observation from π. Propp and Wilson [12] (see also [13] [14] [16]) and Fill [4] have devised algorithms to use Markov chain transitions to produce observations *exactly* from π, without *a priori* estimates of the mixing time of the chain; the applicability of the latter algorithm has recently been extended by Fill, Machida, Murdoch, and Rosenthal [7], and so we will use the terminology "FMMR algorithm." Although the two algorithms are based on different ideas—Propp and Wilson use coupling from the past (CFTP) while FMMR is based on rejection sampling—there is a simple connection between the two, discovered in [7] and reviewed below. For further general discussion of perfect sampling using Markov chains, consult the annotated bibliography maintained on the Web by Wilson [15].

Much of the discussion comparing the two algorithms has focused on the issue of "interruptibility." FMMR has the feature that the output and the running time—when measured in number of Markov chain steps—are independent random variables. Thus, for instance, an impatient user who interrupts a run of the algorithm after any fixed number of steps and restarts the procedure does not introduce any bias into the output. This is not so for CFTP. On the other hand, for many practical applications CFTP is considerably easier to implement, since (see [7]) FMMR requires the user to be able (i) to generate a trajectory from the time-reversal of the basic chain, and (ii) to build couplings "*ex post facto*," i.e., to perform certain imputation steps; CFTP requires neither ability.

Remark 1.1. There *is* a need for time-reversal generation (of an auxiliary chain) and for *ex post facto* coupling in an extension of CFTP known as coupling into and from the past, introduced (under a different name) by Kendall [9]. (See also Section 1.9.3 in [16].)

In this paper we focus on the running time of the two algorithms (but the non-interruptibility of CFTP will turn out to play a key role). In previous case-study comparisons [4] [6], the running times (and memory requirements) have been found to be not hugely different, but CFTP has had the edge. In this paper, by revisiting the case study of [6], we show that, at least in some cases, FMMR can be made to have much smaller running time than CFTP.

The general observation that we exploit—one very closely related to Remark 6.9(c) and Section 8.2 of [7]—is the following. Given a target distribution π, let p_{CFTP} denote the probability that CFTP terminates successfully (coalesces) over a fixed time window (and outputs a sample from π). Similarly, let $p_{\mathrm{CFTP}}(z)$ denote the conditional probability of coalescence over the time window, given that the state (call it Z_{CFTP}) ultimately output by CFTP (after extending the time window into the indefinite past) is z. Let $p_{\mathrm{FMMR}}(z)$ denote the conditional probability that FMMR terminates successfully over the same time window, given that the initial state (call it Z_{FMMR}) is z.

Then, as we show in Theorem 2.3, $p_{\mathrm{CFTP}}(z) \equiv p_{\mathrm{FMMR}}(z)$. That is (now letting the time window vary), if T_{CFTP} and T_{FMMR} denote the respective running times of CFTP and FMMR, then conditional running time distributions agree:

$$\mathcal{L}(T_{\mathrm{CFTP}} \,|\, Z_{\mathrm{CFTP}} = z) \equiv \mathcal{L}(T_{\mathrm{FMMR}} \,|\, Z_{\mathrm{FMMR}} = z),$$

where $\mathcal{L}(X)$ denotes the distribution (law) of the random variable X. As a consequence, $p_{\mathrm{CFTP}} = \mathbf{E}_\pi\,[p_{\mathrm{FMMR}}(Z_{\mathrm{FMMR}})]$; that is, $\mathcal{L}(T_{\mathrm{CFTP}})$ is the π-mixture of the distributions $\mathcal{L}(T_{\mathrm{FMMR}} \,|\, Z_{\mathrm{FMMR}} = z)$.

The important point here is that, except in the rare instance that CFTP is interruptible for the chain of interest (i.e., that T_{CFTP} and Z_{CFTP} are independent), for at least one time window there must exist at least one initial state z for which $p_{\mathrm{FMMR}}(z) > p_{\mathrm{CFTP}}$.

The move-to-front (MTF) process is a nonreversible Markov chain on the permutation group $\mathcal{S}_n$. The two algorithms have been compared for MTF in a previous paper [6]. In that paper, the initial state for FMMR was taken to be the identity permutation, and it was then found, roughly speaking (see Table 1 and Section 5 therein), that T_{CFTP} and T_{FMMR} are of the same size. In this paper, we will revisit that case study and establish a stochastic monotonicity result for $\mathcal{L}(T_{\mathrm{CFTP}} \,|\, Z_{\mathrm{CFTP}} = z)$ in z. It turns out, in particular, that the identity permutation is the *worst* choice of initial state! When we choose instead the reversal permutation, which is the best choice, we obtain a (sometimes huge) speedup for FMMR. (See Table 1, which will be explained more fully in Section 4. Notice that for geometric weights, the change in starting state reduces T_{FMMR} from exponential in n to about n.) The gains obtained by using the optimal z are sufficiently dramatic that, when measured in Markov chain steps, the resulting *worst-case* running time for FMMR (worst over choice of request weights) equals the *best-case* running time for CFTP: see Remark 4.2(b).

We temper our enthusiasm, however, by recognizing that our MTF example is somewhat artificial on two counts. Firstly, as discussed in the introduction to [6], there are algorithms for sampling from the MTF stationary distribution which are both more elementary (in particular, not involving Markov chains) and more efficient than either CFTP or FMMR. So we do not recommend applying either CFTP or FMMR to MTF in practice. Our goal here is to illustrate how judicious choice of starting state for FMMR can greatly improve its performance.

Secondly, MTF has the (evidently rare) property that one can obtain an exact analysis of the running time distribution for FMMR for *every* choice of initial state z. We do not yet know whether our speedup ideas help in any real applications. We hope, however, that the ideas in this paper will stimulate further research on FMMR by pointing to the possibility of speedup of the algorithm. The wise user might at minimum wish to collect data on running times of FMMR applied to the problem of interest, experimenting with a wide variety of initial states, before carrying out the bulk of the simulations.

We briefly review the two perfect sampling algorithms and their general connection in Section 2. The move-to-front rule is reviewed in Section 3. Our new results are given in Section 4. A somewhat different approach to speeding up FMMR is given in Section 5.

2 Perfect sampling

We briefly review the CFTP and FMMR algorithms (omitting a few of the finer measure-theoretic details, which are irrelevant anyway for finite-state chains). We assume that our Markov chain $\mathbf{X}$ can be written in the *stochastic recursive sequence* form

$$\mathbf{X}_s = \phi(\mathbf{X}_{s-1}, \mathbf{U}_s), \tag{2.1}$$

where ϕ is called the *transition rule* and $(\mathbf{U}_s)$ is an i.i.d. sequence. We further assume that our Markov chain has finite state space $\mathcal{X}$ and is irreducible and aperiodic with (unique) stationary distribution π.

