Score Matching Chapter from Köster 2009 PhD Thesis

1 Score Matching

In this chapter, we will describe a novel estimation principle that provides an alternative way to learn the filters in ICA and related models. Score matching has been proposed in 2005 and provides a mechanism for learning in energy based models, where the pdf can be computed only up to a multiplicative normalization constant [1, 2, 3]. It has been used in the two-layer model in *Publications 3 and 4* [4, **?**] and in the Markov Random Field, *Publication 5* [5] of this thesis.

We will start by describing how the score matching optimization proceeds, and then give some intuition for the method by providing a simple example. Consider a distribution defined up to proportionality by the exponential of a non-negative energy function

$$p(\mathbf{x}|\theta) \propto \exp(-E(\mathbf{x}, \theta)) \tag{1}$$

so the normalized distribution is given by

$$p(\mathbf{x}|\theta) = \frac{\exp(-E(\mathbf{x}, \theta))}{\int \exp(-E(\mathbf{x}, \theta)) d\mathbf{x}} = \frac{1}{Z} \exp(-E(\mathbf{x}, \theta)) \tag{2}$$

where the integral is taken over all space. The score function, which we here take to be the derivative w.r.t. the elements of $\mathbf{x}$ is given by

$$\psi_i(\mathbf{x}, \theta) = -\frac{\partial}{\partial x_i} E(\mathbf{x}, \theta). \tag{3}$$

We can now define a new objective function that measures the squared distance of the score functions of the model, denoted by $\psi(\mathbf{x}, \theta)$ and of the data $\psi_{\mathbf{x}}(\mathbf{x})$ as

$$J = \frac{1}{2} \int p_{\mathbf{x}}(\mathbf{x}) ||\psi(\mathbf{x}, \theta) - \psi_{\mathbf{x}}(\mathbf{x})||^2 d\mathbf{x} \tag{4}$$

and we seek the parameter vector θ that minimizes this distance. This stills seems like a difficult problem, since there is no easy way of estimating the data score function $\psi_{\mathbf{x}}(\mathbf{x})$. However, it can be shown whence expansion of the terms and partial integration that minimizing the score matching objective function reduces to

$$J(\mathbf{x}, \theta) = \frac{1}{T} \sum_{i,t} \frac{1}{2} \psi_i(\mathbf{x}(t), \theta)^2 + \frac{\partial}{\partial x_i} \psi_i(\mathbf{x}(t), \theta) + C \tag{5}$$

where we have additionally replaced the expectation by a sample average over T observations, and the constant C does not depend on the parameters. Now supposing that the data follows the model, i.e. there exists a θ^* such that $p_{\mathbf{x}}(\mathbf{x}) = p(\mathbf{x}, \theta^*)$, then under some weak regularity conditions minimizing J gives a consistent estimate of the parameter vector. To show this, consider the case $J(\theta) = 0$ for some θ. Now the non-negativity of the energy implies that $p_{\mathbf{x}}(\mathbf{x}) > 0$ for all $\mathbf{x}$ from which it follows that the score functions $\psi(\mathbf{x}, \theta)$ and $\psi_{\mathbf{x}}(\mathbf{x})$ are equal. This implies that the probabilities are related as $p_{\mathbf{x}}(\mathbf{x}) = cp(\mathbf{x}, \theta)$ but the constant c is necessarily unity since both pdfs have to integrate to zero. From this it follows that $\theta = \theta^*$, showing that the global minimum of the score matching objective corresponds to the true solution.

