A Connection Between Score Matching and Denoising Autoencoders

Pascal Vincent
vincentp@iro.umontreal.ca
Dept. IRO, Université de Montréal,
CP 6128, Succ. Centre-Ville, Montréal (QC) H3C 3J7, Canada.

Département d'Informatique et de Recherche Opérationnelle
December 2010

Keywords: autoencoder, energy based models, score matching, denoising, density estimation.

Abstract

Denoising autoencoders have been previously shown to be competitive alternatives to RBMs for unsupervised pre-training of each layer of a deep architecture. We show that a simple denoising autoencoder training criterion is equiv- alent to matching the score (with respect to the data) of a specific energy based model to that of a non-parametric Parzen density estimator of the data. This yields several useful insights. It defines a proper probabilistic model for the denoising autoencoder technique which makes it in principle possible to sample from them or to rank examples by their energy. It suggests a different way to apply score matching that is related to learning to denoise and does not require computing second derivatives. It justifies the use of tied weights between the encoder and decoder, and suggests ways to extend the success of DAEs to a larger family of EBMs.

1 Introduction

This note uncovers an unsuspected link between the *score matching* technique (Hyvärinen, 2005; Hyvärinen, 2008) for learning the parameters of unnormalized density models over continuous-valued data, and the training of *denoising autoencoders* (Vincent *et al.*, 2008, 2010).

Score matching (SM) is an alternative to the maximum likelihood principle suitable for unnormalized probability density models whose partition function is intractable. Its relationship to maximum likelihood has been investigated by Lyu (2009) who formally relates the Fisher divergence that yields score matching and the Kullback-Leibler divergence that yields maximum likelihood. Interestingly, his formal analysis indicates that score matching searches for parameters that are more robust to small-noise perturbations of the training data (Lyu, 2009). Score matching has also been recast as a special case under the more general frameworks of *generalized score matching* (Lyu, 2009; Marlin *et al.*, 2009) and *minimum probability flow* (Sohl-Dickstein *et al.*, 2009), allowing generalizations of score matching to discrete distributions (Hyvärinen, 2007b; Lyu, 2009; Sohl-Dickstein *et al.*, 2009). The *minimum probability flow* paradigm is particularly interesting as it unifies several recent alternative parameter-estimation methods, both for continuous and discrete data, under a single unified view[1]. Recently, Kingma and LeCun (2010) investigated a *regularized* form of score matching which adds a specific regularization term to the original score matching objective. Its relationship to the present work will be discussed in detail in Section 5.

Denoising Autoencoders (DAE) were proposed by Vincent *et al.* (2008) as a simple and competitive alternative to the CD-trained RBM used by Hinton *et al.* (2006) for pretraining deep networks (Erhan *et al.*, 2010; Vincent *et al.*, 2010). Previous studies have already pointed out connections between SM and CD (Hyvärinen, 2007a; Sohl-Dickstein *et al.*, 2009), have connected SM to optimal denoising for Gaussian noise with infinitesimal variance (Hyvärinen, 2008) and have shown that training Gaussian Binary RBM with SM is equivalent to training a regular (non-denoising) AE with an additional regularization term (Swersky, 2010). The present study however is the first to recast the training of a DAE as a form of regularized score matching. This connection yields insights relevant to both research directions and suggests a novel parameter estimation technique that has its roots in both DAE and SM.

We begin by a very brief presentation of the DAE architecture for continuous-valued inputs in Section 2 and of the SM technique in Section 3. This allows us to introduce our formalism and precise terminology. Then, in section 4, we connect the denoising autoencoder objective to score matching. We conclude by a discussion on how our findings advance our understanding of both approaches.

Notation

$q(\mathbf{x})$	Unknown *true* pdf. $\mathbf{x} \in \mathbb{R}^d$.
$D_n = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}$	Training set: i.i.d. sample from q.

[1]Specifically SM (Hyvärinen, 2005), minimum velocity learning (Movellan, 2008), and certain forms of CD (Hinton, 2002; Welling - Hinton, 2002) are all recast as minimizing the K-L divergence between the data distribution and the distribution obtained after running, for infinitesimal time, a dynamic that would transform it into the model distribution (Sohl-Dickstein *et al.*, 2009).

q0(x) = 1/N ∑i δ_{xi}

$\cancel{q_0(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\delta(\|\mathbf{x}-\mathbf{x}^{(i)}\|)}$	Empirical pdf associated with D_n.
$q_\sigma(\tilde{\mathbf{x}}\|\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\sigma^d}e^{-\frac{1}{2\sigma^2}\|\tilde{\mathbf{x}}-\mathbf{x}\|^2}$	Smoothing kernel or noise model: isotropic Gaussian of variance σ^2.
$q_\sigma(\tilde{\mathbf{x}}) = \frac{1}{n}\sum_{t=1}^{n} q_\sigma(\tilde{\mathbf{x}}\|\mathbf{x}^{(t)})$	Parzen density estimate based on D_n obtainable by marginalizing $q_\sigma(\mathbf{x}\tilde{}, \mathbf{x})$.

