# Hamiltonian Annealed Importance Sampling for partition function estimation

Jascha Sohl-Dickstein and Benjamin J. Culpepper
Redwood Center for Theoretical Neuroscience
University of California, Berkeley
Berkeley, CA

Redwood Technical Report
February 25, 2011

## Abstract

We introduce an extension to annealed importance sampling that uses Hamiltonian dynamics to rapidly estimate normalization constants. We demonstrate this method by computing log likelihoods in directed and undirected probabilistic image models. We compare the performance of linear generative models with both Gaussian and Laplace priors, product of experts models with Laplace and Student's t experts, the mc-RBM, and a bilinear generative model. We provide code to compare additional models.

## 1 Introduction

We would like to use probabilistic models to assign probabilities to data. Unfortunately, this innocuous statement belies an important, difficult problem: many interesting distributions used widely across sciences cannot be analytically normalized. Historically, the training of probabilistic models has been motivated in terms of maximizing the log likelihood of the data under the model or minimizing the KL divergence between the data and the model. However, for most models it is impossible to directly compute the log likelihood, due to the intractability of the normalization constant, or partition function. For this reason, performance is typically measured using a variety of diagnostic heuristics, not directly indicative of log likelihood. For example, image models are often compared in terms of their synthesis, denoising, inpainting, and classification performance. This inability to directly measure the log likelihood has made it difficult to consistently evaluate and compare models.

Recently, a growing number of researchers have given their attention to measures of likelihood in image models. Salakhutdinov & Murray (2008) use annealed importance sampling, and Murray & Salakhutdinov (2009) use a hybrid of annealed importance sampling and a Chib-style estimator to estimate the log likelihood of a variety of MNIST digits and natural image patches modeled using restricted Boltzmann machines and deep belief networks. Bethge (2006) measures the reduction in multi-information, or statistical redundancy, as images undergo various complete linear transformations. Chandler & Field (2007) and Stephens et al. (2008) produce estimates of the entropy inherent in natural scenes, but do not address model evaluation. Karklin (2007) uses kernel density estimates – essentially, vector quantization – to compare different image models, though that technique suffers from severe scaling problems except in specific contexts. Zoran & Weiss (2009) compare the true log likelihoods of a number of image models, but restricts their analysis to the rare cases where the partition function can be solved analytically.

In this work, we merge two existing ideas – annealed importance sampling and Hamiltonian dynamics – into a

single algorithm. To review, Annealed Importance Sampling (AIS) Neal (2001) is a sequential Monte Carlo method Moral et al. (2006) which allows the partition function of a non-analytically-normalizable distribution to be estimated in an unbiased fashion. This is accomplished by starting at a distribution with a known normalization, and gradually transforming it into the distribution of interest through a chain of Markov transitions. Its practicality depends heavily on the chosen Markov transitions. Hamiltonian Monte Carlo (HMC) Neal (2010) is a family of techniques for fast sampling in continuous state spaces, which work by extending the state space to include auxiliary momentum variables, and then simulating Hamiltonian dynamics from physics in order to traverse long iso-probability trajectories which rapidly explore the state space.

The key insight that makes our algorithm more efficient than previous methods is our adaptation of AIS to work with Hamiltonian dynamics. As in HMC, we extend the state space to include auxiliary momentum variables; however, we do this in such a way that the momenta change consistently through the intermediate AIS distributions, rather than resetting them at the beginning of each Markov transition. To make the practical applications of this work clear, we use our method, Hamiltonian Annealed Importance Sampling (HAIS), to measure the log likelihood of holdout data under a variety of directed (generative) and undirected (analysis/feed-forward) probabilistic models of natural image patches.

The source code to reproduce our experiments is available.

# 2 Estimating Log Likelihood

## 2.1 Importance Sampling

Importance sampling Kahn & Marshall (1953) allows an unbiased estimate $\hat{Z}_p$ of the partition function (or normalization constant) $Z_p$ of a non-analytically-normalizable target distribution $p(\mathbf{x})$ over $\mathbf{x} \in \mathbb{R}^M$,

$$p(\mathbf{x}) = \frac{e^{-E_p(\mathbf{x})}}{Z_p} \tag{1}$$

$$Z_p = \int d\mathbf{x}\, e^{-E_p(\mathbf{x})}, \tag{2}$$

to be calculated. This is accomplished by averaging over samples $\mathcal{S}_q$ from a proposal distribution $q(\mathbf{x})$,

$$q(\mathbf{x}) = \frac{e^{-E_q(\mathbf{x})}}{Z_q} \tag{3}$$

$$Z_p = \int d\mathbf{x}\, q(\mathbf{x}) \frac{e^{-E_p(\mathbf{x})}}{q(\mathbf{x})} \tag{4}$$

$$\hat{Z}_p = \frac{1}{|\mathcal{S}_q|} \sum_{x \in \mathcal{S}_q} \frac{e^{-E_p(\mathbf{x})}}{q(\mathbf{x})}, \tag{5}$$

where $|\mathcal{S}_q|$ is the number of samples. $q(\mathbf{x})$ is chosen to be easy both to sample from and to evaluate exactly, and must have support everywhere that $p(\mathbf{x})$ does. Unfortunately, unless $q(\mathbf{x})$ has significant mass everywhere $p(\mathbf{x})$ does, it takes an impractically large number of samples from $q(\mathbf{x})$ for $\hat{Z}_p$ to accurately approximate $Z_p$[1].

