Importance Tempering

Robert B. Gramacy & Richard J. 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

August 16, 2007

Abstract

Simulated tempering (ST) is an established Markov Chain Monte Carlo (MCMC) methodology for sampling from a multimodal density $\pi(\theta)$. The technique involves introducing an auxiliary variable k taking values in a finite subset of $[0, 1]$ and indexing a set of tempered distributions, say $\pi_k(\theta) \propto \pi(\theta)^k$. 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), has tended not work well in practice. 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 this optimal combination has a highly desirable property related to the notion of effective sample size. The methodology is applied in two modelling scenarios requiring reversible-jump MCMC, where the naïve approach to IT fails: model averaging in treed models, and model selection for mark–recapture data.

Key words: simulated tempering, importance sampling, Markov chain Monte Carlo

(MCMC), Metropolis–coupled MCMC, reversible–jump MCMC, treed model, Gaussian process, mark–recapture data, model selection, model averaging

1 Introduction

Markov chain Monte Carlo (MCMC) algorithms, in particular Metropolis–Hastings (MH) (Metropolis et al., 1953; Hastings, 1970) and Gibbs Sampling (GS) (Geman and Geman, 1984), are by now the most widely used methods for simulation–based inference in (particularly Bayesian) statistics. The beauty of MCMC is its simplicity. In order to sample from a density proportional to $\pi(\theta)$, for $\theta \in \Theta \subseteq \mathbb{R}^d$, the MH and GS algorithms require very little user input or expertise to obtain a realisation of a Markov chain whose stationary distribution is proportional to π. 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, classical MCMC methods such as MH and GS can encounter difficulties 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) (Geyer, 1991; Hukushima and Nemoto, 1996) and simulated tempering (ST) (Marinari and Parisi, 1992; Geyer and Thompson, 1995). Both are inspired by the optimisation technique of simulated annealing (SA) (Kirkpatrick et al., 1983). 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. Conversely, large values of k widen/flatten troughs while raising/sharpening peaks. This is illustrated in Figure 1 for a motivating toy

problem described in Section 2.

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 (Geyer, 1991). The disadvantage is that it needs a good choice of pseudo–prior.

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, and indeed it is often the case that an IS estimator constructed with samples from a tempered distribution can have smaller variance than one based on a sample of the same size from π (cf. Example 2.2 in Section 2). For an inclusive review of IS and related methods see Liu (2001, Chapter 2). Jennison (1993) was the first to suggest using a single tempered distribution as a proposal in IS, and Neal (1996, 2001, 2005) has since written several papers which combine IS and tempered distributions. 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 a relatively trivial matter, it may be surprising that successful IT and IMC^3

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

applications have yet to be published. Liu (2001) comes closest, in proposing to augment ST with a dynamic weighting rule (Wong and Liang, 1997) added to guide transitions between adjacent temperatures.

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 obvious way to do this is with

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

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 $\lambda_i = W^{-1} \sum_{t=1}^{T} w(\theta^{(t)}, k^{(t)}) \mathbb{1}_{\{k^{(t)}=k_i\}}$, 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 to how improve this estimator by choosing $\lambda_1, \ldots, \lambda_m$ to maximise the *effective sample size* (see next paragraph), which approximately corresponds to minimising the variance of $\hat{h}$ (Liu, 2001, 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 like the *balance heuristic* (Veach and Guibas, 1995) or MCV (Owen and Zhou, 2000) 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 Liu (2001, Section 2.5.3) we define the *effective sample size* by

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

where $\mathrm{cv}(\mathbf{w})$ 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* (Kass et al., 1998) (due to serially correlated samples coming from a Markov chain as in MCMC). This latter notion is discussed briefly in Section 5.

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 MC^3 is less useful in (Bayesian) model selection/averaging problems which typically involve trans–dimensional Markov chains as in reversible–jump MCMC (RJMCMC) (Richardson and Green, 1997). Since RJMCMC algorithms are particularly prone to slow mixing, and hence an excellent source of applications of our methodology and results (as illustrated in Section 4), 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. Section 2 consists of two toy examples, as motivation and to help fix ideas. We show how IS with tempered distributions can be helpful in these simple problems, but that ultimately the more powerful apparatus of IT is needed. In Section 3, after a quick review of ST, we discuss how IS can be applied to create more efficient estimators. We derive the optimal convex combination of multiple IS estimators, and show that this optimal combination has a particularly attractive property with regard to its effective sample size. Section 4 shows the effectiveness of optimal IT, and the poor performance of the naïve approach, for model selection in treed and mark–recapture models which require RJMCMC. Section 5 concludes with a discussion.

2 Two motivating examples

Example 2.1: Consider the following toy density π, a mixture of two normals:

$$\pi(\theta) = 0.6N(\theta|\mu_1 = -8, \sigma_1^2 = 0.5^2) + 0.4N(\theta|\mu_2 = 8, \sigma_2^2 = 0.9^2). \tag{3}$$

Figure 1 plots $\pi_k(\theta) \propto \pi(\theta)^k$ for various values of k in $[0, 1]$ (scaled appropriately for clarity).

