Bayesian Neural Networks at Finite Temperature

Robert J.N. Baldock, Nicola Marzari
Theory and Simulation of Materials (THEOS)
and National Centre for Computational Design and Discovery of Novel Materials (MARVEL)
École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland
`rjnbaldock@gmail.com`

Abstract

We recapitulate the Bayesian formulation of neural network based classifiers and show that, while sampling from the posterior does indeed typically lead to better generalisation than is obtained by standard optimisation of the cost function, even better performance can in general be obtained by sampling finite temperature distributions derived from the posterior. Taking the example of two different deep (3 hidden layers) classifiers for MNIST data, we find quite different temperatures to be appropriate in each case: for a network with high bias we observe optimal network parameters at $T \ll 1$, whereas for models with high variance a clear minimum of the test error is observed at higher temperature. This is a statistical equivalent to early stopping. As the temperature is increased from the value of minimum test error, classifiers transition from accurate classifiers to classifiers that have higher test loss than assigning equal probability to each class. Efficient studies of these temperature-induced effects are enabled using a replica-exchange Hamiltonian Monte Carlo simulation technique. Finally, we show how thermodynamic integration can be used to perform model selection for deep neural networks. Similar to the Laplace approximation, this approach assumes that the posterior is dominated by a single mode. Crucially, however, no assumption is made about the shape of that mode and it is not required to precisely compute and invert the Hessian.

1 BAYESIAN FORMULATION OF NEURAL NETWORK CLASSIFIERS

The Bayesian formulation of neural network classifiers was introduced by David MacKay [MacKay, 1992b]. In training a classifier one typically uses a minimisation algorithm to find a local minimum of a cost function defined over the parameters of the neural network model. If we use a softmax then the activation of the output units sum to one, and can be interpreted as the probability, assigned by the network, to the assertion that the input belongs to each individual class. We can write this as $\text{prob}(q|\mathbf{x}, \mathbf{w})$ where q is the index of the class, $\mathbf{x}$ is the single data example to be classified and $\mathbf{w}$ represents the parameters of the neural network. For data in the training set $D = \{\mathbf{x}_i, t_i\}$ where $\mathbf{x}_i$ is an input data example and t_i the corresponding class label, then the probability that the network classifies $\mathbf{x}_i$ correctly is $\text{prob}(q = t_i|\mathbf{x}_i, \mathbf{w})$. This probability is a function of the weights $\mathbf{w}$. The probability that the network classifies the complete data set correctly is called the *likelihood* of the data, $\text{prob}(D|\mathbf{w})$, and is given by $\prod_i \text{prob}(q = t_i|\mathbf{x}_i, \mathbf{w})$.

In general, the posterior probability for the weights, which expresses what we have learned about the weights from the data, is given by

$$\text{prob}(\mathbf{w}|D) = \frac{\text{prob}(D|\mathbf{w})\,\text{prob}(\mathbf{w})}{\text{prob}(D)} \tag{1}$$

$$\propto \prod_i \text{prob}(q = t_i|\mathbf{x}_i, \mathbf{w})\text{prob}(\mathbf{w}) \tag{2}$$

where $\text{prob}(\mathbf{w})$ is the prior probability distribution for the weights, and represents our initial state of knowledge before the collection of any data.

If we assign a Gaussian prior with standard devia-

tion σ to the parameters $\mathbf{w}$ then the posterior (2) is given by

$$\text{prob}\left(\mathbf{w}|D\right) \propto \prod_i \text{prob}\left(q = t_i | \mathbf{x}_i, \mathbf{w}\right) e^{-\frac{\mathbf{w}^2}{2\sigma^2}}. \quad (3)$$

The posterior probability of the weights (3) is therefore maximised when we minimise

$$J(\mathbf{w}|D) = -\log\left[\text{prob}\left(D|\mathbf{w}\right)\text{prob}\left(\mathbf{w}\right)\right] \quad (4)$$

$$= -\sum_i \log\left[\text{prob}\left(q = t_i|\mathbf{x}_i, \mathbf{w}\right)\right] + \frac{\mathbf{w}^2}{2\sigma^2}. \quad (5)$$

which can immediately be recognised as the cross entropy loss function with L2 regularisation. The minimum of (5) is called the maximum a posteriori solution.

In this paper we will use a uniform prior for $\mathbf{w}$

$$\text{prob}\left(\mathbf{w}\right) = \begin{cases} \frac{1}{\prod_{i=1}^d \sigma_i}, & |w_i| < \sigma_i/2 \,\forall\, i, \\ 0, & \text{Elsewhere.} \end{cases} \quad (6)$$

Within the bounds of this prior the cost function is given by

$$J(\mathbf{w}|D) = -\sum_i \left(\log\left[\text{prob}\left(q = t_i|\mathbf{x}_i, \mathbf{w}\right)\right] - \log \sigma_i\right). \quad (7)$$

Outside the bounds of the prior the cost function is infinite. Minimising the cost function will simply mean maximising the likelihood within these bounds.

2 ENERGY AND TEMPERATURE

One may define the *potential energy* over the parameters of the network

$$E(D|\mathbf{w}) = -\log \text{prob}\left(D|\mathbf{w}\right) \quad (8)$$

$$= -\sum_i \log\left[\text{prob}\left(q = t_i|\mathbf{x}_i, \mathbf{w}\right)\right] \quad (9)$$

and the *"temperature-adjusted"* posterior distribution

$$\text{prob}\left(\mathbf{w}|T, D\right) \propto e^{-\frac{1}{T}E(D|\mathbf{w})}\text{prob}(\mathbf{w}). \quad (10)$$

where T controls the noise in our posterior. We name T the *"temperature"* in analogy with thermodynamic temperature.

For this paper, it is important to imagine how the temperature-adjusted posterior behaves as we vary

T. The distribution (10) tends towards the prior distribution for $T \to \infty$, and is equal to the true posterior distribution (2) for $T = 1$. At $T < 1$ (low temperature) the temperature-adjusted posterior (10) is concentrated around the maximum a posteriori solution (the minimum of $J(\mathbf{w}|D)$).

2.1 TEMPERATURE VS BATCH SIZE

It is known that mini-batch learning improves generalisation. Recently there has been much interest in how the batch size in mini-batch learning controls the noise level during learning, and under what conditions the optimiser can be said to be approximately sampling from the posterior [Smith & Le, 2017, Welling & Teh, 2011, Mandt *et al.* , 2017, Ahn *et al.* , 2012]. In this paper we instead use temperature to precisely control the noise in full batch training and explore the use of $T > 0$ as an alternative to early stopping.