## 2.2 Annealed Importance Sampling

Annealed importance sampling Neal (2001) extends the state space $\mathbf{x}$ to a series of vectors, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2 \dots \mathbf{x}_N\}$, $\mathbf{x}_n \in \mathbb{R}^M$. It then transforms the proposal distribution $q(\mathbf{x})$ to a forward chain $Q(\mathbf{X})$ over $\mathbf{X}$, by setting $q(\mathbf{x})$ as the distribution over $\mathbf{x}_1$ and then multiplying by a series of Markov transition distributions,

$$Q(\mathbf{X}) = q(\mathbf{x}_1) \prod_{n=1}^{N-1} T_n(\mathbf{x}_{n+1}|\mathbf{x}_n), \tag{6}$$

where $T_n(\mathbf{x}_{n+1}|\mathbf{x}_n)$ represents a *forward* transition distribution from $\mathbf{x}_n$ to $\mathbf{x}_{n+1}$. The target distribution $p(\mathbf{x})$ is similarly transformed to become a reverse chain $P(\mathbf{X})$, starting at $\mathbf{x}_N$, over $\mathbf{X}$,

$$P(\mathbf{X}) = \frac{e^{-E_p(\mathbf{x}_N)}}{Z_p} \prod_{n=1}^{N-1} \widetilde{T}_n(\mathbf{x}_n|\mathbf{x}_{n+1}), \tag{7}$$

where $\widetilde{T}_n(\mathbf{x}_n|\mathbf{x}_{n+1})$ is a *reverse* transition distribution from $\mathbf{x}_{n+1}$ to $\mathbf{x}_n$. The transition distributions are, by definition, normalized (eg, $\int d\mathbf{x}_{\,n+1} T_n(\mathbf{x}_{n+1}|\mathbf{x}_n) = 1$).

In a similar fashion to Equations 4 and 5, samples $\mathcal{S}_Q$ from the forward proposal chain $Q(\mathbf{X})$ can be used to

[1] The expected variance of the estimate $\hat{Z}_p$ is given by an $\alpha$-divergence between $p(\mathbf{x})$ and $q(\mathbf{x})$, times a constant and plus an offset - see Minka (2005).

estimate the partition function $Z_p$ (note that all integrals but the first in Equation 8 go to 1),

$$Z_p = \int d\mathbf{x}_N \, e^{-E_p(\mathbf{x}_N)} \int d\mathbf{x}_{N-1} \, \widetilde{T}_{N-1}\left(\mathbf{x}_{N-1}|\mathbf{x}_N\right) \cdots \int d\mathbf{x}_1 \, \widetilde{T}_1\left(\mathbf{x}_1|\mathbf{x}_2\right) \tag{8}$$

$$= \int d\mathbf{X} \, Q\left(\mathbf{X}\right) \frac{e^{-E_p(\mathbf{x}_N)}}{Q\left(\mathbf{X}\right)} \widetilde{T}_{N-1}\left(\mathbf{x}_{N-1}|\mathbf{x}_N\right) \cdots \widetilde{T}_1\left(\mathbf{x}_1|\mathbf{x}_2\right) \tag{9}$$

$$\hat{Z}_p = \frac{1}{|\mathcal{S}_Q|} \sum_{X \in \mathcal{S}_Q} \frac{e^{-E_p(\mathbf{x}_N)}}{q\left(\mathbf{x}_1\right)} \frac{\widetilde{T}_1\left(\mathbf{x}_1|\mathbf{x}_2\right)}{T_1\left(\mathbf{x}_2|\mathbf{x}_1\right)} \cdots \frac{\widetilde{T}_{N-1}\left(\mathbf{x}_{N-1}|\mathbf{x}_N\right)}{T_{N-1}\left(\mathbf{x}_N|\mathbf{x}_{N-1}\right)}. \tag{10}$$

In order to further define the transition distributions, Neal introduces intermediate distributions $\pi_n\left(\mathbf{x}\right)$ between $q\left(\mathbf{x}\right)$ and $p\left(\mathbf{x}\right)$,

$$\pi_n\left(\mathbf{x}\right) = \frac{e^{-E_{\pi_n}(x)}}{Z_{\pi_n}} \tag{11}$$

$$E_{\pi_n}\left(\mathbf{x}\right) = \left(1-\beta_n\right) E_q\left(\mathbf{x}\right) + \beta_n E_p\left(\mathbf{x}\right), \tag{12}$$

where the mixing fraction $\beta_n = \frac{n}{N}$ for all results reported here. $T_n\left(\mathbf{x}_{n+1}|\mathbf{x}_n\right)$ is then chosen to be any Markov chain transition for $\pi_n\left(\mathbf{x}\right)$, meaning that it leaves $\pi_n\left(\mathbf{x}\right)$ invariant

$$T_n \circ \pi_n = \pi_n. \tag{13}$$

The reverse direction transition distribution $\widetilde{T}_n\left(\mathbf{x}_n|\mathbf{x}_{n+1}\right)$ is set to the reversal of $T_n\left(\mathbf{x}_{n+1}|\mathbf{x}_n\right)$,

$$\widetilde{T}_n\left(\mathbf{x}_n|\mathbf{x}_{n+1}\right) = T_n\left(\mathbf{x}_{n+1}|\mathbf{x}_n\right) \frac{\pi_n\left(\mathbf{x}_n\right)}{\pi_n\left(\mathbf{x}_{n+1}\right)}. \tag{14}$$