2.1 CFTP

For a fixed positive integer t, and a Markov chain with n states, start n copies of the chain at time $-t$ from each of the n states, coupling the transitions by means of the transition rule ϕ, and running the chains until time 0. If all copies of the chain agree at time 0, we say that the trajectories have *coalesced* and return the common value, say $\mathbf{Z}$. If the chains do not agree, then increment t and restart the procedure, using for common values of s the same values of $\mathbf{U}_s$ used in the previous step; again, check for coalescence. The *running time* of the algorithm we define to be the smallest integer t for which coalescence occurs. If we assume the algorithm terminates with probability 1, then $\mathbf{Z} \sim \pi$ exactly.

There is a rich source of papers, primers, and applications of CFTP. The best initial reference is the "Perfectly random sampling with Markov chains" Web site maintained by David Wilson at `http://www.dbwilson.com/exact/`.

2.2 FMMR

Given a Markov chain with transition matrix K, recall that the *time-reversal* chain has transition matrix $\tilde{K}$ which satisfies

$$\pi(x)K(x,y) = \pi(y)\tilde{K}(y,x) \text{ for all } x, y.$$

The FMMR algorithm has two stages: First, choose an initial state $\mathbf{X}_0$. Run the time-reversed chain $\tilde{K}$, obtaining $\mathbf{X}_0, \mathbf{X}_{-1}, \ldots$ in succession. Then (conditionally given the $\mathbf{X}$-values) generate $\mathbf{U}_0, \mathbf{U}_{-1}, \ldots$ independently, with $\mathbf{U}_s$ chosen from its conditional distribution given (2.1) for $s = 0, -1, \ldots$. (One says that the values $\mathbf{U}_s$ are *imputed*.) For $t = 0, 1, \ldots$, and for each state x in the state space, set $\mathbf{Y}_{-t}^{(-t)}(x) := x$ and, inductively,

$$\mathbf{Y}_s^{(-t)}(x) := \phi(\mathbf{Y}_{s-1}^{(-t)}(x), \mathbf{U}_s), \quad -t+1 \le s \le 0.$$

We will sometimes refer to the realization of the chain $\mathbf{X}$ as the *backward* trajectory, and to the realizations of the chains $\mathbf{Y}(x)$ as the *forward* trajectories. The *running time* of the algorithm we define to be the smallest t^* such that $\mathbf{Y}_0^{(-t^*)}(x)$ agree for every x in the state space (and hence all equal $\mathbf{X}_0$). In this case the algorithm reports $\mathbf{X}_{-t^*}$ as an observation from π.

Remark 2.1. The algorithms are presented here in their most general, "vanilla" versions. A large amount of research has gone into improving both algorithms and tailoring them for specific applications. In particular, to improve performance a "doubling trick" is suggested for both algorithms whereby instead of incrementing t by one at each step, t is successively doubled. Since this affects the number of Markov chain steps taken only by constant factors, we shall for our theoretical analysis stick to the "vanilla" versions.

Remark 2.2. For most chains of interest, the state space is very large and the implementations presented here (running copies of the chain from every state in the state space) are not feasible. However, for a large class of cases where a form of monotonicity holds, the algorithms become practical.

Given a Markov chain with transition matrix K, we say that we are in the *(realizably) monotone case* if the following conditions hold. The state space is a partially ordered set $(\mathcal{X}, \leq)$. There exist (necessarily unique) minimum and maximum elements in the state space, denoted $\hat{0}$ and $\hat{1}$, respectively. There exists a *monotone transition rule* ϕ for the chain. Such a rule is a function $\phi : \mathcal{X} \times \mathcal{U} \to \mathcal{X}$ together with a random variable $\mathbf{U}$ taking values in a probability space $\mathcal{U}$ such that (i) $\phi(\mathbf{x}, \mathbf{u}) \leq \phi(\mathbf{y}, \mathbf{u})$ for all $\mathbf{u} \in \mathcal{U}$ whenever $\mathbf{x} \leq \mathbf{y}$; and (ii) for each $\mathbf{x} \in \mathcal{X}$, $P(\phi(\mathbf{x}, \mathbf{U}) = \mathbf{y}) = K(\mathbf{x}, \mathbf{y})$ for all $\mathbf{y} \in \mathcal{X}$.

When in the monotone case, for CFTP one only needs to follow two trajectories of the chain, one started at time $-t$ from $\hat{0}$ and the other from $\hat{1}$, since all other trajectories are sandwiched between these. Likewise, in the second phase of FMMR, one only needs to run the $\mathbf{Y}$-chain from states $\hat{0}$ and $\hat{1}$.

Although the two algorithms are based on different conceptual underpinnings, our first theorem highlights an important connection between them. Roughly, the distribution of the running time for CFTP is equal to the stationary mixture, over initial states, of the distributions of the running time for FMMR. This is given as Remark 6.9(c) in [7], but we wish to emphasize its importance and so recast it as a theorem. We recall our notation from Section 1. For a fixed time window, let $p_{\mathrm{CFTP}}(z)$ denote the probability that CFTP coalesces given that the state (call it Z_{CFTP}) ultimately output by CFTP is z, and let p_{CFTP} denote the corresponding *unconditional* probability. Let $p_{\mathrm{FMMR}}(z)$ denote the conditional probability that FMMR coalesces given that the initial state (call it Z_{FMMR}) is z. Let T_{CFTP} and T_{FMMR} denote the respective running times of CFTP and FMMR (now letting the time window vary).

Theorem 2.3. *We have*

$$p_{\mathrm{CFTP}}(z) = p_{\mathrm{FMMR}}(z) \text{ for } \pi\text{-almost every } z; \tag{2.2}$$

$$p_{\mathrm{CFTP}} = \mathbf{E}_\pi[p_{\mathrm{FMMR}}(Z_{\mathrm{FMMR}})]; \tag{2.3}$$

$$\mathcal{L}(T_{\mathrm{CFTP}} | Z_{\mathrm{CFTP}} = z) = \mathcal{L}(T_{\mathrm{FMMR}} | Z_{\mathrm{FMMR}} = z) \text{ for } \pi\text{-almost every } z; \tag{2.4}$$

and

$$\mathcal{L}(T_{\mathrm{CFTP}}) \text{ is the } \pi\text{-mixture (over } z) \text{ of } \mathcal{L}(T_{\mathrm{FMMR}} \,|\, Z_{\mathrm{FMMR}} = z). \tag{2.5}$$

The result holds in the most general setting, not restricted either to finite-state chains or to monotone transition rules. It is a consequence of the discussion in Sections 6.2 and 8.2 of [7]. For the reader's convenience we give here a simple proof for the discrete case.