3 ON THE OPTIMAL TEMPERATURE FOR BAYESIAN NEURAL NETWORKS

Here we examine the average training and test errors of two deep neural network classifiers with parameters sampled from the temperature-adjusted posterior (10), across a wide range of temperatures. We find that sampling from the posterior ($T = 1$) indeed leads to better test error than standard optimisation of the cost function. However, we also find that in general, improved performance can be obtained at temperatures other than $T = 1$ and that the optimal value of T depends on the neural network and the data set.

3.1 DATA AND NEURAL NETWORK ARCHITECTURES

All calculations reported in this paper are performed using the MNIST data set [LeCun *et al.* , 1998]. For computational speed the MNIST data images are transformed down from 28 x 28 pixels to 16 x 16 (256 input dimensions). Standard normalisation is applied to the full MNIST data set. A random selection of rescaled data is shown in Figure 1.

Calculations are performed on stratified samples from the full MNIST training set. Remaining samples from the MNIST training set, not included in these reduced stratified training sets, are appended to the MNIST test set. We will adopt the convention
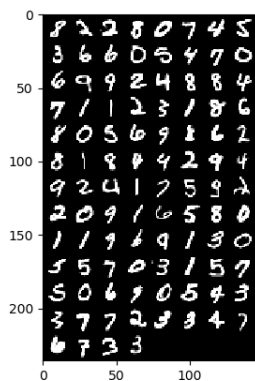
Figure 1: 100 random data samples from the MNIST training set, transformed as described in the text.

that a data set containing n data points is denoted D_n. Thus D_{500} is a data set of 500 images, 50 for each class. Each time a data set of a new size is generated, it is stored. Thus all calculations using data sets of the same size make use of exactly the same images. These data sets are available from an open source repository, together with the code required to run these calculations [Baldock, 2019].

In this section we make use of two neural networks. The first, $M^{(3)}$, has 256 input neurons, 3 hidden layers of 40 logistic neurons, and an output layer of 10 linear neurons to which a softmax is applied. Thus, the loss function for this network is the cross-entropy loss function. A second network $M^{(3*)}$ has the same structure as $M^{(3)}$, except that the final layer of linear neurons is replaced by a layer of *logistic* neurons. A softmax is also applied to the output of these logistic neurons, so the loss function for this network is also the cross entropy loss function. The use of logistic neurons in the final layer drastically reduces the capacity of $M^{(3*)}$ as compared to $M^{(3)}$, since in $M^{(3*)}$ the values entering the softmax are constrained to the interval $(0, 1)$.

3.2 CHOICE OF PRIOR

The standard procedure for initialising the weights and biases of a neural network before training is to choose them uniformly at random inside the region $[-\frac{1}{\sqrt{k_i}}, \frac{1}{\sqrt{k_i}}]$. Here i is an integer labelling the weight or bias, and k_i is the fan in, including the bias, of the neuron to which weight or bias i points.

In this paper we use a (uniform) prior (6) of this same shape, but made 50 times wider for $M^{(3)}$ and 1000 times wider for $M^{(3*)}$. Including the factor $\frac{1}{2}$ from (6) for $M^{(3)}$ we therefore have

$$\sigma_i = \frac{100}{\sqrt{k_i}}. \qquad (11)$$

Similarly, for $M^{(3*)}$ the prior is as in (11), but with a factor 2000 in place of 100.

These factors are chosen to create priors that capture important minima of $E(D|\mathbf{w})$, without being overly specific. We choose the factors 50 and 1000 because, after standard initialisation followed by minimisation we found $\max_i \left(w_i \div \frac{1}{\sqrt{k_i}} \right)$ to be 30.3 for $M^{(3)}$ and 890.0 for $M^{(3*)}$. We repeated this procedure 25 times for each network to be sure these are typical values. Minimisation is performed using Alg. 3 (see Sec. 3.3.2), which is related to the FIRE optimisation algorithm [Bitzek *et al.* , 2006].

3.3 METHODS

We develop a Hamiltonian (also called Hybrid) Monte Carlo (HMC) [Duane *et al.* , 1987, Neal *et al.* , 2011] formulation of a technique called replica-exchange molecular dynamics (REMD) [Swendsen & Wang, 1986, Sugita & Okamoto, 1999] for use with neural network models. This method enables efficient exploration of the temperature-adjusted posterior (10) across a range of temperatures, far faster than sampling those temperatures independently.

In RE-HMC the user specifies a temperature range to study. Normally this range goes from the lowest temperature of interest, up to an artificially high temperature where the sampler moves very quickly. Independent samplers are created at particular temperatures throughout that range. These samplers explore $\mathbf{w}$ space according to (10) at their respective temperatures. Periodic attempts are made to exchange $\mathbf{w}$ vectors between neighbouring temperatures. This accelerates $\mathbf{w}$ space sampling at low temperatures because each parameter vector spends some of the time at high temperature where it explores $\mathbf{w}$ space much more rapidly. Importantly, samplers at each temperature are always sampling from the respective temperature-adjusted posterior (10) even immediately after a parameter swap.

A geometric series of temperatures is often employed for REMD in large parameter spaces. This is known [Kofke, 2002, Sugita & Okamoto, 1999, Sugita *et al.* , 2000] to yield nearly equal acceptance

rates between neighbouring temperatures, unless the problem is pathological.

3.3.1 Hamiltonian Monte Carlo

HMC unifies two schools of statistical sampling: Markov chain Monte Carlo (MCMC) [Metropolis *et al.*, 1953] and Molecular Dynamics (MD) [Alder & Wainwright, 1959]. MD introduces the concept of forward momentum $\mathbf{p}$ and makes use of the gradient of the cost function. Together $\mathbf{p}$ and the gradient information guide $\mathbf{w}$ rapidly through complex $J(\mathbf{w}|D)$ landscapes. Standard MCMC makes no use of $\mathbf{p}$ or the gradient: instead MCMC makes small random moves and diffuses inefficiently through $\mathbf{w}$ space. However, if an intelligent proposal is known, perhaps a move between regions of state space with large weight in the temperature-adjusted posterior distribution, but separated by long distances and barriers of low probability density, then MCMC can be an extremely powerful approach.