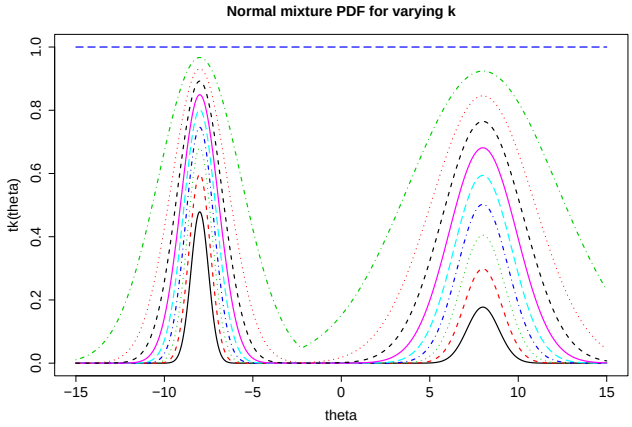


Figure 1: Mixture of two normals with density in (3) at inverse–temperatures $k \in [0, 1]$. The solid black curve corresponds to $k = 1$.

Now consider using the symmetric random–walk Metropolis (RWM) algorithm to sample from π, starting at the mean of the left–hand mode $\theta^{(0)} = \mu_1$ and using proposal distribution $q(\theta^{(t)} \to \theta^{(t+1)}) = N(\theta^{(t+1)}|\theta^{(t)}, \sigma_q^2)$. For good mixing Roberts et al. (1997) suggest choosing σ_q^2 so that the RWM has an acceptance rate of approximately 0.234; pilot tuning suggests $\sigma_q^2 = 6.5$. Standard MCMC diagnostics indicate good mixing has been achieved after 10^5 MCMC iterations, but the Markov chain has not explored the right–hand mode at all!

Following the suggestion of Jennison (1993), we considered the IS estimator obtained from the same RWM algorithm starting at $\theta^{(0)} = \mu_1$, using $\pi'(\theta) = \pi(\theta)^{0.1}$. When using RWM on a tempered distribution, it makes sense to temper the proposal similarly, so that $q(\theta^{(t)} \to \theta^{(t+1)}) = N(\theta^{(t+1)}|\theta^{(t)}, \sigma_q^2)^k \propto N(\theta^{(t+1)}|\theta^{(t)}, \sigma_q^2/k)$. Figure 2 illustrates the improved

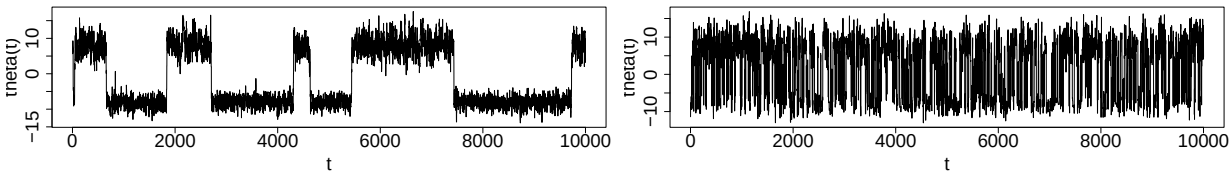


Figure 2: RWM samples $\{\theta^{(t)} : t = 1, \ldots, T\}$ from $\pi'(\theta) = \pi(\theta)^{0.1}$ with original proposal $\sigma_q^2 = 6.5$ *(left)* and tempered proposal $\sigma_q^2/0.1 = 65$ (*right*).

		Mean Squared Error		
	σ_q^2	$\mathbb{P}_\pi(\theta < 0)$	$\mathbb{E}_\pi(\theta)$	$\mathrm{Var}_\pi(\theta)$
Standard RWM for π	6.5	1.6×10^{-1}	40.971	3802.43
IS with RWM for $\pi^{0.1}$	65	6.9×10^{-5}	0.018	0.212

Table 1: MSEs for estimating $\mathbb{P}_\pi(\theta < 0)$, $\mathbb{E}_\pi(\theta)$ and $\mathrm{Var}_\pi(\theta)$ over 100 repeated samples of size 10^5. The true values of these functionals for π in (3) are 0.6, -1.6, and 61.914, respectively.

mixing in the RWM for $\pi^{0.1}$ by showing traces under the original and tempered proposal schemes. Suppose we are interested in estimating $\mathbb{P}_\pi(\theta < 0)$, $\mathbb{E}_\pi(\theta)$ and $\mathrm{Var}_\pi(\theta)$. Table 1 compares RWM sampling from π versus IS with RWM samples from $\pi^{0.1}$ with tempered proposals via the mean squared error (MSE) obtained when estimating these functionals. Jointly, the table and figure attest to the power of combining tempering with IS to improve mixing in the Markov chain between the isolated modes of a distribution. The discussion of this example is continued at the end of Section 3.2.

Example 2.2: 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

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

Note that $k^* \in (1/2, 1)$ for all μ and σ^2. Moreover, there exists $k^- = k^-(\sigma/\mu) < k^*$ 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, $\mathrm{Var}(\hat{\mu}_k) \to \infty$ as $k \to 0$ for all μ and σ^2. Table 2 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 2: Values of k^* and k^- for various values of σ/μ.

