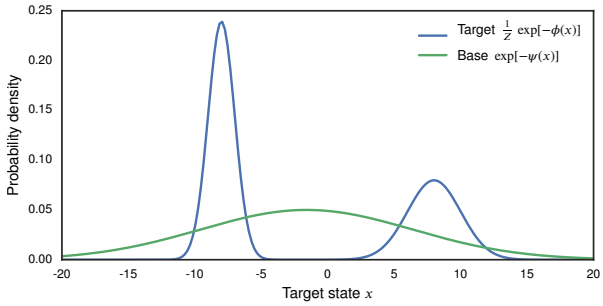
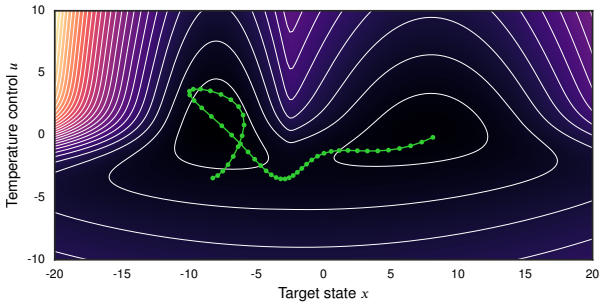
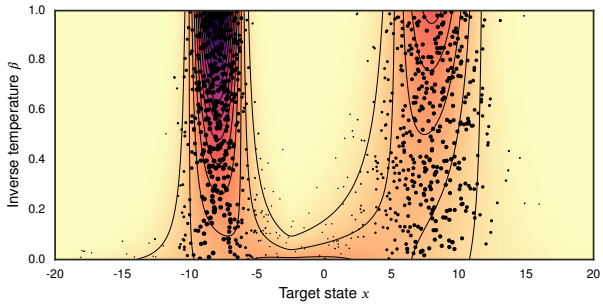
A One-dimensional bimodal target



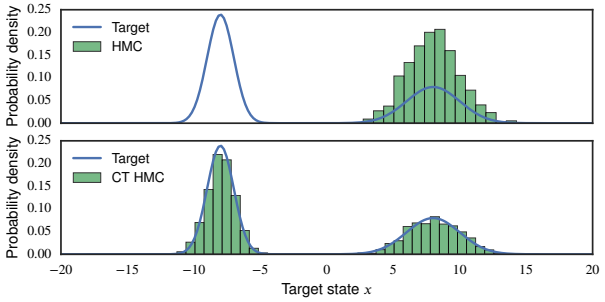
(a) Target (blue curve) & base (green curve) density functions.



(b) Joint energy (contour plot) & example trajectory (green curve).



(c) Joint density (contour plot) and CT HMC samples (circles).



(d) Histograms from HMC (top) and CT HMC (bottom) samples.

Figure 4: Visualisations of *continuous tempering* (CT) in a bimodal univariate target density. (a) A two-component Gaussian mixture target density (blue curve) and Gaussian base density (green curve) with mean and variance matched to the target. (b) The extended potential energy on the target state x and temperature control variable u (contour plot - dark colours indicate low energy) and an example simulated Hamiltonian trajectory in the joint space (green curve). The temperature control variable bridges the base and target densities lowering energy barriers in the target space. (c) Joint density on the target state x and inverse temperature β (17) (contour plot, dark colours indicate high density) and samples from a CT HMC chain run in the joint space (circles, size of each circle is proportional to $p_{\beta|x}(1 \mid x)$ and so larger symbols indicate a greater weighting in estimates of expectations with respect to the target (22)). (d) Example target state sample histograms from running standard HMC in the original target density (top) and running HMC in the extended joint space (bottom).

We give here an illustrative example of the gains of the proposed approach over standard HMC in target densities with isolated modes. We use a one-dimensional Gaussian mixture density with two separated Gaussian components as the target density, shown by the blue curve in Figure 4a. Although performance in this toy univariate model is not necessarily reflective of that in more realistic higher-dimensional models, it has the advantage of allowing the joint density on x and $\beta = \beta(u)$ to be directly visualised.