HMC combines the benefits of MD and MCMC. A short MD trajectory is performed to rapidly displace $\mathbf{w}$. So long as the MD trajectory is reversible and symplectic, then this move can be treated as any other MCMC proposal. Thus HMC enables the combination of intelligent MCMC moves with rapid MD trajectories. HMC has been shown [Neal *et al.*, 2011] to improve the scaling of sampling a $d-$dimensional multivariate Gaussian distribution from d^2 in MCMC to $d^{\frac{5}{4}}$. The cost of HMC can further be reduced by a factor if one avoids backward motion between MD trajectories [Horowitz, 1991, Sohl-Dickstein *et al.*, 2014]. However, in Alg. 1 we present the simplest HMC algorithm since this already improves the scaling, and $M^{(3)}$ and $M^{(3*)}$ have 10690 dimensions each. We use the Velocity Verlet scheme [Swope *et al.*, 1982] to propagate the MD trajectories.

The forward momentum $\mathbf{p}$ is related to the velocity $\frac{d\mathbf{w}}{dt}$ by $p_i = m_i \frac{dw_i}{dt}$ where m_i is called the "mass" of w_i. In this paper $m_i = 1 \, \forall i$. In general these m_i can be set to different values, which may make sampling more efficient. For example, one might set $m_i = k_i$, where k_i is the fan in of the target neuron for w_i as described in Sec. 3.2. This would have the effect of approximately equalising the expected change in the input to each neuron for each time step, subject to the assumption that the activations of neurons in the preceding layer and the weights leading to each neuron are all independent.

Algorithm 1 One Hamiltonian Monte Carlo trajectory. Samples $\text{prob}(\mathbf{w}) \propto e^{-J(\mathbf{w}|D)}$.

---
procedure HMC($\mathbf{w}$)
   $\mathbf{w}_o = \mathbf{w}$ $\triangleright$ Copy initial $\mathbf{w}$
   Draw random $p_{0,i} \sim \mathcal{N}(0, m_i T) \forall i$ $\triangleright$ Random $\mathbf{p}_o$
   $U_o = J(\mathbf{w}_o|D) + \sum_i \frac{p_{o,i}^2}{2m_i}$ $\triangleright$ Initial total energy
   Propagate $(\mathbf{w}_o, \mathbf{p}_o)$ using the Velocity Verlet algorithm for L time steps of length dt. Final coordinates are $(\mathbf{w}_n, \mathbf{p}_n)$.
   $U_n = J(\mathbf{w}_n|D) + \sum_i \frac{p_{n,i}^2}{2m_i}$ $\triangleright$ Final total energy
   Set $\mathbf{w} = \mathbf{w}_n$ with probability
   $\min\left[1, e^{\frac{1}{T}(U_o - U_n)}\right]$.
return $\mathbf{w}$
end procedure

3.3.2 Replica-Exchange Hamiltonian Monte Carlo

Our RE-HMC algorithm proceeds as follows. We used a geometric series of N_T temperatures, as described in Sec. 3.3. The current parameters values for the sampler at temperature T_i are $\mathbf{w}_i$. To save space, we will write $E_i = E(D|\mathbf{w}_i)$.

Algorithm 2 Replica-exchange Hamiltonian Monte Carlo

---
procedure RE-HMC($\{(T_i, \mathbf{w}_i)\}$)
   loop
      for T_i in $\{T\}$ **do**
         Update sample $\mathbf{w}_i$ from (10) by performing N_{traj} HMC trajectories of length L and time step dt.
      end for
      for $i \leftarrow 1, N_T$ **do**
         Choose random adjacent T pair (T_j, T_{j+1}).
         Exchange $(\mathbf{w}_j, \mathbf{w}_{j+1})$ with probability
         $\min\left[1, e^{\left(\frac{1}{T_j} - \frac{1}{T_{j+1}}\right)(E_j - E_{j+1})}\right]$.
      end for
   end loop
end procedure

In this fresh implementation independent time steps dt are set for each temperature by running additional HMC trajectories which are not included in the main simulation, and measuring the acceptance rate of the trajectories. The time step is updated to obtain an acceptance rate inside a the range (0.6–0.7), which is centred around the optimal value for HMC sampling of a multivariate Gaussian distribu-

tion, 0.65 [Neal *et al.* , 2011].

At the start of the simulation the parameters of the neural network are initialised independently for each temperature using the following approach:

1. Draw **w** uniformly at random from the region $w_i \in [-\frac{1}{\sqrt{k_i}}, \frac{1}{\sqrt{k_i}}]$ with k_i as defined in Sec. 3.2.

2. Minimise the cost function (7) using Alg. 3. This minimisation avoids starting the dynamics from parameter values where the gradient is extremely large.

3. Perform a number of burn-in trajectories at the appropriate temperature.

Algorithm 3 Fast minimisation of the cost function (7).

---

procedure RMIN(**w**)
    p = 0   ▷ Set initial momenta to zero.
    for $i = 1, N_{\text{steps}}$ **do**
        $\mathbf{w}_{\text{save}}, J_{\text{save}} = \mathbf{w}, J(\mathbf{w}|D)$ ▷ Save lowest cost state.
        Propagate $(\mathbf{w}, \mathbf{p})$ through 1 Velocity Verlet time step, duration dt.
        if $J(\mathbf{w}|D) < J_{\text{save}}$ **then**▷ Downwards step
            $dt = dt + 0.05.$ ▷ Slowly increase dt.
        else   ▷ Upwards step
            p = 0   ▷ Zero momenta
            $\mathbf{w} = \mathbf{w}_{\text{save}}$ ▷ Return to lowest cost **w**
            $dt = dt * 0.95$ ▷ Rapidly decrease dt.
        end if
    end for
end procedure

Alg. 3 is related to the FIRE optimisation algorithm [Bitzek *et al.* , 2006]. We found Alg. 3 to be extremely efficient for these simple networks. For D_{50} and D_{500} Alg. 3 required just a few hundred steps to reach $E(D|\mathbf{w}) = 0$ in $M^{(3)}$.

