Appendices to
“Resampling Markov chain Monte Carlo algorithms:
Basic analysis and empirical comparisons”

Appendix I: Technical details

Lemma 1. Let $S_1 = \{x_1, \ldots, x_{n_1}\}$ be an asymptotically regular sample from $P_1$, and $S_2 = \{y_1, \ldots, y_k\}$ be a sample of size $k$ for $P_2$ obtained by importance resampling from $S_1$. Then for any function $g(\cdot)$ under suitable conditions, as $n_1 \to \infty$ and $k \to \infty$,

\[
E(\bar{g}_2) = E_2(g) + O\left(\frac{1}{n_1}\right),
\]

\[
\text{var}(\bar{g}_2) = \frac{1}{n_1} v_1 \{g - E_2(g)\} + \frac{1}{k} V_2(g) + o\left(\frac{1}{n_1}\right) + O\left(\frac{1}{n_1 k}\right),
\]

where $\bar{g}_2 = k^{-1} \sum_{i=1}^{k} g(y_i)$, $E_2(g)$ and $V_2(g)$ are the expectation and variance of $g(x)$ for $x$ drawn from $P_2$, and $r(\cdot) = p_2(\cdot)/p_1(\cdot)$. The functions $p_1(\cdot)$ and $p_2(\cdot)$ are the normalized densities for $P_1$ and $P_2$.

Proof of Lemma 1. Define $\delta_1, \ldots, \delta_k$ such that if $\delta_i = j$ then $y_i = x_j$. Then $(\delta_1, \ldots, \delta_k)$ are independent and identically distributed given $S_1$. Let $p_j = P(\delta_i = j | S_1)$. Then $p_j = w(x_j)/\sum_{h=1}^{n_1} w(x_h)$, where $w(x) = q_2(x)/q_1(x)$. Rewrite $\bar{g}_2 = k^{-1} \sum_{i=1}^{k} \sum_{j=1}^{n_1} 1\{\delta_i = j\} g(x_j)$. By direction calculation,

\[
E(\bar{g}_2|S_1) = \sum_{j=1}^{n_1} p_j g(x_j),
\]

\[
\text{var}(\bar{g}_2|S_1) = \sum_{j=1}^{n_1} \text{var}\left(\frac{1}{k} \sum_{i=1}^{k} 1\{\delta_i = j\} | S_1\right) g^2(x_j) +
\]

\[
\sum_{j \neq h} \text{cov}\left(\frac{1}{k} \sum_{i=1}^{k} 1\{\delta_i = j\}, \frac{1}{k} \sum_{i=1}^{k} 1\{\delta_i = h\} \right) g(x_j)g(x_h)
\]

\[
= \sum_{j=1}^{n_1} \frac{p_j(1 - p_j)}{k} g^2(x_j) + \sum_{j \neq h} \frac{-p_j p_h}{k} g(x_j)g(x_h)
\]

\[
= \frac{1}{k} \left[ \sum_{j=1}^{n_1} p_j g^2(x_j) - \left\{ \sum_{j=1}^{n_1} p_j g(x_j) \right\}^2 \right].
\]

Under suitable regularity conditions, it can be shown that

\[
E\left\{ \sum_{j=1}^{n_1} p_j g(x_j) \right\} = E_2(g) + O\left(\frac{1}{n_1}\right),
\]
By simple calculation, according to Proposition 1, the results then follow because
\[ E \left\{ n \sum_{j=1}^{n_1} p_j g(x_j) \right\} = E_2(g) + O \left( \frac{1}{n_1} \right), \]
\[ E \left\{ n \sum_{j=1}^{n_1} p_j g^2(x_j) \right\} = E_2(g^2) + O \left( \frac{1}{n_1} \right), \]
\[ E \left\{ n \sum_{j=1}^{n_1} p_j g(x_j) \right\}^2 = \{ E_2(g) \}^2 + O \left( \frac{1}{n_1} \right). \]

The results then follow by the formulas of iterated expectations and variances, \( E(\tilde{g}_2) = E\{ E(\tilde{g}_2|S_1) \} \) and \( \text{var}(\tilde{g}_2) = \text{var}\{ E(\tilde{g}_2|S_1) \} + E\{ \text{var}(\tilde{g}_2|S_1) \} \).

**Proof of Proposition 1.** If \( y_0 \) is drawn from \( P_2 \), then each \( y_{0:j} \) is also distributed according to \( P_2 \) by the invariance of \( P_2 \) under \( \Psi_2 \) and hence \( E\{ g(y_{0:j}) \} = E_2(g) \) and \( E_2(h) = E\{ \ell^{-1} \sum_{j=1}^{\ell} g(y_{0:j}) \} = E_2(g) \). Let \( \gamma(\cdot) = \text{var}\{ \ell^{-1} \sum_{j=1}^{\ell} g(y_{0:j}) | y_0 = \cdot \} \).

