On Autoencoders and Score Matching for Energy Based Models

Kevin Swersky* KSWERSKY@CS.UBC.CA
Marc'Aurelio Ranzato† RANZATO@CS.TORONTO.EDU
David Buchman* DAVIDBUC@CS.UBC.CA
Benjamin M. Marlin* BMARLIN@CS.UBC.CA
Nando de Freitas* NANDO@CS.UBC.CA
*Department of Computer Science, University of British Columbia, Vancouver, BC V6T 1Z4, Canada
†Department of Computer Science, University of Toronto, Toronto, ON M5S 2G4, Canada

Abstract

We consider estimation methods for the class of continuous-data energy based models (EBMs). Our main result shows that estimating the parameters of an EBM using score matching when the conditional distribution over the visible units is Gaussian corresponds to training a particular form of regularized autoencoder. We show how different Gaussian EBMs lead to different autoencoder architectures, providing deep links between these two families of models. We compare the score matching estimator for the mPoT model, a particular Gaussian EBM, to several other training methods on a variety of tasks including image denoising and unsupervised feature extraction. We show that the regularization function induced by score matching leads to superior classification performance relative to a standard autoencoder. We also show that score matching yields classification results that are indistinguishable from better-known stochastic approximation maximum likelihood estimators.

1. Introduction

In this work, we consider a rich class of probabilistic models called *energy based models* (EBMs) (LeCun et al., 2006; Teh et al., 2003; Hinton, 2002). These models define a probability distribution though an exponentiated energy function. Markov Random Fields (MRFs) and Restricted Boltzmann Machines (RBMs) are the most common instance of such models and have a long history in particular application areas including modeling natural images.

Recently, more sophisticated latent variable EBMs for continuous data including the PoT (Welling et al., 2003), mPoT (Ranzato et al., 2010b), mcRBM (Ranzato & Hinton, 2010), FoE (Schmidt et al., 2010) and others have become popular models for learning representations of natural images as well as other sources of real-valued data. Such models, also called *gated MRFs*, leverage latent variables to represent higher order interactions between the input variables. In the very active research area of deep learning (Hinton et al., 2006), these models been employed as elementary building blocks to construct hierarchical models that achieve very promising performance on several perceptual tasks (Ranzato & Hinton, 2010; Bengio, 2009).

Maximum likelihood estimation is the default parameter estimation approach for probabilistic models due to its optimal theoretical properties. Unfortunately, maximum likelihood estimation is computationally infeasible in many EBM models due to the presence of an intractable normalization term (the partition function) in the model probability. This term arises in EBMs because the exponentiated energies do not automatically integrate to unity, unlike directed models parameterized by products of locally normalized conditional distributions (Bayesian networks). Several alternative methods have been proposed to estimate the parameters of an EBM without the need for computing the partition function. One particularly interesting method is called score matching (SM) (Hyvärinen, 2005). The score matching objective function is constructed from an L2 loss on the difference between the derivatives of the log of the model and empirical distribution functions *with respect to the inputs*. Hyvärinen (2005) showed that this results in a cancellation of the

Appearing in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

partition function. Further manipulation yields an estimator that can be computed analytically and is provably consistent.

Autoencoder neural networks are another class of models that are often used to model high-dimensional real-valued data (Hinton & Zemel, 1994; Vincent et al., 2008; Vincent, 2011; Kingma & LeCun, 2010). Both EBMs and autoencoders are unsupervised models that can be thought of as learning to re-represent input data in a latent space. In contrast to probabilistic EBMs, autoencoders are deterministic and feed-forward. As a result, autoencoders can be trained to reconstruct their input through one or more hidden layers, they have fast feed-forward inference for hidden layer states, and all common training losses lead to computationally tractable model estimation methods. In order to learn better representations, autoencoders are often modified by tying the weights between the input and output layers to reduce the number of parameters, including additional terms in the objective to bias learning toward sparse hidden unit activations, and adding noise to input data to increase robustness (Vincent et al., 2008; Vincent, 2011). Interestingly, Vincent (2011) showed that a particular kind of denoising autoencoder trained to minimize an L2 reconstruction error can be interpreted as Gaussian RBM trained using Hyvärinen's score matching estimator.

In this paper, we apply score matching to a number of latent variable EBMs where the conditional distribution of the visible units given the hidden units is Gaussian. We show that the resulting estimation algorithms can be interpreted as minimizing a regularized L2 reconstruction error on the visible units. For Gaussian-binary RBMs, the reconstruction term corresponds to a standard autoencoder with tied weights. For the mPoT and mcRBM models, the reconstruction terms correspond to new autoencoder architectures that take into account the covariance structure of the inputs. This suggests a new way to derive novel autoencoder training criteria by applying score matching to the free energy of an EBM. We further generalize score matching to arbitrary EBMs with real-valued input units and show that this view leads to an intuitive interpretation for the regularization terms that appear in the score matching objective function.

