Importance Tempering

Robert Gramacy & Richard Samworth
Statistical Laboratory
University of Cambridge
{bobby, rjs57}@statslab.cam.ac.uk

Ruth King
CREEM
University of St Andrews
ruth@mcs.st-and.ac.uk

November 3, 2008

Abstract

Simulated tempering (ST) is an established Markov chain Monte Carlo (MCMC) method for sampling from a multimodal density $\pi(\theta)$. Typically, ST involves introducing an auxiliary variable k taking values in a finite subset of $[0,1]$ and indexing a set of tempered distributions, say $\pi_k(\theta) \propto \pi(\theta)^k$. In this case, small values of k encourage better mixing, but samples from π are only obtained when the joint chain for (θ, k) reaches $k = 1$. However, the entire chain can be used to estimate expectations under π of functions of interest, provided that importance sampling (IS) weights are calculated. Unfortunately this method, which we call importance tempering (IT), can disappoint. This is partly because the most immediately obvious implementation is naïve and can lead to high variance estimators. We derive a new optimal method for combining multiple IS estimators and prove that the resulting estimator has a highly desirable property related to the notion of effective sample size. We briefly report on the success of the optimal combination in two modelling scenarios requiring reversible-jump MCMC, where the naïve approach fails.

Key words: simulated tempering, importance sampling, Markov chain Monte Carlo (MCMC), Metropolis–coupled MCMC

1 Introduction

Markov chain Monte Carlo (MCMC) algorithms, in particular Metropolis–Hastings (MH) and Gibbs Sampling (GS), are by now the most widely used methods for simulation–based inference in Bayesian statistics. The beauty of MCMC is its simplicity. Very little user input or expertise is required in order to establish a Markov chain whose stationary distribution is proportional to $\pi(\theta)$, for $\theta \in \Theta \subseteq \mathbb{R}^d$. As long as the chain is irreducible, the theory of Markov chains guarantees that sample averages computed from this realisation will converge in an appropriate sense to their expectations under π. However, difficulties can arise when π has isolated modes, between which the Markov chain moves only rarely. In such cases convergence is slow, meaning that often infeasibly large sample sizes are needed to obtain accurate estimates.

New MCMC algorithms have been proposed to improve mixing. Two related algorithms are Metropolis–coupled MCMC (MC^3) (**??**) and simulated tempering (ST) (**??**). Both are closely related to the optimisation technique of simulated annealing (SA) (**?**). SA works with a set of *tempered* distributions $\pi_k(\theta)$ indexed by an inverse–temperature parameter $k \in [0, \infty)$. One popular form of tempering is called "powering up", where $\pi_k(\theta) \propto \pi(\theta)^k$. Small values of k have the effect of flattening/widening the peaks and raising troughs in π_k relative to π.

In MC^3 and ST we define a *temperature ladder* $1 = k_1 > k_2 > \ldots > k_m \geq 0$, and call the k_i its *rungs*. Both MC^3 and ST involve simulating from the set of m tempered densities $\pi_{k_1}, \ldots, \pi_{k_m}$. MC^3 runs m parallel MCMC chains, one at each temperature, and regularly proposes swaps of states at adjacent rungs k_i and k_{i+1}. Usually, samples are only saved from the "cold distribution" π_{k_1}. In contrast, ST works with a "pseudo–prior" $p(k_i)$ and uses a single chain to sample from the joint distribution, which is proportional to $\pi_k(\theta)p(k)$. Again, it is only at iterations t for which $k^{(t)} = 1$ that the corresponding realisation of $\theta^{(t)}$ is retained.

ST has an advantage over MC^3 in that only one copy of the process $\{\theta^{(t)} : t = 1, \ldots, T\}$ is needed—rather than m—so the chain uses less storage and also has better mixing (**?**). The disadvantage is that it needs a good choice of pseudo–prior. For further comparison and review, see **?** and **?**.