This example, and the previous one, highlight a trade-off in the choice of tempered IS proposals. On the one hand, low inverse–temperatures k guard against missing modes of π with large support by encouraging better mixing *between* modes, but can yield very inefficient 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 $k = 1$ by exploiting a bias/variance trade-off. Desire to get the best of both worlds will serve as our motivation for combining ST and IS in the next section.

3 Importance tempering

We begin by outlining standard approaches to rung–spacing and pseudo–prior adjustment for ST. Next, an IS estimator is constructed for each rung on the temperature ladder $1 = k_1 > \ldots > k_m \geq 0$, giving a total of m estimators. A new optimal algorithm is given for combining the m IS estimators, where proposals need only be known up to a normalising constant.

3.1 Simulated tempering

The *simulated tempering* (ST) (Geyer and Thompson, 1995) algorithm is an application of the Metropolis–Hastings (MH) algorithm on the product space of parameters and inverse–temperatures. That is, ST uses MH to sample 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 inverse–temperature is allowed to take on a discrete set of values $k \in \{k_1, \ldots, k_m : k_1 = 1,\ k_i > k_{i+1} \geq 0\}$. Typically, ST calls for sampling $(\theta, k)^{(t+1)}$ by first updating $\theta^{(t+1)}$ conditional on $k^{(t)}$ and (possibly) on $\theta^{(t)}$, using MH or GS. Then, for a proposed $k' \sim q(k^{(t)} \to k')$, usually giving equal probability to the nearest inverse–temperatures greater and less than $k^{(t)}$, an acceptance ratio is calculated:

$$A^{(t+1)} = \frac{\pi(\theta^{(t+1)})^{k'} p(k') q(k' \to k^{(t)})}{\pi(\theta^{(t+1)})^{k^{(t)}} p(k^{(t)}) q(k^{(t)} \to k')}.$$

Finally, $k^{(t+1)}$ is determined according to the MH accept/reject rule: set $k^{(t+1)} = k'$ with probability $\alpha^{(t+1)} = \min\{1, A^{(t+1)}\}$, or $k^{(t+1)} = k^{(t)}$ otherwise. Standard theory for MH and GS gives that samples from the marginals π_{k_i} can be obtained by collecting samples $\theta^{(t)}$ where $k^{(t)} = k_i$. Samples from $\pi(\theta)$ are obtained when $k^{(t)} = 1$.

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 increased; (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.) setting the pseudo-prior $p(k)$ for inverse–temperature. Geyer and Thompson (1995) give ways of adjusting the spacing of the rungs of the ladder so that the ST algorithm achieves between–temperature acceptance rates of 20–40%. More recently, authors have preferred to rely on defaults, e.g.,

$$k_i = \begin{cases} (1+\Delta_k)^{1-i} & \text{geometric spacing} \\ \{1+\Delta_k(i-1)\}^{-1} & \text{harmonic spacing} \end{cases} \qquad i = 1, \ldots, m. \tag{5}$$

The rate parameter $\Delta_k > 0$ can be problem specific. Rather than work with Δ_k, we prefer to set the ladder via m and the hottest temperature k_m, thus fixing Δ_k implicitly. I.e., for

the geometric ladder $\Delta_k = (k_m)^{1/(1-m)} - 1$, and for the harmonic ladder $\Delta_k = \frac{(k_m)^{-1}-1}{m-1}$.

Motivation for such default spacings is outlined by Liu (2001, Chapter 10: pp. 213 & 233). Geometric spacing, or uniform spacing of $\log(k_i)$, is also advocated by Neal (1996, 2001) to encourage the Markov chain to traverse rapidly the breadth of the temperature ladder. Harmonic spacing (Geyer, 1991) is more often used for MC^3 (e.g. Altekar et al., 2004).

Once a suitable ladder has been chosen, we follow the suggestions of Geyer and Thompson (1995) in setting the pseudo–prior, starting from a uniform p_0. First, p_0 is adjusted by *stochastic approximation*: add $c_0/[m(t + n_0)]$ to $\log p_0(k)$ for each $k_i \neq k^{(t)}$ and subtract $c_0/(t+n_0)$ from $\log p_0(k^{(t)})$ over $t = 1, \ldots, B$ *burn–in* MCMC rounds sampling from the joint posterior of (θ, k). Then, p_0 is normalised to obtain p_1. Finally, a second MCMC run is used to collect *occupation numbers* $o(k_i) = \sum_{t=1}^{B} \mathbb{1}_{\{k^{(t)}=k_i\}}$, which are used set $p(k_i) \propto p_1(k_i)/o(k_i)$.