For the base density $\exp[-\psi(x)]$ we use a univariate Gaussian with mean and variance matched to those of the target density (corresponding to the Gaussian density minimising the KL divergence from the target to base distribution), shown by the green curve in Figure 4a. We also set $\log \zeta = \log Z$ and so the performance here represents a 'best-case' scenario for the continuous tempering approach.

The resulting potential energy $(-\log p_{x,u})$ on the extended (x, u) space is shown in Figure 4b. For positive temperature control values (and so inverse temperature values close to 1), the energy surface tends increasingly to the double-well potential corresponding to the target distribution, with a high energy barrier between the two modes. For negative temperature control values the energy surface tends towards the single quadratic well corresponding to the Gaussian base density. The resulting joint energy surface allows for paths between the values of the target state x corresponding to the two modes which have much lower potential energy barriers than the potential barrier between the two modes in the original target space, allowing simulated Hamiltonian trajectories such as that shown in green to more easily explore the target state space.

Samples from a HMC chain on the extended joint space are shown in Figure 4c, with the joint density on (x, β) (17) shown in the background as a contoured heat map. It can be seen that the Hamiltonian dynamic is able to explore the joint space well with good coverage of all of the high density regions. The size of the points in 4c is proportional to $w_1(x) = \mathrm{p}_{\beta|\mathrm{x}}(1 \mid x)$ and so reflects the importance weights of the samples in the estimator for expectations with respect to the target in (22). Importantly even points for which β is close to zero can contribute significantly to the expectations if the corresponding x value is probable under the target: this is in contrast to the extended Hamiltonian approach of [19] where only a subset of points corresponding to $\beta = 1$ are used to compute expectations.

The final panel, Figure 4d shows empirical histograms on the target variable x estimated from samples of a chain on the extended space (joint continuous tempering, bottom) and standard HMC on the original target space (top). As can be seen the standard HMC approach gets stuck in one mode thus does not assign any mass to the other mode in the histogram, unlike the tempered chain which identifies both modes and accurately estimates their relative masses.

B Bounding the inverse temperature marginal density

We have a joint density on $(\boldsymbol{x}, \beta)$

$$\mathrm{p}_{\mathbf{x},\beta}(\boldsymbol{x}, \beta) = \frac{1}{C} \exp(-\beta\phi(\boldsymbol{x}) - \beta \log \zeta - (1-\beta)\psi(\boldsymbol{x})). \tag{27}$$

The resulting marginal density on β is

$$\mathrm{p}_\beta(\beta) = \int_{\mathcal{X}} \mathrm{p}_{\mathbf{x},\beta}(\boldsymbol{x}, \beta)\, \mathrm{d}\boldsymbol{x} = \frac{1}{C\zeta^\beta} \int_{\mathcal{X}} \exp(-\beta\phi(\boldsymbol{x}) - (1-\beta)\psi(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x}. \tag{28}$$

To derive an upper-bound on $\mathrm{p}_\beta(\beta)$ we use Hölder's inequality

$$\int_{\mathcal{X}} g(\boldsymbol{x})h(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \leq \left(\int_{\mathcal{X}} |g(\boldsymbol{x})|^{\frac{1}{a}}\, \mathrm{d}\boldsymbol{x} \right)^a \left(\int_{\mathcal{X}} |h(\boldsymbol{x})|^{\frac{1}{1-a}}\, \mathrm{d}\boldsymbol{x} \right)^{1-a} \tag{29}$$

where $a \in [0, 1]$ and g and h are measurable functions. We also use the definitions

$$\int_{\mathcal{X}} \exp(-\phi(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x} = Z \quad \text{and} \quad \int_{\mathcal{X}} \exp(-\psi(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x} = 1. \tag{30}$$

From (28) we have that

$$\mathrm{p}_\beta(\beta) = \frac{1}{C\zeta^\beta} \int_{\mathcal{X}} \left(\exp(-\phi(\boldsymbol{x}))^\beta\right) \left(\exp(-\psi(\boldsymbol{x}))^{1-\beta}\right) \mathrm{d}\boldsymbol{x}. \tag{31}$$