Both MC^3 and ST suffer from inefficiency because they discard all samples from π_k for $k \neq 1$. The discarded samples could be used to estimate expectations under π if they were given appropriate importance sampling (IS) weights. For an inclusive review of IS and related methods see **?**, Chapter 2. Moreover, it may be the case that an IS estimator constructed with samples from a tempered distribution has smaller variance than one based on a sample of the same size from π. As a simple motivating example, let $\pi(\theta) = N(\theta|\mu, \sigma^2)$, and consider estimating $\mu = \mathbb{E}_\pi(\theta)$ by IS from a tempered distribution $\pi_k(\theta) \propto \pi(\theta)^k$. A straightforward calculation shows that the value of k which minimises the variance of the IS estimator is

$$k^* = \begin{cases} 1/2 & \text{if } \mu = 0 \\ \frac{3}{2} + \left(\frac{\sigma}{\mu}\right)^2 - \frac{1}{2}\left\{1 + 8\left(\frac{\sigma}{\mu}\right)^2 + 4\left(\frac{\sigma}{\mu}\right)^4\right\}^{1/2} & \text{otherwise.} \end{cases} \qquad (1)$$

Note that $k^* \in (1/2, 1)$ for all μ and σ^2. Moreover, one can compute (numerically) $k^- = k^-(\sigma/\mu) < k^*$ such that for all $k \in (k^-, 1)$, the variance of the IS estimator $\hat{\mu}_k$ based on samples from π_k is smaller than that of one based on a sample of the same size from π. However, $\text{Var}(\hat{\mu}_k) \to \infty$ as $k \to 0$ for all μ and σ^2. Table 29 gives k^* and k^- for various values of σ/μ.

σ/μ	1/16	1/4	1	4	16
k^*	1.00	0.95	0.70	0.52	0.50
k^-	0.99	0.89	0.42	0.18	0.16

Table 1: Values of k^* and k^- for various values of σ/μ.

Therefore, there is a trade-off in the choice of tempered IS proposals. On the one hand, low inverse–temperatures k in ST can guard against missing modes of π with large support by encouraging better mixing *between* modes, but can yield very inefficient (IS) estimators overall. On the other hand, "lukewarm" temperatures k, especially $k \in (1/2, 1)$, can yield more efficient estimators *within* modes than those obtained from samples at $k = 1$.

? was the first to suggest using a single tempered distribution as a proposal in IS, and **???** has since written several papers combining IS and tempering. Indeed, in the discussion of the 1996 paper on *tempered transitions*, Neal writes "simulated tempering allows data associated with p_i other than p_0 [the cold distribution] to be used to calculate expectations with respect to ... p_0 (using an importance sampling estimator)"[1]. It is this natural extension that we call *importance tempering* (IT), with IMC^3 defined similarly. Given the work of the above-mentioned authors, and the fact that calculating importance weights is relatively trivial, it may be surprising that successful IT and IMC^3 applications have yet to be published. **?** comes close in proposing to augment ST with dynamic weighting (**?**) and in applying the Wang–Landau algorithm (**?**) to ST.

This paper addresses why the straightforward methodology described above has tended not to work well in practice, primarily due to a lack of a principled way of combining the importance weights collected at each temperature to obtain an overall estimator. If we are interested in estimating $\mathbb{E}_\pi\{h(\theta)\}$, one way to do this is with

$$\hat{h} = W^{-1} \sum_{t=1}^{T} w(\theta^{(t)}, k^{(t)}) h(\theta^{(t)}), \qquad \text{where} \qquad W = \sum_{t=1}^{T} w(\theta^{(t)}, k^{(t)}), \tag{2}$$

and $w(\theta, k) = \pi(\theta)/\pi(\theta)^k = \pi(\theta)^{1-k}$. Observe that this estimator is of the form $\hat{h} = \sum_{i=1}^m \lambda_i \hat{h}_i$, where $0 \leq \lambda_i \leq \sum_{i=1}^m \lambda_i = 1$, with $\lambda_i = W^{-1} \sum_{t=1}^T w(\theta^{(t)}, k^{(t)}) \mathbb{I}_{\{k^{(t)} = k_i\}}$, and where each $\hat{h}_i$ is an IS estimator of $\mathbb{E}_\pi\{h(\theta)\}$ constructed using only the observations at the inverse–temperature k_i. We show how to improve this estimator by choosing $\lambda_1, \ldots, \lambda_m$ to maximise the *effective sample size* (see next paragraph), which approximately corresponds

[1] A similar note is made in the 2001 paper with regard to *annealed importance sampling.*

to minimising the variance of $\hat{h}$ (**?**, Section 2.5.3). For the applications that we have in mind, it is important that our estimator can be constructed without knowledge of the normalising constants of $\pi_{k_1}, \ldots, \pi_{k_m}$. It is for this reason that methods motivated by the *balance heuristic* (**???**) cannot be applied.