Fig. 2 shows the trajectory of a RE-HMC simulation with $M^{(3)}$ and D_{500}. The parameters used are given in Table 1. The typical value of E explores a narrow range which is specified by the temperature. At high temperature $M^{(3)}$ has higher training loss than the test loss of an uninformed classifier that assigns equal probability to each digit class for every image. From (9) it is apparent that such high training losses can be achieved by assigning low probability to the correct class label for just fraction of training examples.
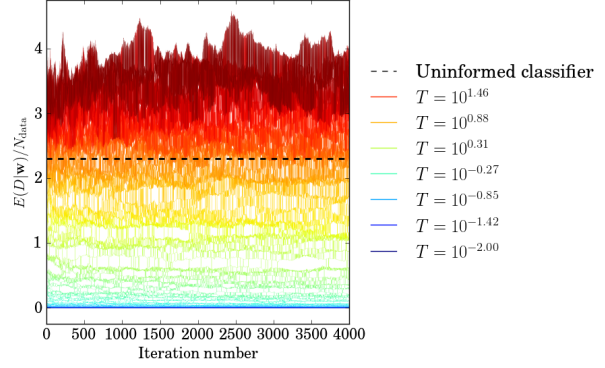

Figure 2: Trajectory of a replica-exchange HMC Simulation for $M^{(3)}$ with D_{500}.

Table 1: RE-HMC Parameters in Figs. 2, 3 and 4.

PARAMETER	FIGS. 2 AND 3	FIG. 4
N_T	112	84
$T_{\min}$	10^{-2}	$10^{-\frac{3}{2}}$
$T_{\max}$	10^2	$10^{\frac{3}{2}}$
N_{traj}	10	10
L	100	100

3.3.3 Relationship to Previous Work

HMC was first introduced to Bayesian learning for neural networks by Radford Neal [Neal, 2012]. Researchers have also applied a MD algorithm called Langevin dynamics [Welling & Teh, 2011] to sample the posterior for neural networks. In contrast to HMC, samples drawn using Langevin dynamics are only drawn from the true posterior asymptotically in the limit of vanishing dt. In [Chandra *et al.* , 2018] the authors use both replica-exchange with MCMC (also called "parallel tempering") and REMD with Langevin dynamics to accelerate exploration of neural network parameter spaces. However, they do not investigate temperature effects. Instead, after performing replica-exchange simulations, they set $T = 1$ for all samplers, irrespective of their temperatures during replica-exchange, and draw further MCMC samples from the posterior. Parallel tempering has also been successfully applied to the training of restricted Boltzmann machines (e.g. [Cho *et al.* , 2011, Cho *et al.* , 2010, Desjardins *et al.* , 2010, Desjardins *et al.* , 2014].)

To our knowledge, this is the first time a replica-exchange HMC (RE-HMC) approach has been applied to study neural networks. Different RE-HMC algorithms adapted to other scientific

domains can be found in [Jenkins *et al.* , 2013, Venditto *et al.* , 2015].

3.4 RESULTS: THE BEST GENERALISATION IS NOT ALWAYS FOUND AT $T = 1$

There is broad agreement in the Bayesian learning community that sampling from the posterior leads to better generalisation than minimising the cost function. Our results strongly support this assertion. However, we also observe that improved performance can in general be obtained at temperatures other than $T = 1$, and that the optimal value of T depends on the network and the data set.

Fig. 3 shows the sampled average of the potential energy $\langle E(D|\mathbf{w})\rangle_T$ for $M^{(3)}$ across a wide range of T, and using different sized training sets. The parameters of this calculation are shown in Table 1. In all data sets sampling from the posterior yields much improved generalisation as compared to standard optimisation. Here we apply the conventional initialisation described in Sec. 3.2 and minimise $E(D|\mathbf{w})$ using Alg. 3. For D_{50} and D_{500} we repeat the procedure until we obtain 100 vectors $\{\mathbf{w}\}$ with $E(D|\mathbf{w}) = 0$ for each data set (200 in total). Fig. 3 reports the mean test errors of these parameter sets. It is far more difficult to obtain $E(D_{5000}|\mathbf{w}) = 0$ by this optimisation method, so instead we repeated this optimisation 4000 times, keeping the 100 $\mathbf{w}$ with the lowest training error. The reported test error for D_{5000} is the mean test error of those 100 $\mathbf{w}$.

At low temperatures $M^{(3)}$ achieves perfect performance on all training sets. Conversely, clear minima are observed in the test error. This corresponds to the temperature at which $M^{(3)}$ begins overfitting the data, exhibiting high variance. For the smaller data sets, D_{50} and D_{500}, the minimum test error does indeed occur in the region of $T \sim 1$. However, for the larger data set D_{5000} (which is still relatively small) that minimum occurs at $T > 1$.

Sampling at a range of temperatures to identify that with the lowest test error is a statistical equivalent of early stopping: at low temperatures the temperature-adjusted posterior lies close to the global minimum of $E(D|\mathbf{w})$ and overfits the noise in these small data sets. At very high T the network is essentially untrained, and between these two extremes the network learns, but not so much as to overfit the data. This leads to the observed minima.

At high T and for all data sets, $M^{(3)}$ has higher test loss than both $\mathbf{w}$ obtained by standard optimi-
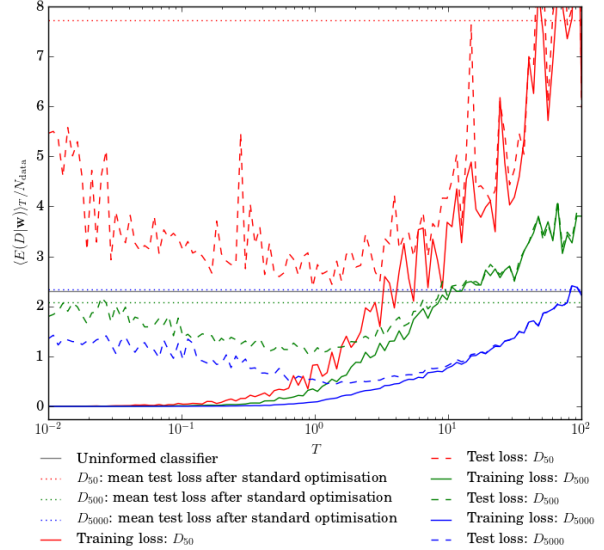

Figure 3: Empirical $\langle E(D|\mathbf{w})\rangle_T$ for $M^{(3)}$ at different temperatures T. Sharp variations are due to sampling noise. "Standard optimisation" of $M^{(3)}$ is described in the text.

sation and an "uninformed" classifier, which assigns an equal probability of 0.1 to each digit class, independent of the image. As the data set is increased, the temperature at which the network test loss exceeds that of the "uninformed" classifier becomes higher. For the smallest data set D_{50} the network is in this state for all temperatures.