By simple calculation, \( E(\tilde{g}_2|R_2) = k^{-1} \sum_{i=1}^{k} h_i(y_i) \) and \( \text{var}(\tilde{g}_2|R_2) = k^{-2} \sum_{i=1}^{k} \gamma(y_i) \).

Consider the case where \( \ell \) is bounded and \( k \to \infty \). By the rules of iterated expectations and variances and by Lemma 1,
\[ E(\tilde{g}_2) = E\{ E(\tilde{g}_2|R_2) \} = E_2(h) + O \left( \frac{1}{n_1} \right), \]
\[ \text{var}(\tilde{g}_2) = \text{var}\{ E(\tilde{g}_2|R_2) \} + E\{ \text{var}(\tilde{g}_2|R_2) \} \]
\[ = \frac{1}{n_1} v_1 \left[ r\{ h - E_2(h) \} \right] + \frac{1}{k} V_2(h) + \frac{1}{k} E_2(\gamma) + o \left( \frac{1}{n_1} \right) + O \left( \frac{1}{n_1 k} \right). \]

The results then follow because \( E_2(g) = E_2(h) \) and, by the formula of iterated variances, \( \text{var}_{y_0 \sim P_2} \{ \ell^{-1} \sum_{j=1}^{\ell} g(y_{0:j}) \} = V_2(h) + E_2(\gamma) \).

Now consider the case where \( k \) is bounded and \( \ell \to \infty \). By asymptotic theory of Markov chains, \( h(y_i) = E_2(g) + O_p(\ell^{-1}) \) and \( \gamma(y_i) = \text{var}_{y_0 \sim P_2} \{ \ell^{-1} \sum_{j=1}^{\ell} g(y_{0:j}) \} + o_p(\ell^{-1}) \) for each \( i = 1, \ldots, k \). Under suitable regularity conditions, it can then be shown that \( E\{ E(\tilde{g}_2|R_2) \} = E_2(g) + O(n_2^{-1}) \), \( \text{var}\{ E(\tilde{g}_2|R_2) \} = o(n_2^{-1}) \), and \( E\{ \text{var}(\tilde{g}_2|R_2) \} = k^{-1} \text{var}_{y_0 \sim P_2} \{ \ell^{-1} \sum_{j=1}^{\ell} g(y_{0:j}) \} + o(n_2^{-1}) \).

Therefore, the desired results hold.

**Derivation of tempering distributions in Section 3.2.** Write \( p(x) = \lambda_1 \phi_1(x) + \lambda_2 \phi_2(x) \), where \( \phi_i(x) = (2\pi)^{-d/2} \sigma_i^{-d} \exp\{-\|x - \mu_i\|^2/(2\sigma_i^2)\} \). We show that for \( T \) in a certain range, \( \{ p(x) \}^{1/T} \) can be approximated by \( p_T(x) = \{ \lambda_1 \phi_1(x) \}^{1/T} + \{ \lambda_2 \phi_2(x) \}^{1/T} \), which is an unnormalized density of the Gaussian mixture described in the text. The normalizing constant for \( p_T(x) \) is \( C_1 + C_2 \), where \( C_i = \lambda_i^{1/T} \{(2\pi)^{d/2}(\sigma_i \sqrt{T})^d\}^{1/T} \).
{\{(2\pi)^{d/(2T)}\sigma_i^{d/T}\}}. For 0 < \alpha < 1, let \( A_i = \{x : \|x - \mu_i\|^2 \leq \sigma_i^2 T \lambda_{d,\alpha}^2\} \). Assume that \( A_1 \cap A_2 = \emptyset \). Then the following inequality holds:

\[
|\{p(x)\}^{1/T} - p_T(x)| \leq \begin{cases} 
(\epsilon_1/T + \epsilon_1^{1/T})\{\lambda_1 \phi_1(x)\}^{1/T}, & \text{if } x \in A_1, \\
(\epsilon_2/T + \epsilon_2^{1/T})\{\lambda_2 \phi_2(x)\}^{1/T}, & \text{if } x \in A_2, \\
\{\lambda_1 \phi_1(x)\}^{1/T} + \{\lambda_2 \phi_2(x)\}^{1/T}, & \text{if } x \notin A_1 \cup A_2,
\end{cases}
\]