The notion of *effective sample size* plays an important role in the study of IS estimators. Suppose we are interested in estimating $\mathbb{E}_\pi\{h(\theta)\}$ using a vector of observations $\boldsymbol{\theta} = (\theta^{(1)}, \ldots, \theta^{(T)})$ from a density π'. Define the vector of importance weights $\mathbf{w} \equiv \mathbf{w}(\boldsymbol{\theta}) = (w(\theta^{(1)}), \ldots, w(\theta^{(T)}))$, where $w(\theta) = \pi(\theta)/\pi'(\theta)$. Following **?**, Section 2.5.3 we define the *effective sample size* by

$$\mathrm{ESS}\big(\mathbf{w}(\boldsymbol{\theta})\big) \equiv \mathrm{ESS}(\mathbf{w}) = \frac{T}{1 + \mathrm{cv}^2(\mathbf{w})}, \tag{3}$$

where $\mathrm{cv}^2(\mathbf{w})$ is the *coefficient of variation* of the weights, given by

$$\mathrm{cv}^2(\mathbf{w}) = \frac{\sum_{t=1}^{T}(w(\theta^{(t)}) - \bar{w})^2}{(T-1)\bar{w}^2}, \qquad \text{where} \qquad \bar{w} = T^{-1}\sum_{t=1}^{T} w(\theta^{(t)}).$$

This should not be confused with the concept of *effective sample size due to autocorrelation* (**?**) (due to serially correlated samples from a Markov chain). This latter notion is discussed briefly in Section 60.

Observe that the swap operations in MC^3 require that the state space Θ be common for all m tempered distributions. This is not a requirement for ST, as the state stays fixed when changes in temperature are proposed. Thus applying MC^3 is less straightforward in (Bayesian) model selection/averaging problems which typically involve trans–dimensional Markov chains as in reversible–jump MCMC (RJMCMC) (**?**), though it is possible (**?**). Since RJMCMC algorithms are particularly prone to slow mixing, and hence are an excellent source of applications of our idea (as illustrated in Section 59), the rest of the paper will focus on

IT. Most of our results apply equally to IMC^3 by ignoring the pseudo–prior.

The outline of the paper is as follows. In Section 58 we derive the optimal convex combination of multiple IS estimators, and show how this estimator has a particularly attractive property with regard to its effective sample size. In Section 59 we briefly report on the effectiveness of optimal IT, and the poor performance of the naïve approach, on several real and synthetic examples. Section 60 concludes with a discussion.

2 Importance tempering

The *simulated tempering* (ST) (**?**) algorithm is an application of MH on the product space of parameters and inverse–temperatures. That is, samples are obtained from the joint chain $\pi(\theta, k) \propto \pi(\theta)^k p(k)$. This is only possible if $\pi(\theta)^k$ is integrable, but Hölder's inequality may be used to show that this is indeed the case provided that $\mathbb{E}_\pi(\|\theta\|^{\frac{1-k}{k}+\delta}) < \infty$ for some $\delta > 0$, where $\|\cdot\|$ denotes the Euclidean norm. The success of ST depends crucially on the ability of the Markov chain frequently to: (a) visit high temperatures (low k) where the probability of escaping local modes is high; (b) visit $k = 1$ to obtain samples from π. The algorithm can be tuned by: (i.) adjusting the number and location of the rungs of the temperature ladder; or (ii.) adjusting the pseudo-prior $p(k)$. **?** give some automated ways of adjusting the spacing of the rungs of the ladder. **?** reviews similar techniques from the physics literature. A recent alternative—and very promising—approach involves the Wang–Landau algorithm (**?**). However, many authors prefer to rely on defaults, e.g.,

$$k_i = \begin{cases} (1+\Delta_k)^{1-i} & \text{geometric spacing} \\ \{1+\Delta_k(i-1)\}^{-1} & \text{harmonic spacing} \end{cases} \qquad i = 1, \ldots, m. \tag{4}$$

The rate parameter $\Delta_k > 0$ can be problem specific. Motivation for such default spacings is outlined by **?**, Chapter 10: pp. 213 & 233. Geometric spacing, or uniform spacing of $\log(k_i)$, is also advocated by **??**.

Once a suitable ladder has been chosen, the goal is typically to choose the pseudo–prior so that the posterior over temperatures is uniform. The best way to accomplish this is to set $p(k_i) = 1/Z_i$, where $Z_i = \int_\Theta \pi(\theta)^{k_i} d\theta$ is the normalising constant in $\pi_{k_i} = \pi^{k_i}/Z_i$, which is generally unknown. So while normalising constants are not a prerequisite for ST, it can certainly be useful to know them. We follow the suggestions of **?** in setting the pseudo–prior by a method that roughly approximates the Z_i in two–stages: first by stochastic approximation (**?**), and then by observation counts accumulated through pilot runs. To some extent, a non-uniform posterior on the temperatures is less troublesome in the context of IT than ST. So long as the chain still visits the heated temperatures often enough to get good mixing in Θ, and if the ESS of the IS estimators at some temperature(s) is not too low, useful samples can be obtained without ever visiting the cold distribution.