Proof. Each iteration of FMMR is an implementation of rejection sampling (see, e.g., Devroye [2] for background). The goal is to use an observation from $\tilde{K}^t(z,\cdot)$ to simulate one from π. One obtains an upper bound c on $\max_x \pi(x)/\tilde{K}^t(z,x)$, generates x with probability $\tilde{K}^t(z,x)$, and accepts x as an observation from π with probability $c^{-1}\pi(x)/\tilde{K}^t(z,x)$. The unconditional probability of acceptance is then $1/c$. Observe that, for every x,

$$\frac{\pi(x)}{\tilde{K}^t(z,x)} = \frac{\pi(z)}{K^t(x,z)} \le \frac{\pi(z)}{P(\text{coalescence to } z)} := c,$$

where "coalescence to z" refers, of course, to coalescence over the given time window of length t. Thus for the desired conditional acceptance probability given x we can use

$$\frac{P(\text{coalescence to } z)}{K^t(x,z)} = P(\text{coalescence to } z \,|\, \text{trajectory from } x \text{ ends at } z),$$

and the FMMR algorithm is designed precisely to implement this. Thus $p_{\mathrm{FMMR}}(z) = 1/c$ and hence

$$p_{\mathrm{FMMR}}(z) = \frac{P(\text{coalescence to } z)}{\pi(z)} = p_{\mathrm{CFTP}}(z). \tag{2.6}$$

Thus, (2.2) is immediate. Taking expectations with respect to π gives (2.3). And (2.4) [from which (2.5) is immediate] follows from (2.2) since, for a fixed time window of length t, $p_{\mathrm{FMMR}}(z)$ [respectively, $p_{\mathrm{CFTP}}(z)$] is the value at t of the conditional distribution function of T_{FMMR} given that $Z_{\mathrm{FMMR}} = z$ (respectively, of T_{CFTP} given that $Z_{\mathrm{CFTP}} = z$). □

Corollary 2.4. *If T_{CFTP} and Z_{CFTP} are not independent random variables, then there exist at least one time window and at least one initial state z for which $p_{\mathrm{FMMR}}(z) > p_{\mathrm{CFTP}}$.*

Proof. This is immediate from (2.3). □

The following simple examples are artificial, but they give a first demonstration that judicious choice of starting state can lead to dramatic speedup. First, consider a three-state Markov chain with states labeled 0, 1, and 2. Let

$$K = \begin{bmatrix} \epsilon & (1-\epsilon)/2 & (1-\epsilon)/2 \\ \epsilon & 1-\epsilon & 0 \\ \epsilon & 0 & 1-\epsilon \end{bmatrix},$$

where $\epsilon > 0$ is small. One checks that this chain is reversible. Let $\mathbf{U} = 0, 1, 2$ with respective probabilities $(1-\epsilon)/2, \epsilon, (1-\epsilon)/2$ and use the monotone transition rule

$$\phi(x,1) = 1 \text{ for all } x, \quad \phi(1,0) = \phi(0,0) = \phi(0,2) = 0, \quad \phi(1,2) = \phi(2,0) = \phi(2,2) = 2.$$

Coalescence occurs over a given time window of length t if and only if the value of some $\mathbf{U}_s$ in that window is 1; thus $p_{\mathrm{CFTP}} = 1 - (1-\epsilon)^t$, which requires t of order $1/\epsilon$ to become nonnegligible. On the other hand, if FMMR is started in state 1, then with high probability ($= 1 - \epsilon$) we'll see (going backward in time) one of the transitions $0 \leftarrow 1$ or $2 \leftarrow 1$. If we do, then (whichever we see) in the forward phase we impute $\mathbf{U} = 1$ and hence get coalescence (to state 1) in one step.

For our second example, consider a Gibbs sampler on an attractive spin system with n sites arranged in a row and left-to-right site-update sweeps. (Consult, e.g., [4] or [10] for background on attractive spin systems.) This gives a monotone, nonreversible chain where $\hat{0}$ is the state consisting of all $-$'s and $\hat{1}$ is the state of all $+$'s. Suppose that we deal with an Ising model where the Gibbs distribution is such that there is (i) a strong external field for spin $+$ at sites 1 through $n-1$, (ii) a much stronger effect of nearest-neighbor attractiveness throughout the system, and (iii) a very much stronger yet external field for spin $+$ at site n (the rightmost site). First consider CFTP. The state $\hat{1}$ is a state of very high probability and so the chain won't budge out of that state for a long time. On the other hand, from $\hat{0}$, in one sweep (a full left-to-right update), we obtain [with high probability, because of (ii) and (iii)] $--\cdots--+$. In the next sweep we obtain [with high probability, because of (ii) and (i)] $--\cdots-++$. Continuing, in about n sweeps we obtain $++\cdots++$; that is, with high probability we obtain coalescence in n sweeps. On the other hand, consider FMMR started in $\hat{0}$. Here, the reversed chain is Gibbs sampling with right-to-left updates. The reversed chain, started in $\hat{0}$, [with high probability, because of (iii) and then (ii)] flips each site from $-$ to $+$ as it moves from right to left. Hence, we obtain $++\cdots+++$; that is, with high probability there is coalescence *in one sweep*.

(Of course, were we to use right-to-left sweeps or reversible sweeps as the sampler, the relative disadvantage of CFTP would disappear.)

Remark 2.5. In general, we know of no simpler expression for $p_{\mathrm{FMMR}}(z)$ than the ratio in (2.6). In the monotone case, however, when $z = \hat{0}$ or $z = \hat{1}$ we obtain significant simplification. Indeed, then

$$p_{\mathrm{FMMR}}(\hat{0}) = \frac{K^t(\hat{1},\hat{0})}{\pi(\hat{0})} = \min_z \frac{K^t(z,\hat{0})}{\pi(\hat{0})} \quad \text{and} \quad p_{\mathrm{FMMR}}(\hat{1}) = \frac{K^t(\hat{0},\hat{1})}{\pi(\hat{1})} = \min_z \frac{K^t(z,\hat{1})}{\pi(\hat{1})}.$$

Recall that for a Markov chain with transition matrix K and stationary distribution π, the *separation* at time t given that the chain starts in state x is

$$\mathrm{sep}_x(t) := 1 - \min_z \frac{K^t(x,z)}{\pi(z)}.$$

Thus, $p_{\mathrm{FMMR}}(z) = 1 - \widetilde{\mathrm{sep}}_z(t)$ for $z = \hat{0}, \hat{1}$, where $\widetilde{\mathrm{sep}}$ refers to separation for the transition matrix $\tilde{K}$. See (e.g.) [1] for more on separation.

3 Move-to-front

Let $(w_1, \ldots, w_n)$ be a probability mass function on $\{1, \ldots, n\}$ with $w_i > 0$ for each i. In this study we are concerned with generating an observation from the distribution

$$\pi(z) := \prod_{r=1}^{n} \frac{w_{z_r}}{\sum_{j=r}^{n} w_{z_j}}, \quad z \in \mathcal{S}_n, \tag{3.1}$$

where $\mathcal{S}_n$ is the group of permutations of $\{1, \ldots, n\}$. Consider sampling without replacement from a population of n items, where item i has probability w_i of being chosen, $1 \le i \le n$. Then the probability of drawing the n items in the order z is given by (3.1).