2. Score Matching for Latent Energy Based Models

A latent variable energy based model defines a probability distribution over real valued data vectors $v \in \mathcal{V} \subseteq \mathbb{R}^{n_v}$ as follows:

$$P(v, h; \theta) = \frac{\exp(-E_\theta(v, h))}{Z(\theta)}, \tag{1}$$

where $h \in \mathcal{H} \subseteq \mathbb{R}^{n_h}$ are the latent variables, $E_\theta(v, h)$ is an *energy* function parameterized by $\theta \in \Theta$, and $Z(\theta)$ is the partition function. We refer to these models as latent energy based models. This general latent energy based model subsumes many specific models for real-valued data such as Boltzmann machines, exponential-family harmoniums (Welling et al., 2005), factored RBMs and Product of Student's T (PoT) models (Memisevic & Hinton, 2009; Ranzato & Hinton, 2010; Ranzato et al., 2010a;b).

The marginal distribution in terms of the *free energy* $F_\theta(v)$ is obtained by integrating out the hidden variables as seen below. Typically, but not always, this marginalization can be carried out analytically.

$$P(v; \theta) = \frac{\exp(-F_\theta(v))}{Z(\theta)}. \tag{2}$$

Maximum likelihood parameter estimation is difficult when $Z(\theta)$ is intractable. In EBMs the intractability of $Z(\theta)$ arises due to the fact that it is a very high-dimensional integral that often lacks a closed form solution. In such cases, stochastic algorithms can be applied to approximately maximize the likelihood and a variety of algorithms have been described and evaluated (Swersky et al., 2010; Marlin et al., 2010) in the literature including contrastive divergence (CD) (Hinton, 2002), persistent contrastive divergence (PCD) (Younes, 1989; Tieleman, 2008), and fast persistent contrastive divergence (FPCD) (Tieleman & Hinton, 2009). However, these methods often require very careful hand-tuning of optimization-related parameters like step size, momentum, batch size and weight decay, which is complicated by the fact that the objective function can not be computed.

The *score matching* estimator was proposed by Hyvärinen (2005) to overcome the intractability of $Z(\theta)$ when dealing with continuous data. The score matching objective function is defined through a *score function* applied to the empirical $\widetilde{p}(v)$ and model $p_\theta(v)$ distributions. The score function for a generic distribution $p(v)$ is given by $\psi_i(p(v)) = \frac{\partial \log p(v)}{\partial v_i} = -\frac{\partial F_\theta(v)}{\partial v_i} = \int_h -\frac{\partial E_\theta(v,h)}{\partial v_i} p_\theta(h|v) dh$. The full objective function is given below.

$$J(\theta) = \mathbb{E}_{\widetilde{p}(v)} \left[\sum_{i=1}^{n_v} \left(\psi_i(\widetilde{p}(v)) - \psi_i(p_\theta(v)) \right)^2 \right]. \tag{3}$$

The benefit of optimizing $J(\theta)$ is that $Z(\theta)$ cancels off in the derivative of $\log p_\theta(v)$ since it is constant with respect to each v_i. However, in the above form, $J(\theta)$ is still intractable due to the dependence on $\widetilde{p}(v)$. Hyvärinen, shows that under weak regularity conditions $J(\theta)$ can be expressed in the following form, which can be tractably approximated by replacing the expectation over the empirical distribution by an empirical average over the training set:

$$J(\theta) = \mathbb{E}_{\widetilde{p}(v)}\left[\sum_{i=1}^{n_v}\frac{1}{2}\left(\psi_i(p_\theta(v))\right)^2 + \frac{\partial\psi_i(p_\theta(v))}{\partial v_i}\right]. \quad (4)$$

In theoretical situations where the regularity conditions on the derivatives of the empirical distribution are not satisfied, or in practical situations where a finite sample approximation to the expectation over the empirical distribution is used, a smoothed version of the score matching estimator may be of interest. Consider smoothing $\widetilde{p}(v)$ using a probabilistic kernel $q_\beta(v|v')$ with bandwidth parameter $\beta > 0$. We obtain a new distribution $q_\beta(v) = \int q_\beta(v|v')\widetilde{p}(v')dv'$. Vincent (2011) showed that applying score matching to $q_\beta(v)$ is equivalent to the following objective function where $q_\beta(v, v') = q_\beta(v|v')\widetilde{p}(v')$:

$$Q(\theta) = \mathbb{E}_{q_\beta(v,v')}\left[\sum_{i=1}^{n_v}\left(\psi_i(q_\beta(v|v')) - \psi_i(p_\theta(v))\right)^2\right]. \quad (5)$$

For the case where $q_\beta(v|v') = \mathcal{N}\left(v|v', \beta^2\right)$ i.e. a Gaussian smoothing kernel with variance β^2, this is equivalent to the regularized score matching objective proposed in (Kingma & LeCun, 2010). We refer to the objective given by Equation 5 as denoising score matching (SMD). Although SMD is intractable to evaluate analytically, we can again replace the integral over v' by an empirical average over a finite sample of training data. We can then replace the integral over v by an empirical average over samples v, which can be easily drawn from $q_\beta(v|v')$ for each training sample v'.