2.1 A new optimal way to combine IS estimators

ST provides us with $\{(\theta^{(t)}, k^{(t)}) : t = 1, \dots, T\}$, where $\theta^{(t)}$ is an sample from $\pi_{k^{(t)}}$. Write $\mathcal{T}_i = \{t : k^{(t)} = k_i\}$ for the index set of observations at the i^{th} temperature, and let $T_i = |\mathcal{T}_i|$. Let the vector of observations at the i^{th} temperature collect in $\boldsymbol{\theta}_i = (\theta_{i1}, \dots, \theta_{iT_i})$, so that $\{\theta_{ij}\}_{j=1}^{T_i} \sim \pi_{k_i}$. Similarly, the vector of IS weights at the i^{th} temperature is $\mathbf{w}_i = \mathbf{w}_i(\boldsymbol{\theta}_i) = (w_i(\theta_{i1}), \dots, w_i(\theta_{iT_i}))$, where $w_i(\theta) = \pi(\theta)/\pi_{k_i}(\theta)$.

Each vector $\boldsymbol{\theta}_i$ can be used to construct an IS estimator of $\mathbb{E}_\pi\{h(\theta)\}$ by setting

$$\hat{h}_i = \frac{\sum_{j=1}^{T_i} w_i(\theta_{ij}) h(\theta_{ij})}{\sum_{j=1}^{T_i} w_i(\theta_{ij})} \equiv \frac{\sum_{j=1}^{T_i} w_{ij} h(\theta_{ij})}{W_i}.$$

It is natural to consider an overall estimator of $\mathbb{E}_\pi\{h(\theta)\}$ defined by a convex combination:

$$\hat{h}_\lambda = \sum_{i=1}^{m} \lambda_i \hat{h}_i, \qquad \text{where} \qquad 0 \leq \lambda_i \leq \sum_{i=1}^{m} \lambda_i = 1. \tag{5}$$

Unfortunately, if $\lambda_1, \ldots, \lambda_m$ are not chosen carefully, $\mathrm{Var}(\hat{h}_\lambda)$, can be nearly as large as the largest $\mathrm{Var}(\hat{h}_i)$ (**?**). Notice that ST is recovered as a special case when $\lambda_1 = 1$ and $\lambda_2 = \cdots = \lambda_m = 0$. It may be tempting to choose $\lambda_i = W_i/W$, where $W = \sum_{i=1}^{m} W_i$, recovering the estimator in Eq. (156). This can lead to a very poor estimator, even compared to ST, which is demonstrated empirically in Section 59.

Observe that we can write

$$\hat{h}_\lambda = \sum_{i=1}^{m} \sum_{j=1}^{T_i} w_{ij}^{\lambda} h(\theta_{ij}), \tag{6}$$

where $w_{ij}^{\lambda} = \lambda_i w_{ij}/W_i$. Let $\mathbf{w}^\lambda = (w_{11}^\lambda, \ldots, w_{1T_1}^\lambda, w_{21}^\lambda, \ldots, w_{2T_2}^\lambda, \ldots, w_{m1}^\lambda, \ldots, w_{mT_m}^\lambda)$. Attempting to choose $\lambda_1, \ldots, \lambda_m$ to minimise $\mathrm{Var}(\hat{h}_\lambda)$ directly can be difficult. In the balance heuristic, **?** explore combinations of IS estimators of the form (160), where $w_i(\theta) = \pi(\theta)/g_i(\theta)$ for a family of proposal densities g_i, with

$$\lambda_{ij} = \frac{c_i g_i(\theta_{ij})}{\sum_{r=1}^{m} c_r g_r(\theta_{ij})}, \tag{7}$$

and where $0 \leq c_i \leq \sum_{i=1}^{m} c_i = 1$ is the proportion of samples taken from g_i. It turns out that this is equivalent to IS with the mixture proposal $\tilde{\pi}(\theta) = \sum_{r=1}^{m} c_r g_r(\theta)$:

$$\hat{h}_{\mathrm{bal}} \equiv \frac{1}{T} \sum_{t=1}^{T} w(\theta_t) h(\theta_t), \qquad \text{where} \qquad w(\theta) = \frac{\pi(\theta)}{\sum_{r=1}^{m} c_r g_r(\theta)}. \tag{8}$$

The balance heuristic has since been generalised by **?**; it was reinvented by (**?**, Section 4) in the context of applied probability.