2 Denoising Autoencoders

DAEs are a simple modification of classical autoencoder neural networks that are trained, not to reconstruct their input, but rather to denoise an artificially corrupted version of their input (Vincent *et al.*, 2008, 2010). Whereas an over-complete regular autoencoder can easily learn a useless identity mapping, a DAE must extract more useful features in order to solve the much harder denoising problem. DAEs have proven to be an empirically successful alternative to RBM for pre-training deep networks (Vincent *et al.*, 2008, 2010; Erhan *et al.*, 2010). DAEs have also been used in different contexts in the earlier works of LeCun (1987); Gallinari *et al.* (1987); Seung (1998).

In this study, we will consider the denoising version of a simple classical autoencoder that uses a single sigmoidal hidden layer. Since data points originate from a continuous real valued distribution it is natural to use a linear decoder with a squared reconstruction loss[3]. We will be using *tied weights* whereby encoder and decoder share the same linear transformation parameters. The considered corruption is additive isotropic Gaussian noise. A detailed description of the architecture follows:

AE: dec(enc(x)) ≈ x

DAE: train AE by data (x_i~, xi)

- A training input $\mathbf{x} \in D_n$ is first corrupted by additive Gaussian noise of covariance $\sigma^2\mathbf{I}$ yielding corrupted input $\tilde{\mathbf{x}} = \mathbf{x} + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$. This corresponds to conditional density $q_\sigma(\tilde{\mathbf{x}}|\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\sigma^d}e^{-\frac{1}{2\sigma^2}\|\tilde{\mathbf{x}}-\mathbf{x}\|^2}$.

- The corrupted version $\tilde{\mathbf{x}}$ is *encoded* into a hidden representation $\mathbf{h} \in \mathbb{R}^{d_h}$ through an affine mapping followed by a nonlinearity:
$\mathbf{h} = \text{encode}(\tilde{\mathbf{x}}) = \text{sigmoid}(\mathbf{W}\tilde{\mathbf{x}} + \mathbf{b})$, where $\tilde{\mathbf{x}} \in \mathbb{R}^d$, $\mathbf{h} \in (0,1)^{d_h}$, $\mathbf{W}$ is a $d_h \times d$ matrix and $\mathbf{b} \in \mathbb{R}^{d_h}$.

- The hidden representation $\mathbf{h}$ is *decoded* into reconstruction $\mathbf{x}^{\mathrm{r}}$ through affine mapping: $\mathbf{x}^{\mathrm{r}} = \text{decode}(\mathbf{h}) = \mathbf{W}^T\mathbf{h} + \mathbf{c}$, where $\mathbf{c} \in \mathbb{R}^d$.

- The parameters $\theta = \{\mathbf{W}, \mathbf{b}, \mathbf{c}\}$ are optimized so that the expected squared reconstruction error $\|\mathbf{x}^{\mathrm{r}} - \mathbf{x}\|^2$ is minimized, i.e. the objective function being minimized by such a DAE is:

$$\begin{aligned} J_{DAE\sigma}(\theta) &= \mathbb{E}_{q_\sigma(\tilde{\mathbf{x}},\mathbf{x})}\left[\|\text{decode}(\text{encode}(\tilde{\mathbf{x}})) - \mathbf{x}\|^2\right] \\ &= \mathbb{E}_{q_\sigma(\tilde{\mathbf{x}},\mathbf{x})}\left[\left\|\mathbf{W}^T\text{sigmoid}(\mathbf{W}\tilde{\mathbf{x}} + \mathbf{b})) + \mathbf{c} - \mathbf{x}\right\|^2\right]. \end{aligned} \quad (1)$$

3 Score Matching

3.1 Explicit Score Matching (ESM)

Score Matching was introduced by Hyvärinen (2005) as a technique to learn the parameters θ of probability density models $p(\mathbf{x};\theta)$ with intractable partition function $Z(\theta)$,

$$p(\mathbf{x};\theta) = \frac{1}{Z(\theta)} \exp(-E(\mathbf{x};\theta)).$$

E is called the energy function. Following Hyvärinen (2005), we will call *score* the gradient of the log density with respect to the data vector: $\psi(\mathbf{x};\theta) = \frac{\partial \log p(\mathbf{x};\theta)}{\partial \mathbf{x}}$. Beware that this usage differs slightly from traditional statistics terminology where *score* usually refers to the derivative of the log likelihood with respect to *parameters,* whereas here we are talking about a *score* with respect to the *data*. The core principle of *score matching* (Hyvärinen, 2005) is to learn θ so that $\psi(\mathbf{x};\theta) = \frac{\partial \log p(\mathbf{x};\theta)}{\partial \mathbf{x}}$ best matches the corresponding *score* of the true distribution, i.e. $\frac{\partial \log q(\mathbf{x})}{\partial x}$. The corresponding objective function to be minimized is the expected squared error between these two vectors, i.e.