Compared to PCD and CD, SM and SMD give tractable objective functions that can be used to monitor training progress. While SMD is not consistent, it does have significant computational advantages relative to SM (Vincent, 2011).

3. Applying and Interpreting Score Matching For Latent EBMs

We now derive score matching objectives for several commonly used EBMs. In order to apply score matching to a particular EBM, one simply needs an expression for the corresponding free energy.

Example 1 ***Score Matching for Gaussian-binary RBMs:*** *Here, the energy $E_\theta(v, h)$ is given by:*

$$-\sum_{i=1}^{n_v}\sum_{j=1}^{n_h}\frac{v_i}{\sigma_i}W_{ij}h_j - \sum_{j=1}^{n_h}b_j h_j + \frac{1}{2}\sum_{i=1}^{n_v}\frac{(c_i - v_i)^2}{\sigma_i^2}, \quad (6)$$

where the parameters are $\theta = (W, \sigma, b, c)$ and $h_j \in \{0, 1\}$. This leads to the free energy $F_\theta(v)$:

$$\frac{1}{2}\sum_{i=1}^{n_v}\frac{(c_i - v_i)^2}{\sigma_i^2} - \sum_{j=1}^{n_h}\log\left(1 + \exp\left(\sum_{i=1}^{n_v}\frac{v_i}{\sigma_i}W_{ij} + b_j\right)\right), \quad (7)$$

The corresponding score matching objective is:

$$J(\theta) = \frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{n_v}\left[\frac{1}{2}\left(\frac{v_{in}}{\sigma_i^2} - \frac{c_i}{\sigma_i^2} - \sum_{j=1}^{n_h}\frac{W_{ij}}{\sigma_i}\hat{h}_{jn}\right)^2 - \frac{1}{\sigma_i^2} + \sum_{j=1}^{n_h}\frac{W_{ij}^2}{\sigma_i^2}\hat{h}_{jn}\left(1 - \hat{h}_{jn}\right)\right], \quad (8)$$

where $\hat{h}_{jn} := sigm\left(\sum_{i=1}^{n_v}\frac{v_{in}}{\sigma_i}W_{ij} + b_j\right)$ and $sigm(x) := \frac{1}{1+\exp(-x)}$.

For a standardized Normal model, with $c = 0$ and $\sigma = 1$, this objective reduces to:

$$J(\theta) = \frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{n_v}\left[\frac{1}{2}\left(v_{in} - \sum_{j=1}^{n_h}W_{ij}\hat{h}_{jn}\right)^2 - 1 + \sum_{j=1}^{n_h}W_{ij}^2\hat{h}_{jn}\left(1 - \hat{h}_{jn}\right)\right], \quad (9)$$

The first term corresponds to the quadratic reconstruction error of an autoencoder with tied weights. From this we can see that this type of of autoencoder, which researchers have previously treated as a different model, can in fact be explained by the application of the score matching estimation principle to Gaussian RBMs.

Example 2 ***Score matching for mcRBM:*** *The energy $E_\theta(v, h^m, h^c)$ of the mcRBM model for each data point includes mean Bernoulli hidden units $h_j^m \in \{0, 1\}$ and covariance Bernoulli hidden units $h_k^c \in \{0, 1\}$. The latter allow one to model correlations in the data v (Ranzato & Hinton, 2010; Ranzato et al., 2010a). To ease the notation, we will ignore the index* *n over the data. The energy for this model is:*

$$\begin{aligned}&-\frac{1}{2}\sum_{f=1}^{n_f}\sum_{k=1}^{n_{hc}} P_{fk}h^c_k(\sum_{i=1}^{n_v} C_{if}v_i)^2 - \sum_{j=1}^{n_{hm}}\sum_{i=1}^{n_v} W_{ij}h^m_j v_i\\ &- \sum_{j=1}^{n_{hm}} b^m_j h^m_j - \sum_{k=1}^{n_{hc}} b^c_k h^c_k - \sum_{i=1}^{n_v} b^v_i v_i + \frac{1}{2}\sum_{i=1}^{n_v} v_i^2,\end{aligned} \tag{10}$$

where $\theta = (b^v, b^m, b^c, P, W, C)$. This leads to the free energy $F_\theta(v)$:

$$\begin{aligned}&-\sum_{k=1}^{n_{hc}} \log(1 + e^{\phi^c_k})\\ &-\sum_{j=1}^{n_{hm}} \log(1 + e^{\phi^m_j}) - \sum_{i=1}^{n_v} b^v_i v_i + \frac{1}{2}\sum_{i=1}^{n_v} v_i^2,\end{aligned} \tag{11}$$