This distribution arises as the limiting distribution of the much-studied move-to-front (MTF) process. The MTF heuristic is used to "self-organize" a linear list of data records in a computer file. Let $\{1, \ldots, n\}$ be a set of records (or rather the set of *keys*, or identifying labels for the records), where record i has probability w_i of being requested. At discrete units of time, and independent of past requests, item i is requested (with probability w_i) and brought to the front of the list, leaving the relative order of the other records unchanged. The successive orders of the list of records forms an ergodic Markov chain on the permutation group $\mathcal{S}_n$ with stationary distribution π.

We will assume that the records have been labeled so that $w_1 \ge \cdots \ge w_n > 0$ and refer to $\mathbf{w} = (w_1, \ldots, w_n)$ as the *weight vector* of the chain. For extensive treatment of MTF, see [5], which contains pointers to the sizable literature on the subject. Hendricks [8] was the first to show that the stationary distribution of the MTF Markov chain is given by (3.1).

Fill [6] used MTF as a case study to compare CFTP and FMMR. We omit many details but for completeness describe the set-up briefly. Partially order the symmetric group $\mathcal{S}_n$ by declaring $z \le z'$ if z' can be obtained from z by a sequence of adjacent transpositions which switch records out of order (that begin in natural order). This is the *weak Bruhat order*. With

$$\hat{0} := \mathrm{id} = (1, 2, \ldots, n), \qquad \hat{1} := \mathrm{rev} = (n, n-1, \ldots, 1),$$

we have $\hat{0} \le z \le \hat{1}$ for all $z \in \mathcal{S}_n$. (For the definition of the *Bruhat order*, used later, delete the word "adjacent.") The MTF chain possesses the following monotone transition rule with respect to the weak Bruhat order. Let U be a random variable satisfying $P(U = i) = w_i$ for $1 \le i \le n$. Define

$$\phi(z, i) = \mathrm{move}_i(z) \text{ for } z \in \mathcal{S}_n \text{ and } 1 \le i \le n,$$

where $\mathrm{move}_i(z)$ is defined to be the permutation resulting from the list z by requesting record i and applying the MTF rule. It is easily checked (see Lemma 2.2 in [6]) that this gives a monotone transition rule for M.

MTF, of course, is not a reversible Markov chain; however, it is relatively straightforward to generate transitions from the time-reversed chain. We refer the reader to [6] for further details on implementing MTF both using CFTP and using FMMR.

Our first result (Theorem 3.1) exhibits explicitly the dependence of T_{FMMR} on the initial state Z_{FMMR}. In what follows, given $z \in \mathcal{S}_n$, let $y_r := w_{z_r}$ for $1 \le r \le n$. In this notation, (3.1) can be written in the form

$$\pi(z) = \prod_{r=1}^{n} \frac{y_r}{1 - y_{r-1}^+},$$

where we have also introduced the notation

$$y_r^+ := \sum_{j=1}^{r} y_j, \quad 0 \le r \le n,$$

for any vector $(y_1, \ldots, y_n)$.

Theorem 3.1. (a) *The conditional distribution of* $\mathcal{L}(T_{\mathrm{FMMR}})$ *given the initial state* Z_{FMMR} *satisfies*

$$\mathcal{L}(T_{\mathrm{FMMR}} \,|\, Z_{\mathrm{FMMR}} = z) = \mathcal{L}(T_z),$$

where the law of T_z *is the convolution of* $\mathrm{Geometric}(1 - y_r^+)$ *distributions,* $0 \le r \le n-2$. *We write*

$$T_z \sim \oplus_{r=0}^{n-2} \mathrm{Geom}(1 - y_r^+).$$

(b) *The random variables* T_z *decrease stochastically in the Bruhat order for* z.

(c) *The distribution* $\mathcal{L}(T_z)$ *is stochastically minimized (respectively, maximized) by choosing* $z = \mathrm{rev}$ *(resp.,* $z = \mathrm{id}$*). In that case we find*

$$T_{\mathrm{rev}} \sim \oplus_{r=2}^{n} \mathrm{Geom}(w_r^+) \quad [\text{resp., } T_{\mathrm{id}} \sim \oplus_{r=0}^{n-2} \mathrm{Geom}(1 - w_r^+)]. \tag{3.2}$$

Proof. Part (a) is a consequence of (2.4) in our Theorem 2.3 and Lemma 3.7 in [6]; indeed, that lemma states that $\mathcal{L}(T_{\mathrm{CFTP}} \,|\, Z_{\mathrm{CFTP}} = z) = \oplus_{r=0}^{n-2} \mathrm{Geom}(1 - y_r^+)$. For the *weak* Bruhat order, Lemma 3.9 in [6] gives part (b); but one need only compare $\mathcal{L}(T_z)$ and $\mathcal{L}(T_{z'})$ when z and z' differ by any transposition to see that part (b) holds for the Bruhat order. Part (c) is an immediate consequence of part (b). □

Remark 3.2. Theorem 3.1 for the special case of MTF belies the general Remark 2.5. Indeed, for *every* initial state for FMMR, we know exactly the distribution of T_{FMMR}. The *worst* starting state for coalescence is the identity permutation, and the result for $\mathcal{L}(T_{\mathrm{id}})$ in Theorem 3.1(c) recaptures Theorem 4.2 in [6]. The comparison of FMMR and CFTP in [6] was based on starting FMMR in this worst state. In the next section we will discuss how much speedup can be achieved by instead starting in the *best* state, the permutation rev.

4 Speedup results for MTF

4.1 General weight vectors

From now on, we abbreviate T_{rev} of (3.2) as T. We first consider how $\mathcal{L}(T)$ varies with the weight vector $\mathbf{w}$. For an understanding of the terminology used in Theorem 4.1 and a thorough treatment of majorization, see [11].

Theorem 4.1. *The distribution* $\oplus_{r=2}^{n}\mathrm{Geom}(w_r^+)$ *of* T *is stochastically strictly Schur-concave in the weight vector* $\mathbf{w}$*. In particular, the distribution is stochastically maximized (respectively, minimized), over all vectors* $\mathbf{w}$ *with* $w_1 \geq w_2 \geq \cdots \geq w_n \geq 0$*, at the uniform distribution* $\mathbf{w} = (1/n, \ldots, 1/n)$ *(resp., at any distribution* $\mathbf{w}$ *with* $w_1 + w_2 = 1$*).*

Proof. The result follows simply from the fact that the Geometric(p) distribution is stochastically strictly decreasing in p. □

Remark 4.2. (a) The possibility $w_1 + w_2 = 1$ is ruled out for MTF (for $n > 2$) by our assumption that all weights are positive. Nevertheless it is a limiting case. In this limiting case, $T = n - 1$ with probability 1. At the other extreme of uniform weights, asymptotics for $\mathcal{L}(T)$ are well known (since this is a slight modification of the standard coupon collector's problem). (The distribution of $\mathcal{L}(T)$ for uniform weights is treated in detail in Theorem 4.3(a) and Section 4.2 of [5]. Roughly put, the distribution of T is concentrated tightly about $n \ln n$.) Thus, for *any* sequence $\mathbf{w}^{(1)} = (w_{11}), \mathbf{w}^{(2)} = (w_{21}, w_{22}, \ldots), \ldots$ of weight vectors, writing $T \equiv T_n$ for the T corresponding to weight vector $\mathbf{w}^{(n)}$ we have

$$P(T \geq n - 1) = 0 \qquad \text{and} \qquad \liminf_{c \to -\infty} \liminf_{n \to \infty} P(T \leq n \ln n + cn) = 1.$$

So the distribution of T is always tightly sandwiched between n and about $n \ln n$, in sharp contrast (cf. Table 1 of [6] or Table 1 below) to the distribution of T_{id} or of T_{CFTP}.