$$\boxed{J_{ESMq}(\theta) = \mathbb{E}_{q(\mathbf{x})}\left[\frac{1}{2}\left\|\psi(\mathbf{x};\theta) - \frac{\partial \log q(\mathbf{x})}{\partial \mathbf{x}}\right\|^2\right].}$$

We refer to this formulation as *explicit score matching* (ESM).

Note that the score $\psi(\mathbf{x};\theta)$ does not depend on troublesome $Z(\theta)$. But since q is unknown, we do not have explicit regression targets $\frac{\partial \log q(\mathbf{x})}{\partial \mathbf{x}}$. Hyvärinen (2005) mentions in passing that non-parametric methods might be used to estimate those, and we shall later pay closer attention to this possibility.

3.2 Implicit Score Matching (ISM)

Hyvärinen (2005) instead proceeds by proving the following remarkable property:

$$\underbrace{\mathbb{E}_{q(\mathbf{x})}\left[\frac{1}{2}\left\|\psi(\mathbf{x};\theta)-\frac{\partial \log q(\mathbf{x})}{\partial \mathbf{x}}\right\|^2\right]}_{J_{ESMq}(\theta)} = \underbrace{\mathbb{E}_{q(\mathbf{x})}\left[\frac{1}{2}\|\psi(\mathbf{x};\theta)\|^2+\sum_{i=1}^{d}\frac{\partial \psi_i(\mathbf{x};\theta)}{\partial \mathbf{x}_i}\right]}_{J_{ISMq}(\theta)} + C_1 \quad (2)$$

where $\psi_i(\mathbf{x};\theta) = \psi(\mathbf{x};\theta)_i = \frac{\partial \log p(\mathbf{x};\theta)}{\partial \mathbf{x}_i}$, and C_1 is a constant that does not depend on θ. This yields an *implicit score matching* objective J_{ISMq} that no longer requires having an explicit score target for q but is nevertheless equivalent to J_{ESMq}. Hyvärinen (2005) formally shows that, provided $q(\mathbf{x})$ and $\psi(\mathbf{x};\theta)$ satisfy some weak regularity conditions[4], we have

$$\boxed{J_{ESMq} \smile J_{ISMq}} \quad (3)$$

3.3 Finite Sample Version of ISM

Since we only have samples D_n from q, Hyvärinen proposes to optimize the fi

nite sample version of J_{ISMq}:

$$J_{ISMq_0}(\theta) = \frac{1}{n}\sum_{t=1}^{n}\left(\frac{1}{2}\left\|\psi(\mathbf{x}^{(t)};\theta)\right\|^2+\sum_{i=1}^{d}\frac{\partial \psi_i(\mathbf{x}^{(t)};\theta)}{\partial \mathbf{x}_i}\right). \quad (4)$$

$\to J_{ISMq}$ when $n \to \infty$

What happens in the transition from J_{ISMq} to finite sample version J_{ISMq_0} is however not entirely clear. Concerns regarding the stability of the resulting criterion were raised by Kingma and LeCun (2010), who propose instead to optimize a regularized version of J_{ISMq_0}:

$$J_{ISMreg}(\theta) = J_{ISMq_0}(\theta) + \lambda\sum_{i=1}^{d}\left(\frac{\partial \psi_i(\mathbf{x}^{(t)};\theta)}{\partial \mathbf{x}_i}\right)^2, \quad (6)$$

where the strength of the additional regularization term is controlled by hyperparameter λ. The relationship between this criterion and the criteria we propose below will be further discussed in Section 5.

[4] $q(\mathbf{x})$ and $\psi(\mathbf{x};\theta)$ are differentiable, $\mathbb{E}_{q(\mathbf{x})}\left[\left\|\frac{\partial \log q(\mathbf{x})}{\partial \mathbf{x}}\right\|^2\right]$ is finite, and for any θ, $\mathbb{E}_{q(\mathbf{x})}\left[\|\psi(\mathbf{x};\theta)\|^2\right]$ is finite and $\lim_{\|\mathbf{x}\|\to\infty} q(\mathbf{x})\psi(\mathbf{x};\theta) = 0$.