where $\phi^c_k = \frac{1}{2}\sum_{f=1}^{n_f} P_{fk}(\sum_{i=1}^{n_v} C_{if}v_i)^2 + b^c_k$ and $\phi^m_j = \sum_{i=1}^{n_v} W_{ij}v_i + b^m_j$. The corresponding score matching objective is:

$$\begin{aligned}J(\theta) =& \Bigg[\sum_{i=1}^{n_v} \frac{1}{2}\psi_i(p_\theta(v))^2\\ &+ \sum_{k=1}^{n_{hc}} \Big(\rho(\hat{h}^c_k)D^2_{ik} + \hat{h}^c_k K_{ik}\Big)\\ &+ \sum_{j=1}^{n_{hm}} \Big(\hat{h}^m_j(1-\hat{h}^m_j)W^2_{ij}\Big) - 1\Bigg] \qquad (12)\\ \psi_i(p_\theta(v)) =& \sum_{k=1}^{n_{hc}} \hat{h}^c_k D_{ik} + \sum_{j=1}^{n_{hm}} \hat{h}^m_j W_{ij} + b^v_i - v_i\\ K_{ik} =& \sum_{f=1}^{n_f} P_{fk}C^2_{if}\\ D_{ik} =& \sum_{f=1}^{n_f} \left(P_{fk}\left(\sum_{i'=1}^{n_v} C_{i'f}v_{i'}\right)C_{if}\right)\\ \hat{h}^c_k =& sigm\left(\phi^c_k\right)\\ \hat{h}^m_j =& sigm\left(\phi^m_j\right)\\ \rho(x) :=& x(1-x).\end{aligned}$$

Example 3 ***Score matching for mPoT*** *The energy $E_\theta(v, h^m, h^c)$ of the mPoT model is:*

$$\begin{aligned}&\sum_{k=1}^{n_{hc}}\left[h^c_k(1 + \frac{1}{2}(\sum_{i=1}^{n_v} C_{ik}v_i)^2) + (1-\gamma)\log(h^c_k)\right]\\ &+ \frac{1}{2}\sum_{i=1}^{n_v} v_i^2 - \sum_i b^v_i v_i - \sum_{i=1}^{n_v}\sum_{j=1}^{n_{hm}} h^m_j W_{ij} v_i - \sum_{j=1}^{n_{hm}} b^m_j h^m_j,\end{aligned} \tag{13}$$

where $\theta = (\gamma, W, C, b^v, b^m)$ and h^c is a vector of Gamma covariance latent variables, C is a filter bank and γ is a scalar parameter. This leads to the free energy $F_\theta(v)$:

$$\begin{aligned}&\sum_{k=1}^{n_{hc}} \gamma \log(1 + \frac{1}{2}(\phi^c_k)^2)\\ &-\sum_{j=1}^{n_{hm}} \log(1 + e^{\phi^m_j}) - \sum_{i=1}^{n_v} b^v_i v_i + \frac{1}{2}\sum_{i=1}^{n_v} v_i^2,\end{aligned} \tag{14}$$

where $\phi^c_k = \sum_{i=1}^{n_v} C_{ik}v_i$ and $\phi^m_j = \sum_{i=1}^{n_v} W_{ij}v_i + b^m_j$. The corresponding score matching objective $J(\theta)$ is equivalent to the objective given in Equation 12 with the following redefinition of terms:

$$\begin{aligned}P =& -I_{n_{hc}}\\ \hat{h}^c_k =& \gamma\varphi(\phi^c_k) \qquad (15)\\ \hat{h}^m_j =& sigm(\phi^m_j) \qquad (16)\\ \varphi(x) :=& \frac{1}{1 + \frac{1}{2}(x)^2}\\ \rho(x) :=& x^2,\end{aligned}$$

where $I_{n_{hc}}$ is the $n_{hc} \times n_{hc}$ identity matrix.

In each of these examples, we see that an objective emerges which seeks to minimize a form of regularized reconstruction error, and that the forms of these regularizers can end up being quite different. Rather than trying to interpret score matching on a case by case basis, we provide a general theorem for all latent EBMs on which score matching can be applied:

Theorem 1 *The score matching objective, Equation (4), for a latent energy based model can be expressed succinctly in terms of either the free energy or expectations of the energy with respect to the conditional distribution $p(h|v)$. Specifically,*

$$\begin{aligned}J(\theta) =& \mathbb{E}_{\widetilde{p}(v)}\left[\sum_{i=1}^{n_v}\frac{1}{2}\left(\psi_i(p_\theta(v))\right)^2 + \frac{\partial\psi_i(p_\theta(v))}{\partial v_i}\right]\\ =& \mathbb{E}_{\widetilde{p}(v)}\Bigg[\sum_{i=1}^{n_v}\frac{1}{2}\left(\mathbb{E}_{p_\theta(h|v)}\left[\frac{\partial E_\theta(v,h)}{\partial v_i}\right]\right)^2\\ &+ var_{p_\theta(h|v)}\left[\frac{\partial E_\theta(v,h)}{\partial v_i}\right] - \mathbb{E}_{p_\theta(h|v)}\left[\frac{\partial^2 E_\theta(v,h)}{\partial v_i^2}\right]\Bigg].\end{aligned}$$