where \( \epsilon_1 = \sup_{x \in A_1}\{\{\lambda_2 \phi_2(x)\}/\{\lambda_1 \phi_1(x)\}\} \) and \( \epsilon_2 = \sup_{x \in A_2}\{\{\lambda_1 \phi_1(x)\}/\{\lambda_2 \phi_2(x)\}\} \). If \( x \in A_1 \), then \( \{p(x)\}^{1/T} = \{\lambda_1 \phi_1(x)\}^{1/T}[1+\{\lambda_2 \phi_2(x)\}/\{\lambda_1 \phi_1(x)\}]^{1/T} \) and the inequality follows because \( 0 \leq (1+y)^c - 1 \leq cy \) for \( y \geq 0 \) and \( 0 < c < 1 \). If \( x \notin A_1 \cup A_2 \), then the inequality follows because \( 0 \leq (y+z)^c \leq 2(y^c + z^c) \) for \( y, z \geq 0 \) and \( 0 < c < 1 \). Therefore, if \( \alpha \approx 1, \epsilon_1^{1/T} \approx 0 \), and \( \epsilon_2^{1/T} \approx 0 \), then \( \{p(x)\}^{1/T} \) can be well approximated by \( p_T(x) \). By simple calculation, \( \epsilon_1 = (\lambda_2/\sigma_2^d)/(\lambda_1/\sigma_1^d) \exp\{-\|\mu_1 + \sigma_1 \sqrt{T \lambda_{d,\alpha}^2}/d(1,\ldots,1)-\mu_2\|^2/(2\sigma_2^2)+T \lambda_{d,\alpha}^2/2\} \) and \( \epsilon_2 \) is obtained by symmetry. Take, for example, \( \alpha = 95\% \). Then \( \epsilon_1^{1/T} = 4 \times 10^{-13} \) and \( \epsilon_2^{1/T} = 6 \times 10^{-6} \) for \( T = 10 \), and \( \epsilon_1^{1/T} = 0.0002 \) and \( \epsilon_2^{1/T} = 0.01 \) for \( T = 15 \).

**Appendix II: Multimodal bivariate Gaussian mixture**

**Setup and difficulty.** Consider sampling from a multimodal bivariate Gaussian mixture

\[
p(x) = \sum_{i=1}^{20} \frac{\lambda_i}{2\pi \sigma_i^2} \exp\left\{ -\frac{1}{2\sigma_i^2}(x - \mu_i)^T(x - \mu_i) \right\}
\]

where \( (\mu_1, \mu_2, \ldots, \mu_{20}) \) are taken from Liang & Wong (2001) and Kou et al. (2006) and \( \{(\sigma_i, \lambda_i) : i = 1, \ldots, 20\} \) are specified in three ways, depending on \( d_i \) the Euclidean distance from \( \mu_i \) to \((5,5)\):

- \( \sigma_1 = \cdots = \sigma_{20} = 0.1 \) and \( \lambda_1 = \cdots = \lambda_{20} = 0.05 \),
- \( \sigma_i = d_i/20 \) and \( \lambda_i \propto 1/\sqrt{d_i} \) for \( i = 1, \ldots, 20 \),
- \( \sigma_i = 0.25/d_i \) and \( \lambda_i \propto \sqrt{d_i} \) for \( i = 1, \ldots, 20 \).

The three cases are referred to as HOM, HET1, and HET2. For HOM, the Gaussian components have equal variances and weights. For HET1, Gaussian components

\[3\]
closer to \((5, 5)\) have larger weights and smaller variances. The smallest \(\sigma_i\) is 0.036 and the largest 0.29. For HET2, Gaussian components closer to \((5, 5)\) have smaller weights and larger variances. The smallest \(\sigma_i\) is 0.043 and the largest 0.34. For each case, there are clusters of local modes, well separated from each other.

As in Kou et al. (2006), we take \(m = 5\) and use the sequence of tempering distributions, \(q_j(x) = \{p(x)\}^{1/T_j}\), with \((T_1, T_2, T_3, T_4, T_5) = (60.0, 21.6, 7.7, 2.8, 1.0)\). Figure A1 presents the scatterplots of \(P_1\), \(P_2\), and \(P_3\) for the cases HET1 and HET2, based on single runs of the generalized resample-move algorithm. There are more pronounced differences between the two cases at a high temperature than at a low temperature. For example, consider the scatterplots for \(P_2\) at temperature \(T_2 = 21.6\). For the case HET1, the Gaussian components with small variances lie in the center area, surrounded by those with large variances. For the case HET2, the Gaussian components with small variances are scattered in the tail area of those with large variances. Such differences seem to make it more difficult to properly sample from
Table A1: Sizes of energy rings (in percentages)