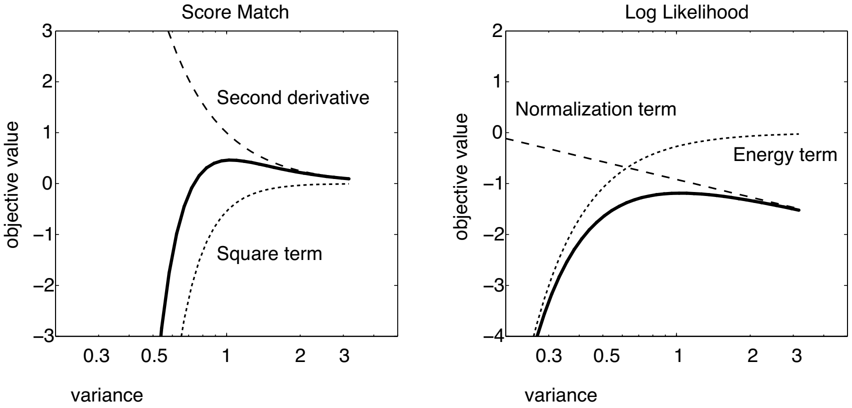

Figure 1. Illustration of score matching, applied to infer the variance of a univariate normal distribution. Both the score matching objective and the log-likelihood (solid curves) have the optimum at the correct position, but the functions differ significantly in shape. Comparing the two terms of the score matching objective with the energy and normalization term of the log-likelihood reveals the similarity between the second derivative term in the score matching objective and the normalization term: both penalize unspecific models with unnecessarily high variance.

1.1 A Simple Example

Consider the simple problem of fitting the mean and variance of a univariate gaussian to observed data samples $x(t)$. We have the log-likelihood

$$\log p(x(t)|\mu,\sigma^2) = -\frac{1}{2\sigma^2}(x(t)-\mu)^2 - \log Z \tag{6}$$

where the partition function Z is treated as unknown. The score function of the model is

$$\psi = \frac{\partial}{\partial x}\log p = -\frac{1}{\sigma^2}(x(t)-\mu) \tag{7}$$

and the derivative of the score function

$$\psi' = \frac{\partial^2}{\partial x^2}\log p = -\frac{1}{\sigma^2} \tag{8}$$

so the sample version of the score matching objective for T observations is

$$J = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{2}\left(\sigma^{-2}(x(t)-\mu)\right)^2 - \sigma^{-2} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{2}\sigma^{-4}(x(t)-\mu)^2 - \sigma^{-2}. \tag{9}$$

To obtain the score matching estimate of the mean μ we take the derivative and set it to zero to obtain

$$\frac{\partial J}{\partial \mu} = \frac{1}{T}\sum_{t=1}^{T}\sigma^{-4}(\mathbf{x}(t)-\mu) = 0 \tag{10}$$

$$\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}(t) \tag{11}$$

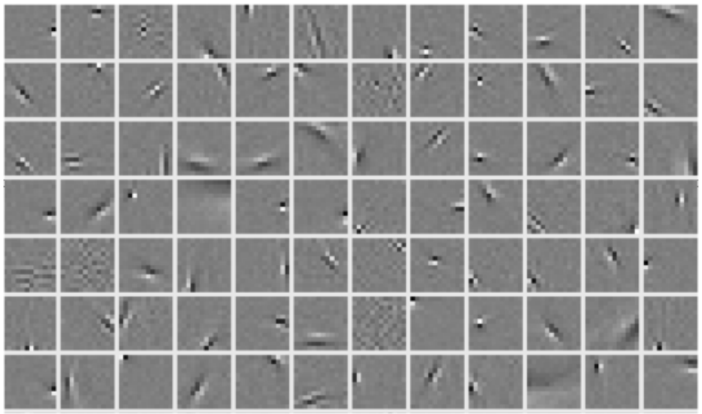

Figure 2. Overcomplete ICA model estimated with score matching. Only a subset of the 2304 filters from the 16 times overcomplete model are shown. Note that in contrast to the sparse coding model, we estimate an overcomplete set of filters rather than basis functions. This leads to some very different properties of the model.

which is the sample mean, in agreement with the maximum likelihood estimate. Similarly, to fit the variance σ^2 we take the derivative

$$\frac{\partial J}{\partial \sigma} = \frac{1}{T} \sum_{t=1}^{T} \left(-2\sigma^{-5} (\mathbf{x}(t) - \mu)^2 \right) + 2\sigma^{-3} = 0 \tag{12}$$

$$\hat{\sigma} = \sqrt{\sum_{t=1}^{T} (\mathbf{x}(t) - \mu)^2} \tag{13}$$

which is the sample variance, again in accordance with maximum likelihood. In general, there is no closed form solution for the optimal parameters, so gradient methods have to be used for optimization.