Applying Hölder's inequality (29) with $g(\boldsymbol{x}) = \exp(-\phi(\boldsymbol{x}))^\beta$, $h(\boldsymbol{x}) = \exp(-\psi(\boldsymbol{x}))^{1-\beta}$ and $a = \beta$

$$\mathrm{p}_\beta(\beta) \leq \frac{1}{C\zeta^\beta} \left(\int_{\mathcal{X}} \left|\exp(-\phi(\boldsymbol{x}))^\beta\right|^{\frac{1}{\beta}}\, \mathrm{d}\boldsymbol{x} \right)^\beta \left(\int_{\mathcal{X}} \left|\exp(-\psi(\boldsymbol{x}))^{1-\beta}\right|^{\frac{1}{1-\beta}}\, \mathrm{d}\boldsymbol{x} \right)^{1-\beta} \tag{32}$$

$$= \frac{1}{C\zeta^\beta} \left(\int_{\mathcal{X}} \exp(-\phi(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x} \right)^\beta \left(\int_{\mathcal{X}} \exp(-\psi(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x} \right)^{1-\beta}. \tag{33}$$

Substituting the definitions in (30) gives

$$\mathrm{p}_\beta(\beta) \leq \frac{1}{C} \left(\frac{Z}{\zeta} \right)^\beta. \tag{34}$$

To derive a lower-bound on $\mathrm{p}_\beta(\beta)$, we use Jensen's inequality

$$\varphi\left(\int_{\mathcal{X}} g(\boldsymbol{x})q(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \right) \geq \int_{\mathcal{X}} \varphi(g(\boldsymbol{x}))q(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}, \tag{35}$$

for a concave function φ, normalised density $q : \int_{\mathcal{X}} q(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = 1$ and measurable g. The logarithm of (28) gives

$$\log \mathrm{p}_\beta(\beta) + \beta \log \zeta + \log C = \log\left(\int_{\mathcal{X}} \exp(-\beta(\phi(\boldsymbol{x}) - \psi(\boldsymbol{x}))) \exp(-\psi(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x} \right). \tag{36}$$

Applying Jensen's inequality (35) with $\varphi = \log$, $q = \exp(-\psi)$ and $g = \exp(-\beta(\phi - \psi))$

$$\log \mathrm{p}_\beta(\beta) + \log C + \beta \log \zeta \geq \beta \int_{\mathcal{X}} (\psi(\boldsymbol{x}) - \phi(\boldsymbol{x})) \exp(-\psi(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x} \tag{37}$$

$$= \beta \int_{\mathcal{X}} (\log Z - \log Z - \log \exp(-\psi(\boldsymbol{x}) + \phi(\boldsymbol{x}))) \exp(-\psi(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x} \tag{38}$$

$$= \beta \log Z - \beta \int_{\mathcal{X}} \exp(-\psi(\boldsymbol{x})) \log\left(\frac{\exp(-\psi(\boldsymbol{x}))}{\exp(-\phi(\boldsymbol{x}))/Z} \right) \mathrm{d}\boldsymbol{x}. \tag{39}$$

Recognising the integral in the last line as the *Kullback–Leibler* (KL) divergence $d^{b\to t}$ from the base density $\exp(-\psi(\boldsymbol{x}))$ to the target density $\exp(-\phi(\boldsymbol{x}))/Z$

$$d^{b\to t} = \int_{\mathcal{X}} \exp(-\psi(\boldsymbol{x})) \log\left(\frac{\exp(-\psi(\boldsymbol{x}))}{\exp(-\phi(\boldsymbol{x}))/Z} \right) \mathrm{d}\boldsymbol{x}, \tag{40}$$

and taking the exponential of both sides and rearranging we have

$$\mathrm{p}_\beta(\beta) \geq \frac{1}{C} \left(\frac{Z}{\zeta} \right)^\beta \exp\left(-\beta d^{b\to t} \right). \tag{41}$$