4 Linking SM to the DAE Objective

4.1 Matching the Score of a Non-Parametric Estimator

As previously stated, the possibility of matching the score $\psi(\mathbf{x};\theta)$ with an explicit target score for q obtained through non-parametric estimation was mentioned but not pursued in Hyvärinen (2005). We now examine this possibility more closely. Explicitly matching $\psi(\mathbf{x};\theta)$ with the score of Parzen windows density estimator $q_\sigma(\tilde{\mathbf{x}})$ yields the following objective:

$$J_{ESMq_\sigma}(\theta) = \mathbb{E}_{q_\sigma(\tilde{\mathbf{x}})}\left[\frac{1}{2}\left\|\psi(\tilde{\mathbf{x}};\theta) - \frac{\partial \log q_\sigma(\tilde{\mathbf{x}})}{\partial \tilde{\mathbf{x}}}\right\|^2\right]. \tag{7}$$

For $\sigma > 0$, q_σ is differentiable, decreases to 0 at infinity, and $\mathbb{E}_{q_\sigma(\tilde{\mathbf{x}})}\left[\left\|\frac{\partial \log q_\sigma(\tilde{\mathbf{x}})}{\partial \tilde{\mathbf{x}}}\right\|^2\right]$ is finite. All regularity conditions are satisfied, so the same equivalence with ISM as in Eq. 3 holds, i.e.

$$J_{ESMq_\sigma} \smile J_{ISMq_\sigma}. \tag{8}$$

Note that this equivalence however breaks in the limit $\sigma \to 0$, because q_σ no longer satisfies these regularity conditions, and J_{ESMq_σ} can no longer be computed (whereas J_{ISMq_σ} remains well-behaved).

4.2 Denoising Score Matching (DSM)

Let us now consider a slightly different objective, that is inspired by both the Score Matching principle and by the Denoising Autoencoder approach of using pairs of clean and corrupted examples $(\mathbf{x}, \tilde{\mathbf{x}})$. For joint density $q_\sigma(\tilde{\mathbf{x}}, \mathbf{x}) = q_\sigma(\tilde{\mathbf{x}}|\mathbf{x})q_0(\mathbf{x})$, we define the following *denoising score matching* (DSM) objective:

CSM(x) = SM(q(y|x))
CSM = exp_x CSM(x)
= sum_i CSM(x_i)/N

E x̃|x

$$J_{DSMq_\sigma}(\theta) = \mathbb{E}_{q_\sigma(\mathbf{x},\tilde{\mathbf{x}})}\left[\frac{1}{2}\left\|\psi(\tilde{\mathbf{x}};\theta) - \frac{\partial \log q_\sigma(\tilde{\mathbf{x}}|\mathbf{x})}{\partial \tilde{\mathbf{x}}}\right\|^2\right]. \tag{9}$$

The underlying intuition is that following the gradient ψ of the log density at some corrupted point $\tilde{\mathbf{x}}$ should ideally move us towards the clean sample $\mathbf{x}$. Note that with the considered Gaussian kernel we have

$$\frac{\partial \log q_\sigma(\tilde{\mathbf{x}}|\mathbf{x})}{\partial \tilde{\mathbf{x}}} = \frac{1}{\sigma^2}(\mathbf{x} - \tilde{\mathbf{x}}). \tag{10}$$

Direction $\frac{1}{\sigma^2}(\mathbf{x} - \tilde{\mathbf{x}})$ clearly corresponds to moving from $\tilde{\mathbf{x}}$ back towards clean sample $\mathbf{x}$, and we want ψ to match that as best it can.

Now this alternate objective, inspired by DAE, == ESM. Formally,

$$J_{ESMq_\sigma} \smile J_{DSMq_\sigma} \tag{11}$$

4.3 An Energy Function that Yields the DAE Objective

Let

$$
\begin{aligned}
p(\mathbf{x};\theta) &= \frac{1}{Z(\theta)}\exp(-E(x;\theta)) \\
E(\mathbf{x};\underbrace{\mathbf{W},\mathbf{b},\mathbf{c}}_{\theta}) &= -\frac{\langle \mathbf{c},\mathbf{x}\rangle - \frac{1}{2}\|\mathbf{x}\|^2 + \sum_{j=1}^{d_h}\operatorname{softplus}\left(\langle \mathbf{W}_j,\mathbf{x}\rangle + \mathbf{b}_j\right)}{\sigma^2}. \qquad (12)
\end{aligned}
$$