Corollary 1 *If the energy function of a latent EBM $E_\theta(v,h)$ takes the following form:*

$$E_\theta(v,h) = \frac{1}{2}(v - \mu(h))^T\Omega(h)(v - \mu(h)) + g(h),$$

where $\mu(h)$ is an arbitrary vector-valued function of length n_v, $g(h)$ is an arbitrary scalar function, and

$\Omega(h)$ is an $n_v \times n_v$ positive-definite matrix-valued function, then the vector-valued score function $\psi(p_\theta(v))$ will be: $\mathbb{E}_{p_\theta(h|v)}\left[\Omega(h)(v-\mu(h))\right]$. As a result, the score matching objective can be expressed as:

$$
\begin{aligned}
J(\theta) =& \mathbb{E}_{\widetilde{p}(v)}\Bigg[\sum_{i=1}^{n_v} \frac{1}{2}\left(\mathbb{E}_{p_\theta(h|v)}\left[\Omega(h)(v-\mu(h))\right]_i\right)^2 \\
& + var_{p_\theta(h|v)}\left[\Omega(h)(v-\mu(h))\right]_i \\
& - \mathbb{E}_{p_\theta(h|v)}\left[\Omega(h)\right]_{ii}\Bigg].
\end{aligned}
$$

The proofs of Theorem 1 and Corollary 1 are straightforward, and can be found in an online appendix to this paper.[1] Corollary 1 states that score matching applied to a Gaussian latent EBM will always result in a quadratic reconstruction term with penalties to minimize the variance of the reconstruction and to maximize the expected curvature of the energy with respect to v. This shows that we can develop new autoencoder architectures in a principled way by simply starting with an EBM and applying score matching.

One further connection between the two models is that one step of gradient descent on the free energy $F_\theta(v)$ of an EBM corresponds to one feed-forward step of an autoencoder. To see this, consider the mPoT model. If we start at some visible configuration v and update a single dimension i:

$$
\begin{aligned}
v_i^{(t+1)} &= v_i^{(t)} - \eta \frac{\partial F_\theta(v)}{\partial v_i} \\
&= v_i^{(t)} + \eta \left(\sum_{k=1}^{n_{hc}} \hat{h}_k^c D_{ik} \right. \\
&\left. + \sum_{j=1}^{n_{hm}} \hat{h}_j^m W_{ij} + b_i^v - v_i^{(t)} \right).
\end{aligned}
$$

Then setting $\eta = 1$, the $v_i^{(t)}$ terms cancel and we get:

$$
v_i^{(t+1)} = \sum_{k=1}^{n_{hc}} \hat{h}_k^c D_{ik} + \sum_{j=1}^{n_{hm}} \hat{h}_j^m W_{ij} + b_i^v. \qquad (17)
$$

This corresponds to the reconstruction produced by mPoT in its score matching objective. In general, an autoencoder reconstruction can be produced by taking a single step of gradient descent along the free energy of its corresponding EBM.

4. Experiments

In this section, we study several estimation methods applied to the mPoT model including SM, SMD, CD, PCD, and FPCD with the goal of uncovering differences in the characteristics of trained models due to variations in training methods. For our experiments, we used two datasets of images.

The first dataset consists of 128,000 color image patches of size 16x16 pixels randomly extracted from the Berkeley segmentation dataset[2]. We subtracted the per-patch means and applied PCA whitening. We retained 99% of the variance, corresponding to 105 eigenvectors. All estimation methods were applied to the mPoT model by training on mini-batches of size 128 for 100 epochs of stochastic gradient descent.

The second dataset, named CIFAR 10 (Krizhevsky, 2009), consists of color images of size 32x32 pixels belonging to one of 10 categories. The task is to classify a set of 10,000 test images. CIFAR 10 is a subset of a larger dataset of tiny images (Torralba et al., 2008). Using a protocol established in previous work (Krizhevsky, 2009; Ranzato & Hinton, 2010) we built a training dataset of 8x8 color image patches from this larger dataset, ensuring there was no overlap with CIFAR 10. The preprocessing of the data is exactly the same as for the Berkeley dataset, but here we use approximately 800,000 image patches and perform only 10 epochs of training. For our experiments, we used the Theano package[3], and mPoT[4] code from (Ranzato et al., 2010b).