Finally we remark that in $M^{(3)}$, which is a small neural network, the transition from accurate to inaccurate classifier is gradual, and does not exhibit a sharp jump. A statistical physicist would say that no "first-order phase transition" is observed. If a sharp transition between accurate and inaccurate classifiers were observed then accurate and efficient sampling of the temperature-adjusted posterior would require specialised techniques including Metadynamics [Barducci *et al.* , 2011] or an adaptive temperature replica-exchange algorithm such as [Katzgraber *et al.* , 2006]. From this perspective, the absence of a first-order phase transition is an important result, since it makes sampling the temperature-adjusted posterior distribution relatively simple and inexpensive.

In order to check the generality of these observations we repeated the calculations shown in Fig. 3 for the neural network $M^{(3*)}$. The results for $M^{(3*)}$ are shown in Fig. 4. The parameters for this calculation can be found in Table 1.
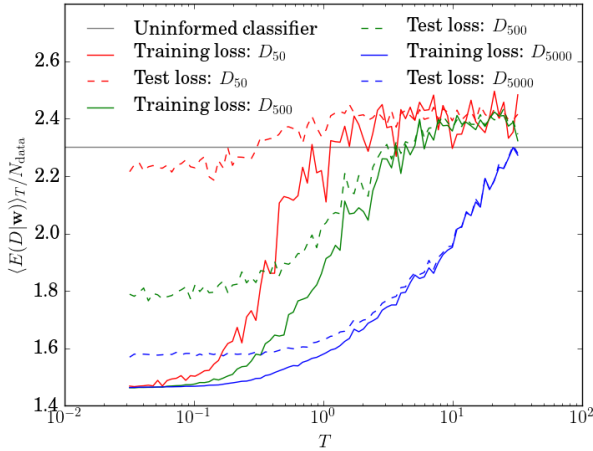
Figure 4: Empirical $\langle E(D|\mathbf{w}) \rangle_T$ for $M^{(3*)}$ across a range of temperatures T. Sharp variations are due to sampling noise.

From Fig. 4 we see that the same trends are observed for the training error in both $M^{(3)}$ and $M^{(3*)}$. In particular, for $M^{(3*)}$ $T = 1$ does not lead to optimal generalisation. Rather the best generalisation is achieved at low T: no minimum in the test error is apparent across the temperatures studied. This is an expression of high bias in $M^{(3*)}$. As in $M^{(3)}$, a transition is observed for $M^{(3*)}$ from relatively accurate to inaccurate classification as T is increased. Overall $M^{(3*)}$ generalises approximately as well as $M^{(3)}$ for the smallest data sets D_{50} and D_{500}, but for D_{5000} $M^{(3)}$ shows far better performance than $M^{(3*)}$. We anticipate that this trend will continue, since the minimum training errors for any data set obtained by $M^{(3)}$ and $M^{(3*)}$ are 0 and 1.46 respectively: while $M^{(3)}$ exhibits high variance, $M^{(3*)}$ exhibits high bias on this problem.

3.5 SUMMARY

We find that, for $M^{(3)}$, which is a typical neural network classifier, sampling the posterior distribution does indeed give lower test error than standard optimisation of the cost function. However, even better performance can often be obtained at temperatures other than $T = 1$. Furthermore, quite different values of T are appropriate to different neural networks and data sets. The optimal temperature can be located by sampling the temperature-adjusted posterior across a wide range of temperatures to identify that with the lowest test error. For networks with high bias optimal network parameters are found at $T \ll 1$, whereas for models with high variance a clear minimum of the test error is observed at higher

temperature. This is a statistical equivalent to early stopping. As temperature is increased we observe a gradual transition from accurate to inaccurate classifiers. The absence of a sharp transition is significant for the choice of sampling method used. It implies that no special temperature schedule is required for sampling methods such as the RE-HMC scheme presented here, and that more specialised methods such as Metadynamics, which are designed to enhance mixing across such sharp transitions, are not required.

4 BAYESIAN MODEL SELECTION FOR DEEP (AND SHALLOW) NEURAL NETWORKS WITH THERMODYNAMIC INTEGRATION

In this section we demonstrate how thermodynamic integration (TI), can be used to perform Bayesian model selection between two neural network models: one deep and one shallow network. We begin with a quick overview of Bayesian model selection before describing TI, the details of our calculation, and presenting the results.

TI is related to annealed importance sampling [Neal, 2001]. Encouragingly, annealed importance sampling has successfully been applied to Boltzmann machines [Salakhutdinov, 2008], deep belief nets [Salakhutdinov & Murray, 2008] and deep generative models [Wu *et al.*, 2016].

4.1 NEURAL NETWORK ARCHITECTURES AND PRIORS

In this section we reuse the neural network $M^{(3)}$, described in Sec. 3.1. We shall compare $M^{(3)}$ to a second neural network, $M^{(1)}$, which is identical to $M^{(3)}$ except that, in place of the 3 hidden layers in $M^{(3)}$, $M^{(1)}$ contains only a single hidden layer.

The priors of these two networks are different because $M^{(1)}$ has fewer parameters than $M^{(3)}$. Standard training of $\mathbf{w}$ for $M^{(1)}$ using Alg. 3 gives $\max_i \left(w_i \div \frac{1}{\sqrt{k_i}} \right) = 33.3$ at $E(D_{500}|\mathbf{w}) = 0$. This is similar to the value 30.3 obtained for $M^{(3)}$. Consequently we used the same width factor 50 to define the prior for $M^{(1)}$, as in $M^{(3)}$, thereby assigning the bounds of the prior σ according to (11).

4.2 INTRODUCTION TO BAYESIAN MODEL SELECTION

Bayesian model selection for neural networks was first formulated by David MacKay [MacKay, 1992a, MacKay, 1991]. Imagine we have two different machine learning models M_1 and M_2, and a data set $D = \{\mathbf{x}, t_i\}$, or in vector notation $D = (\mathbf{x}, \mathbf{t})$ where $\mathbf{x}$ is a matrix of all the input data and $\mathbf{t}$ a vector of the corresponding class labels.

In Bayesian model selection one computes and compares the probabilities of the models given the data, $\mathrm{prob}\,(M|\mathbf{x}, \mathbf{t})$. A true Bayesian would believe both models to different extents, according to their probabilities. However it is common practice to choose the model with highest probability, discarding all others.