Equation 10 thus reduces to

$$\hat{Z}_p = \frac{1}{|\mathcal{S}_Q|} \sum_{X \in \mathcal{S}_Q} \frac{e^{-E_p(\mathbf{x}_N)}}{q\left(\mathbf{x}_1\right)} \frac{\pi_1\left(\mathbf{x}_1\right)}{\pi_1\left(\mathbf{x}_2\right)} \cdots \frac{\pi_{N-1}\left(\mathbf{x}_{N-1}\right)}{\pi_{N-1}\left(\mathbf{x}_N\right)} \tag{15}$$

$$= \frac{1}{|\mathcal{S}_Q|} \sum_{X \in \mathcal{S}_Q} \frac{e^{-E_p(\mathbf{x}_N)}}{q\left(\mathbf{x}_1\right)} \frac{e^{-E_{\pi_1}(x_1)}}{e^{-E_{\pi_1}(x_2)}} \cdots \frac{e^{-E_{\pi_{N-1}}(x_{N-1})}}{e^{-E_{\pi_{N-1}}(x_N)}}. \tag{16}$$

If the number of intermediate distributions $N$ is large, and the transition distributions $T_n\left(\mathbf{x}_{n+1}|\mathbf{x}_n\right)$ and $\widetilde{T}_n\left(\mathbf{x}_n|\mathbf{x}_{n+1}\right)$ mix effectively, then the distributions over intermediate states $\mathbf{x}_n$ will be nearly identical to $\pi_n\left(\mathbf{x}_n\right)$ in both the forward and backward chains. $P\left(\mathbf{X}\right)$ and $Q\left(\mathbf{X}\right)$ will then be extremely similar to one another, and the variance in the estimate $\hat{Z}_p$ will be extremely low[2]. If the transitions $T_n\left(\mathbf{x}_{n+1}|\mathbf{x}_n\right)$ do a poor job mixing, then the marginal distributions over $\mathbf{x}_n$ under $P\left(\mathbf{X}\right)$ and $Q\left(\mathbf{X}\right)$ will look different from $\pi_n\left(\mathbf{x}_n\right)$. The estimate $\hat{Z}_p$ will still be unbiased, but with a potentially larger variance. Thus, to make AIS practical, it is important to choose Markov transitions $T_n\left(\mathbf{x}_{n+1}|\mathbf{x}_n\right)$ for the intermediate distributions $\pi_n\left(\mathbf{x}\right)$ that mix quickly.

## 2.3 Hamiltonian Annealed Importance Sampling

Hamiltonian Monte Carlo Neal (2010) uses an analogy to the physical dynamics of particles moving with momentum under the influence of an energy function to propose Markov chain transitions which rapidly explore the state space. It does this by expanding the state space to include auxiliary momentum variables, and then simulating Hamiltonian dynamics to move long distances along iso-probability contours in the expanded state space. A similar technique is powerful in the context of annealed importance sampling. Additionally, by retaining the momenta variables across the intermediate distributions, significant momentum can build up as the proposal distribution is transformed into the target. This provides a mixing benefit that is unique to our formulation.

The state space $\mathbf{X}$ is first extended to $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2 \ldots \mathbf{y}_N\}$, $\mathbf{y}_n = \{\mathbf{x}_n, \mathbf{v}_n\}$, where $\mathbf{v}_n \in \mathbb{R}^M$ consists of a momentum associated with each position $\mathbf{x}_n$. The momenta associated with both the proposal and target distributions is taken to be unit norm isotropic gaussian. The proposal and target distributions $q\left(\mathbf{x}\right)$ and $p\left(\mathbf{x}\right)$ are extended to corresponding distributions $q_\cup\left(\mathbf{y}\right)$ and $p_\cup\left(\mathbf{y}\right)$

[2] There is a direct mapping between annealed importance sampling and the Jarzynski equality in non-equilibrium thermodynamics - see Jarzynski (1997). It follows from this mapping, and the reversibility of quasistatic processes, that the variance in $\hat{Z}_p$ can be made to go to 0 if the transition from $q\left(\mathbf{x}_1\right)$ to $p\left(\mathbf{x}_N\right)$ is sufficiently gradual.

over position and momentum $\mathbf{y} = \{\mathbf{x}, \mathbf{v}\}$,

$$p_{\cup}(\mathbf{y}) = p(\mathbf{x})\ \Phi(\mathbf{v}) = \frac{e^{-E_{p_{\cup}}(\mathbf{y})}}{Z_{p_{\cup}}} \tag{17}$$

$$q_{\cup}(\mathbf{y}) = q(\mathbf{x})\ \Phi(\mathbf{v}) = \frac{e^{-E_{q_{\cup}}(\mathbf{y})}}{Z_{q_{\cup}}} \tag{18}$$

$$\Phi(\mathbf{v}) = \frac{e^{-\frac{1}{2}\mathbf{v}^T\mathbf{v}}}{(2\pi)^{\frac{M}{2}}} \tag{19}$$

$$E_{p_{\cup}}(\mathbf{y}) = E_p(\mathbf{x}) + \frac{1}{2}\mathbf{v}^T\mathbf{v} \tag{20}$$

$$E_{q_{\cup}}(\mathbf{y}) = E_q(\mathbf{x}) + \frac{1}{2}\mathbf{v}^T\mathbf{v}. \tag{21}$$