4.1. Objective Function Analysis

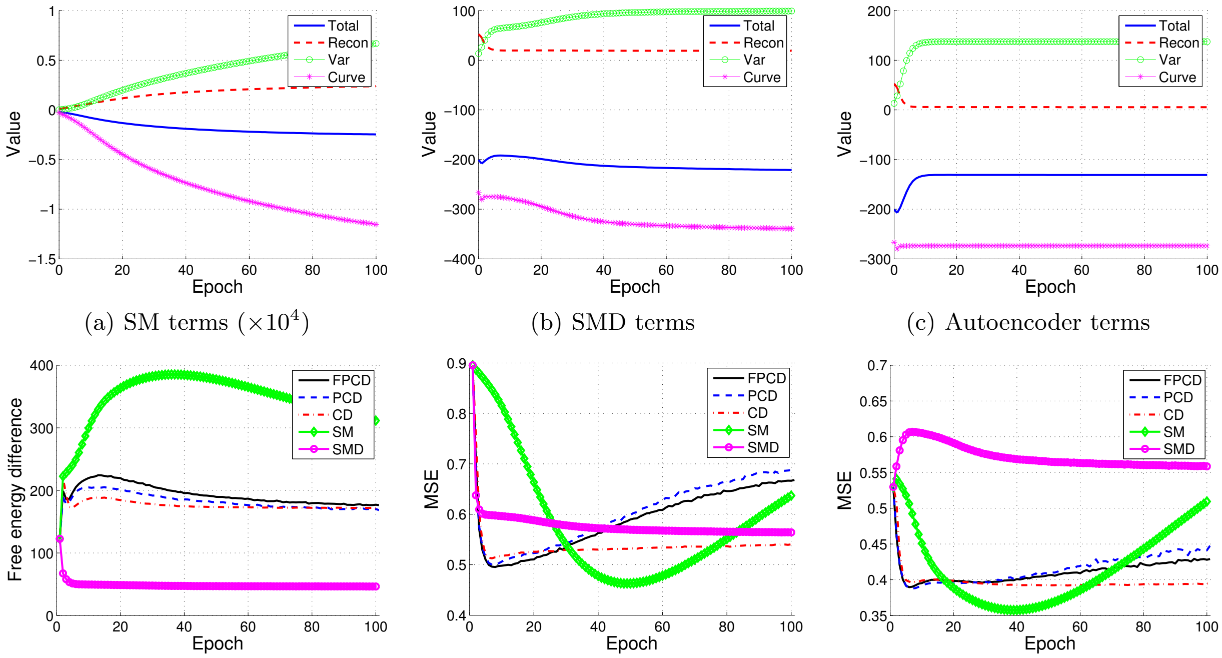
From Corollary 1, we know that we can interpret score matching for mPoT as trading off reconstruction error, reconstruction variance and the expected curvature of the energy function with respect to the visible units. This experiment, using the Berkeley dataset, is designed to determine how these terms evolve over the course of training and to what degree their changes impact the final model. Figures 1(a) and 1(b) show the values of the three terms using non-noisy inputs on each training epoch, as well as the overall objective function (the sum of the 3 terms). Surprisingly, these results show that most of the training is involved with maximizing the expected curvature (corresponding to a lower negative curvature). In SM, each point

[1] `http://www.cs.ubc.ca/~nando/papers/smpaper-appendix.pdf`

[2] `http://www.cs.berkeley.edu/projects/vision/grouping/segbench/`

[3] `http://deeplearning.net/software/theano/`

[4] `http://www.cs.toronto.edu/~ranzato/publications/mPoT/mPoT.html`



(a) SM terms ($\times 10^4$)

(b) SMD terms

(c) Autoencoder terms

(d) Free energy difference

(e) Mean-field denoising

(f) Bayesian denoising

Figure 1. (a), (b), (c) Expected reconstruction error, reconstruction variance, and energy curvature for SM, SMD, and AE. Total represents the sum of these terms. (d) Difference of free energy between noisy and test images. (e) MSE of denoised test images using mean-field. (f) MSE of denoised test images using Bayesian MAP.

is relatively isolated in v-space meaning that the objective will try to make the distribution very peaked. In SMD, each point exists near a cloud of points and so the distribution must be broader. From this perspective, SMD can be seen as a regularized version of SM that puts less emphasis on changing the expected curvature. This also seems to give SMD some room to reduce the reconstruction error.

To examine the impact of regularization, we trained an autoencoder (AE) based on the mPoT model using the reconstruction given by Equation 17, which corresponds to SM without the variance and curvature terms. Figure 1(c) shows that simply optimizing the reconstruction leaves the curvature almost invariant, which agrees with the findings of (Ranzato et al., 2007).