By instead noting (28) can be rearranged into the form

$$\log \mathrm{p}_\beta(\beta) + \log C + \beta \log \zeta - \log Z = \log\left(\int_{\mathcal{X}} \exp(-(1-\beta)(\psi(\boldsymbol{x}) - \phi(\boldsymbol{x}))) \frac{1}{Z} \exp(-\phi(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x} \right), \tag{42}$$

by an equivalent series of steps we can also derive a bound using the reversed form of the KL divergence

$$d^{t\to b} = \int_{\mathcal{X}} \frac{1}{Z} \exp(-\phi(\boldsymbol{x})) \log\left(\frac{\exp(-\phi(\boldsymbol{x}))/Z}{\exp(-\psi(\boldsymbol{x}))} \right) \mathrm{d}\boldsymbol{x}. \tag{43}$$

from the target to the base distribution, giving that

$$\mathrm{p}_\beta(\beta) \geq \frac{1}{C} \left(\frac{Z}{\zeta} \right)^\beta \exp\left(-(1-\beta) d^{t\to b} \right). \tag{44}$$

C Gaussian mixture Boltzmann machine relaxations

We define a *Boltzmann machine distribution* on a signed binary state $\mathbf{s} \in \{-1, +1\}^{D_B} = S$ as

$$\mathrm{p}_{\mathbf{s}}(\boldsymbol{s}) = \frac{1}{Z_B} \exp\left(\frac{1}{2} \boldsymbol{s}^\mathsf{T} \boldsymbol{W} \boldsymbol{s} + \boldsymbol{s}^\mathsf{T} \boldsymbol{b} \right) \qquad Z_B = \sum_{\boldsymbol{s} \in S} \left(\exp\left(\frac{1}{2} \boldsymbol{s}^\mathsf{T} \boldsymbol{W} \boldsymbol{s} + \boldsymbol{s}^\mathsf{T} \boldsymbol{b} \right) \right). \tag{45}$$

We introduce an auxiliary real-valued vector random variable $\mathbf{x} \in \mathbb{R}^D$ with a Gaussian conditional distribution

$$\mathrm{p}_{\mathbf{x}|\mathbf{s}}(\boldsymbol{x} \mid \boldsymbol{s}) = \frac{1}{(2\pi)^{D/2}} \exp\left(-\frac{1}{2} \left(\boldsymbol{x} - \boldsymbol{Q}^\mathsf{T} \boldsymbol{s} \right)^\mathsf{T} \left(\boldsymbol{x} - \boldsymbol{Q}^\mathsf{T} \boldsymbol{s} \right) \right) \tag{46}$$

with $\boldsymbol{Q}$ a $D_B \times D$ matrix such that $\boldsymbol{Q}\boldsymbol{Q}^\mathsf{T} = \boldsymbol{W} + \boldsymbol{D}$ for some diagonal $\boldsymbol{D}$ which makes $\boldsymbol{W} + \boldsymbol{D}$ positive semi-definite. In our experiments, based on the observation in [44] that minimising the maximum eigenvalue of $\boldsymbol{W} + \boldsymbol{D}$ decreases the maximal separation between the Gaussian components in the relaxation, we set $\boldsymbol{D}$ as the solution to the semi-definite programme

$$\min_{\boldsymbol{D}} \left(\lambda_{\mathrm{MAX}}(\boldsymbol{W} + \boldsymbol{D}) \right) : \boldsymbol{W} + \boldsymbol{D} \succeq 0 \tag{47}$$

where λ_{MAX} denotes the maximal eigenvalue. In general the optimised $\boldsymbol{W} + \boldsymbol{D}$ lies on the semi-definite cone and so has rank less than D_B hence a $\boldsymbol{Q}$ can be found such that $D < D_B$. The resulting joint distribution on $(\mathbf{x}, \mathbf{s})$ is

$$\mathrm{p}_{\mathbf{x},\mathbf{s}}(\boldsymbol{x}, \boldsymbol{s}) = \frac{1}{(2\pi)^{D/2} Z_B} \exp\Big(-\frac{1}{2}\boldsymbol{x}^\mathsf{T}\boldsymbol{x} + \boldsymbol{s}^\mathsf{T}\boldsymbol{Q}\boldsymbol{x} - \frac{1}{2}\boldsymbol{s}^\mathsf{T}\boldsymbol{Q}\boldsymbol{Q}^\mathsf{T}\boldsymbol{s} + \frac{1}{2}\boldsymbol{s}^\mathsf{T}\boldsymbol{W}\boldsymbol{s} + \boldsymbol{s}^\mathsf{T}\boldsymbol{b}\Big) \tag{48}$$