The remaining distributions are extended to cover both position and momentum in a nearly identical fashion: the forward and reverse chains $Q(\mathbf{X}) \to Q_{\cup}(\mathbf{Y})$, $P(\mathbf{X}) \to P_{\cup}(\mathbf{Y})$, the intermediate distributions and energy functions $\pi_n(\mathbf{x}) \to \pi_{\cup\, n}(\mathbf{y})$, $E_{\pi_n}(\mathbf{x}) \to E_{\pi_{\cup\, n}}(\mathbf{y})$,

$$E_{\pi_{\cup\, n}}(\mathbf{y}) = (1-\beta_n) E_{q_{\cup}}(\mathbf{y}) + \beta_n E_{p_{\cup}}(\mathbf{y}) \tag{22}$$

$$= (1-\beta_n) E_q(\mathbf{x}) + \beta_n E_p(\mathbf{x}) + \frac{1}{2}\mathbf{v}^T\mathbf{v}, \tag{23}$$

and the forward and reverse Markov transition distributions $T_n(\mathbf{x}_{n+1}|\mathbf{x}_n) \to T_{\cup n}(\mathbf{y}_{n+1}|\mathbf{y}_n)$ and $\widetilde{T}_n(\mathbf{x}_n|\mathbf{x}_{n+1}) \to \widetilde{T}_{\cup n}(\mathbf{y}_n|\mathbf{y}_{n+1})$. Similarly, the samples $\mathcal{S}_{Q_{\cup}}$ now each have both position $\mathbf{X}$ and momentum $\mathbf{V}$, and are drawn from the forward chain described by $Q_{\cup}(\mathbf{Y})$.

The annealed importance sampling estimate $\hat{Z}_p$ given in Equation 16 remains *unchanged*, except for a replacement of $\mathcal{S}_Q$ with $\mathcal{S}_{Q_{\cup}}$ – all the terms involving the momentum $\mathbf{V}$ conveniently cancel out, since the same momentum distribution $\Phi(\mathbf{v})$ is used for the proposal $q_{\cup}(\mathbf{y}_1)$ and target $p_{\cup}(\mathbf{y}_N)$,

$$\hat{Z}_p = \frac{1}{|\mathcal{S}_{Q_{\cup}}|} \sum_{Y \in \mathcal{S}_{Q_{\cup}}} \frac{e^{-E_p(\mathbf{x}_N)}\Phi(\mathbf{v}_N)}{q(\mathbf{x}_1)\Phi(\mathbf{v}_1)} \frac{e^{-E_{\pi_1}(x_1)+\frac{1}{2}\mathbf{v}_1^T\mathbf{v}_1}}{e^{-E_{\pi_1}(x_2)+\frac{1}{2}\mathbf{v}_2^T\mathbf{v}_2}} \cdots \frac{e^{-E_{\pi_{N-1}}(x_{N-1})+\frac{1}{2}\mathbf{v}_{N-1}^T\mathbf{v}_{N-1}}}{e^{-E_{\pi_{N-1}}(x_N)+\frac{1}{2}\mathbf{v}_N^T\mathbf{v}_N}} \tag{24}$$

$$= \frac{1}{|\mathcal{S}_{Q_{\cup}}|} \sum_{Y \in \mathcal{S}_{Q_{\cup}}} \frac{e^{-E_p(\mathbf{x}_N)}}{q(\mathbf{x}_1)} \frac{e^{-E_{\pi_1}(x_1)}}{e^{-E_{\pi_1}(x_2)}} \cdots \frac{e^{-E_{\pi_{N-1}}(x_{N-1})}}{e^{-E_{\pi_{N-1}}(x_N)}}. \tag{25}$$

Thus, the momentum only matters when generating the samples $\mathcal{S}_{Q_{\cup}}$, by drawing from the initial proposal distribution $p_{\cup}(\mathbf{y}_1)$, and then applying the series of Markov transitions $T_{\cup n}(\mathbf{y}_{n+1}|\mathbf{y}_n)$.

For the transition distributions, $T_{\cup n}(\mathbf{y}_{n+1}|\mathbf{y}_n)$, we propose a new location by integrating Hamiltonian dynamics for a short time using a single leapfrog step, accept or reject the new location via Metropolis rules, and then partially corrupt the momentum. That is, we generate a sample from $T_{\cup n}(\mathbf{y}_{n+1}|\mathbf{y}_n)$ by following the procedure:

1. $\left\{\mathbf{x}_H^0, \mathbf{v}_H^0\right\} = \{\mathbf{x}_n, \mathbf{v}_n\}$

2. leapfrog: $\mathbf{x}_H^{\frac{1}{2}} = \mathbf{x}_H^0 + \frac{\epsilon}{2}\mathbf{v}_H^0$

   $$\mathbf{v}_H^1 = \mathbf{v}_H^0 - \epsilon \frac{\partial E_{\pi_n}(\mathbf{x})}{\partial \mathbf{x}}\Big|_{\mathbf{x}=\mathbf{x}_H^{\frac{1}{2}}}$$

   $$\mathbf{x}_H^1 = \mathbf{x}_H^{\frac{1}{2}} + \frac{\epsilon}{2}\mathbf{v}_H^1$$

   where the step size $\epsilon = 0.2$ for all experiments in this paper.

3. accept/reject: $\{\mathbf{x}', \mathbf{v}'\} = \left\{\mathbf{x}_H^1, -\mathbf{v}_H^1\right\}$ with probability $P_{accept} = \min\left[1, \frac{e^{-E_{\pi_n}\left(\mathbf{x}_H^1\right) - \frac{1}{2}{\mathbf{v}_H^1}^T\mathbf{v}_H^1}}{e^{-E_{\pi_n}\left(\mathbf{x}_H^0\right) - \frac{1}{2}{\mathbf{v}_H^0}^T\mathbf{v}_H^0}}\right]$, otherwise $\{\mathbf{x}', \mathbf{v}'\} = \left\{\mathbf{x}_H^0, \mathbf{v}_H^0\right\}$

4. partial momentum refresh: $\tilde{\mathbf{v}}' = -\sqrt{1-\gamma}\mathbf{v}' + \gamma\mathbf{r}$, where $r \sim \mathcal{N}(0, \mathbf{I})$, and $\gamma \in (0, 1]$ is chosen so as to randomize half the momentum power per unit simulation time Culpepper et al. (2011).