3.2 A new optimal way to combine IS estimators

ST provides us with $\{(\theta^{(t)}, k^{(t)}) : t = 1, \ldots, T\}$, where $\theta^{(t)}$ is an observation from $\pi_{k^{(t)}}$. It is convenient to 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 $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iT_i})$, so that $\{\theta_{ij}\}_{j=1}^{T_i} \sim \pi_{k_i}$. Similarly, the vector of IS weights at the i^{th} temperature is $\mathbf{w}_i = \mathbf{w}_i(\boldsymbol{\theta}_i) = (w_i(\theta_{i1}), \ldots, w_i(\theta_{iT_i}))$, 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

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

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

tion:

$$\hat{h}_\lambda = \sum_{i=1}^{m} \lambda_i \hat{h}_i, \qquad \text{where} \qquad 0 \leq \lambda_i \leq \sum_{i=1}^{m} \lambda_i = 1. \tag{6}$$

Unfortunately, if $\lambda_1, \ldots, \lambda_m$ are not chosen carefully, $\mathrm{Var}(\hat{h}_\lambda)$, can be nearly as large as the largest $\mathrm{Var}(\hat{h}_i)$ (Owen and Zhou, 2000). Notice that ST is recovered as a special case when $\lambda_1 = 1$ and $\lambda_2, \ldots, \lambda_m = 0$. It may be tempting to choose $\lambda_i = W_i/W$, where $W = \sum_{i=1}^{m} W_i$, recovering the estimator in Eq. (1). This can lead to a very poor estimator, even compared to ST, which is demonstrated empirically in Section 4.

Observe that we can write

$$\hat{h}_\lambda = \sum_{i=1}^{m} \sum_{j=1}^{T_i} w_{ij}^{\lambda} h(\theta_{ij}), \tag{7}$$

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, Veach and Guibas (1995) explore combinations of IS estimators of the form in (7), except that $w_{ij}^{\lambda} = \lambda_{ij} w_{ij}/W_i$, where

$$\lambda_{ij} = \frac{c_i g_i(\theta_{ij})}{\sum_{r=1}^{m} c_r g_r(\theta_{ij})}, \tag{8}$$

and $0 \leq c_i \leq \sum_{i=1}^{m} c_i = 1$ is the proportion of samples taken from g_i.[2] It turns out that this is equivalent to IS with the mixture proposal $\tilde{\pi}(\theta) = \sum_{r=1}^{m} c_r g_r(\theta)$:

$$\hat{h}_{\mathrm{bal}} \equiv \frac{1}{T} \sum_{t=1}^{T} w(\theta_t) h(\theta_t), \qquad \text{where} \qquad w(\theta) = \frac{\pi(\theta)}{\sum_{r=1}^{m} c_r g_r(\theta)}. \tag{9}$$

The balance heuristic is a special case of a more general approach called MCV (Owen and Zhou, 2000) that combines *defensive Mixture* importance sampling (Hesterberg, 1995) with

[2] g_i may be any IS proposal density, not necessarily a tempered version of π

Control Variates (Ripley, 1987) to yield a pooled estimator with variance nearly equal to that of the best estimator in the pool—even when some $\hat{h}_i$ have infinite variance. The method involves regressing $h(\theta_t)\pi(\theta_t)/\sum_{r=1}^m g_r(\theta_t)$ on $g_i(\theta_t)/\sum_{r=1}^m g_r(\theta_t)$, and the special case of $\hat{h}_{\text{bal}}$ arises when the regression coefficients are set to zero.

Note that due to the denominator in the definition of $w(\theta)$ in Eq. (9), 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. Moreover, the balance heuristic 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 seek maximise the effective sample size of $\hat{h}_\lambda$ in (6), and look for an $O(T)$ operation to determine the optimal λ^*.

Theorem 3.1. *Among estimators of the form (6),* $\text{ESS}(\mathbf{w}^\lambda)$ *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

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

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

In the following discussion and in Theorem 3.2 below, we assume that $T_i \ge 2$ for $i = 1, \ldots, m$. The efficiency of each IS estimator $\hat{h}_i$ can be measured through $\text{ESS}(\mathbf{w}_i)$.

Intuitively, we hope that with a good choice of λ, the ESS (2) 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

$$\mathrm{ESS}(\mathbf{w}_i) = \frac{T_i(T_i-1)\ell_i}{T_i^2 - \ell_i}. \tag{10}$$

Theorem 3.2 below shows that this is indeed the case for $\hat{h}_{\lambda^*}$.

Theorem 3.2. *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 (10) that $\ell_i \leq T_i$. Thus

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

since $x(1-x)$ is maximised at $x = 1/4$ and $\sum \ell_i \leq \sum T_i = T$. □

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

Example 2.1 (cont.): Returning to our toy example of the mixture of normals (3), Table 3 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

		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 3: Summary of K–S distances to the true mixture of normals (3) 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.

optimally–combined IT estimator has both the largest ESS and the smallest variance of the three estimators. Also notice that on average $\text{ESS}(\mathbf{w}^{\lambda^*}) > \sum_i \text{ESS}(\mathbf{w}_i)$. Naïve IT improves upon ST in this example, but has higher variance than $\hat{h}_{\lambda^*}$. In the next section we show how the naïve method can fail spectacularly, yielding lower ESS than vanilla ST.

4 Applications to reversible-jump MCMC

Bayesian model selection, and averaging, with RJMCMC (Richardson and Green, 1997) is notoriously prone to poor mixing. The trans–dimensional Markov chains can easily become stuck in a local mode of the posterior of model space, failing to "back–out" to find other equivalent, or better, models with disparate parameterisation. Prime examples include treed models (e.g., Chipman et al., 1998), covariate selection (e.g., George and McCulloch, 1997), choosing the order of an auto–regressive or moving–average (ARMA) process (e.g., Ehlers and Brooks, 2006), etc.