==>

$$
\begin{aligned}
\psi_i(\mathbf{x};\theta) &= -\frac{\partial E}{\partial \mathbf{x}_i} \\
&= \frac{1}{\sigma^2}\left(\mathbf{c}_i - \mathbf{x}_i + \sum_{j=1}^{d_h}\operatorname{softplus}'\left(\langle \mathbf{W}_j,\mathbf{x}\rangle + \mathbf{b}_j\right)\frac{\partial\left(\langle \mathbf{W}_j,\mathbf{x}\rangle + \mathbf{b}_j\right)}{\partial \mathbf{x}_i}\right) \\
&= \frac{1}{\sigma^2}\left(\mathbf{c}_i - \mathbf{x}_i + \sum_{j=1}^{d_h}\overset{\text{expit}}{\operatorname{sigmoid}}\left(\langle \mathbf{W}_j,\mathbf{x}\rangle + \mathbf{b}_j\right)\mathbf{W}_{ji}\right)
\end{aligned}
$$

write

$$
\psi(\mathbf{x};\theta) = \frac{1}{\sigma^2}\left(\mathbf{W}^T\operatorname{sigmoid}(\mathbf{W}\mathbf{x}+\mathbf{b}) + \mathbf{c} - \mathbf{x}\right). \qquad (13)
$$

Substituting Eq. (10) and (13) in the expression for J_{DSMq_σ} (9) we get, for this choice of Parzen kernel and density model,

$$
\begin{aligned}
J_{DSMq_\sigma}(\theta) &= \mathbb{E}_{q_\sigma(\mathbf{x},\tilde{\mathbf{x}})}\left[\frac{1}{2}\left\|\psi(\tilde{\mathbf{x}};\theta) - \frac{\partial \log q_\sigma(\tilde{\mathbf{x}}|\mathbf{x})}{\partial \tilde{\mathbf{x}}}\right\|^2\right] \\
&= \mathbb{E}_{q_\sigma(\mathbf{x},\tilde{\mathbf{x}})}\left[\frac{1}{2}\left\|\frac{1}{\sigma^2}\left(\mathbf{W}^T\operatorname{sigmoid}(\mathbf{W}\tilde{\mathbf{x}}+\mathbf{b}) + \mathbf{c} - \tilde{\mathbf{x}}\right) - \frac{1}{\sigma^2}(\mathbf{x}-\tilde{\mathbf{x}})\right\|^2\right] \\
&= \frac{1}{2\sigma^4}\mathbb{E}_{q_\sigma(\mathbf{x},\tilde{\mathbf{x}})}\left[\left\|\mathbf{W}^T\operatorname{sigmoid}(\mathbf{W}\tilde{\mathbf{x}}+\mathbf{b}) + \mathbf{c} - \mathbf{x}\right\|^2\right] \\
&= \frac{1}{2\sigma^4}J_{DAE\sigma}(\theta).
\end{aligned}
$$

==>

$$
J_{DSMq_\sigma} \smile J_{DAE\sigma} \qquad (14)
$$

5 Discussion

Putting together (8), (11) and (14), we can write, for $\sigma > 0$,

$$J_{ISMq_\sigma} \smile J_{ESMq_\sigma} \smile J_{DSMq_\sigma} \smile J_{DAE\sigma}. \qquad (15)$$

In summary, **training the denoising autoencoder defined in section 2 is equivalent to performing score matching (explicit or implicit) with the energy function of Eq. 12 on Parzen density estimate** q_σ. Such a training would typically use SGD, whereby samples from q_σ are obtained by corrupting samples from D_n. And it may be carried out with any of these four optimization objective formulations[5].

We introduced the kernel-smoothed empirical distribution q_σ to show a connection between SM and a simple DAE. Interestingly, the regularized SM criterion J_{ISMreg} (6) that Kingma and LeCun (2010) recently introduced with the very different motivation of curing possible instabilities, was derived by approximating[6] what amounts to J_{ISMq_σ}. From this perspective our four q_σ-based criteria in (15), including the DAE, may be seen as alternative approximation-free forms of regularized SM. A key difference is that, as is done with DAE training, we would optimize *stochastic* versions of these approximation-free regularized criteria by corrupting training examples (i.e. sampling form q_σ), whereas Kingma and LeCun (2010) optimize an *approximation* of J_{ISMq_σ} centered on the training samples only (i.e. sampling from q_0). Also, whereas J_{ISMreg}, like the other ISM criteria, requires computing second derivatives, the stochastic version of our novel J_{DSMq_σ} criterion does not, and thus appears much simpler to implement.

Note that the energy function in (12) is particular in that its scaling, which we may call its *temperature*, is chosen to match the corrupting noise level σ^2. This is required only to establish the last equivalence with the specific DAE we considered. But regarding the generic objectives $J_{ISMq_\sigma} \smile J_{ESMq_\sigma} \smile J_{DSMq_\sigma}$, their σ may in principle be chosen irrespective of the form or temperature of whatever energy function is to be learnt. Interestingly, the energy function in (12), that we designed to yield the equivalence with our DAE objective, happens to be very similar to the free energy of a RBM with binary hidden units and Gaussian visible units (Welling *et al.*, 2005; Bengio *et al.*, 2007; Swersky, 2010). The major difference is that this latter free energy does not have a *global* temperature scaling of the whole expression [7]. We designed (12) to exactly yield the denoising version of the classic AE described in Section 2. But with *tied* weights, it may be preferable to allow for an extra positive scaling parameter α for the reconstruction, so that there at least exists an equivalent reparametrization of the model for scaled input

[5] Note however that while these q_σ-based objectives are formally equivalent, their SGD optimization, based on sampling a limited number of corrupted examples, is likely to behave differently for each objective.