$$= \frac{1}{(2\pi)^{D/2} Z_B} \exp\Big(-\frac{1}{2}\boldsymbol{x}^\mathsf{T}\boldsymbol{x} + \boldsymbol{s}^\mathsf{T}(\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{b}) - \frac{1}{2}\boldsymbol{s}^\mathsf{T}\boldsymbol{D}\boldsymbol{s}\Big) \tag{49}$$

$$= \frac{1}{(2\pi)^{D/2} Z_B \exp\Big(\frac{1}{2}\operatorname{Tr}(\boldsymbol{D})\Big)} \exp\Big(-\frac{1}{2}\boldsymbol{x}^\mathsf{T}\boldsymbol{x}\Big) \prod_{i=1}^{D_B} \big(\exp\big(s_i\big(\boldsymbol{q}_i^\mathsf{T}\boldsymbol{x} + b_i\big)\big)\big), \tag{50}$$

where $\big\{\boldsymbol{q}_i^\mathsf{T}\big\}_{i=1}^{D_B}$ are the D_B rows of $\boldsymbol{Q}$. We can marginalise over the binary state $\mathbf{s}$ as each s_i is conditionally independent of all the others given $\mathbf{x}$ in the joint distribution. This gives the *Boltzmann machine relaxation* density on $\mathbf{x}$

$$\mathrm{p}_{\mathbf{x}}(\boldsymbol{x}) = \frac{2^{D_B}}{(2\pi)^{D/2} Z_B \exp\Big(\frac{1}{2}\operatorname{Tr}(\boldsymbol{D})\Big)} \exp\Big(-\frac{1}{2}\boldsymbol{x}^\mathsf{T}\boldsymbol{x}\Big) \prod_{i=1}^{D_B} \big(\cosh\big(\boldsymbol{q}_i^\mathsf{T}\boldsymbol{x} + b_i\big)\big), \tag{51}$$

which is a structured Gaussian mixture density with 2^{D_B} components. If we define $\mathrm{p}_{\mathbf{x}}(\boldsymbol{x}) = \frac{1}{Z}\exp(-\phi(\boldsymbol{x}))$ with

$$\phi(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\mathsf{T}\boldsymbol{x} - \sum_{i=1}^{D_B} \big(\log\cosh\big(\boldsymbol{q}_i^\mathsf{T}\boldsymbol{x} + b_i\big)\big), \tag{52}$$

then the normalisation constant Z of the relaxation density can be related to the normalising constant of the corresponding Boltzmann machine distribution by

$$\log Z = \log Z_B + \frac{1}{2}\operatorname{Tr}(\boldsymbol{D}) + \frac{D}{2}\log(2\pi) - D_B \log 2. \tag{53}$$

It can also be shown that the first and second moments of the relaxation distribution are related to the first and second moments of the corresponding Boltzmann machine distribution by

$$\mathbb{E}[\mathbf{x}] = \int_{\mathcal{X}} \boldsymbol{x} \sum_{\boldsymbol{s}\in\mathcal{S}} \big(\mathrm{p}_{\mathbf{x}|\mathbf{s}}(\boldsymbol{x} \mid \boldsymbol{s})\, \mathrm{p}_{\mathbf{s}}(\boldsymbol{s})\big)\, \mathrm{d}\boldsymbol{x} = \sum_{\boldsymbol{s}\in\mathcal{S}} \Big(\int_{\mathcal{X}} \boldsymbol{x}\, \mathcal{N}\big(\boldsymbol{x}; \boldsymbol{Q}^\mathsf{T}\boldsymbol{s},\, \boldsymbol{I}\big)\, \mathrm{d}\boldsymbol{x}\, \mathrm{p}_{\mathbf{s}}(\boldsymbol{s})\Big) = \mathbb{E}\big[\boldsymbol{Q}^\mathsf{T}\mathbf{s}\big] = \boldsymbol{Q}^\mathsf{T}\,\mathbb{E}[\mathbf{s}], \tag{54}$$