(b) According to Remark 2.6 and the sentence following (3.2) in [6], $\mathcal{L}(T_{\mathrm{CFTP}})$ is strictly Schur-*convex* in $\mathbf{w}$. In particular, the *best-case* $\mathcal{L}(T_{\mathrm{CFTP}})$, corresponding to uniform weights, equals the *worst-case* $\mathcal{L}(T)$, also corresponding to uniform weights.

4.2 Specific examples of weight vectors

We now measure quantitatively, for certain standard examples of weight vectors, the speedup gained for FMMR by using the best choice of initial permutation, rev. Given a triangular array of weights $\mathbf{w}^{(n)} = (w_{ni}, i = 1, \ldots, n)$, $n \geq 1$, we say that k_n steps are necessary and sufficient for convergence of $\mathcal{L}(T)$ to mean that

$$\frac{T_n}{k_n} \to 1 \quad \text{in probability.}$$

Here T_n denotes T_{rev} for the weight vector $\mathbf{w}^{(n)}$; when there is no danger of confusion, we will sometimes drop the subscript n.

For some examples of $\mathbf{w}^{(n)}$ we can obtain results of sharper form than provided by "k_n steps are necessary and sufficient". However, for simplicity and for uniformity of presentation, we stick to the above definition.

Let $H_n^{(\alpha)} := \sum_{i=1}^{n} i^{-\alpha}$ for $\alpha > 0$, and let $\zeta(\alpha) := \sum_{i=1}^{\infty} i^{-\alpha}$, $\alpha > 1$, denote the Riemann zeta function. We consider the following choices of weights, where (now suppressing dependence on n in our notation) each weight vector $\mathbf{w}$ is listed up to a constant of proportionality. The numbers of steps necessary and sufficient for convergence of $\mathcal{L}(T)$ for these examples are stated in Theorem 4.3 and collected in Table 1. The second and

third columns of Table 1 are taken from [6]. [In these columns, the meaning of "ck_n steps are necessary and sufficient" is that, for some h and H,

$$h(c) \le \liminf_{n\to\infty} P(T_n \le \lfloor ck_n \rfloor) \le \limsup_{n\to\infty} P(T_n \le \lfloor ck_n \rfloor) \le H(c),$$

where $0 < h(c) \le H(c) < 1$ for all $c \in (0,\infty)$, $h(c) \to 0$ as $c \to 0$, and $H(c) \to 1$ as $c \to \infty$.] The fourth column in Table 1 is the content of our next theorem.

Weights	$w_i \propto$
Uniform	1
Zipf's law	i^{-1}
Generalized Zipf's law (GZL)	$i^{-\alpha}$, $\alpha > 0$ fixed
Power law	$(n-i+1)^s$, $s > 0$ fixed
Geometric	θ^i, $0 < \theta < 1$ fixed

Table 1. Rates of convergence for $\mathcal{L}(T_{\mathrm{CFTP}})$ and $\mathcal{L}(T_{\mathrm{FMMR}})$.

Weights	$\mathcal{L}(T_{\mathrm{CFTP}})$	$\mathcal{L}(T_{\mathrm{FMMR}})(worst)$	$\mathcal{L}(T_{\mathrm{FMMR}})(best)$
Uniform	$n \ln n$	$n \ln n$	$n \ln n$
Zipf's law	$n(\ln n)^2$	$n(\ln n)^2$	n
GZL			
$0 < \alpha < 1$	$\frac{n}{1-\alpha} \ln n$	$\frac{n}{1-\alpha} \ln n$	$\frac{n}{\alpha}$
$\alpha > 1$	$\zeta(\alpha) n^\alpha \ln n$	$\zeta(\alpha) n^\alpha \ln n$	n
Power law	cn^{s+1}	cn^{s+1}	$\frac{n \ln n}{s+1}$
Geometric	$c\theta^{-n}$	$c\theta^{-n}$	n

Theorem 4.3. (a) (Uniform weights.) *If $w_i = 1/n$ for all i, then $n \ln n$ steps are necessary and sufficient.*

(b) (Zipf's law.) *If $w_i = (H_n i)^{-1}$, with $H_n := H_n^{(1)} = \sum_{k=1}^n k^{-1}$, then n steps are necessary and sufficient.*

(c) (Generalized Zipf's law.) *When $w_i = \left(i^\alpha H_n^{(\alpha)}\right)^{-1}$,* (i) *if $0 < \alpha < 1$, then n/α steps are necessary and sufficient, and* (ii) *if $\alpha > 1$, then n steps are necessary and sufficient.*

(d) (Power law.) *Fix $s > 0$. If $w_i = (n-i+1)^s / f(n,s)$, with $f(n,s) := \sum_{j=1}^n j^s$, then $\frac{n \ln n}{s+1}$ steps are necessary and sufficient.*

(e) (Geometric weights.) *Fix $0 < \theta < 1$. If $w_i = (1-\theta)\theta^{i-1}$ for $i = 1, \ldots, n-1$ and $w_n = \theta^{n-1}$, then n steps are necessary and sufficient.*

Proof. We shall ignore the trivialities induced by the need to consider integer parts in various arguments, leaving these to the meticulous reader.

(a) (Uniform weights.) The asymptotics here are well known, as this is essentially the standard coupon collector's problem. A very sharp asymptotic result is that

$$P(T > \lfloor n \ln n + cn \rfloor) \to 1 - (1 + e^{-c}) e^{-e^{-c}}, \qquad c \in \mathbf{R}.$$

A thorough treatment of the uniform-weights case is provided by Diaconis et al. [3].