[6] A first order Taylor expansion and a diagonal Hessian approximation are used.

[7] Specifically, in the free energy of a Gaussian-binary RBM, the softplus terms are *not* divided by σ^2 nor scaled in any way.

data[8]. This is easily obtained in the energy function by multiplying the sum of $\mathrm{softplus}$ terms in Eq. 12 by α. We may even allow an arbitrary rescaling factor α_j for each hidden layer dimension independently by multiplying each $\mathrm{softplus}$ term by its own rescaling parameter α_j, which yields

$$E(\mathbf{x}; \underbrace{\mathbf{W}, \mathbf{b}, \mathbf{c}, \alpha, \sigma_m}_{\theta}) = -\frac{1}{\sigma_m^2}\left(\langle \mathbf{c}, \mathbf{x}\rangle - \frac{1}{2}\|\mathbf{x}\|^2 + \sum_{j=1}^{d_h} \alpha_{\mathrm{j}} \mathrm{softplus}\left(\langle \mathbf{W}_j, \mathbf{x}\rangle + \mathbf{b}_j\right)\right).$$

Here we have also included, as model parameter, a σ_m (where m stands for model) distinct from the noise σ of the training objective[9].

Our q_σ-based objectives J_{ISMq_σ}, J_{ESMq_σ}, or J_{DSMq_σ} can be used as alternatives to the finite sample objective J_{ISMq_0} (Eq. 4) advocated in Hyvärinen (2005) for learning unnormalized densities. Note that J_{ISMq_0} is a special case of J_{ISMq_σ} obtained in the limit of $\sigma \to 0$. Also, since Kingma and LeCun (2010) showed that it may be preferable to use a regularized criterion (that they derived form smoothed empirical distribution q_σ), it is likely that our q_σ-based criteria may, for $\sigma > 0$, yield better generalization performance than the J_{ISMq_0} advocated in Hyvärinen (2005)[10]. It seems that σ could allow one to choose an optimal bias-variance tradeoff for the finite-sample estimation of the true score matching gradient of interest $\nabla_\theta J_{ESMq} = \nabla_\theta J_{ISMq}$. While $\nabla_\theta J_{ISMq_0}$ is an unbiased estimator of it, $\nabla_\theta J_{ISMq_\sigma} = \nabla_\theta J_{ESMq_\sigma} = \nabla_\theta J_{DSMq_\sigma}$ will generally be biased when $\sigma > 0$ but are also likely to have a lower variance.

Among the three equivalent score matching objectives based on q_σ, objective J_{DSMq_σ} appears particularly interesting as a novel alternative formulation. It was motivated by both the SM principle and the DAE principle. From DAE it borrows the idea of learning to denoise artificially corrupted samples, and from SM it borrows the idea of learning a score function derived from an unnormalized density. J_{DSMq_σ} may prove simpler and more efficient in practice than the mathematically equivalent J_{ISMq_σ} , as it does not require computing second derivatives.

Our result is also a significant advance for DAEs. First, we have defined a proper energy function for the considered DAE through Eq. 12. This will enable many previously impossible or ill-defined operations on a trained DAE, for example deciding which is the more likely among several inputs, or sampling from a trained DAE using Hybrid Monte-Carlo (Duane *et al.*, 1987). Second, whereas using the same weight matrix ("tied weights") for the encoder and decoder is justified for RBMs, the encoder-decoder framework does not constrain that choice. Previous work on DAEs (Vincent *et al.*, 2008; Erhan *et al.*, 2010; Vincent *et al.*, 2010) explored both options, often finding tied weights to yield better empirical results. Within the SM framework presented here, using tied weights between encoder and decoder now has a proper justification,

---

[8] If for example one multiplies the input values by 100, one can obtain the same hidden representation as before by dividing $\mathbf{W}$ by 100. But because of the tied weights this means that the reconstruction would also be divided by 100 (i.e. there is no equivalent reparametrization), unless it can be compensated by an additional scaling of the reconstruction by a parameter α.

[9] We would however have to set $\sigma_m = \sigma$ to recover a recognizable denoising autoencoder objective.

[10] It is also noteworthy that the experimental results of Vincent *et al.* (2008, 2010) on DAE showed that the best models, judged by their ability to extract useful features, were obtained for non negligible values of the noise parameters. Moreover this way of controlling the model's capacity worked much better than either reducing the hidden layer size or than traditional weight decay.

since it follows naturally from differentiating the energy. Third, this framework opens the door to new variants that would naturally fall out from other choices of the energy function.