The ratio of the probabilities for M_1 and M_2 can be found as follows

$$\mathrm{prob}\,(M, \mathbf{t}|\mathbf{x}) = \mathrm{prob}\,(M|\mathbf{x}, \mathbf{t})\,\mathrm{prob}\,(\mathbf{t}|\mathbf{x}) \quad (12)$$
$$= \mathrm{prob}\,(\mathbf{t}|\mathbf{x}, M)\,\mathrm{prob}\,(M|\mathbf{x}) \quad (13)$$
$$\Rightarrow \frac{\mathrm{prob}\,(M_1|\mathbf{x}, \mathbf{t})}{\mathrm{prob}\,(M_2|\mathbf{x}, \mathbf{t})} = \frac{\mathrm{prob}\,(\mathbf{t}|\mathbf{x}, M_1)}{\mathrm{prob}\,(\mathbf{t}|\mathbf{x}, M_2)} \frac{\mathrm{prob}\,(M_1)}{\mathrm{prob}\,(M_2)} \quad (14)$$

The term furthest to the right in (14) is the ratio of our priors for the models and represents our initial bias towards either model. Typically we have no initial preference between M_1 and M_2, in which case this term is equal to 1.

In theory one can calculate $\mathrm{prob}\,(\mathbf{t}|\mathbf{x}, M)$ as follows

$$\mathrm{prob}\,(\mathbf{t}|\mathbf{x}, M) = \int d\mathbf{w}\,\mathrm{prob}\,(\mathbf{t}, \mathbf{w}|\mathbf{x}, M) \quad (15)$$

$$= \int d\mathbf{w}\,\mathrm{prob}\,(\mathbf{t}|\mathbf{w}, \mathbf{x}, M)\,\mathrm{prob}\,(\mathbf{w}) \quad (16)$$

$$= \int d\mathbf{w}\, e^{-J(\mathbf{w}|D)} \quad (17)$$

In (17) we have made a substitution from (4).

4.3 METHOD: THERMODYNAMIC INTEGRATION

In this section we describe how (17) can be calculated using TI. The TI algorithm we describe here was originally introduced in [Frenkel & Ladd, 1984]. This approach asserts that the integrand of (17) is approximately unimodal, and localised in the region surrounding a particular local minimum of $J(\mathbf{w}|D)$, at $\mathbf{w}_0$.

We then sample $e^{-J(\mathbf{w}|D)}$ in the region of $\mathbf{w}_0$ and fit an approximate quadratic form for $J(\mathbf{w}|D)$

$$\frac{1}{2}(\mathbf{w} - \mathbf{w}_0)^\mathsf{T}\,\mathbf{k}\,(\mathbf{w} - \mathbf{w}_0) \simeq J(\mathbf{w}|D). \quad (18)$$

The evidence for that quadratic form

$$Z_0 = \int_{|\mathbf{w}| < \frac{\sigma}{2}} e^{-(\mathbf{w} - \mathbf{w}_0)^\mathsf{T} \frac{\mathbf{k}}{2}(\mathbf{w} - \mathbf{w}_0)} \quad (19)$$

is known exactly, provided $\mathbf{k}$ is diagonal. For this reason we fit a diagonal matrix $\mathbf{k}$ with

$$k_{ii} = \frac{1}{\langle (w_i - w_{0,i})^2 \rangle}. \quad (20)$$

The average $\langle (w_i - w_{0,i})^2 \rangle$ is taken by sampling from the posterior distribution (2). If one is using a uniform prior then it is important that, for this averaging alone, we should instead sample from $\mathrm{prob}(\mathbf{w}) \propto e^{-E(D|\mathbf{w})}$ without applying hard boundaries to $\mathbf{w}$. This ensures that the approximation (18) is a good fit to $J(\mathbf{w}|D)$.

Next in TI one sequentially samples a series of N_{bridge} bridging potentials constructed to interpolate between $J(\mathbf{w}|D)$ and the approximation (18).

$$J_{\mathrm{bridge}}(\mathbf{w}|D, \lambda) = (1 - \lambda)(J(\mathbf{w}|D) - J(\mathbf{w}_0|D))$$

$$+ \lambda(\mathbf{w} - \mathbf{w}_0)^\mathsf{T}\frac{\mathbf{k}}{2}(\mathbf{w} - \mathbf{w}_0) + J(\mathbf{w}_0|D). \quad (21)$$

Thus $J_{\mathrm{bridge}}(\mathbf{w}|D, \lambda = 0) = J(\mathbf{w}|D)$ and $J_{\mathrm{bridge}}(\mathbf{w}|D, \lambda = 1) = (\mathbf{w} - \mathbf{w}_0)^\mathsf{T}\frac{\mathbf{k}}{2}(\mathbf{w} - \mathbf{w}_0)$.

Defining

$$F = -\log \int_{|\mathbf{w}| < \frac{\sigma}{2}} e^{-J(\mathbf{w}|D)} \quad (22)$$

$$F_0 = -\log Z_0 \quad (23)$$

it can be shown [Frenkel & Smit, 2001] that

$$F = F_0 + \int_{\lambda=1}^{\lambda=0} d\lambda \left\langle \frac{\partial J_{\mathrm{bridge}}(\mathbf{w}|D, \lambda)}{\partial \lambda} \right\rangle_\lambda, \quad (24)$$

where

$$\left\langle \frac{\partial J_{\mathrm{bridge}}(\mathbf{w}|D, \lambda)}{\partial \lambda} \right\rangle_\lambda =$$

$$\left\langle \sum_{i=1}^{d} \frac{k_{ii}}{2}(w_i - w_{0,i})^2 - [J(\mathbf{w}|D) - J(\mathbf{w}_0|D)] \right\rangle_\lambda. \quad (25)$$

Averaging in (25) occurs over the distribution $\mathrm{prob}(\mathbf{w}) \propto e^{-J_{\mathrm{bridge}}(\mathbf{w}|D, \lambda)}$.

In this way, TI evaluates the evidence for a model,

$$\mathrm{prob}\,(\mathbf{t}|\mathbf{x}, M) = e^{-F} \quad (26)$$

as a correction to Z_0.