Note that due to the denominator in the definition of $w(\theta)$ in Eq. (162), the g_i must be normalised densities. This precludes us from using the balance heuristic with $g_i \propto \pi_{k_i}$. When MCMC is necessary to sample from π, the normalisation constant of π, and therefore π_{k_i}, is generally unknown. The method also requires evaluations of $\pi_{k_i}(\theta^{(t)})$, $i = 1, \ldots, m$, at all T rounds, an $O(mT)$ operation that trivialises any computational advantage ST has over MC^3. Instead, we consider maximising the ESS of $\hat{h}_\lambda$ in (159).

Proposition 2.1. *Among estimators of the form (159),* $\mathrm{ESS}(\mathbf{w}^\lambda)$ *is maximised by* $\lambda = \lambda^*$*, where, for* $i = 1, \ldots, m$*,*

$$\lambda_i^* = \frac{\ell_i}{\sum_{i=1}^m \ell_i}, \qquad \textit{and} \qquad \ell_i = \frac{W_i^2}{\sum_{j=1}^{T_i} w_{ij}^2}.$$

Proof. Since $\sum_{i=1}^m \sum_{j=1}^{T_i} w_{ij}^\lambda = 1$, the problem of maximising the effective sample size is the same as

$$\min_{\lambda_1,\ldots,\lambda_m} \sum_{i=1}^m \sum_{j=1}^{T_i} \left(\lambda_i \frac{w_{ij}}{W_i} - \frac{1}{T} \right)^2, \qquad \text{subject to} \qquad 0 \leq \lambda_i \leq \sum_{i=1}^m \lambda_i = 1.$$

The result then follows by a straightforward Lagrange multiplier argument. □

In the following discussion and in Remark 58.2 below, we assume that for $i = 1, \ldots, m$, $T_i \geq 2$. The efficiency of each IS estimator $\hat{h}_i$ can be measured through $\mathrm{ESS}(\mathbf{w}_i)$. Intuitively, we hope that with a good choice of λ, the ESS of $\hat{h}_\lambda$, given by

$$\mathrm{ESS}(\mathbf{w}^\lambda) = \frac{T(T-1)}{T^2 \sum_{i=1}^m \lambda_i^2/\ell_i - 1},$$

would be close to the sum over i of the effective sample sizes of $\hat{h}_i$, namely

$$\mathrm{ESS}(\mathbf{w}_i) = \frac{T_i(T_i - 1)\ell_i}{T_i^2 - \ell_i}. \tag{9}$$

The remark below shows that this is indeed the case for $\hat{h}_{\lambda^*}$.

Remark 2.2. *We have*

$$\mathrm{ESS}(\mathbf{w}^{\lambda^*}) \geq \sum_{i=1}^{m} \mathrm{ESS}(\mathbf{w}_i) - \frac{1}{4} - \frac{1}{T}.$$

Proof. Since $\mathrm{ESS}(\mathbf{w}_i) \leq T_i$, it follows from (163) that $\ell_i \leq T_i$. Thus

$$\begin{aligned}\mathrm{ESS}(\mathbf{w}^{\lambda^*}) = \frac{(1-T^{-1})\sum_{i=1}^{m}\ell_i}{1-\sum_{i=1}^{m}\frac{\ell_i}{T_i^2}} &\geq \Big(1-\frac{1}{T}\Big)\Big(1+\frac{1}{T^2}\sum_{i=1}^{m}\ell_i\Big)\sum_{i=1}^{m}\ell_i \\ &= \sum_{i=1}^{m}\ell_i - \frac{\sum_{i=1}^{m}\ell_i}{T}\Big(1-\frac{\sum_{i=1}^{m}\ell_i}{T}\Big) - \frac{(\sum_{i=1}^{m}\ell_i)^2}{T^3} \\ &\geq \sum_{i=1}^{m}\ell_i - \frac{1}{4} - \frac{1}{T},\end{aligned}$$

since $x(1-x)$ attains its maximum of $1/4$ at $x = 1/2$ and $\sum \ell_i \leq \sum T_i = T$. □

In practice we have found that this bound is slightly conservative and that often it is the case that $\mathrm{ESS}(\mathbf{w}^{\lambda^*}) \geq \sum_{i=1}^{m} \mathrm{ESS}(\mathbf{w}_i)$. Thus our optimally–combined IS estimator has a highly desirable and intuitive property in terms of its effective sample size.