Acknowledgements

Yoshua Bengio ,Olivier Delalleau and the other members of the Lisa Lab .NSERC , MITACS and FQRNT.

References

Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. (2007). Greedy layer-wise training of deep networks. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19 (NIPS'06)*, pages 153–160. MIT Press.

Duane, S., Kennedy, A., Pendleton, B., and Roweth, D. (1987). Hybrid Monte Carlo. *Phys. Lett. B*, **195**, 216–222.

Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., and Bengio, S. (2010). Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, **11**, 625–660.

Gallinari, P., LeCun, Y., Thiria, S., and Fogelman-Soulie, F. (1987). Memoires associatives distribuees. In *Proceedings of COGNITIVA 87*, Paris, La Villette.

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

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

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

Hyvärinen, A. (2007a). Connections between score matching, contrastive divergence, and pseudolikelihood for continuous-valued variables. *IEEE Transactions on Neural Networks*, **18**, 1529–1531.

Hyvärinen, A. (2007b). Some extensions of score matching. *Computational Statistics and Data Analysis*, **51**, 2499–2512.

Hyvärinen, A. (2008). Optimal approximation of signal priors. *Neural Computation*, **20**(12), 3087–3110.

Kingma, D. and LeCun, Y. (2010). Regularized estimation of image statistics by score matching. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23 (NIPS'10)*, pages 1126–1134. MIT Press.

LeCun, Y. (1987). *Modèles connexionistes de l'apprentissage*. Ph.D. thesis, Université de Paris VI.

Lyu, S. (2009). Interpretation and generalization of score matching. In *Proceedings of the 25th Conference in Uncertainty in Artificial Intelligence (UAI'09)*.

Marlin, B., Swersky, K., Chen, B., and de Freitas, N. (2009). Inductive principles for restricted boltzmann machine learning. In *Proceedings of The Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS'10)*, volume 9, pages 509–516.

Movellan, J. R. (2008). Contrastive divergence in gaussian diffusions. *Neural Computation*, **20**(9), 2238–2252.

Seung, S. H. (1998). Learning continuous attractors in recurrent networks. In M. Jordan, M. Kearns, and S. Solla, editors, *Advances in Neural Information Processing Systems 10 (NIPS'97)*, pages 654–660. MIT Press.

Sohl-Dickstein, J., Battaglino, P., and DeWeese, M. R. (2009). Minimum probability flow learning. Technical report, arXiv:0906.4779.

Swersky, K. (2010). *Inductive Principles for Learning Restricted Boltzmann Machines*. Master's thesis, University of British Columbia.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In W. W. Cohen, A. McCallum, and S. T. Roweis, editors, *ICML'08*, pages 1096–1103. ACM.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. (2010). Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, **11**, 3371–3408.

Welling, M. and Hinton, G. E. (2002). A new learning algorithm for mean field Boltzmann machines. In *ICANN '02: Proceedings of the International Conference on Artificial Neural Networks*, pages 351–357, London, UK. Springer-Verlag.

Welling, M., Rosen-Zvi, M., and Hinton, G. E. (2005). Exponential family harmoniums with an application to information retrieval. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17 (NIPS'04)*, pages 1481–1488, Cambridge, MA. MIT Press.

Appendix

Proof $\quad J_{ESM_\sigma} \smile J_{DSM_{q_\sigma}}$ (11)

q The ESM criterion using the Parzen density estimator is defined in Eq. (7) as

$$J_{ESM_{q_\sigma}}(\theta) = \mathbb{E}_{q_\sigma(\tilde{\mathbf{x}})}\left[\frac{1}{2}\left\|\psi(\tilde{\mathbf{x}};\theta) - \frac{\partial \log q_\sigma(\tilde{\mathbf{x}})}{\partial \tilde{\mathbf{x}}}\right\|^2\right]$$

$$= \mathbb{E}_{q_\sigma(\mathbf{x}^{\sim})}\left[\frac{1}{2}\|\psi(\tilde{\mathbf{x}};\theta)\|^2\right] - S(\theta) + C_2 \tag{16}$$

where $C_2 = \mathbb{E}_{q_\sigma(\tilde{\mathbf{x}})}\left[\frac{1}{2}\left\|\frac{\partial \log q_\sigma(\tilde{\mathbf{x}})}{\partial \tilde{\mathbf{x}}}\right\|^2\right]$ is a constant that does not depend on θ, and

Fact.