<table>
<thead>
<tr>
<th></th>
<th>Case HET1</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Case HET2</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$D_1$</td>
<td>$D_2$</td>
<td>$D_3$</td>
<td>$D_4$</td>
<td>$D_5$</td>
<td>$D_1$</td>
<td>$D_2$</td>
<td>$D_3$</td>
</tr>
<tr>
<td>$T_5 = 1$</td>
<td>65.3</td>
<td>34.2</td>
<td>0.5</td>
<td>0.0</td>
<td>0.0</td>
<td>85.2</td>
<td>14.4</td>
<td>0.4</td>
</tr>
<tr>
<td>$T_4 = 2.8$</td>
<td>14.8</td>
<td>66.0</td>
<td>19.1</td>
<td>0.1</td>
<td>0.0</td>
<td>36.2</td>
<td>45.2</td>
<td>18.5</td>
</tr>
<tr>
<td>$T_3 = 7.7$</td>
<td>5.4</td>
<td>42.5</td>
<td>45.0</td>
<td>7.1</td>
<td>0.0</td>
<td>13.5</td>
<td>35.6</td>
<td>46.4</td>
</tr>
<tr>
<td>$T_2 = 21.6$</td>
<td>2.2</td>
<td>22.0</td>
<td>42.9</td>
<td>30.1</td>
<td>2.8</td>
<td>5.6</td>
<td>19.8</td>
<td>47.2</td>
</tr>
<tr>
<td>$T_1 = 60.0$</td>
<td>1.0</td>
<td>11.5</td>
<td>28.3</td>
<td>38.8</td>
<td>20.4</td>
<td>2.3</td>
<td>9.8</td>
<td>29.5</td>
</tr>
</tbody>
</table>

Note: The energy rings are $D_1 = (0.2, 2.0)$, $D_2 = [2.0, 6.3)$, $D_3 = [6.3, 20.0)$, $D_4 = [20.0, 63.2)$, and $D_5 = [63.2, \infty)$.

the sequence of distributions in the case HET2 than in HET1.

**Simulation details.** For our simulations, the Markov kernel $\Psi_j$ corresponds to 1 iteration of random-walk Metropolis sampling, with a bivariate Gaussian proposal with mean at the current observation and standard deviation $0.25\sqrt{T_j}$. The parameter $\ell_j$ is set to 10 for each $j$ for generalized resample-move, and $\alpha_j$ is set to 10% for each $j$ for algorithms labeled 4–11. The initial value $x_{10}$ is drawn uniformly from $[0, 1]^2$. The sample sizes are set with $n = 50,000$ and $b = 5,000$.

For equi-energy sampling, the potential function $u(x)$ is defined as $-\log\{p(x)\}$ and the energy levels $(U_1, U_2, \ldots, U_5)$ are set to $(0.2, 2.0, 6.3, 20.0, 63.2)$ for the three cases HOM, HET1, and HET2, although these levels are used only for the case HOM in Kou et al. (2006). Table A1 shows the sizes of energy rings from single runs of static equi-energy sampling for HET1 and HET2. There appear to be adequate differentiation in the relative ring sizes across different temperatures.

**Summary of results.** Estimates are obtained for the means and variances of the components of $x = (x_{(1)}, x_{(2)})$. Figure A2 shows the squared biases and mean squared errors (MSE) for the estimates of $E(x_{(2)})$ and $\text{var}(x_{(2)})$ under $P_5$ based on 200 repeated simulations. For comparison with Kou et al. (2006), the standard deviations for the estimates of $E(x_{(2)})$ and $E(x_{(2)}^2)$ are, respectively, 0.15 and 1.45 for static equi-energy sampling, 0.20 and 2.0 for dynamic equi-energy sampling, and 0.16 and 1.7 for parallel tempering in the case HOM. The results for the latter two algorithms differ noticeably from those in Kou et al. (2006, Table 1), possibly due to various
implementation details. For example, Kou et al. (2006) employed tuning of the step size in the Metropolis proposal and energy truncation in the definition of tempering distributions.

A number of remarks can be drawn from Figure A2. First, generalized resample-move has overall the lowest MSEs among all the algorithms across the three cases. The MSE ratios of resample-move versus generalized one are slightly below 1 in the case HET1, but about 2 in the case HET2. This difference can be attributed to different configurations of distributions in the two cases discussed earlier (see Figure A1). For the case HET1, even sequential importance resampling performs almost the best of all the algorithms, similarly to generalized resample-move. But for the case HET2, sequential importance resampling performs poorly, with MSEs 2–3 times as large as those of generalized resample-move. Second, the three static resampling MCMC algorithms, labelled 4–6, yield MSEs about 2–3 times as large as those of gen-
eralized resample-move across the three cases. The MSEs of importance resampling MCMC and equi-energy sampling are mostly closer to each other than to those of independence Metropolis resampling MCMC, in agreement with Remark 4 in Section 2.4. These three algorithms are substantially better than their dynamic counterparts, labeled 8–10. Third, parallel tempering performs similarly to or slightly worse than static resampling MCMC algorithms labelled 4–6. Finally, independence Metropolis tempering performs rather unsatisfactorily, sometimes with large MSEs due to large biases, particularly for estimating \( \text{var}(x_{(2)}) \).