3 Empirical Results

Here we briefly report on the success of optimal IT, relative to the naïve approach and ST, on one simple example and two involving RJMCMC.

3.1 A simple mixture of normals

Consider the following toy density π, a mixture of two normals:

$$\pi(\theta) = 0.6N(\theta|\mu_1 = -8, \sigma_1^2 = 0.5^2) + 0.4N(\theta|\mu_2 = 8, \sigma_2^2 = 0.9^2). \tag{10}$$

Table 30 summarises Kolmogorov–Smirnov distances obtained under three IT estimators: ST ($\lambda_1 = 1$), naïve IT ($\lambda_i = W_i/W$) and the optimally–combined IT estimator ($\hat{h}_{\lambda^*}$). Observe that the optimally–combined IT estimator has both the largest ESS and the smallest variance of the three estimators, and that $\text{ESS}(\mathbf{w}^{\lambda^*}) > \sum_i \text{ESS}(\mathbf{w}_i)$. Naïve IT improves upon ST in this example, but has higher variance than $\hat{h}_{\lambda^*}$.

		K–S distance	
Method	ESS($\mathbf{w}^{\lambda}$)	mean	var
ST	2535	0.0938	8.5×10^{-4}
naïve IT	17779	0.0849	1.4×10^{-4}
$\hat{h}_{\lambda^*}$	22913	0.0836	5.2×10^{-5}
$\sum_i \text{ESS}(\mathbf{w}_i)$	22910		

Table 2: Summary of K–S distances to the true mixture of normals (164) for ST ($\lambda_1 = 1$), naïve IT ($\lambda_i = W_i/W$), the optimally–combined IT estimator ($\hat{h}_{\lambda^*}$). We used 100 repeated samples of size 10^5, with tempered RWM proposals.

3.2 Bayesian treed Gaussian process models

Bayesian treed models extend classification and regression tree (CART) models (**?**), by putting a prior on the tree structure. We focus on the implementation of **?** who fit Gaussian Process (GP) models at the leaves of the tree, specify the tree prior through a process that limits its depth, and then define the tree operations *grow*, *prune*, *change*, and *swap*, to allow inference to proceed by RJMCMC. The RJMCMC chain usually identifies the correct *maximum a posteriori* (MAP) tree, but consistently and significantly over estimates the posterior probability of deep trees.

To guard against the transdimensional chain getting stuck in local modes of the posterior, **?** resorted regularly restarting the chain from the null tree. ST provides an alternative by increasing the rate of accepted tree operations in higher temperatures. In particular, we find that ST can increase the rate of accepted *prune* operations by an order of magnitude, thus enabling the chain to escape the local modes of deep trees. To demonstrate IT we fit a treed

GP model with ST using a geometric ladder with $m = 40$ and $k_m = 0.1$ to two datasets first explored by ?: the 1-d motorcycle accident data and 2-d exponential data. We refer to that paper for details about the data and models.

For the motorcycle accident data the ST chain was run for $T = 1.5 \times 10^5$ iterations, where a total of $T_1 = 3732$ ($\approx T/m = 3750$) samples were obtained from the cold distribution. That $\mathrm{ESS}(\mathbf{w}^{\lambda^*}) = 9338 \approx 2.5 T_1$ shows the considerable improvement of IT over ST. Moreover, we have $\mathrm{ESS}(\mathbf{w}^{\lambda^*}) > \sum_i \mathrm{ESS}(\mathbf{w}_i) = 9334$. The naïve combination $\lambda_i = \frac{W_i}{W}$ in (156) yields $\mathrm{ESS}(\mathbf{w}^{\lambda}) = 285 < \frac{1}{10} T_1$, undermining the very motivation of IT. For the exponential data the ST chain was run for a total of $T = 5 \times 10^5$ iterations. A total of $T_1 = 12436$ ($\approx T/m = 12500$) samples were obtained from the cold distribution. We found that $\mathrm{ESS}(\mathbf{w}^{\lambda^*}) = 21778 \approx 1.75 T_1$, illustrating how IT improves on ST. Moreover, we have $\mathrm{ESS}(\mathbf{w}^{\lambda^*}) > \sum_i \mathrm{ESS}(\mathbf{w}_i) = 21776$. The naïve combination $\lambda_i = \frac{W_i}{W}$ in (156) yields $\mathrm{ESS}(\mathbf{w}^{\lambda^*}) = 654 \approx \frac{1}{18} T_1$—worse than ST.