5. $\mathbf{y}_{n+1} = \{\mathbf{x}_{n+1}, \mathbf{v}_{n+1}\} = \{\mathbf{x}', \tilde{\mathbf{v}}'\}$

This combines the advantages of many intermediate distributions, which can lower the variance in the estimated $\hat{Z}_p$, with the improved mixing which occurs when momentum is maintained over many update steps. For details on Hamiltonian Monte Carlo sampling techniques, and a discussion of why the specific steps above leave $\pi_n(\mathbf{x})$ invariant, we recommend Culpepper et al. (2011); Neal (2010).

Some of the models discussed below have linear constraints on their state spaces. These are dealt with by negating the momentum $\mathbf{v}$ and reflecting the position $\mathbf{x}$ across the constraint boundary every time a leapfrog half-step violates the constraint.

### 2.4 Log likelihood of analysis models

Analysis models are defined for the purposes of this paper as those which have an easy to evaluate expression for $E_p(\mathbf{x})$ when they are written in the form of Equation 1. The average log likelihood $\mathcal{L}$ of an analysis model $p(\mathbf{x})$ over a set of testing data $\mathcal{D}$ is

$$\mathcal{L} = \frac{1}{|\mathcal{D}|}\sum_{\mathbf{x}\in\mathcal{D}} \log p(\mathbf{x}) = -\frac{1}{|\mathcal{D}|}\sum_{\mathbf{x}\in\mathcal{D}} E_p(\mathbf{x}) - \log Z_p \tag{26}$$

where $|\mathcal{D}|$ is the number of samples in $\mathcal{D}$, and the $Z_p$ in the second term can be directly estimated by Hamiltonian annealed importance sampling.

### 2.5 Log likelihood of generative models

Generative models are defined here to be those which have a joint distribution,

$$p(\mathbf{x},\mathbf{a}) = p(\mathbf{x}|\mathbf{a})\,p(\mathbf{a}) = \frac{e^{-E_{x|a}(\mathbf{x},\mathbf{a})}}{Z_{x|a}}\frac{e^{-E_a(\mathbf{a})}}{Z_a}, \tag{27}$$

over visible variables $\mathbf{x}$ and auxiliary variables $\mathbf{a} \in \mathbb{R}^L$ which is easy to exactly evaluate and sample from, but for which the marginal distribution over the visible variables $p(\mathbf{x}) = \int d\mathbf{a}\, p(\mathbf{x},\mathbf{a})$ is intractable to compute. The average log likelihood $\mathcal{L}$ of a model of this form over a testing set $\mathcal{D}$ is

$$\mathcal{L} = \frac{1}{|\mathcal{D}|}\sum_{\mathbf{x}\in\mathcal{D}} \log Z_{a|x} \tag{28}$$

$$Z_{a|x} = \int d\mathbf{a}\, e^{-E_{x|a}(\mathbf{x},\mathbf{a}) - \log Z_{x|a} - E_a(\mathbf{a}) - \log Z_a}, \tag{29}$$

where each of the $Z_{a|x}$ can be estimated using HAIS. Generative models take significantly longer to evaluate than analysis models, as a separate HAIS chain must be run for each test sample.

## 3 Models

The probabilistic forms for all models whose log likelihood we evaluate are given below. In all cases, $\mathbf{x} \in \mathbb{R}^M$ refers to the data vector.

1. linear generative:

$$p(\mathbf{x}|\mathbf{a}) = \frac{\exp\left[-\frac{1}{2\sigma_n^2}(\mathbf{x}-\Phi\mathbf{a})^T(\mathbf{x}-\Phi\mathbf{a})\right]}{(2\pi)^{\frac{M}{2}}\,\sigma_n^M} \tag{30}$$

   parameters: $\Phi \in \mathbb{R}^{M\times L}$
   auxiliary variables: $\mathbf{a} \in \mathbb{R}^L$
   constant: $\sigma_n = 0.1$
   Linear generative models were tested with a two priors, as listed:

   (a) Gaussian prior:

$$p(\mathbf{a}) = \frac{\exp\left[-\frac{1}{2}\mathbf{a}^T\mathbf{a}\right]}{(2\pi)^{\frac{L}{2}}} \tag{31}$$

   (b) Laplace prior Olshausen & Field (1997):

$$p(\mathbf{a}) = \frac{\exp\left[-||\mathbf{a}||_1^1\right]}{2} \tag{32}$$

2. bilinear generative Culpepper et al. (2011): The form is the same as for the linear generative model, but with the coefficients $\mathbf{a}$ decomposed into 2 multiplicative factors,

$$\mathbf{a} = (\Theta\mathbf{c}) \odot (\Psi\mathbf{d}) \tag{33}$$

$$p(\mathbf{c}) = \frac{\exp\left[-||\mathbf{c}||_1^1\right]}{2} \tag{34}$$

$$p(\mathbf{d}) = \exp\left[-||\mathbf{d}||_1^1\right], \tag{35}$$

where $\odot$ indicates element-wise multiplication.
parameters: $\Phi \in \mathbb{R}^{M \times L}$, $\Theta \in \mathbb{R}^{L \times K_c}$, $\Psi \in \mathbb{R}^{L \times K_d}$
auxiliary variables: $\mathbf{c} \in \mathbb{R}^{K_c}$, $\mathbf{d} \in \mathbb{R}_+^{K_d}$