**Appendix III: Additional results for Section 3.2**

*Experiment on subsampling strategies.* As mentioned in Section 3.2, Figure A3 shows the squared biases and MSEs based on 200 repeated simulations. For the strategy ITER, generalized resample-move has the smallest MSEs. Compared with the simple case before the increase of iterations, the performance of generalized resample-move is substantially improved whereas those of independence Metropolis resampling MCMC and parallel tempering are only slightly improved. The relative performances of these three algorithms under the two choices of \( \Psi_j \) with 1 or 10 Metropolis iterations are in the same direction as under the two choices of \( \Psi_j \) with proposal standard deviation \( 0.2 \sqrt{T_j} \) or \( 0.1 \sqrt{T_j} \) discussed earlier. Generalized resample-move tends to perform better than independence Metropolis resampling MCMC and parallel tempering for a fast-mixing kernel \( \Psi_j \), whereas the latter two algorithms seem advantageous for a slow-mixing kernel \( \Psi_j \).

For the strategy SUB1, independence Metropolis resampling MCMC and parallel tempering have the smallest MSEs. In this case, generalized resample-move is reduced to resample-move \( (\ell_j = 1) \) under the strategy ITER. Compared with the simple case before the increase of iterations, the performances of independence Metropolis resampling MCMC and parallel tempering are substantially improved whereas that of generalized resample-move becomes slightly worse. These changes are in the opposite direction to those in the comparison between ITER and the simple case.

For the strategy SUB2, independence Metropolis resampling MCMC has the small-
est MSEs. The performances of resampling MCMC algorithms labeled 3–6 are comparable to or better than under SUB1, whereas the MSEs of parallel tempering (algorithm 7) are increased by at least a factor of 2 from those under SUB1. Another interesting observation is that for algorithms labeled 4–7, SUB2 leads to considerably more accurate results than ITER, even though the expected rates of resampling or swapping per Markov iteration are the same for the two strategies.

**Experiment on tempering distributions.** For the proposal standard deviation \(0.2\sqrt{T_j}\), we study the effects of using sequences of more finely spaced tempering distributions than in the previous experiments, with a comparable total number of iterations:

- 9 distributions at temperatures \((t_1, t_3, t_5, t_7, t_9, t_{11}, t_{13}, t_{15}, t_{17})\), with the number of iterations \((n + 9b, n + 8b, \ldots, n + b)\) for static algorithms 3–6, where \(n = 10b\) and \(b = 1930\) is the nearest integer that is divisible by 10 and no smaller than

---

Figure A3: Squared bias (×) and MSE (○) for estimates of \(E(x_{(d)})\) in top plots and \(\text{var}(x_{(d)})\) in bottom plots under the target \(P_5\), with longer iterations and subsampling under the choice of proposal standard deviation \(0.2\sqrt{T_j}\). For comparison, the results for SUB1 are also shown as + (squared bias) and □ (MSE) within the limits of the plots for SUB2.
Figure A4: Cumulative relative variances (left) and effective sample fractions (right) for sequences of distributions at 5 (○), 9 (□), 11 (+), and 17 (△) temperatures.

4000 × ∑^5_j=1(10 + j)/∑^9_j=1(10 + j) ≈ 1926.

- 11 distributions at temperatures \(t_1, t_3, t_5, t_7, t_9, t_{11}, t_{13}, t_{14}, t_{15}, t_{16}, t_{17}\), with the number of iterations \((n+11b, n+10b, \ldots, n+b)\) for static algorithms 3–6, where \(n = 10b = 14800\), determined as in the first case.

- 17 distributions at temperatures \(t_1, t_2, \ldots, t_{17}\), with the number of iterations \((n+17b, n+16b, \ldots, n+b)\) for static algorithms 3–6, where \(n = 10b = 8100\), determined as in the first case.