In this section we apply the IT methodology to two recent applications of RJMCMC. The first is a treed Gaussian process (GP) (Gramacy and Lee, 2006), which uses RJMCMC to integrate out the tree component. Standard and importance–tempered treed GP models are implemented in the (beta version 1.3) `R` package called `tgp`. The second is an application to

model selection for integrated recovery/recapture data (King and Brooks, 2002). Applying the balance heuristic in either application is infeasible since calculating the normalising constants is intractable.

4.1 Bayesian treed Gaussian process models

Bayesian treed models extend classification and regression tree (CART) models (Breiman et al., 1984), by putting a prior on the tree structure and using RJMCMC to integrate out the treed partitions (Denison et al., 1998; Chipman et al., 1998). We focus on the implementation of Chipman et al. (1998), who specify the tree prior through a process that limits how deep trees can be grown, and then define the tree operations *grow*, *prune*, *change*, and *swap*, which are accepted or rejected following the reversible jump acceptance ratio. "Constant" models (i.e., i.i.d. normal with mean and variance independent of other partitions) are used at the leaves. Chipman et al. (2002) extended this idea to fitting linear models at the leaves of the trees, and Gramacy and Lee (2006) to full GP models.

To ensure that the RJMCMC for the tree process did not get stuck in local modes of the posterior, Chipman et al. (1998, 2002) recommended regularly restarting the chain from the null tree. Though treed GP models tend not to grow as deeply, similar tactics are necessary.

We apply ST to the treed GP model using a geometric ladder with $m = 40$ and $k_m = 0.1$. Only the multivariate normal likelihood (for the GP) and the tree prior were tempered, by powering up; hierarchical priors were left untempered. Proposals for most parameters in the model are naturally tempered via GS with a tempered likelihood. Parameters to the GP correlation function, which require MH steps, use tempered random–walk proposals.

4.1.1 Univariate Motorcycle accident data

With one–dimensional input–data, the treed GP model is a regression model with change–points. Consider a fit of the motorcycle accident data (Silverman, 1985) shown in Figure 3.

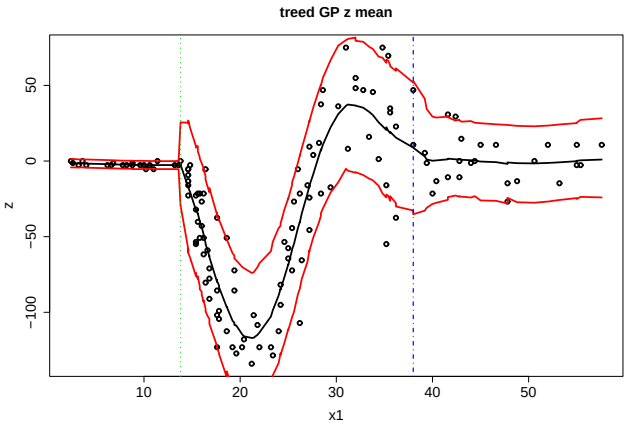
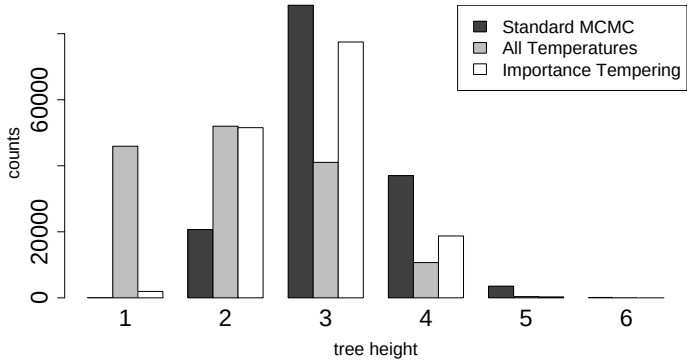


Figure 3: Treed GP fit to the motorcycle accident data.

The main advantage of treed partitioning on this data is to enable the estimation to take into account the heteroscedasticity. Even in this simple case the mixing in tree space can be poor. Figure 4 summarises the mixing in terms of tree heights and acceptance rates for tree operations under the original RJMCMC algorithm, and those obtained from all temperatures under the entire ST chain. Both chains were run for $T = 1.5 \times 10^5$ iterations. A total of $T_1 = 3732$ ($\approx T/m = 3750$) samples were obtained from the cold distribution. Notice in the figure how the original RJMCMC algorithm never returns to trees of height



	%accept	
tree op	vanilla	ST
grow	3.3	4.4
prune	**3.2**	**6.5**
change	40.0	46.3
swap	89.3	95.6

Figure 4: Tree heights encountered for the motorcycle accident data, and tree operations accepted, under vanilla RJMCMC and with ST. *All temperatures* refers to $\{k^{(t)}\}_{t=1}^{T}$ obtained from the ST chain on (θ, k), whereas the IT temperatures are weighted according to $\mathbf{w}^{\lambda^*}$.