3. product of experts Hinton (2002): This is the analysis model analogue of the linear generative model,

$$p(\mathbf{X}) = \frac{1}{Z_{POE}} \prod_{l=1}^{L} \exp\left(-E_{POE}\left(\Phi_l \mathbf{x}; \lambda_l\right)\right). \tag{36}$$

parameters: $\Phi \in \mathbb{R}^{L \times M}$, $\lambda \in \mathbb{R}_+^L$,
Product of experts models were tested with two experts, as listed:

   (a) Laplace expert:

$$E_{POE}(u; \lambda_l) = \lambda_l |u| \tag{37}$$

   (changing $\lambda_l$ is equivalent to changing the length of the row $\Phi_l$, so it is fixed to $\lambda_l = 1$)

   (b) Student's t expert:

$$E_{POE}(u; \lambda_l) = \lambda_l \log\left(1 + u^2\right) \tag{38}$$

4. Mean and covariance restricted Boltzmann machine (mcRBM) Ranzato & Hinton (2010): This is an analysis model analogue of the bilinear generative model. The exact marginal energy function $E_{mcR}$ is taken from the released code rather than the paper.

$$p(\mathbf{x}) = \frac{\exp\left[-E_{mcR}(\mathbf{x})\right]}{Z_{mcR}} \tag{39}$$

$$\begin{aligned} E_{mcR}(\mathbf{x}) = &- \sum_{k=1}^{K} \log\left[1 + e^{\frac{1}{2}\sum_{l=1}^{L} P_{lk} \frac{(\mathbf{c}_l \mathbf{x})^2}{||\mathbf{x}||_2^2 + \frac{1}{2}} + b_k^c}\right] \\ &- \sum_{j=1}^{J} \log\left[1 + e^{\mathbf{W}_j \mathbf{x} + b_j^m}\right] \\ &+ \frac{1}{2\sigma^2} \mathbf{x}^T \mathbf{x} - \mathbf{x}^T \mathbf{b}^v \end{aligned} \tag{40}$$

parameters: $P \in \mathbb{R}^{L \times K}$, $C \in \mathbb{R}^{L \times M}$, $W \in \mathbb{R}^{J \times M}$, $b^m \in \mathbb{R}^J$, $b^c \in \mathbb{R}^K$, $b^v \in \mathbb{R}^K$, $\sigma \in \mathbb{R}$

# 4 Training

All models were trained on 10,000 $16x16$ pixel image patches taken at random from 4,112 linearized images of natural scenes from the van Hateren dataset van Hateren & van der Schaaf (1998). The extracted image patches were first logged, and then mean subtracted. They were then projected onto the top $M$ PCA components, and whitened by rescaling each dimension to unit norm.

All generative models were trained using Expectation Maximization over the full training set, with a Hamiltonian Monte Carlo algorithm used during the expectation step to maintain samples from the posterior distribution. See Culpepper et al. (2011) for details. All analysis models were trained using LBFGS on the minimum probability flow learning objective function for the full training set, with a transition function $\Gamma$ based on Hamiltonian dynamics. See Sohl-Dickstein et al. (2009) for details. Figures 2, 3 and 4 compare the log likelihood $\mathcal{L}$ of models learned in this fashion (blue dotted line labeled "true") to the maximum likelihood solution (solid black line labeled "oracle"), for 3 models where the true log likelihood and it's gradient can be analytically computed. The discrepancy between the "true" and "oracle" lines in Figure 2 reflects the fact that the complete POE model with Student's t experts is difficult to train by means other than maximum likelihood. In particular, a similar discrepancy also results when contrastive divergence is used. Note that no regularization or decay terms were required on any of the model parameters.

# 5 Results

100 images from the van Hateren dataset were chosen at random and reserved as a test set for evaluation of log likelihood. The test data was preprocessed in an identical fashion to the training data. Unless otherwise noted, log likelihood is estimated on the same set of 100 patches drawn from the test images, using Hamiltonian annealed importance sampling with $N = 100,000$ intermediate distributions, and 200 particles. This procedure takes about 170 seconds for the 36 PCA component analysis models tested below. The generative models take approximately 4 hours, because models with unmarginalized auxiliary variables require one full HAIS run for each test


Figure 1: A subset of the basis functions and filters learned by each model. *(a)* Bases $\Phi$ for the linear generative model with Gaussian prior and *(b)* Laplace prior; *(c)* filters $\Phi$ for the product of experts model with Laplace experts, and *(d)* Student's t experts; *(e)* Bases $\Phi$ for the bilinear generative model and *(f)* the basis elements making up a single grouping from $\Psi$, ordered by and contrast modulated according to the strength of the corresponding $\Psi$ weight (decreasing from left to right); mcRBM *(g)* $C$ filters, *(h)* $W$ means, and *(i)* a single $P$ grouping, showing the pooled filters from $C$, ordered by and contrast modulated according to the strength of the corresponding $P$ weight (decreasing from left to right).

datapoint.

## 5.1 Validating Hamiltonian annealed importance sampling

The log likelihood of the test data can be analytically computed for three of the models outlined above: linear generative with Gaussian prior (Section 3, model 1a), and product of experts with a complete representation ($M = L$) for both Laplace and Student's t experts (Section 3, model 3). Figures 2, 3 and 4 show the convergence of Hamiltonian annealed importance sampling, with 200 particles, for each of these three models as a function of the number $N$ of intermediate distributions. Note that the Student's t expert is a pathological case for sampling based techniques, as for several of the learned $\lambda_l$ even the first moment of the Student's t-distribution was infinite.