4.3.1 Experimental Setup

We reuse our HMC implementation for these calculations. HMC trajectories are all performed using $J_{\text{bridge}}(\mathbf{w}|D, \lambda)$. Before fitting $\mathbf{k}$, dt is adjusted to obtain an acceptance rate in the range (0.6–0.7). For fitting $\mathbf{k}$ we set $\lambda = 0$ and perform 1000 trajectories of burn-in each $L = 100$ steps long, then a further 1000 trajectories of the same length to obtain an estimate of $\mathbf{k}$ according to (20).

We use 100 bridging distributions (102 distributions in total). These distributions are sampled sequentially, from $\lambda = 0$ to $\lambda = 1$. To evaluate each expectation (25) 100 trajectories of HMC with $L = 100$ are performed as burn-in, before a further 100 trajectories of the same length to collect samples. We update dt after every 10 bridging distributions to recover an acceptance rate in the range (0.6–0.7). Additional trajectories, not included in any averaging or burn-in, are performed in order to set dt. The integration in (24) is performed using Simpson's rule.

4.4 RESULTS: BAYESIAN SELECTION BETWEEN $M^{(1)}$ AND $M^{(3)}$

In this section we perform a Bayesian model selection between the neural networks $M^{(3)}$ which has three hidden layers, and $M^{(1)}$ which has one. These networks and their priors are fully described in Secs. 3.1 and 4.1.

To obtain $\mathbf{w}_0$ for each model we followed the standard training procedure: weights and biases are initialised uniformly at random inside the region $w_i \in [-\frac{1}{\sqrt{k_i}}, \frac{1}{\sqrt{k_i}}]$ where k_i is the corresponding fan in, as described in Sec. 3.2. All calculations are performed with data set D_{500} (see Sec. 3.1).

For the uniform prior (6)

$$\text{prob}\,(\mathbf{t}|\mathbf{x}, M) = \int d\mathbf{w}\, e^{-E(D|\mathbf{w})} \times \frac{1}{\prod_{i=1}^{d} \sigma_i}. \quad (27)$$

Using TI we obtain

$$\log\left(\int_{|\mathbf{w}|<\frac{\sigma}{2}} d\mathbf{w}\, e^{-E^{(3)}(\mathbf{w}|D)}\right) = (26.48 \pm 0.17) \times 10^3$$

$$\log\left(\int_{|\mathbf{w}|<\frac{\sigma}{2}} d\mathbf{w}\, e^{-E^{(1)}(\mathbf{w}|D)}\right) = (19.79 \pm 0.01) \times 10^3$$

The uncertainties shown correspond to the standard deviations of 28 independent calculations for each

model. Additionally, for our prior we have

$$\log\left(\prod_{i=1}^{d} \sigma_i^{(3)}\right) = 28.960 \times 10^3$$

$$\log\left(\prod_{i=1}^{d} \sigma_i^{(1)}\right) = 19.946 \times 10^3$$

Thus we obtain

$$\frac{\text{prob}\left(M^{(3)}|\mathbf{x}, \mathbf{t}\right)}{\text{prob}\left(M^{(1)}|\mathbf{x}, \mathbf{t}\right)} = \frac{e^{(26.475 \pm 0.173) \times 10^3}}{e^{(19.793 \pm 0.013) \times 10^3}} \times \frac{e^{19946}}{e^{28960}} \tag{28}$$

$$= e^{6682} \times e^{-9014} \tag{29}$$

$$\simeq e^{-2332}$$

Bayes theorem clearly favours the shallow network $M^{(1)}$ over the deep network $M^{(3)}$. For larger data sets, this preference may well be reversed, but for D_{500} the accuracy of $M^{(3)}$ is not enough to justify its larger parameter space.

5 CONCLUSION

We have seen that for a number of small MNIST training sets sampling from the posterior does indeed lead to markedly improved generalisation than standard optimisation of the cost function. Even better generalisation can often be obtained by sampling the temperature-adjusted posterior (10) at $T \neq 1$. The optimal temperature, with the lowest test error, depends on the network and the data set. In particular, our results suggest that networks with high bias generalise best at $T \ll 1$, whereas networks with high variance exhibit a clear minimum of the test error at higher temperature: a statistical equivalent to early stopping. We saw that for network parameters sampled from the temperature-adjusted posterior across a range of temperatures, neural network based classifiers exhibit a transition between accurate and inaccurate classifiers, although no sharp "phase transition" occurs between these states. The absence of such a sharp transition makes sampling the temperature-adjusted posterior relatively simple and inexpensive. Finally we showed how thermodynamic integration can be used to perform model selection on deep and shallow neural networks, avoiding the need to precisely calculate or invert the Hessian of the cost function.

References

[Ahn *et al.* , 2012] Ahn, Sungjin, Korattikara, Anoop, & Welling, Max. 2012. Bayesian posterior

sampling via stochastic gradient fisher scoring. *arxiv preprint arxiv:1206.6380.*

[Alder & Wainwright, 1959] Alder, Berni J, & Wainwright, Thomas Everett. 1959. Studies in molecular dynamics. i. general method. *The journal of chemical physics*, **31**(2), 459–466.

[Baldock, 2019] Baldock, Robert J.N. 2019. *nn_sample.* https://github.com/rjnbaldock/nn_sample.

[Barducci *et al.* , 2011] Barducci, Alessandro, Bonomi, Massimiliano, & Parrinello, Michele. 2011. Metadynamics. *Wiley interdisciplinary reviews: Computational molecular science*, **1**(5), 826–843.

[Bitzek *et al.* , 2006] Bitzek, Erik, Koskinen, Pekka, Gähler, Franz, Moseler, Michael, & Gumbsch, Peter. 2006. Structural relaxation made simple. *Phys. rev. lett.*, **97**(Oct), 170201.

[Chandra *et al.* , 2018] Chandra, Rohitash, Jain, Konark, Deo, Ratneel V, & Cripps, Sally. 2018. Langevin-gradient parallel tempering for bayesian neural learning. *arxiv preprint arxiv:1811.04343.*

[Cho *et al.* , 2010] Cho, KyungHyun, Raiko, Tapani, & Ilin, Alexander. 2010. Parallel tempering is efficient for learning restricted boltzmann machines. *Pages 1–8 of: The 2010 international joint conference on neural networks (ijcnn).* Citeseer.

[Cho *et al.* , 2011] Cho, KyungHyun, Ilin, Alexander, & Raiko, Tapani. 2011. Improved learning of gaussian-bernoulli restricted boltzmann machines. *Pages 10–17 of: International conference on artificial neural networks.* Springer.