The relation between score matching and maximum likelihood are further illustrated in Fig. 1. Comparing the terms in the score matching objective with the log-likelihood allows an intuitive interpretation of score matching. The square term in the score matching objective acts similar to the energy (i.e. the non-normalized log-likelihood) and tries to make the model general enough to cover all of the data points. The second derivative has a similar effect to a normalization term, penalizing the unspecific model and forcing it to focus probability mass to where the observed data lies.

1.2 Overcomplete ICA Example

Score matching can easily be applied to estimate ICA models, and because the normalization constant is not required to be known, it is an obvious choice for overcomplete ICA. In this case the model is

$$-\log p(\mathbf{x}) \propto E(\mathbf{x}) = \sum_{i=1}^{M} g(\mathbf{w}_i^{\mathrm{T}} \mathbf{x}) \tag{14}$$

where the number of filters M is larger than the dimensionality of the data. Models of this kind, also known as *Products of Experts* (PoE) have been studied extensively [6], where they were estimated with

Contrastive Divergence (CD) [7], a method that shares some similarities with score matching and is based on Monte Carlo methods.

Estimating this model for natural image patches again leads to familiar Gabor-like filters, which is shown in Fig. 2 for a 16 times overcomplete model. Here we show the filters **w** rather than basis functions, because in the overcomplete model the filter matrix cannot be inverted, and thus no basis functions are defined.

The complete ICA model which we considered in the previous section can be seen as a bridge between overcomplete sparse coding models on one hand, and overcomplete PoE or ICA models on the other hand, which have some very important differences. In an energy-based model, the internal representation can be computed by a fast, simple feed-forward computation, whereas in the generative sparse coding model, the optimal pattern of activities **s** is the solution to an optimization problem. This implicitly nonlinear mapping between the data **x** and the components **s** has some attractive properties for neuroscience: since the basis functions inhibit each other, behavior such as end-stopping and nonclassical surround effects has been observed in overcomplete sparse coding models [8]. On the other hand, a recent study [9] has shown some advantages in energy based models over generative models in denoising applications, and it is not clear at the time how the two approaches are related and which provides a better model for natural images, and therefore ultimately for visual processing.

References

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

[2] A. Hyvärinen. Connections between score matching, contrastive divergence, and pseudolikelihood for continuous-valued variables. *IEEE Transactions on Neural Networks*, 18(5):1529–1531, 2007.

[3] A. Hyvärinen. Some extensions of score matching. *Computational Statistics and Data Analysis*, 51:2499–2512, 2007.

[4] U. Köster and A. Hyvärinen. A two-layer ICA-like model estimated by score matching. In *Artificial Neural Networks - ICANN 2007, Lecture Notes in Computer Science*, pages 798–807. Springer Berlin / Heidelberg, 2007.

[5] U. Köster, A. Hyvärinen, and J. T. Lindgren. Estimating Markov random field potentials for natural images. In *ICA 2007, Lecture Notes in Computer Science*, pages 515–522. Springer Berlin / Heidelberg, 2009.

[6] G. E. Hinton. Products of experts. In *Proceedings of the Ninth International Conference on Artificial Neural Networks (ICANN)*, volume 1, pages 1–6, 1999.

[7] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

[8] H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. In *Advances in Neural Information Processing Systems 19*, pages 801–808. MIT Press, 2006.

[9] M. Elad, P. Milanfar, and R. Rubinstein. Analysis versus synthesis in signal priors. In *Inverse Problems 23 (2007*, pages 947–968, 2005.