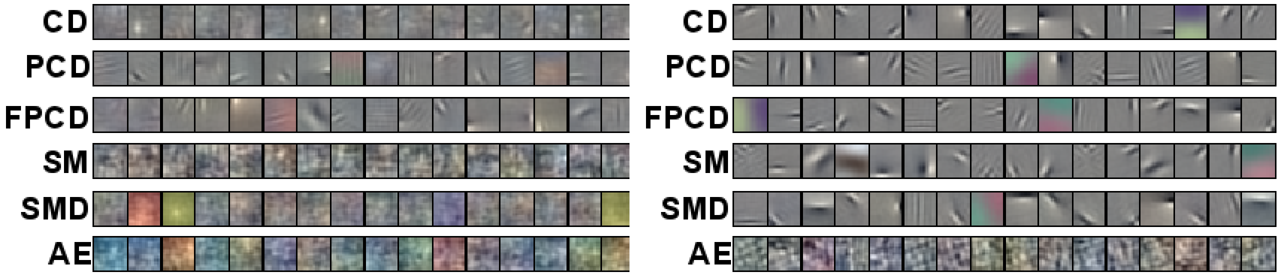
4.2. Denoising

In our next set of experiments, we compare models learned by each of the score matching estimators with models learned by the more commonly used stochastic estimators. For these experiments, we trained mPoT models corresponding to SM, SMD, FPCD, PCD, and CD. We compare the models in terms of the average free energy difference between natural image patches and patches corrupted by Gaussian noise. We also consider denoising natural image patches.[5]

During training, we hope that the probability of natural images will increase while that of other images decreases. The free energy difference between natural and other images is equivalent to the log of their probability ratio, so we expect the free energy difference to increase during training as well. Figure 1(d) shows the difference in free energy between a test set of 10,000 image patches from the Berkeley dataset, and the energy of the same images corrupted by noise. For most estimators, the free energy difference improves as training proceeds, as expected. Interestingly, SM and SMD exhibit completely opposite behaviors. SM seems to significantly increase the free energy difference relative to nearby noisy images, corresponding to a distribution that is peaked around natural images. SMD, on the other hand, actually *decreases* the free energy difference relative to nearby noisy images.

In the next experiment, we consider an image denoising task. We take an image patch v and add Gaussian white noise, obtaining a noisy patch v'. We then ap-

[5] Note that for convenience, both tasks were performed in the PCA domain. We use a standard deviation of 1 for the Gaussian noise in all cases.



(a) Mean filters

(b) Covariance filters

Figure 2. mPoT filters learned using different estimation methods: (a) "mean" filters, (b) "covariance" filters.

ply each model to denoise each patch v', obtaining a reconstruction $\hat{v}$. The first denoising method, shown in Figure 1(e), computes a reconstruction $\hat{v}$ by simulating one step of a Markov chain using a mean-field approximation. That is, we first compute h_k^c and h_j^m by Equations 15 and 16 using v' as the input. The reconstruction is the expectation of the conditional distribution $P_\theta(v|h_k^c, h_j^m)$. The second method, shown in Figure 1(f), is the Bayesian MAP estimator:

$$\hat{v} = \arg\min_v F_\theta(v) + \frac{\lambda}{2} \|v - v'\|^2 , \qquad (18)$$

where λ is a scalar representing how close the reconstruction should remain to the noisy input. We select λ by cross-validation. The results show that score matching achieves the minimum error using both denoising approaches, however it quickly overfits as training proceeds. FPCD and PCD do not match the minimum error of SM and also overfit, albeit to a lesser extent. CD and SMD do not appear to overfit. However, we note that the minimum error obtained by SMD is significantly higher than the minimum error obtained by SM using both denoising methods. This is quite intuitive since SMD is equivalent to estimating the model using a smoothed training distribution that shifts mass onto nearby noisy images.

4.3. Feature Extraction and Classification

One of the primary uses for latent EBMs is to generate discriminative features. Table 1 shows the result of using each method to extract features on the benchmark CIFAR 10 dataset. We follow the protocol of (Ranzato & Hinton, 2010) with early stopping. We use a validation set to select regularization parameters. With the exception of AE, all methods appear to do well and the differences between them are not statistically significant. AE, on the other hand, does significantly worse.

Table 1. Recognition accuracy on CIFAR 10.

CD	PCD	FPCD	SM	SMD	AE
64.6%	64.7%	65.5%	65.0%	64.7%	57.6%

Finally, we show examples of filters learned by each method. Figure 2(a) shows a random subset of "mean" filters corresponding to the columns of W, while Figure 2(b) shows a random subset of "covariance" filters corresponding to the columns of C. Interestingly, only FPCD and PCD show structure in the learned mean filters. In the covariance units, all methods except AE learn localized Gabor-like filters. It is well known that obtaining nice looking filters will usually correlate with good performance, but it is not always clear what leads to these filters.

We have shown here that one way to obtain good qualitative and quantitative performance is to focus on appropriately modeling the curvature of the energy with respect to v. In this context, the SM reconstruction and variance terms serve to ensure that the peaks of the distribution occur around the training cases.