3.3 Mark-Recapture-Recovery Data

We now consider a Bayesian model selection problem with data relating to the mark-recapture and recovery of shags on the Isle of May (?). The three demographic parameters of interest are: survival rates, recapture rates and recovery rates. The models considered for each of the demographic parameters allowed a possible age– and/or time–dependence, where the time dependence was conditional on the age structure of the parameters. Typically, movement between the different possible models—by adding/removing time dependence for a given age group, or updating the age structure of the parameters—is slow, with small acceptance probabilities. For further details of the data, model structure, and RJMCMC algorithm see ?.

Using the same ST setup as above, we ran $T = 10^7$ iterations and discarded the first 10% as burn-in. As with the treed examples, higher temperatures yielded higher acceptance rates and an order of magnitude better exploration of model space compared to (untempered) RJMCMC. A total of $T_1 = 248158$ ($\approx T/m = 225000$) realisations were obtained from the cold distribution. By comparison, for optimal IT we have $\mathrm{ESS}(\mathbf{w}^{\lambda^*}) = 612026 \approx 2.5T_1$ and $\mathrm{ESS}(\mathbf{w}^{\lambda^*}) > \sum_i \mathrm{ESS}(\mathbf{w}_i) = 612020$. The corresponding naïve IT approach (using $\lambda_i = \frac{W_i}{W}$) performed exceptionally poorly, with $\mathrm{ESS}(\mathbf{w}^{\lambda})$ of only 5.43, due to a few large weights obtained at hot temperatures.

4 Discussion

This paper has addressed the inefficiencies and wastefulness of simulated tempering (ST), and related algorithms that are designed to improve mixing in the Markov chain using tempered distributions. We argued that importance sampling (IS) from tempered distributions can produce estimators that are more efficient than ones based on independent sampling, provided that the temperature is chosen carefully. This motivated augmenting the ST algorithm by calculating importance weights to salvage discarded samples—a technique which we have called *importance tempering* (IT). This idea has been suggested before, but to our knowledge little exploration has been carried out for real, complex, applications. We have derived optimal combination weights for the resulting collection of IS estimators, which can be calculated even when the normalisation constants of the tempered distributions are unknown. The weights are essentially proportional to the effective sample size (ESS) of the individual estimators, and we found that the resulting combined ESS in this case would be approximately equal to their sum.

We note that the overall success of the optimal IT estimator depends crucially on a successful implementation of ST, i.e., having a good temperature ladder and pseudo–prior. However, it is also important to recognise that the optimal combination, as a resource–efficient post-processing step, is equally applicable in other contexts, i.e., within MC^3, or even outside of the domain of tempered MCMC to combine any collection IS estimators. Sequential Monte Carlo samplers (**?**) may facilitate a natural extension. We have illustrated IT on several examples which benefit from the improved mixing ST provides. For example, the optimal IT methodology can increase the resulting ESS compared to retaining samples only from the cold distribution by roughly a factor of two.

Since IT involves sampling from a Markov chain, ideally one would take into account the serial correlation in the objective criteria for combining the individual estimators. The *effective sample size due to autocorrelation* is defined (**?**) by

$$\mathrm{ESS}_\rho(\boldsymbol{\theta}) = \frac{T}{1 + 2\sum_{\ell=1}^{T-1} \hat{\rho}(\ell, \boldsymbol{\theta})}, \tag{11}$$

where $\hat{\rho}(\ell, \boldsymbol{\theta})$ is the sample autocorrelation in $\boldsymbol{\theta}$ at lag ℓ; thus for scalar θ we have that $\hat{\rho}(\ell, \boldsymbol{\theta}) = \hat{\gamma}(\ell, \boldsymbol{\theta})/\hat{\gamma}(0, \boldsymbol{\theta})$, where $\hat{\gamma}(\ell, \boldsymbol{\theta}) = (T - \ell)^{-1} \sum_{t=1}^{T-\ell}(\theta^{(t)} - \bar{\theta})(\theta^{(t+\ell)} - \bar{\theta})$, and $\bar{\theta} = T^{-1}\sum_{t=1}^{T} \theta^{(t)}$. The results from the previous section suggest that, when the temperature ladder is fixed, a sensible heuristic might be to consider combining the individual estimators with weights λ_i^* proportional to product of $T_i^{-1}\mathrm{ESS}_\rho(\boldsymbol{\theta}_i)$ and $\mathrm{ESS}(\mathbf{w}_i)$, say. However, when considering modifications to the number (m) and spacing of inverse temperatures $\mathbf{k} = \{k_1, \ldots, k_m\}$, there is clearly a conflict of interest between the two measures of effective sample size. Adding more inverse temperatures near one may increase $\mathrm{ESS}(\mathbf{w}^{\lambda^*})$, but may also increase autocorrelation in the marginal chain for k. Therefore it may be sensible to factor $\mathrm{ESS}_\rho(\mathbf{k})$ into the objective as well. Searching for temperature ladders that maximise a hybrid of ESS and ESS_ρ would represent a natural extension of this work.