The total number of iterations for parallel tempering is set to match that of the other algorithms. The 17 temperatures \((t_1, t_2, \ldots, t_{17})\) are (50.0, 39.15, 30.67, 24.01, 18.80, 14.72, 11.53, 9.03, 7.07, 5.54, 4.34, 3.40, 2.66, 2.08, 1.63, 1.28, 1.0), evenly spaced in the logarithm between 50 and 1. The 9 temperatures \((t_1, t_3, t_5, t_7, t_9, t_{11}, t_{13}, t_{15}, t_{17})\) are also evenly spaced between 50 and 1. The 11 temperatures in the second case are obtained by interpolation in the intervals \((t_{13}, t_{15})\) and \((t_{15}, t_{17})\). As seen later in Figure A4, there appears to be more substantial distances between the successive distributions at \((t_{13}, t_{15}, t_{17})\) than at the other temperatures. Following Section 3.2.3, we increase the number of iterations by 10 times and then perform subsampling, using the strategy SUB1 for parallel tempering and SUB2 for resampling MCMC algorithms (including ITER for generalized resample-move).

Figure A4, similarly to Figure 5, shows the cumulative relative variances \(RV_2 + \cdots + RV_j\) and the effective sample fractions \(ESS_j/n_{j-1}\), where \(RV_j\) is averaged over
Figure A5: Squared bias (×) and MSE (○) for estimates of $E(x(d))$ in top plots and $\text{var}(x(d))$ in bottom plots under the target $P_m$ with the choice of proposal standard deviation $0.2 \sqrt{T_j}$, for $m = 5, 9, 11, 17$ tempering distributions.

200 repeated simulations. For the 5 temperatures used earlier, there are substantial distances between successive distributions, indicated by effective sample fractions 20% or lower. Using 9, 11, or 17 temperatures significantly reduces the distances between successive distributions and leads to much higher effective sample fractions.

Figure A5 shows the squared biases and MSEs based on 200 repeated simulations. For $m = 5$ tempering distributions, Metropolis resampling MCMC and parallel tempering have the smallest MSEs, as seen in Section 3.2.3. But for $m = 9, 11, 17$ tempering distributions, generalized resample-move and parallel tempering yield the most accurate estimates. For generalized resample-move, the MSEs are significantly reduced when $m$ is increased from 5 to 9, and they are only slightly higher when $m$ is further increased to 11 or 17 (even though the sample size used for estimation under $P_m$ is decreased from $n = 19300$ to $n = 8100$). For parallel tempering, the MSEs are consistently small for the choices of $m$ from 5 to 17. See Appendix IV for further discussion on the effects of using more tempering distributions.
Figure A6: Generalized resample-move at temperature $T = 2.66$ (Gaussian mixture): top left is the scatter plot of $(y_i, y_{i:10})$; top right are the smoothed histograms of $y_i$ (solid) and $y_{i:10}$ (dashed), separately among $\{i : y_i \leq .5\}$ and $\{i : y_i > .5\}$, and the densities of $N(-1, .2^2T)$ and $N(1, .1^2T)$ (dotted), with a vertical line at $.5$; bottom left are the plots of the correlations between $y_i$ and $y_{i:l}$ (•) and those separately among $\{i : y_i \leq .5\}$ (−) and $\{i : y_i > .5\}$ (+) for $l = 1, \ldots, 10$; bottom right are the boxplots of the averages $l^{-1} \sum_{j=1}^l y_{i:j}$ separately among $\{i : y_i \leq .5\}$ and $\{i : y_i > .5\}$ for $l = 1, \ldots, 10$.

Appendix IV: Additional discussion

**Bimodal Gaussian mixtures.** We illustrate various features discussed in Section 2.2 on the generalized-resample algorithm.

Figure A6 shows $x_{(d)}$ from a sample, $S_4$, of size $n + 2b = 48000$ at temperature $T_4 = 2.66$, obtained by a single run of the generalized resample-move algorithm with the ITER strategy using 5 tempering distributions and the proposal standard deviation $0.2 \sqrt{T_j}$ as in Section 3.2.3. By the notation of Section 2.1, $S_4$ consists of 4800 chains $(y_{4i:1}, \ldots, y_{4i:10})$, each of length 10, with initial values $\{y_{4i} : i = 1, \ldots, 4800\}$ obtained by importance resampling from $S_3$. For simplicity, we drop the temperature index 4.
in $y_{ti}$ and $y_{4t,j}$, and use $y$ to indicate its $d$th component $y_{(d)}$.