We establish the remaining results [as we could also have established (a)] by showing, in each case, (i) that the number of steps k_n claimed to be necessary and sufficient is the lead order term of $\mathbf{E}[T]$, that is, that $\mathbf{E}[T_n] \sim k_n$ as $n \to \infty$, and (ii) that the standard deviation of T_n is $o(\mathbf{E}[T_n])$. The result then follows by application of Chebyshev's inequality. Showing (ii) for each of the weight examples is easy since

$$\begin{aligned} \mathbf{Var}[T] &= \sum_{r=2}^{n} \frac{1-w_r^+}{(w_r^+)^2} = \sum_{r=2}^{n} \frac{1}{(w_r^+)^2} - \mathbf{E}[T] \\ &\le \frac{1}{w_2^+} \sum_{r=2}^{n} \frac{1}{w_r^+} - \mathbf{E}[T] = \mathbf{E}[T] \left(\frac{1}{w_2^+} - 1 \right), \end{aligned}$$

and it is easy to check in each case that $1/w_2^+ = o(\mathbf{E}[T])$. The remainder of the proof thus consists of showing (i). In each case we give explicit upper and lower bounds for $\mathbf{E}[T]$.

(b) (Zipf's law weights.) Here

$$\begin{aligned} n-1 \le \mathbf{E}[T] &= H_n \sum_{r=2}^{n} (H_r)^{-1} \le (\ln n + 1) \sum_{r=2}^{n} \frac{1}{\ln(r+1)} \\ &\le (\ln n + 1) \int_2^{n+1} \frac{dx}{\ln x}. \end{aligned} \tag{4.1}$$

Observe that

$$\int_2^{n+1} \frac{dx}{\ln x} = \frac{n+1}{\ln(n+1)} - \frac{2}{\ln 2} + \int_2^{n+1} \frac{dx}{(\ln x)^2},$$

and that

$$\int_2^{n/(\ln n)^2} \frac{dx}{(\ln x)^2} \le \frac{n/(\ln n)^2}{(\ln 2)^2}$$

and

$$\int_{n/(\ln n)^2}^{n+1} \frac{dx}{(\ln x)^2} \le \frac{1}{\ln[n/(\ln n)^2]} \int_{n/(\ln n)^2}^{n+1} \frac{dx}{\ln x} \le \frac{1}{\ln[n/(\ln n)^2]} \int_2^{n+1} \frac{dx}{\ln x}.$$

Thus,

$$\int_2^{n+1} \frac{dx}{\ln x} \le \frac{n+1}{\ln(n+1)} - \frac{2}{\ln 2} + \frac{n}{(\ln n)^2 (\ln 2)^2} + \frac{1}{\ln[n/(\ln n)^2]} \int_2^{n+1} \frac{dx}{\ln x},$$

i.e.,

$$\begin{aligned} \int_2^{n+1} \frac{dx}{\ln x} &\le \left[1 - \frac{1}{\ln[n/(\ln n)^2]} \right]^{-1} \left[\frac{n+1}{\ln(n+1)} + \frac{n}{(\ln n)^2 (\ln 2)^2} - \frac{2}{\ln 2} \right] \\ &= \frac{n+1}{\ln(n+1)} + O\left(\frac{n}{(\ln n)^2} \right). \end{aligned}$$

Continuing now from (4.1) we have

$$\mathbf{E}[T] \le (\ln n + 1)\left(\frac{n+1}{\ln(n+1)}\right) + O\left(\frac{n}{\ln n}\right) = n + O\left(\frac{n}{\ln n}\right).$$

(c) (Generalized Zipf's law.) For $0 < \alpha < 1$,

$$\begin{aligned}
\mathbf{E}[T] &= H_n^{(\alpha)} \sum_{r=2}^{n} \frac{1}{H_r^{(\alpha)}} \le (n-1)^{1-\alpha} \sum_{r=2}^{n} \frac{1}{(r+1)^{1-\alpha} - 1} \\
&\le n^{1-\alpha} \sum_{r=3}^{n+1} \frac{1}{r^{1-\alpha} - 1} = n^{1-\alpha} \sum_{r=3}^{n+1} \frac{1}{r^{1-\alpha}} \left(1 + O\left(r^{\alpha-1}\right)\right) \\
&= n^{1-\alpha} \sum_{r=3}^{n+1} \left(r^{\alpha-1} + O(r^{2\alpha-2})\right) = n^{1-\alpha} \sum_{r=3}^{n+1} r^{\alpha-1} + O(c_n) \\
&= \frac{n}{\alpha} + o(n),
\end{aligned}$$

where c_n is defined as $n^{1-\alpha}$ if $0 < \alpha < 1/2$, as $n^{1/2} \ln n$ if $\alpha = 1/2$, and as $n^{-\alpha}$ if $1/2 < \alpha < 1$. Also,

$$\begin{aligned}
\mathbf{E}[T] &\ge \left((n+1)^{1-\alpha} - 1\right) \sum_{r=2}^{n} \frac{1}{(r-1)^{1-\alpha}} \\
&= (1 + o(1)) n^{1-\alpha} \frac{n^{\alpha}}{\alpha} \\
&= \frac{n}{\alpha} + o(n).
\end{aligned}$$

For $\alpha > 1$, we use the fact that

$$H_n^{(\alpha)} = \zeta(\alpha) - \frac{n^{-(\alpha-1)}}{\alpha - 1} + O(n^{-\alpha}). \tag{4.2}$$

Now

$$\begin{aligned}
\sum_{r=2}^{n} \frac{1}{H_r^{(\alpha)}} &= \sum_{r=2}^{n} \left[\zeta(\alpha) - \frac{r^{-(\alpha-1)}}{\alpha - 1} + O(r^{-\alpha})\right]^{-1} \\
&= \frac{1}{\zeta(\alpha)} \sum_{r=2}^{n} \left[1 - \frac{r^{-(\alpha-1)}}{(\alpha-1)\zeta(\alpha)} + O(r^{-\alpha})\right]^{-1} \\
&= \frac{1}{\zeta(\alpha)} \sum_{r=2}^{n} \left[1 + \frac{r^{-(\alpha-1)}}{(\alpha-1)\zeta(\alpha)} + O(r^{-\alpha}) + O(r^{-(2\alpha-2)})\right] \\
&= \frac{n}{\zeta(\alpha)} + o(n).
\end{aligned}$$

Together with (4.2) this gives

$$\begin{aligned}\mathbf{E}[T] &= H_n^{(\alpha)} \sum_{r=2}^{n} \frac{1}{H_r^{(\alpha)}} \\ &= \left[\zeta(\alpha) + O(n^{-(\alpha-1)})\right] \left[\frac{n}{\zeta(\alpha)} + o(n)\right] \\ &= n + o(n).\end{aligned}$$

(d) (Power law.) Here

$$\begin{aligned}\mathbf{E}[T] &= \sum_{r=2}^{n} \frac{f(n,s)}{(n-r+1)^s + \cdots + n^s} \le f(n,s) \sum_{r=2}^{n} \frac{s+1}{n^{s+1} - (n-r)^{s+1}} \\ &= \frac{f(n,s)(s+1)}{n^{s+1}} \sum_{r=2}^{n} \frac{1}{1-(1-\frac{r}{n})^{s+1}}.\end{aligned}$$