References

Atchadé, Y. and Liu, J. (2007). "The Wang–Landau algorithm in general state spaces: applications and convergence analysis." Tech. rep., University of Harvard.

Breiman, L., Friedman, J. H., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth.

Del Moral, P., Doucet, A., and Jasra, A. (2006). “Sequential Monte Carlo Samplers.” *Journal of the Royal Statistical Society, Series B*, 68, 411–436.

Geyer, C. (1991). “Markov chain Monte Carlo Maximum Likelihood.” In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, 156–163.

Geyer, C. and Thompson, E. (1995). “Annealing Markov chain Monte Carlo with applications to ancenstral inference.” *Journal of the American Statistical Association*, 90, 909–920.

Gramacy, R. B. and Lee, H. K. H. (2006). “Bayesian treed Gaussian process models.” Tech. rep., Dept. of Applied Math & Statistics, University of California, Santa Cruz.

Green, P. (1995). “Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination.” *Biometrika*, 82, 711–732.

Hukushima, K. and Nemoto, K. (1996). “Exchange Monte Carlo Method and Application to Spin Glass Simulations.” *Journal of the Physical Society of Japan*, 65, 4, 1604–1608.

Iba, Y. (2001). “Extended ensemble Monte Carlo.” *International Journal of Modern Physics*, 12, 5, 623–656.

Jasra, A., Stephens, D., and Holmes, C. (2007a). “On Population-based Simulation for Static Inference.” *Statistics and Computing*, 17, 3, 263–279.

— (2007b). “Population-based reversible jump Markov chain Monte Carlo.” *Biometrica*. (to appear).

Jennison, C. (1993). “Discussion on the meeting on the Gibbs sampler and other Markov chain Monte Carlo methods.” *Journal of the Royal Statistical Society, Series B*, 55, 54–56.

Kass, R. E., Carlin, B. P., Gelman, A., and Neal, R. M. (1998). “Markov Chain Monte Carlo in Practice: A Roundtable Discussion.” *The American Statistician*, 52, 2, 93–100.

King, R. and Brooks, S. (2002). "Model Selection for Integrated Recovery/Recapture Data." *Biometrics*, 58, 841–851.

Kirkpatrick, S., Gelatt, C., and Vecci, M. (1983). "Optimization by simulated annealing." *Science*, 220, 671–680.

Kushner, H. and Lin, G. (1997). *Stochastic Approximation Algorithms and Applications*. New York: Springer.

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. New York: Springer.

Madras, N. and Picconi, M. (1999). "Importance sampling for families of distributions." *Annals of Applied Probability*, 9, 1202–1225.

Marinari, E. and Parisi, G. (1992). "Simulated tempering: A new Monte Carlo scheme." *Europhysics Letters*, 19, 451–458.

Neal, R. M. (1996). "Sampling from multimodal distributions using tempered transition." *Statistics and Computing*, 6, 353–366.

— (2001). "Annealed Importance Sampling." *Statistics and Computing*, 11, 125–129.

— (2005). "Estimating ratios of normalizing constants using Linked Importance Sampling." Tech. Rep. 0511, Department of Statistics, University of Toronto. 37 pages.

Owen, A. and Zhou, Y. (2000). "Safe and Effective Importance Sampling." *Journal of the American Statstical Association*, 95, 449, 135–143.

Veach, E. and Guibas, L. J. (1995). "Optimally combining sampling techniques for Monte Carlo rendering." In *SIGGRAPH '95 Conference Proceedings*, 419–428. Reading, MA: Addison–Wesley.

Wong, W. and Liang, F. (1997). "Dynamic weighting in Monte Carlo and optimization." In *Proceedings of the National Academy of Sciences of USA*, vol. 94(26), 14220–14224.