Additionally, for all of the generative models, if $\mathbf{\Phi} = \mathbf{0}$

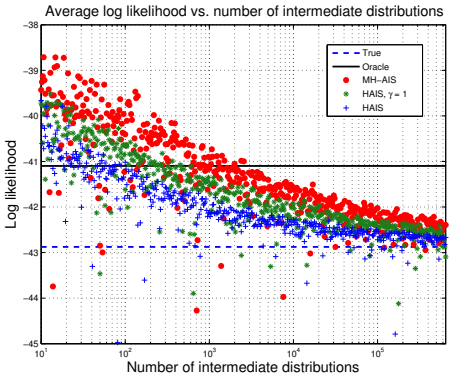


Figure 2: Comparison of HAIS with alternate AIS algorithms in a complete ($M = L = 36$) POE Student's t model. The scatter plot shows estimated log likelihoods of the test data under the POE model for different numbers of intermediate distributions $N$. The blue crosses indicate HAIS. The green stars indicate AIS with a single Hamiltonian dynamics leapfrog step per distribution, but no continuity of momentum. The red dots indicate AIS with a Gaussian proposal distribution. The dashed blue line indicates the true log likelihood of the minimum probability flow trained model. The solid black line indicates the log likelihood of the maximum likelihood model. This product of Student's t distribution is an extremely difficult distribution to normalize numerically, because many of the shape parameters $\lambda$ have adjusted their values to the regime where the distribution has infinite variance.

then the statistical model reduces to,

$$p\left(\mathbf{x}|\mathbf{a}\right) = \frac{\exp\left[-\frac{1}{2\sigma_n^2}\mathbf{x}^T\mathbf{x}\right]}{\left(2\pi\right)^{\frac{M}{2}}\sigma_n^M}, \tag{41}$$

and the log likelihood $\mathcal{L}$ has a simple form that can be used to directly verify the estimate computed via HAIS. We performed this sanity check on all generative models, and found the HAIS estimated log likelihood converged to the true log likelihood in all cases.

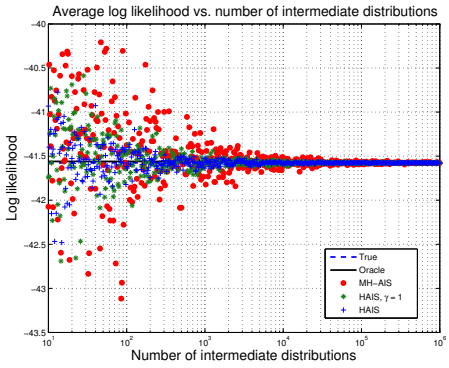


Figure 3: Comparison of HAIS with alternate AIS algorithms in a complete ($M = L = 36$) POE Laplace model. Format as in Figure 2, but for a Laplace expert.

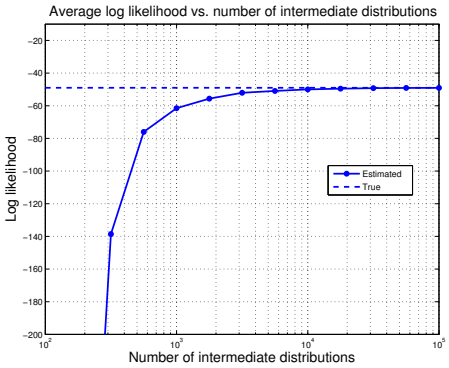


Figure 4: Convergence of HAIS for the linear generative model with a Gaussian prior. The dashed blue line indicates the true log likelihood of the test data under the model. The solid blue line indicates the HAIS estimated log likelihood of the test data for different numbers of intermediate distributions $N$.

## 5.2 Speed of convergence

In order to demonstrate the improved performance of HAIS, we compare against two alternate AIS learning methods. First, we compare to AIS with transition distributions $T_n(\mathbf{x}_{n+1}|\mathbf{x}_n)$ consisting of a Gaussian ($\sigma_{diffusion} = 0.1$) proposal distribution and Metropolis-Hastings rejection rules. Second, we compare to AIS with a single Hamiltonian leapfrog step per intermediate distribution $\pi_n(\mathbf{x}_n)$, and unit norm isotropic Gaussian momentum. Unlike in HAIS however, in this case we randomize the momenta before each update step, rather than allowing them to remain consistent across intermediate transitions. As can be seen in Figures 2 and 3, HAIS requires fewer intermediate distributions by an order of magnitude or more.

## 5.3 Model size

By training models of different sizes and then using HAIS to compute their likelihood, we are able to explore how each model behaves in this regard, and find that three have somewhat different characteristics, shown in Figure 5. The POE model with a Laplace expert has relatively poor performance and we have no evidence that it is able to overfit the training data; in fact, due to the relatively weak sparsity of the Laplace prior, we tend to think the only thing it can learn is oriented, band-pass functions that more finely tile the space of orientation and frequency. In contrast, the Student-t expert model rises quickly to a high level of performance, then overfits dramatically. Surprisingly, the mcRBM performs poorly with a number of auxiliary variables that is comparable to the best performing POE model. One explanation for this is that we are testing it in a regime where the major structures designed into the model are not of great benefit. That is, the mcRBM is primarily good at capturing long range image structures, which are not sufficiently present in our data because we use only 36 PCA components. Although for computational reasons we do not yet have evidence that the mcRBM can overfit our dataset, it likely does have that power. We expect that it will fare better against other models as we scale up to more sizeable images. Finally, we are excited by the superior performance of the bilinear generative model, which outperforms all other models with only a small number of auxiliary variables. We suspect this is mainly due to the high degree of flexibility of the sparse prior, whose parameters (through $\Theta$ and $\Psi$) are learned from the data. The fact that for a comparable number of "hidden units" it outperforms the mcRBM, which can be thought of as the bilinear generative model's

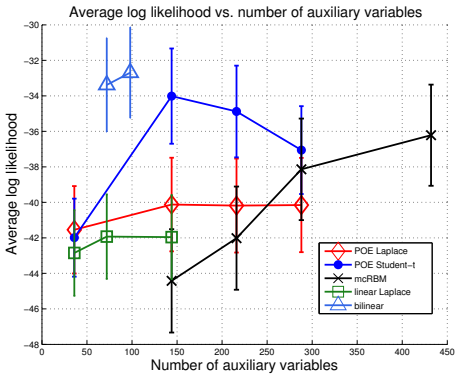


Figure 5: Increasing the number of auxiliary variables in a model increases the likelihood it assigns to the test data until it saturates, or overfits.