5. Conclusion

By applying score matching to the energy space of a latent EBM, as opposed to the free energy space, we gain an intuitive interpretation of the score matching objective. We can always break the objective down into three terms corresponding to expectations under the conditional distribution of the hidden units: reconstruction, reconstruction variance, and curvature. We have determined that for the Gaussian-binary RBM, the reconstruction term will always correspond to an autoencoder with tied weights. While autoencoders and RBMs were previously considered to be related, but separate models, this analysis shows that they can be interpreted as different estimators applied to the same underlying model. We also showed that one can derive novel autoencoders by applying score matching to more complex EBMs. This allows us to think about models in terms of EBMs before creating a corresponding autoencoder to leverage fast inference. Furthermore, this framework provides guidance on selecting principled regularization functions for autoencoder training, leading to improved representations.

Our experiments show that not only does score matching yield similar performance to existing estimation methods when applied to classification, but that shaping the curvature of the energy appropriately may be important for generating good features. While this seems obvious for probabilistic EBMs, it has previously been difficult to apply to autoencoders because they were not thought of as having a corresponding energy function. Now that we know which statistics may be important to monitor during training, it would be interesting to see what happens when other heuristics, such as sparsity, are applied to help generate interpretable features.

References

Bengio, Y. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127, 2009.

Hinton, G.E. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14:1771–1800, 2002.

Hinton, G.E. and Zemel, R.S. Autoencoders, minimum description length and Helmholtz free energy. In *Advances in Neural Information Processing Systems*, pp. 3–10, 1994.

Hinton, G.E., Osindero, S., and Teh, Y.W. A fast learning algorithm for deep belief nets. *Neural Computation*, 18 (7):1527–1554, 2006.

Hyvärinen, A. Estimation of non-normalized statistical models using score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

Kingma, D. and LeCun, Y. Regularized estimation of image statistics by score matching. In *Advances in Neural Information Processing Systems*, 2010.

Krizhevsky, A. Learning multiple layers of features from tiny images, 2009. MSc Thesis, Dept. of Comp. Science, Univ. of Toronto.

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., and Huang, F.J. A tutorial on energy-based learning. In *Predicting Structured Data*. MIT Press, 2006.

Marlin, B.M., Swersky, K., Chen, B., and de Freitas, N. Inductive principles for restricted Boltzmann machine learning. In *Artificial Intelligence and Statistics*, pp. 509–516, 2010.

Memisevic, R. and Hinton, G.E. Learning to represent spatial transformations with factored higher-order Boltzmann machines. *Neural Computation*, 22:1473–1492, 2009.

Ranzato, M. and Hinton, G.E. Modeling pixel means and covariances using factorized third-order Boltzmann machines. In *IEEE Computer Vision and Pattern Recognition*, pp. 2551–2558, 2010.

Ranzato, M., Boureau, Y.L., Chopra, S., and LeCun, Y. A unified energy-based framework for unsupervised learning. In *Artificial Intelligence and Statistics*, 2007.

Ranzato, M., Krizhevsky, A., and Hinton, G.E. Factored 3-way restricted Boltzmann machines for modeling natural images. In *Artificial Intelligence and Statistics*, pp. 621–628, 2010a.

Ranzato, M., Mnih, V., and Hinton, G.E. How to generate realistic images using gated MRF's. In *Advances in Neural Information Processing Systems*, pp. 2002–2010, 2010b.

Schmidt, U., Gao, Q., and Roth, S. A generative perspective on MRFs in low-level vision. In *IEEE Computer Vision and Pattern Recognition*, 2010.

Swersky, K., Chen, B., Marlin, B.M., and de Freitas, N. A tutorial on stochastic approximation algorithms for training restricted Boltzmann machines and deep belief nets. In *Information Theory and Applications Workshop*, pp. 1 –10, 2010.

Teh, Y.W., Welling, M., Osindero, S., and Hinton, G.E. Energy-based models for sparse overcomplete representations. *Journal of Machine Learning Research*, 4:1235–1260, 2003.

Tieleman, T. Training restricted Boltzmann machines using approximations to the likelihood gradient. In *International Conference on Machine Learning*, pp. 1064–1071, 2008.

Tieleman, T. and Hinton, G.E. Using fast weights to improve persistent contrastive divergence. In *International Conference on Machine Learning*, 2009.

Torralba, A., Fergus, R., and Freeman, W.T. 80 million tiny images: A large dataset for non-parametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30:1958–1970, 2008.

Vincent, P. A connection between score matching and denoising autoencoders. *Neural Computation*, To appear, 2011.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.A. Extracting and composing robust features with denoising autoencoders. In *International Conference on Machine Learning*, pp. 1096–1103, 2008.

Welling, M., Hinton, G.E., and Osindero, S. Learning sparse topographic representations with products of student-t distributions. In *Advances in Neural Information Processing Systems*, 2003.

Welling, M., Rosen-Zvi, M., and Hinton, G.E. Exponential family harmoniums with an application to information retrieval. In *Advances in Neural Information Processing Systems*, 2005.

Younes, L. Parametric inference for imperfectly observed Gibbsian fields. *Probability Theory and Related Fields*, 82(4):625–645, 1989.