Ep(x) < ψ, D log p(x) > =
Ep(x,z) < ψ, D log p(x | z) >

$$\begin{aligned}
S(\theta) &:= \mathbb{E}_{q_\sigma(\mathbf{x}^{\sim})}\left[\left\langle \psi(\tilde{\mathbf{x}};\theta), \frac{\partial \log q_\sigma(\tilde{\mathbf{x}})}{\partial \tilde{\mathbf{x}}}\right\rangle\right] \\
&= \int_{\tilde{\mathbf{x}}} \left\langle \psi(\tilde{\mathbf{x}};\theta), \frac{\partial}{\partial \tilde{\mathbf{x}}} q_\sigma(\tilde{\mathbf{x}})\right\rangle d\tilde{\mathbf{x}} \\
&= \int_{\tilde{\mathbf{x}}} \left\langle \psi(\tilde{\mathbf{x}};\theta), \frac{\partial}{\partial \tilde{\mathbf{x}}} \int_{\mathbf{x}} q_0(\mathbf{x}) q_\sigma(\tilde{\mathbf{x}}|\mathbf{x}) d\mathbf{x}\right\rangle d\tilde{\mathbf{x}} \\
&= \int_{\tilde{\mathbf{x}}} \left\langle \psi(\tilde{\mathbf{x}};\theta), \int_{\mathbf{x}} q_0(\mathbf{x}) \frac{\partial q_\sigma(\tilde{\mathbf{x}}|\mathbf{x})}{\partial \tilde{\mathbf{x}}} d\mathbf{x}\right\rangle d\tilde{\mathbf{x}} \\
&= \int_{\tilde{\mathbf{x}}} \left\langle \psi(\tilde{\mathbf{x}};\theta), \int_{\mathbf{x}} q_0(\mathbf{x}) q_\sigma(\tilde{\mathbf{x}}|\mathbf{x}) \frac{\partial \log q_\sigma(\tilde{\mathbf{x}}|\mathbf{x})}{\partial \tilde{\mathbf{x}}} d\mathbf{x}\right\rangle d\tilde{\mathbf{x}} \\
&= \int_{\tilde{\mathbf{x}}} \int_{\mathbf{x}} q_0(\mathbf{x}) q_\sigma(\tilde{\mathbf{x}}|\mathbf{x}) \left\langle \psi(\tilde{\mathbf{x}};\theta), \frac{\partial \log q_\sigma(\tilde{\mathbf{x}}|\mathbf{x})}{\partial \tilde{\mathbf{x}}}\right\rangle d\mathbf{x} d\tilde{\mathbf{x}} \\
&= \int_{\tilde{\mathbf{x}}} \int_{\mathbf{x}} q_\sigma(\tilde{\mathbf{x}}, \mathbf{x}) \left\langle \psi(\tilde{\mathbf{x}};\theta), \frac{\partial \log q_\sigma(\tilde{\mathbf{x}}|\mathbf{x})}{\partial \tilde{\mathbf{x}}}\right\rangle d\mathbf{x} d\tilde{\mathbf{x}} \\
&= \mathbb{E}_{q_\sigma(\tilde{\mathbf{x}},\mathbf{x})}\left[\left\langle \psi(\tilde{\mathbf{x}};\theta), \frac{\partial \log q_\sigma(\tilde{\mathbf{x}}|\mathbf{x})}{\partial \tilde{\mathbf{x}}}\right\rangle\right].
\end{aligned}$$

==> $$\begin{aligned} J_{ESM_{q_\sigma}}(\theta) &= \mathbb{E}_{q_\sigma(\tilde{\mathbf{x}})}\left[\frac{1}{2}\|\psi(\tilde{\mathbf{x}};\theta)\|^2\right] \\ &\quad - \mathbb{E}_{q_\sigma(\mathbf{x},\tilde{\mathbf{x}})}\left[\left\langle \psi(\tilde{\mathbf{x}};\theta), \frac{\partial \log q_\sigma(\tilde{\mathbf{x}}|\mathbf{x})}{\partial \tilde{\mathbf{x}}}\right\rangle\right] + C_2. \end{aligned} \tag{17}$$

$$\begin{aligned} J_{DSM_{q_\sigma}}(\theta) &= \mathbb{E}_{q_\sigma(\mathbf{x}^{\sim})}\left[\frac{1}{2}\|\psi(\tilde{\mathbf{x}};\theta)\|^2\right] \\ &\quad - \mathbb{E}_{q_\sigma(\mathbf{x},\tilde{\mathbf{x}})}\left[\left\langle \psi(\tilde{\mathbf{x}};\theta), \frac{\partial \log q_\sigma(\tilde{\mathbf{x}}|\mathbf{x})}{\partial \tilde{\mathbf{x}}}\right\rangle\right] + C_3 \end{aligned} \tag{18}$$

==> $J_{ESM_{q_\sigma}}(\theta) = J_{DSM_{q_\sigma}}(\theta) + C_2 - C_3.$

Q.E.D.