'analysis counterpart', highlights the power of this model.

### 5.4 Comparing model classes

As illustrated in Table 1, we used HAIS to compute the log likelihood of the test data under each of the image models in Section 3. The model sizes are indicated in the table – for both POE models and the mcRBM they were chosen from the best performing datapoints in Figure 5. In linear models, the use of sparse priors or experts leads to a large ($> 6\ nat$) increase in the log likelihood over a Gaussian model. The choice of sparse prior was similarly important, with the POE model with Student's t experts performing more than 7 $nats$ better than the POE or generative model with Laplace prior or expert. Although previous work Ranzato & Hinton (2010); Culpepper et al. (2011) has suggested bilinear models outperform their linear counterparts, our experiments show the Student's t POE performing within the noise of the more complex models. One explanation is the relatively small dimensionality (36 PCA components) of the data – the advantage of bilinear models over linear is expected to increase with dimensionality. Another is that Student's t POE models are in fact better than previously believed. Further investigation is underway. The surprising performance of the Student's t POE, however, highlights the power and usefulness of being able to directly compare the log likelihoods of probabilistic models.

Table 1: Average log likelihood for the test data under each of the models. The model 'size' column denotes the number of experts in the POE models, the sum of the mean and covariance units for the mcRBM, and the total number of latent variables in the generative models.

| MODEL | SIZE | AVG. LOG LIKELIHOOD |
|---|---|---|
| LINEAR GENERATIVE, GAUSSIAN | 36 | -49.15± 2.31 |
| LINEAR GENERATIVE, LAPLACE | 36 | -42.85± 2.41 |
| POE, LAPLACE EXPERTS | 144 | -41.54± 2.46 |
| MCRBM | 432 | -36.01± 2.57 |
| POE, STUDENT'S T EXPERTS | 144 | -34.01± 2.68 |
| BILINEAR GENERATIVE | 98 | -32.69± 2.56 |

## 6 Conclusion

By improving upon the available methods for partition function estimation, we have made it possible to directly compare large probabilistic models in terms of the likelihoods they assign to data. This is a fundamental measure of the quality of a model – especially a model trained in terms of log likelihood – and one which is frequently neglected due to practical and computational limitations. It is our hope that the Hamiltonian annealed importance sampling technique presented here will lead to better and more relevant empirical comparisons between models.

## References

Bethge, M. Factorial coding of natural images: how effective are linear models in removing higher-order dependencies? *JOSA A*, Jan 2006.

Chandler, Damon M and Field, David J. Estimates of the information content and dimensionality of natural scenes from proximity distributions. *JOSA A*, Jan 2007.

Culpepper, Benjamin J., Sohl-Dickstein, Jascha, and Olshausen, Bruno. Learning higher-order features of nat-

ural images via factorization. *Redwood Technical Report*, 2011.

Hinton, Geoffrey E. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, Aug 2002. doi: 10.1162/089976602760128018.

Jarzynski, C. Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach. *Physical Review E*, Jan 1997.

Kahn, H and Marshall, A. Methods of reducing sample size in monte carlo computations. *Journal of the Operations Research Society of America*, 1:263–278, Jan 1953.

Karklin, Y. Hierarchical statistical models of computation in the visual cortex. *School of Computer Science, Carnegie Melon University*, Thesis, Jan 2007.

Minka, T. Divergence measures and message passing. *Microsoft Research*, TR-2005-173, Jan 2005.

Moral, Pierre Del, Doucet, Arnaud, and Jasra, Ajay. Sequential monte carlo samplers. *Journal Of The Royal Statistical Society*, 68(3):1–26, Jan 2006.

Murray, Iain and Salakhutdinov, Ruslan. Evaluating probabilities under high-dimensional latent variable models. *Advances in Neural Information Processing Systems*, 21, Jan 2009.

Neal, Radford M. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, Jan 2001.

Neal, Radford M. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, Jan 2010. sections 5.2 and 5.3 for langevin dynamics.

Olshausen, Bruno A and Field, David J. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, Jan 1997.

Ranzato, Marc'Aurelio and Hinton, Geoffrey E. Modeling pixel means and covariances using factorized third-order boltzmann machines. *IEEE Conference on Computer Vision and Pattern Recognition*, Jan 2010.

Salakhutdinov, Ruslan and Murray, Iain. On the quantitative analysis of deep belief networks. *International Conference on Machine Learning*, 25, Jan 2008.

Sohl-Dickstein, Jascha, Battaglino, Peter, and DeWeese, Michael R. Minimum probability flow learning. *arXiv*, cs.LG, Jan 2009. 10 pages, 7 figures.

Stephens, Greg J, Mora, Thierry, Tkacik, Gasper, and Bialek, William. Thermodynamics of natural images. *Arxiv preprint arXiv:0806.2694*, Jan 2008.

van Hateren, J H and van der Schaaf, A. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 265(1394):359–366, Jan 1998.

Zoran, Daniel and Weiss, Yair. The" tree-dependent components" of natural images are edge filters. *Neural and Information Processing Systems*, Jan 2009.