$$\text{and } \mathbb{E}\big[\mathbf{x}\mathbf{x}^\mathsf{T}\big] = \sum_{\boldsymbol{s}\in\mathcal{S}} \Big(\int_{\mathcal{X}} \boldsymbol{x}\boldsymbol{x}^\mathsf{T}\, \mathcal{N}\big(\boldsymbol{x}; \boldsymbol{Q}^\mathsf{T}\boldsymbol{s},\, \boldsymbol{I}\big)\, \mathrm{d}\boldsymbol{x}\, \mathrm{p}_{\mathbf{s}}(\boldsymbol{s})\Big) = \mathbb{E}\big[\boldsymbol{Q}^\mathsf{T}\mathbf{s}\mathbf{s}\boldsymbol{Q} + \boldsymbol{I}\big] = \boldsymbol{Q}^\mathsf{T}\,\mathbb{E}\big[\mathbf{s}\mathbf{s}^\mathsf{T}\big]\boldsymbol{Q} + \boldsymbol{I}. \tag{55}$$

The weight parameters $\boldsymbol{W}$ of the Boltzmann machine distributions used in the experiments in Section 7.1 were generated using an eigendecomposition based method. A uniformly distributed (with respect to the Haar measure) random orthogonal matrix $\boldsymbol{R}$ was sampled. A vector of eigenvalues $\boldsymbol{e}$ was generated by sampling independent zero-mean unit-variance normal variates $n_i \sim \mathcal{N}(\cdot;\, 0,\, 1)\ \forall i \in \big\{1, \ldots D_B\big\}$ and then setting $e_i = s_1 \tanh(s_2 n_i)\ \forall i \in \big\{1, \ldots D_B\big\}$, with $s_1 = 6$ and $s_2 = 2$ in the experiments. This generates eigenvalues concentrated near $\pm s_1$ with this empirically observed to lead to systems which tended to be highly multimodal. A symmetric matrix $\boldsymbol{V} = \boldsymbol{R}\operatorname{diag}(\boldsymbol{e})\boldsymbol{R}^\mathsf{T}$ was then computed and the weights $\boldsymbol{W}$ set such that $W_{i,j} = V_{i,j}\ \forall i \neq j$ and $W_{i,i} = 0\ \forall i \in \big\{1, \ldots D_B\big\}$. The biases $\boldsymbol{b}$ where generated using $b_i \sim \mathcal{N}\big(\cdot;\, 0, 0.1^2\big)\ \forall i \in \big\{1, \ldots D_B\big\}$. An example of a two-dimensional projection of independent samples from a Boltzmann machine relaxation density with $D = 27$ ($D_B = 28$), and $\boldsymbol{W}$ and $\boldsymbol{b}$ generated as just described in shown in figure 5. As can be seen even when projected down to two-dimensions the resulting density shows multiple separated modes.