As seen from the scatter plot of $(y_i, y_{i:10})$, each $y_i$, after 100 ($= 10 \times 10$) random-walk Metropolis iterations, remains in the same region (to the left or right of 0.5) as the initial value $y_i$ is located. The random-walk Metropolis sampler, when initialized from any fixed point, fails to explore the essentially mixture distribution $P_4$ in the entire sample space. On the other hand, as indicated by the histograms and correlation plots, the sampler converges fast to the mixture distribution restricted to the region of the initial value, either $N(-1, .2^2T)$ or $N(1, .1^2T)$. Particularly, the set of particles after Markov moving, $\{y_{i:10} : y_i > .5\}$, is a much better approximation to $N(1, .1^2T)$ than the set of initial particles $\{y_i : y_i > .5\}$, both of size 233 here. Given such a small size, the accuracy of the approximation is near the mode 1 has a substantial impact on the quality of particles for the next distribution.

The foregoing observations underlie the concept of local geometric convergence introduced in Section 2.2. The convergence of $h_t(\cdot)$ to a piecewise constant function is seen from the boxplots: the averages $l^{-1}\sum_{j=1}^{l} y_{t,j,i}$ converge to either $-1$ or 1, depending on whether the initial values $y_i$ are smaller or greater than 0.5. Finally, the correlation between $y_i$ and $y_{i:l}$ does not decrease to 0 but remains higher than .6 as $l$ increases, even though the correlations between $y_i$ and $y_{i:l}$ within $\{i : y_i \leq .5\}$ and within $\{i : y_i > .5\}$ both decrease fast to 0.

**Potts model.** We first examine various aspects of the generalized resample-move algorithm, in parallel to those in the Gaussian mixture example.

Figure A7 shows $u(x)/K$ from a sample, $S_3$, of size $n+3b = 520000$ at temperature $T_3 = 1.413^{-1}$, obtained by a single run of the generalized resample-move algorithm with the ITER strategy and $j = 1000$ using 5 tempering distributions as in Section 3.3.3. By the notation of Section 2.1, $S_3$ consists of 520 chains $(y_{3i:1}, \ldots, y_{3i:1000})$, each of length 1000, with initial values $\{y_{3i} : i = 1, \ldots, 520\}$ obtained by importance resampling from $S_2$. For simplicity, we drop the temperature index 3 in $y_{3i}$ and $y_{3i:j}$, and treat $y$ the same as the per-spin energy $u(y)/K$.

Figure A7 makes clear fundamental differences between the Potts model and the bimodal Gaussian mixture example. As seen from the scatter plot and histograms, $y_i$ near one mode, after $10000 (= 10 \times 10000)$ Metropolis sweeps, can move to near the
Figure A7: Generalized resample-move at temperature $T = 1.413^{-1}$ (Potts model): top left is the scatter plot of $(y_i, y_{i:1000})$; top right are the plots of the correlations between $y_i$ and $y_{i:l}$ (○) and those separately among $\{i: y_i \leq -1.275\}$ (−) and $\{i: y_i > -1.275\}$ (+) for $l = 1, \ldots, 1000$; bottom left are the smoothed histograms of $y_i$ (solid) and $y_{i:1000}$ (dashed) among $\{i: y_i \leq -1.275\}$, and bottom right are those of $y_i$ (solid) and $y_{i:1000}$ (dashed) among $\{i: y_i > -1.275\}$, with a vertical line at $-1.275$.

Other mode. Here 50 out of 393 particles $y_i$ moved from above $-1.275$ to below, and 84 out of 127 particles $y_i$ moved from below $-1.275$ to above. Moreover, the correlation between $y_i$ and $y_{i:l}$ decreases slowly in $l$, overall or separately among $\{i: y_i \leq -1.275\}$ or $\{i: y_i > -1.275\}$. For the Potts model, the Metropolis sampler exhibits very high auto-correlations, whether initialized from the stationary distribution or randomly from any region of the sample space. Therefore, the Metropolis sampler does not satisfy local geometric convergence defined in Section 2.2.

Next, we discuss two possible reasons why the part of Monte Carlo variability from importance resampling, represented by $v_1[r\{h_\ell - E_2(g)\}]$ in (1), can be large, even though the relative variances are low or, equivalently, the effective sample fractions are high, for assessing the efficiency of importance sampling from $P_{j-1}$ to $P_j$,
represented by $P_1$ to $P_2$ in (1). First, $v_{j-1}[r\{h_\ell - E_j(g)\}]$ can be much larger than $V_{j-1}[r\{h_\ell - E_j(g)\}]$ in the presence of high auto-correlations in the sample $S_{j-1}$ from $P_{j-1}$. Second, the use of relative variance or effective sample fraction for assessing the efficiency of importance sampling is based on approximating