The inequality follows from an integral comparison. Another integral comparison shows that the last sum above is bounded above by

$$\int_1^n \frac{dx}{1-(1-\frac{x}{n})^{s+1}} = n \int_0^{1-\frac{1}{n}} \frac{dy}{1-y^{s+1}} =: n \times I.$$

Now

$$\begin{aligned}I &= \int_0^{1-\frac{1}{n}} \sum_{k=0}^{\infty} (y^{s+1})^k \, dy \\ &= \sum_{k=0}^{\infty} \frac{1}{(s+1)k+1} \left(1-\frac{1}{n}\right)^{(s+1)k+1} \\ &\le \left(1-\frac{1}{n}\right) + \frac{1-\frac{1}{n}}{s+1} \sum_{k=1}^{\infty} \frac{\left[(1-\frac{1}{n})^{s+1}\right]^k}{k} \\ &= \left(1-\frac{1}{n}\right) + \frac{1-\frac{1}{n}}{s+1} \left| \ln\left(1-\left(1-\frac{1}{n}\right)^{s+1}\right)\right| \\ &\le \left(1-\frac{1}{n}\right) + \frac{1-\frac{1}{n}}{s+1} \left| \ln\left(\frac{s+1}{n} - \frac{(s+1)s}{2n^2}\right)\right| \\ &\le \left(1-\frac{1}{n}\right) + \frac{1-\frac{1}{n}}{s+1} \ln n \\ &\le \frac{\ln n}{s+1} + 1,\end{aligned}$$

where the penultimate inequality holds for all sufficiently large n [in particular, for $n \ge (s+1)/2$]. We then have that

$$\begin{aligned}\mathbf{E}[T] &\le \frac{f(n,s)(s+1)}{n^s} \left(\frac{\ln n}{s+1} + 1\right) \\ &\le \frac{n \ln n}{s+1} + n + \ln n + s + 1 = \frac{n \ln n}{s+1} + O(n) = (1+o(1)) \frac{n \ln n}{s+1},\end{aligned}$$

using

$$f(n,s) \le \frac{n^{s+1}}{s+1} + n^s.$$

For the lower bound,

$$\begin{aligned}\mathbf{E}[T] &\ge f(n,s)\sum_{r=2}^{n} \frac{s+1}{(n+1)^{s+1}-(n+1-r)^{s+1}} \\ &= \frac{f(n,s)(s+1)}{(n+1)^{s+1}} \sum_{r=2}^{n} \left[1-\left(1-\frac{r}{n+1}\right)^{s+1}\right]^{-1}.\end{aligned}$$

But

$$\begin{aligned}\sum_{r=2}^{n} \left[1-\left(1-\frac{r}{n+1}\right)^{s+1}\right]^{-1} &\ge \int_2^{n+1} \left[1-\left(1-\frac{x}{n+1}\right)^{s+1}\right]^{-1} dx \\ &= (n+1)\int_0^{1-\frac{2}{n+1}} \frac{dy}{1-y^{s+1}}.\end{aligned}$$

We can show, but omit the details, that

$$\int_0^{1-\frac{2}{n+1}} \frac{dy}{1-y^{s+1}} = \frac{\ln n}{s+1} - O(1),$$

so

$$\mathbf{E}[T] \ge \left(\frac{n}{n+1}\right)^{s+1} \frac{(n+1)\ln n}{s+1} - O(n) \ge \frac{n\ln n}{s+1} - O(n),$$

where we have used

$$f(n,s) \ge \frac{n^{s+1}}{s+1}.$$

(e) (Geometric weights.) We have

$$\begin{aligned}n-1 \le \mathbf{E}[T] &= n-1+\sum_{r=2}^{n-1} \frac{\theta^r}{1-\theta^r} \\ &\le n-1+\sum_{r=2}^{\infty} \frac{\theta^r}{1-\theta^r} \\ &= n+O(1) = (1+o(1))n.\end{aligned}$$

□

5 Coalescence into a set

Here we present a different approach to speeding up the FMMR algorithm, which has the same spirit as the other results in this paper. In brief, recall that FMMR starts in a user-chosen state and then, in the second phase of the algorithm, checks whether

there is coalescence to that state. In the generalization we consider here, one starts the algorithm in some subset of the state space (not necessarily a singleton) and then checks if there is coalescence back to that set (but not necessarily to the state in which the algorithm began).

Theorem 5.1. *In the general setting for FMMR described in Section 2.2, let S be a subset of the state space and define $\pi_0(\cdot) := \pi(\cdot|S)$. Consider the modified algorithm which starts in a state $\mathbf{X}_0$ distributed according to π_0 and outputs $W := \mathbf{X}_{-T}$, where T is defined to be the smallest t such that all the forward trajectories from time $-t$ coalesce into S, i.e., such that $\mathbf{Y}^{(-t)}(x) \in S$ for every state x. Then W has the stationary distribution π. Further, the algorithm is interruptible (i.e., T and W are independent random variables).*

Proof. For simplicity we consider only the discrete case. It suffices to show that

$$P(T \le t, \mathbf{X}_{-t} = x) = P(T \le t)\pi(x) \tag{5.1}$$

for every t and x, for then

$$\begin{aligned} P(T = t,\ W = x) &= P(T = t,\ \mathbf{X}_{-T} = x) = P(T = t,\ \mathbf{X}_{-t} = x) \\ &= P(T \le t,\ \mathbf{X}_{-t} = x) - P(T \le t-1,\ \mathbf{X}_{-t} = x) \\ &= P(T \le t)\pi(x) - P(T \le t-1)\pi(x) \\ &= P(T = t)\pi(x), \end{aligned}$$

as desired. Here we have used the fact that π is stationary for the time-reversed kernel $\tilde{K}$, so that

$$\begin{aligned} &P(T \le t-1,\ \mathbf{X}_{-t} = x) \\ &= \sum_y P(T \le t-1, \mathbf{X}_{-(t-1)} = y)\, P(\mathbf{X}_{-t} = x \,|\, T \le t-1,\ \mathbf{X}_{-(t-1)} = y) \\ &= \sum_y P(T \le t-1)\pi(y)\tilde{K}(y,x) \quad \text{by (5.1) and the Markov property for } \tilde{K} \\ &= P(T \le t-1)\pi(x). \end{aligned}$$

To establish (5.1), we first observe that

$$\begin{aligned} P(\mathbf{X}_{-t} = x) &= \sum_z \pi_0(z)\tilde{K}^t(z,x) = \pi(x)\sum_z \frac{\pi_0(z)}{\pi(z)}K^t(x,z) \\ &= \frac{\pi(x)}{\pi(S)}\sum_{z\in S} K^t(x,z) = \frac{\pi(x)}{\pi(S)}K^t(x,S). \end{aligned} \tag{5.2}$$

One can check that, conditionally given $\mathbf{X}_{-t} = x$, the forward trajectory $(\mathbf{X}_{-t}, \ldots, \mathbf{X}_0)$ has the same distribution as a K-trajectory conditioned to start at x and end in S. Therefore, by the algorithm's design,

$$\begin{aligned} &P(T \le t \,|\, \mathbf{X}_{-t} = x) \\ &= P(\text{forward coalescence into } S \text{ over a time-interval of length } t)/K^t(x,S). \end{aligned}$$

Combining this with (5.2) we conclude (5.1), and the additional result

$$P(T \le t) = P(\text{forward coalescence into } S \text{ over a time-interval of length } t)/\pi(S).$$

□

Remark 5.2. In the monotone case, if S is a down-set (meaning: $z \in S$ and $y \le z$ implies $y \in S$), then the computational problem of determining whether or not there is coalescence into S is eased considerably: we need only determine whether the terminal state (call it y) of the forward trajectory started in $\hat{1}$ belongs to S. And so if S is a *principal* down-set, that is, if $S = \{z : z \le z_0\}$ for some z_0, the problem is even easier: we need only check whether $y \le z_0$.