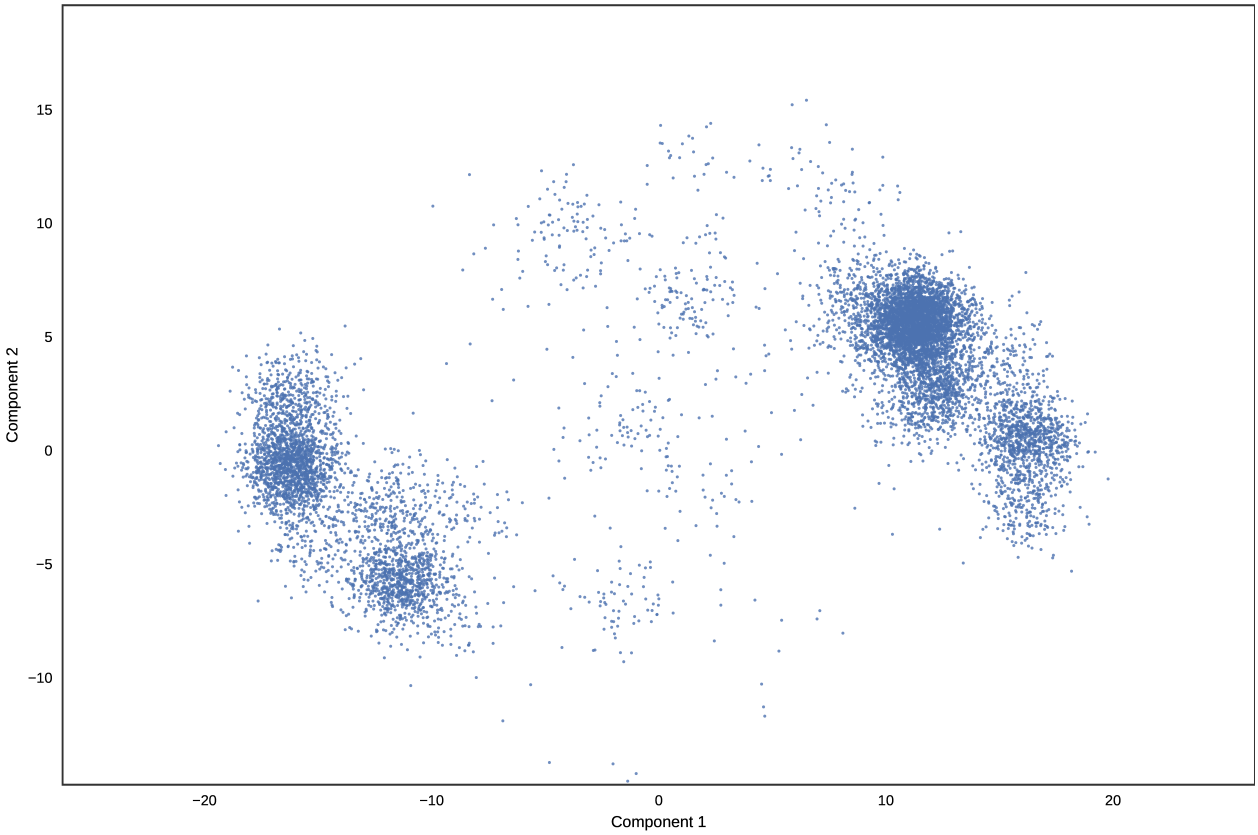


Figure 5: Two-dimensional projection of 10000 independent samples from a Gaussian mixture relaxation of a Boltzmann machine distribution. The parameters $\boldsymbol{W}$ and $\boldsymbol{b}$ of the Boltzmann machine distribution where generated as described in Section C, with here $D_B = 28$ (rather than $D_B = 30$ as in the experiments) as independent sampling from larger systems exceeded the memory available on the workstation used. The two components shown correspond to the two eigenvectors of the generated basis $\boldsymbol{R}$ with the largest corresponding eigenvalues.