$$V_{j-1}[r\{h_\ell - E_j(g)\}] = E_j[r\{h_\ell - E_j(g)\}^2]$$

by $E_j(r)E_j[{h_\ell - E_j(g)}^2]$, where $E_j(r) = 1 + V_{j-1}(r)$ and $V_{j-1}(r)$ is the relative variance of $q_j(\cdot)/q_{j-1}(\cdot)$ under $P_{j-1}$. As remarked in Liu (2001, Section 2.5.3), this approximation, while convenient, can be substantially off. For example, take $g(x) = u(x)/K$ and $r(x) \propto \exp\{-\left(T_3^{-1} - T_2^{-1}\right)u(x)\}$ for importance sampling from $P_2$ to $P_3$. Then $E_3[r\{g - E_3(g)\}^2]$ is about 2.6 times the approximation $E_3(r)E_3[{g - E_3(g)}^2]$, both estimated using the sample $S_3$ shown in Figure A7.

**Comparison of algorithms.** For 5 tempering distributions, the relative performances of the algorithms for the Potts model (Section 3.3.3) resemble those for the bimodal Gaussian mixture using the Markov kernel $\Psi_j$ with proposal standard deviation $0.2\sqrt{T_j}$ (Sections 3.2.2–3.2.3), where the best-performing algorithms are independence Metropolis resampling MCMC and parallel tempering. The resemblance can be attributed to not only how the major mode changes at different temperatures in Figures 1 and 3, but also the fact the Markov kernel leads to slow mixing in the entire sample space or at least in the region where the major mode is eventually located as the temperature decreases. (Recall that for the choice of $\Psi_j$ with proposal standard deviation $0.1\sqrt{T_j}$, the best performing algorithms are generalized resample-move and independence Metropolis resampling MCMC.)

The similarity between the two problems should not be over-interpreted. As mentioned in Section 3.3.1, the histograms in Figures 1 and 3 are of different nature. In fact, there are substantial differences between the two problems as seen from Figures A6–A7. The single-spin-flip Metropolis sampler is slow mixing, but eventually explores both of the two modes, for the Potts model near the critical temperature on the $20 \times 20$ lattice. In contrast, for the bimodal Gaussian mixture, the random walk Metropolis chain is easily trapped in the region where an initial value is located, but converges fast to the mixture distribution restricted to the region, so that the sampler
is locally fast mixing as defined in Section 2.2.

These differences seem to partially underlie the different results of using more tempering distributions for the two examples (Appendix III and Section 3.3.4). For the bimodal Gaussian mixture, the Markov chains are locally fast mixing, and hence the distances between successive distributions are the main source of Monte Carlo errors. Using more finely spaced distributions leads to more accurate results for generalized resample-move. For the Potts model, the Markov chains are slow mixing, without converging fast to any restricted distribution, and hence high-autocorrelations are the main source of Monte Carlo errors. Using more finely spaced distributions appears not worthwhile, at the cost of a smaller sample size per distribution.

Appendix V: Additional figures

Figures A8–A9 shows the smoothed histograms from Figure 1 in separate plots at different temperatures. Figures A10–A11 shows the smoothed histograms from Figure 3 in separate plots at different temperatures.
Figure A8: Smoothed histograms of the $d$th component $x_{(d)}$ at temperatures $T_1$ (dotted), $T_2$ (dotdash), $T_3$ (dash), $T_4$ (longdash), and $T_5$ (solid) and the true density at $T_5$ ($\times$), for the proposal standard deviation $\tau_j = 0.1\sqrt{\bar{T}_j}$.
Figure A9: Smoothed histograms of the $d$th component $x_{(d)}$ at temperatures $T_1$ (dotted), $T_2$ (dotdash), $T_3$ (dash), $T_4$ (longdash), and $T_5$ (solid) and the true density at $T_5$ ($\times$), for the proposal standard deviation $\tau_j = 0.2\sqrt{T_j}$. 

\[
\begin{align*}
\tau_j &= 0.2\sqrt{T_j} & (T_1 &= 50.0) \\
\tau_j &= 0.2\sqrt{T_j} & (T_2 &= 18.80) \\
\tau_j &= 0.2\sqrt{T_j} & (T_3 &= 7.07) \\
\tau_j &= 0.2\sqrt{T_j} & (T_4 &= 2.66) \\
\tau_j &= 0.2\sqrt{T_j} & (T_5 &= 1.0)
\end{align*}
\]
Figure A10: Smoothed histograms of $u(x)/K$ at the temperatures $(T_1, T_2, \ldots, T_5)$ labeled as 1, 2, …, 5 under the Potts model, obtained by generalized resample-move.
Figure A11: Smoothed histograms of $u(x)/K$ at the temperatures $(T_1, T_2, \ldots, T_5)$ labeled as 1, 2, \ldots, 5 under the Potts model, obtained by parallel tempering.