one (after burn–in) in contrast to the many visits by the ST chain. The improved mixing of the ST chain is also evident in the increased rate of accepted tree operations, also shown in the figure. In particular, the doubled rate of accepted *prune* operations explains how the tempered distributions helped the chain escape the local modes of deep trees. Re–weighting with the optimally–combined importance weights from Theorem 3.1 facilitates a comparison between IT and vanilla RJMCMC in terms of posterior tree heights. The original RJMCMC chain correctly identified a *maximum a posteriori* (MAP) tree height of three, but overestimated the posterior of height four. Consequently, it also underestimated the posterior probability of trees with height two and one. That $\mathrm{ESS}(\mathbf{w}^{\lambda^*}) = 9338 \approx 2.5T_1$ shows the considerable improvement of IT over ST. The naïve combination $\lambda_i = \frac{W_i}{W}$ in (1) yields $\mathrm{ESS}(\mathbf{w}^{\lambda}) = 285 < \frac{1}{10}T_1$, undermining the very motivation of IT.

4.1.2 Simulated 2–d data

Figure 5 shows a fit, using a treed GP, to 400 observations from a model for independent realisations of (x_1, x_2, z) in which $(x_1, x_2) \in [-6, 6]^2$ are chosen D–optimally (Santner et al., 2003) and

$$z = x_1 \exp(-x_1^2 - x_2^2) + \epsilon, \qquad \text{where} \qquad \epsilon \sim N(0, 0.001^2).$$

There are many equivalent treed partitionings that separate the interesting region near the origin from the flat region on the periphery. One possible partitioning—the MAP found in a single RJMCMC run—is shown graphically in the centre panel, and diagrammatically on the right. Partitioning is useful since it allows for a calculation of region–specific predictive variance (centre panel).

Once the RJMCMC chain finds a tree like the one in Figure 5 it can become stuck, unable to prune back to find other high posterior trees which split up the (x_1, x_2) space in

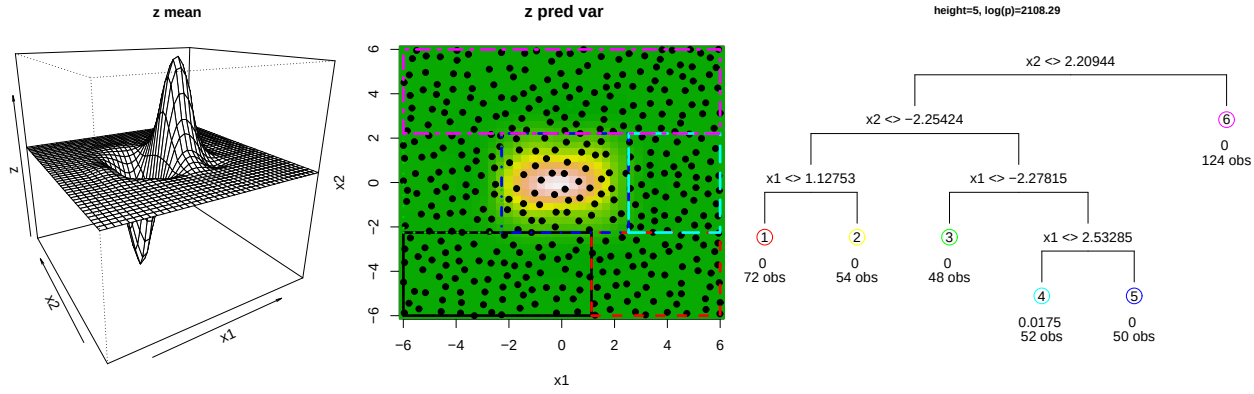
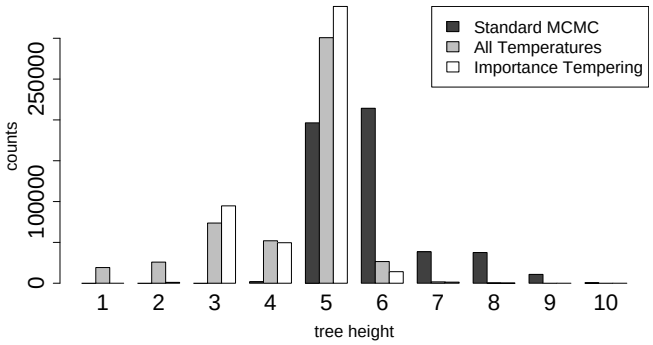
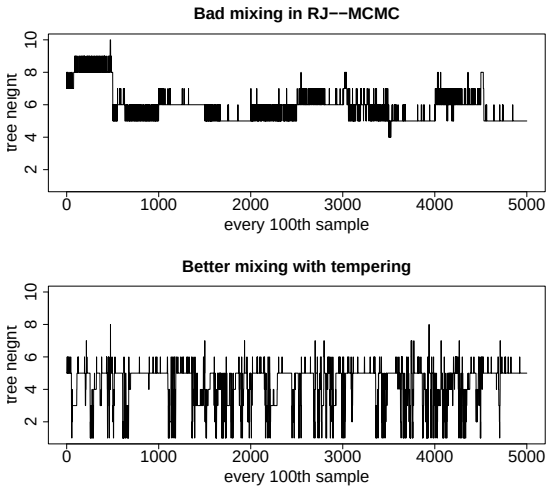


Figure 5: Treed GP fit to the 2–d simulated bivariate data in terms of posterior mean *(left)* and variance *(centre)* surfaces with MAP tree *(centre and right)*.

a different order (say, starting with x_1 rather than x_2). The split `x_1 <> 1.127` is spurious; it is not clear how this split is supported by the data. Deeper trees, i.e. those with height larger than five, should be viewed with even greater scepticism.