D Importance weighted autoencoder test images

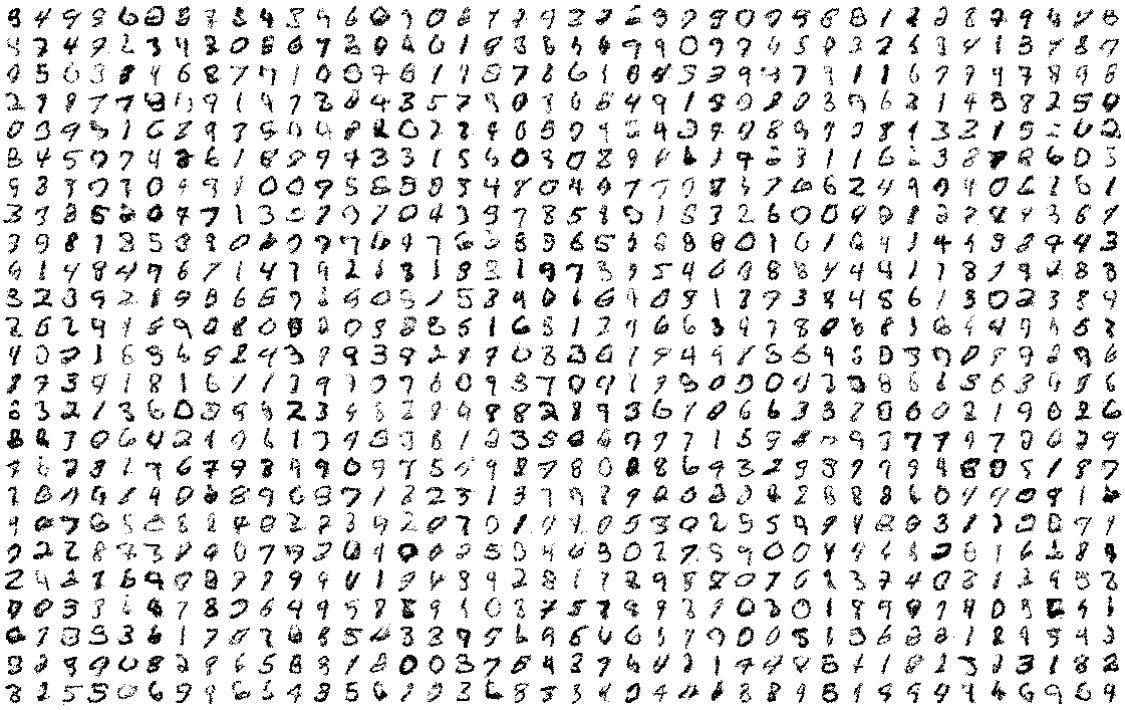


Figure 6: MNIST test images.

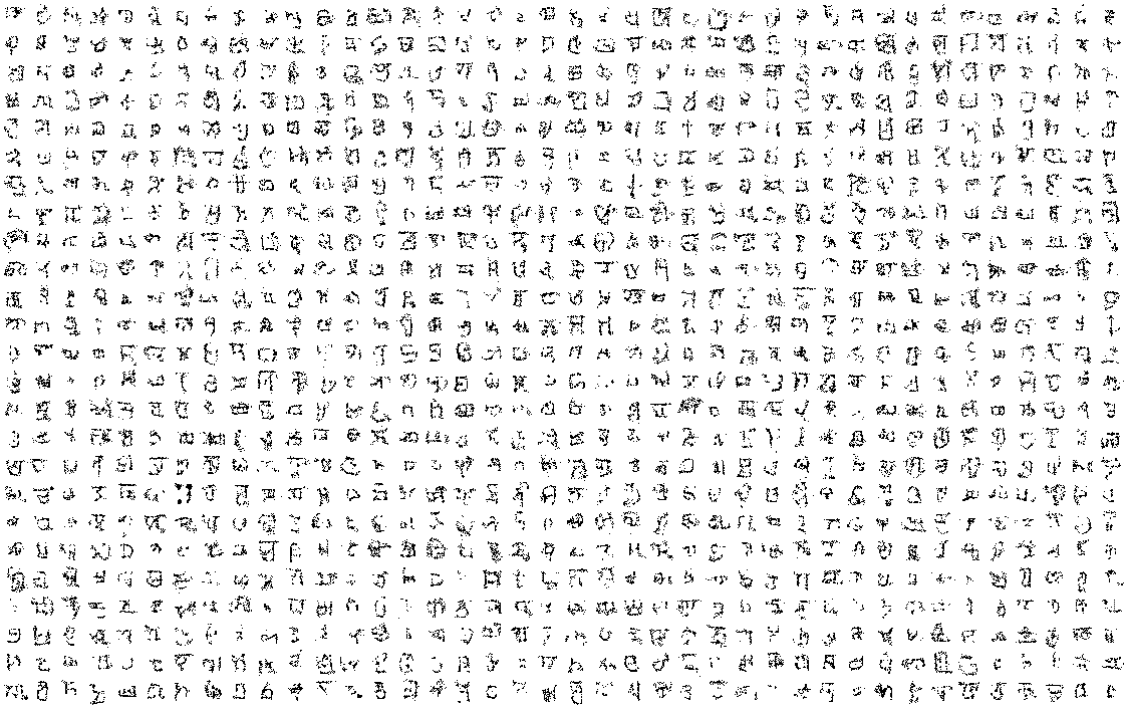


Figure 7: Omniglot test images.