We will now give a "toy" application of these ideas to MTF by describing an algorithm to build up a stationary observation in just $n-1$ steps, regardless of the weights $w_1, \ldots, w_n$. Let π_k denote the MTF stationary distribution on $\mathcal{S}_k$ restricted to the (normalized) weights $w_1, \ldots, w_k$; that is, to the weights $w_1/w_k^+, \ldots, w_k/w_k^+$. Let MTF_k denote the MTF process on $\mathcal{S}_k$, and let S_k denote the set of permutations of $\{1, \ldots, k\}$ that begin with k. Observe that S_k is the principal order ideal $\{z : z \le z_k\}$ in the dual (i.e., "upside-down") Bruhat order of $\mathcal{S}_k$, where z_k is the permutation $(k\ 1\ \cdots\ k-1)$. (We will not refer to the symmetric group $\mathcal{S}_k$ any further; thus there will be no notational confusion with its special subset S_k.) Inductively, after k steps of our algorithm we will have a permutation distributed according to π_{k+1}; thus, after $n-1$ steps we will have an observation from π.

Initialize the algorithm (step 0) with the permutation (1) on $\{1\}$. Suppose that after $k-1$ steps we have the permutation $x = (x_1, \ldots, x_k)$ distributed according to π_k. For the next (kth) step, we first get immediately an observation from $\pi_{k+1}(\cdot \mid S_{k+1})$, namely, $(k+1, x_1, \ldots, x_k)$. Then we apply the "coalesce into S" routine of Theorem 5.1, taking $S = S_{k+1}$. We claim that that routine will terminate in a single step! Indeed, in one time-reversed MTF_{k+1} transition we obtain the permutation

$$x' = (x_1, \ldots, x_{j-1}, k+1, x_j, \ldots, x_k)$$

for some j. In the forward phase of the routine, record $k+1$ is brought to the front of every trajectory, giving coalescence into the set S_{k+1}. We thus conclude from Theorem 5.1 that $x' \sim \pi_{k+1}$, completing the induction.

6 Acknowledgements

This research was carried out in part while the first-listed author was a member of the Department of Mathematics and Computer Science at Clarkson University, and while the second-listed author was Visiting Researcher, Theory Group, Microsoft Research.

References

[1] Aldous, David and Fill, James Allen. *Reversible Markov Chains and Random Walks on Graphs.* Book in preparation. Draft of manuscript available via `http://stat-www.berkeley.edu/users/aldous.`

[2] Devroye, Luc. Nonuniform random variate generation. *Springer–Verlag, New York,* 1986.

[3] Diaconis, Persi; Fill, James Allen; and Pitman, Jim. Analysis of top-to-random shuffles. *Combin. Probab. Comput.* **1** (1992), no. 2, 135–155.

[4] Fill, James Allen. An interruptible algorithm for perfect sampling via Markov chains. *Ann. Appl. Probab.* **8** (1998), no. 1, 131–162.

[5] Fill, James Allen. An exact formula for the move-to-front rule for self-organizing lists. *J. Theoret. Probab.* **9** (1996), no. 1, 113–160.

[6] Fill, James Allen. The move-to-front rule: a case study for two perfect sampling algorithms. *Probab. Engrg. Inform. Sci.* **12** (1998), no. 3, 283–302.

[7] Fill, James Allen; Machida, Motoya; Murdoch, Duncan J.; Rosenthal, Jeffrey S. Extension of Fill's perfect rejection sampling algorithm to general chains: Proceedings of the Ninth International Conference "Random Structures and Algorithms" (Poznan, 1999). *Random Structures and Algorithms* **17** (2000), no. 3-4, 290–316.

[8] Hendricks, W. J. The stationary distribution of an interesting Markov chain. *J. Appl. Probability* **9** (1972), 231–233.

[9] Kendall, Wilfrid S. Perfect simulation for the area-interaction point process. *Probability towards 2000 (New York, 1995)*, 218–234, Lecture Notes in Statist., 128, *Springer, New York,* 1998.

[10] Liggett, Thomas M. Interacting particle systems. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], 276, *Springer-Verlag, New York,* 1985.

[11] Marshall, Albert W.; Olkin, Ingam. Inequalities: theory of majorization and its applications. Mathematics in Science and Engineering, 143. *Academic Press, Inc.* New York-London, 1979.

[12] Propp, James Gary; Wilson, David Bruce. Exact sampling with coupled Markov chains and applications to statistical mechanics. Proceedings of the Seventh International Conference on Random Structures and Algorithms (Atlanta, GA, 1995). *Random Structures and Algorithms* **9** (1996), no. 1-2, 223–252.

[13] Propp, James Gary; Wilson; David Bruce. Coupling from the past: a user's guide. *Microsurveys in discrete probability (Princeton, NJ, 1997),* 181–192, DIMACS Ser. Discrete Math. Theoret. Comput. Sci., 41, *Amer. Math. Soc., Providence, RI,* 1998.

[14] Propp, James Gary; Wilson, David Bruce. How to get a perfectly random sample from a generic Markov chain and generate a random spanning tree of a directed graph. 7th Annual ACM-SIAM Symposium on Discrete Algorithms (Atlanta, GA, 1996). *J. Algorithms* **27** (1998), no. 2, 170–217.

[15] Wilson, David Bruce. Annotated bibliography of perfectly random sampling with Markov chains. *Microsurveys in discrete probability (Princeton, NJ, 1997)*, 209–220, DIMACS Ser. Discrete Math. Theoret. Comput. Sci., 41, *Amer. Math. Soc., Providence, RI,* 1998. Latest updated version is posted at `http://www.dbwilson.com/exact/`.

[16] Wilson, David Bruce. Layered multishift coupling for use in perfect sampling algorithms (with a primer on CFTP). *Monte Carlo methods (Toronto, ON, 1998)*, 143–179, Fields Inst. Commun., 26, *Amer. Math. Soc., Providence, RI,* 2000.