[Desjardins *et al.* , 2010] Desjardins, Guillaume, Courville, Aaron, & Bengio, Yoshua. 2010. Adaptive parallel tempering for stochastic maximum likelihood learning of rbms. *arxiv preprint arxiv:1012.3476.*

[Desjardins *et al.* , 2014] Desjardins, Guillaume, Luo, Heng, Courville, Aaron, & Bengio, Yoshua. 2014. Deep tempering. *arxiv preprint arxiv:1410.0123.*

[Duane *et al.* , 1987] Duane, Simon, Kennedy, Anthony D, Pendleton, Brian J, & Roweth, Duncan. 1987. Hybrid monte carlo. *Physics letters b*, **195**(2), 216–222.

[Frenkel & Ladd, 1984] Frenkel, Daan, & Ladd, Anthony JC. 1984. New monte carlo method to compute the free energy of arbitrary solids. application to the fcc and hcp phases of hard spheres. *The journal of chemical physics*, **81**(7), 3188–3193.

[Frenkel & Smit, 2001] Frenkel, Daan, & Smit, Berend. 2001. *Understanding molecular simulation: from algorithms to applications.* Vol. 1. Elsevier.

[Horowitz, 1991] Horowitz, Alan M. 1991. A generalized guided monte carlo algorithm. *Physics letters b*, **268**(2), 247 – 252.

[Jenkins *et al.* , 2013] Jenkins, R., Curotto, E., & Mella, Massimo. 2013. Replica exchange with smart monte carlo and hybrid monte carlo in manifolds. *Chemical physics letters*, **590**, 214 – 220.

[Katzgraber *et al.* , 2006] Katzgraber, Helmut G, Trebst, Simon, Huse, David A, & Troyer, Matthias. 2006. Feedback-optimized parallel tempering monte carlo. *Journal of statistical mechanics: Theory and experiment*, **2006**(03), P03018.

[Kofke, 2002] Kofke, David A. 2002. On the acceptance probability of replica-exchange monte carlo trials. *The journal of chemical physics*, **117**(15), 6911–6914.

[LeCun *et al.* , 1998] LeCun, Yann, Bottou, Léon, Bengio, Yoshua, Haffner, Patrick, *et al.* . 1998. Gradient-based learning applied to document recognition. *Proceedings of the ieee*, **86**(11), 2278–2324.

[MacKay, 1991] MacKay, David J.C. 1991. Bayesian interpolation. *Neural computation*, **4**, 415–447.

[MacKay, 1992a] MacKay, David JC. 1992a. The evidence framework applied to classification networks. *Neural computation*, **4**(5), 720–736.

[MacKay, 1992b] MacKay, David JC. 1992b. A practical bayesian framework for backpropagation networks. *Neural computation*, **4**(3), 448–472.

[Mandt *et al.* , 2017] Mandt, Stephan, Hoffman, Matthew D, & Blei, David M. 2017. Stochastic gradient descent as approximate bayesian inference. *The journal of machine learning research*, **18**(1), 4873–4907.

[Metropolis *et al.* , 1953] Metropolis, Nicholas, Rosenbluth, Arianna W, Rosenbluth, Marshall N, Teller, Augusta H, & Teller, Edward. 1953. Equation of state calculations by fast computing

machines. *The journal of chemical physics*, **21**(6), 1087–1092.

[Neal, 2001] Neal, Radford M. 2001. Annealed importance sampling. *Statistics and computing*, **11**(2), 125–139.

[Neal, 2012] Neal, Radford M. 2012. *Bayesian learning for neural networks.* Vol. 118. Springer Science & Business Media.

[Neal *et al.* , 2011] Neal, Radford M, *et al.* . 2011. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, **2**(11), 2.

[Salakhutdinov, 2008] Salakhutdinov, Ruslan. 2008. Learning and evaluating boltzmann machines. *Tech. rep., technical report utml tr 2008-002, department of computer science, university of toronto.*

[Salakhutdinov & Murray, 2008] Salakhutdinov, Ruslan, & Murray, Iain. 2008. On the quantitative analysis of deep belief networks. *Pages 872–879 of: Proceedings of the 25th international conference on machine learning.* ACM.

[Smith & Le, 2017] Smith, Samuel L, & Le, Quoc V. 2017. A bayesian perspective on generalization and stochastic gradient descent. *arxiv preprint arxiv:1710.06451.*

[Sohl-Dickstein *et al.* , 2014] Sohl-Dickstein, Jascha, Mudigonda, Mayur, & DeWeese, Michael R. 2014. Hamiltonian monte carlo without detailed balance. *arxiv preprint arxiv:1409.5191.*

[Sugita & Okamoto, 1999] Sugita, Yuji, & Okamoto, Yuko. 1999. Replica-exchange molecular dynamics method for protein folding. *Chemical physics letters*, **314**(1), 141 – 151.

[Sugita *et al.* , 2000] Sugita, Yuji, Kitao, Akio, & Okamoto, Yuko. 2000. Multidimensional replica-exchange method for free-energy calculations. *The journal of chemical physics*, **113**(15), 6042–6051.

[Swendsen & Wang, 1986] Swendsen, Robert H., & Wang, Jian-Sheng. 1986. Replica monte carlo simulation of spin-glasses. *Phys. rev. lett.*, **57**(Nov), 2607–2609.

[Swope *et al.* , 1982] Swope, William C, Andersen, Hans C, Berens, Peter H, & Wilson, Kent R. 1982. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *The journal of chemical physics*, **76**(1), 637–649.

[Venditto *et al.* , 2015] Venditto, J.G., Wolf, S., Curotto, E., & Mella, Massimo. 2015. Replica exchange hybrid monte carlo simulations of the ammonia dodecamer and hexadecamer. *Chemical physics letters*, **635**, 127 – 133.

[Welling & Teh, 2011] Welling, Max, & Teh, Yee W. 2011. Bayesian learning via stochastic gradient langevin dynamics. *Pages 681–688 of: Proceedings of the 28th international conference on machine learning (icml-11).*

[Wu *et al.* , 2016] Wu, Yuhuai, Burda, Yuri, Salakhutdinov, Ruslan, & Grosse, Roger. 2016. On the quantitative analysis of decoder-based generative models. *arxiv preprint arxiv:1611.04273.*