Figure 6 summarises mixing in terms of tree heights under the untempered RJMCMC algorithm, and those obtained from all temperatures under ST. Both chains were run for a total of $T = 5 \times 10^5$ iterations. The untempered RJMCMC was restarted every 5×10^4 iterations, following the recommendation of Chipman et al. Under ST, a total of $T_1 = 12436$ ($\approx T/m = 12500$) samples were obtained from the cold distribution. The left–hand side of the figure compares tree heights observed in the two chains and those obtained by weighting according to $\mathbf{w}^{\lambda^*}$. The right–hand figures show traces of tree–heights captured every 100 iterations. Notice how the untempered RJMCMC chain (top–right panel) gets trapped until the tree is re–started, producing ten visually detectable regimes. After the burn–in period briefly following each tree restart, the Markov chain never visits trees below height four. Unable to prune successfully, the chain makes periodic and lengthy excursions into trees as deep as height ten, which is clearly not warranted by the data. The ST chain, by contrast, moves much more freely between trees, frequently pruning back to null trees (bottom panels). It spends most of its time at height five or lower, making periodic but





	tree operation: % accept			
method	grow	**prune**	change	swap
vanilla	0.9	**0.8**	65.9	44.3
ST	9.7	**10.56**	83.1	63.1

Figure 6: Tree heights in the untempered RJMCMC and ST chains via histogram *(top–left)*; accepted tree operations *(bottom–left)*; and traces *(right)* taken every 100 iterations *(top: RJMCMC; bottom: ST)*.

short excursions into deeper trees. Compared to IT, it is clear that the RJMCMC chain has missed many more parsimonious models, particularly those at height three. Finally, $\text{ESS}(\mathbf{w}^{\lambda^*}) = 21778 \approx 1.75T_1$, illustrating how IT improves on ST. The naïve combination $\lambda_i = \frac{W_i}{W}$ in (1) yields $\text{ESS}(\mathbf{w}^{\lambda^*}) = 654 \approx \frac{1}{18}T_1$—worse than ST.

4.2 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, off the coast of Scotland. The three demographic parameters of interest are: survival rates, recapture rates and recovery rates. Models in which each of these parameters was allowed to be age–and/or time–dependent were considered, where the time dependence was conditional on the age structure of the parameters. For further details of the data, model structure, and RJMCMC algorithm see King and Brooks (2002). 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.

In implementing ST, we use a geometric ladder with $m = 40$ and $k_m = 0.1$. We use a random–walk MH algorithm to update the parameters within each model. The proposal distribution was chosen to be uniform within $\pm\ \delta$ of the current parameter value— for the recapture rate $\delta = 0.05$; $\delta = 0.1$ for the recovery and survival parameters. Between model moves (modifying time or age dependence) are as described by King and Brooks (2002).

The simulations are run for 10^7 iterations with the first 10% discarded as burn-in. Table 4 tallies the total number of (marginal) models that are visited along the ST chain compared to the (untempered) vanilla RJMCMC algorithm. It can be clearly seen that the number of models visited is significantly improved under the ST chain (i.e., ST & IT). Overall, the total numbers of different models visited by the Markov chain (i.e. combinations of the marginal models for all of the demographic parameters), were 3080 for IT, compared to 233 for the standard RJMCMC and only 177 for ST. Table 5 summarises the between–model acceptance rates under each model move. All rates are increased in the ST chain, drastically in the cases of *merge age* and *add time*.

Marginal model	Survival rates			Recapture rates			Recovery rates		
	IT	ST	RJ	IT	ST	RJ	IT	ST	RJ
Age and time	51	26	16	12	5	3	75	25	28
Age only	7	4	3	7	4	2	15	11	12

Table 4: Number of models visited in an untempered RJMCMC, compared to the ST chain. For survival rates there are 54 possible marginal models with age– and time–dependence and 10 with age dependence only; for recapture and recovery rates there are 94 and 15 respectively.

For the ST approach, from the $T = 9 \times 10^6$ iterations following burn-in, a total of $T_1 = 248158$ ($\approx T/m = 225000$) realisations were obtained from the cold distribution. By comparison, for optimal IT we have $\mathrm{ESS}(\mathbf{w}^{\lambda^*}) = 612026 \approx 2.5 \times T_1$. The corresponding

	Model Moves: % accept			
Approach	split age	merge age	add time	remove time
RJMCMC	1.30	0.50	0.01	0.14
ST chain	1.32	1.21	0.30	1.45

Table 5: Acceptance rates for model moves under the ST chain and untempered RJMCMC: "split" and "merge" age—increasing or decreasing the number of age groups by one; "add" and "remove" time—for the given age group, propose to add or remove time dependence.

naïve IT approach (using $\lambda_i = \frac{W_i}{W}$) performed exceptionally poorly, with ESS($\mathbf{w}^\lambda$) of only 5.43. This can be blamed on a few very large weights obtained at hot temperatures.

5 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 a successful application has never been published in the literature.

One reason for the lack of IT applications has been the absence of a principled way of combining the IS estimators obtained at each temperature. We have therefore derived optimal combination weights, which can be calculated even when the normalisation constants of the tempered distributions are unknown—as is usually the case in practice. This framework is equally applicable within MC^3, and can certainly be applied outside of the domain of tempered MCMC to combine IS estimators obtained from any proposal distributions. We proved that the effective sample size of the optimally–combined estimator is guaranteed to be almost as large as the sum of the corresponding effective sample sizes at each temperature,

and in fact it is often larger in practice.

In addition to using a motivating example, IT has been applied to two different model selection problems which require the use of reversible–jump MCMC, and are thus prone to poor mixing. The first involved Bayesian treed models with model averaging, and the second involved a model selection task for mark–recapture data. In both cases we showed that ST yields improved mixing in model space, and that, when applied carefully, the IT methodology can indeed increase the resulting ESS compared to retaining samples only from the cold distribution. The computational cost required of this post-processing operation is $O(T)$, and its implementation requires minimal programming effort.

Since IT involves sampling from a Markov chain, ideally one would take into account the serial correlation in an appropriate notion of effective sample size. The *effective sample size due to autocorrelation* is defined (Kass et al., 1998) by

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

where $\hat{\rho}(\ell)$ is the sample autocorrelation at lag ℓ; thus $\hat{\rho}(\ell) = \hat{\gamma}(\ell)/\hat{\gamma}(0)$, where $\hat{\gamma}(\ell) = (T-\ell)^{-1}\sum_{t=1}^{T-\ell}(\theta^{(t)} - \bar{\theta})(\theta^{(t+\ell)} - \bar{\theta})$, and $\bar{\theta} = T^{-1}\sum_{t=1}^{T}\theta^{(t)}$. It is natural to try to combine these two concepts of effective sample size. In the light of Theorem 3.2, one way to do this might be to define an *overall effective sample size* (OESS) by

$$\mathrm{OESS}(\boldsymbol{\theta}) = T^{-1}\mathrm{ESS}_c(\mathbf{k}) \sum_{i=1}^{m} \frac{\mathrm{ESS_w}\big(\mathbf{w}_i(\boldsymbol{\theta}_i)\big) \times \mathrm{ESS_c}(\boldsymbol{\theta}_i)}{T_i}, \tag{12}$$

where $\mathbf{k} = (k^{(1)}, \ldots, k^{(T)})$. For example, when modifying the temperature ladder, adding more inverse–temperatures near one will increase $\mathrm{ESS}(\mathbf{w}^{\lambda^*})$, but will also increase autocorrelation in the marginal chain for k. Searching for temperature ladders that maximise (12) would represent a natural next step in the development of importance tempering.

References

Altekar, G., Dwarkadas, S., Huelsenbeck, J. P., and Ronquist, F. (2004). "Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference." *Bioinformatics*, 20, 3, 407–415.

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

Chipman, H., George, E., and McCulloch, R. (1998). "Bayesian CART Model Search (with discussion)." *Journal of the American Statistical Association*, 93, 935–960.

— (2002). "Bayesian Treed Models." *Machine Learning*, 48, 303–324.

Denison, D., Mallick, B., and Smith, A. (1998). "A Bayesian CART Algorithm." *Biometrika*, 85, 363–377.

Ehlers, R. and Brooks, S. (2006). "Bayesian Analysis of Order Uncertainty for ARIMA Models." Tech. rep., Statistical Laboratory, University of Cambridge.

Geman, S. and Geman, D. (1984). "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

George, E. and McCulloch, R. (1997). "Approaches to bayesian Variable Selection." *Statistica Sinica*, 7, 339–373.

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.

Hastings, W. (1970). "Monte Carlo Sampling Methods using Markov Chains and their Applications." *Biometrika*, 57, 97–109.

Hesterberg, T. (1995). "Weighted Average Importance Sampling and Defensive Mixture Distributions." *Technometrics*, 37, 2, 185–194.

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.

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.

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

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

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, R. (1953). "Equations of State Calculations by Fast Computing Machine." *Journal of Chemical Physics*, 21, 1087–1091.

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.

Richardson, S. and Green, P. J. (1997). "On Bayesian Analysis of Mixtures With An Unknown Number of Components." *Journal of the Royal Statistical Society, Series B, Methodological*, 59, 731–758.

Ripley, B. D. (1987). *Stochastic Simulation*. New York: Wiley.

Roberts, G., Gelman, A., and Gilks, W. (1997). "Weak convergence and optimal scaling of random walk Metropolis algorithms." *Annals of Applied Probability*, 7, 110–120.

Santner, T. J., Williams, B. J., and Notz, W. I. (2003). *The Design and Analysis of Computer Experiments*. New York, NY: Springer-Verlag.

Silverman, B. W. (1985). "Some Aspects of the Spline Smoothing Approach to Non-Parametric Curve Fitting." *Journal of the Royal Statistical Society Series B*, 47, 1–52.

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.