Learning Doubly Intractable Latent Variable Models via Score Matching

Eszter Vértes
Gatsby Unit, UCL
25 Howland St, London W1T 4JG
eszter@gatsby.ucl.ac.uk

Maneesh Sahani
Gatsby Unit, UCL
25 Howland St, London W1T 4JG
maneesh@gatsby.ucl.ac.uk

Abstract

Most current approaches for learning latent variable models, such as variational methods, require access to a normalized joint distribution. However, for models that do not belong to the standard, widely-studied parametric classes or that are derived from an undirected graphical model, the normalizer is often not readily available. Here we generalize the score matching approach [1] to learn a wide class of latent variable models based on joint exponential family (or maximum-entropy) distributions with arbitrary sufficient statistic vectors. We derive a stochastic gradient based optimization scheme that does not depend on the computation of normalizing constants for either of the joint or the posterior densities.

1 Background

Latent variable modelling is a powerful tool for learning about the underlying structure of a data set in an unsupervised setting. However, inference and learning are difficult in most complex (e.g. non-Gaussian) models. Intractability often arises because normalising functions cannot be computed: either for the posterior density, or the joint density itself, or both (a situation often referred to as "double intractability").

In the following, we consider models where the joint distribution over observed (x) and latent variables (z) is in the exponential family:

$$p(x, z) = \exp(\theta^T S(x, z) - A(\theta)) \,, \tag{1}$$

with a sufficient statistic vector $S(x, z)$, natural parameters θ and log-partition function $A(\theta)$.

While the exponential family is a special class of latent variable models, it is also a very flexible one that can be used to approximate many densities that are not themselves within this family. However, if the sufficient statistic vector is of a non-standard form the log-partition function $A(\theta)$ often cannot be expressed in closed form, making the model doubly intractable. For this reason, they are unsuitable for variational methods such as expectation maximization (EM) where the maximization (M) step optimizes the expectation of the normalized log joint likelihood.

Even though we do not have access to the likelihood function in these models, we can still compute the gradient of the log-likelihood with respect to the parameters:

$$\nabla_\theta \log p(x) = \nabla_\theta \log \int p(x, z) dz = \frac{\int \nabla_\theta p(x, z) dz}{\int p(x, z) dz} = \langle S(x, y) \rangle_{z|x} - \langle S(x, y) \rangle_{x,z} \,, \tag{2}$$

where angle brackets represent expectations. To approximate the above expectations, one typically has to sample from the corresponding distributions. However, sampling from the joint density may be inefficient in practice (in part, because this joint is more often multimodal than is the posterior), which leads to high variance in the gradient. Here we show how score matching makes it possible to learn in such doubly intractable exponential family models, without the need to sample from the joint distribution.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

2 Score matching for fully observed models

Score matching (SM) is an algorithm originally developed to fit statistical models without latent variables that are not easily normalized [1], i.e. when the model density is only available up to proportionality. SM circumvents the inaccessibility of the normaliser by minimizing a cost function based only on the gradients of the log density.

In principle, SM matches the gradients of the true and model densities, according to the objective function:

$$J(\theta) = E_x \left[\|\partial_x \log p^*(x) - \partial_x \log p_\theta(x)\|^2 \right] ,$$

where $p^*(x)$ is the true density and $p_\theta(x)$ is the model density.

Hyvarinen [1] showed that minimizing the discrepancy between these *score functions* ($\partial_x \log p(x)$) for the model and the true density is equivalent to minimizing the following cost function:

$$\tilde{J}(\theta) = E_x \left[\partial_x^2 \log p_\theta(x) + \frac{1}{2} \left(\partial_x \log p_\theta(x) \right)^2 \right] \tag{3}$$

$$\approx \frac{1}{N} \sum_n \partial_x^2 \log p_\theta(x^{(n)}) + \frac{1}{2} \left(\partial_x \log p_\theta(x^{(n)}) \right)^2 . \tag{4}$$

Note that $\tilde{J}(\theta)$ depends on the true density $p^*(x)$ only through the data points $\{x^{(n)}\}_{n=1\ldots N}$ and that $\partial_x \log p_\theta(x)$ does not depend on the normalizer of $p_\theta(x)$.

3 Score matching for latent variable models

3.1 Defining the score matching objective

Previous work analyzing the formal relationship between Gaussian restricted Boltzmann machines (RBMs) and autoencoders [2] has shown how the SM cost function can be defined for energy based models that incorporate latent variables z:

$$p(x, z) \propto \exp(-E(x, z; \theta)) . \tag{5}$$

The score function for such model can be expressed using an expectation with respect to the posterior density, $p(z|x)$:

$$\partial_x \log p_\theta(x) = - \int p(z|x) \; \partial_x E(x, z; \theta) dz , \tag{6}$$

and thus the SM objective can be rewritten as follows [2]:

$$J^{LV}(\theta) = \frac{1}{N} \sum_x \sum_i -\frac{1}{2} \left\langle \partial_{x_i} E(x, z; \theta) \right\rangle_{z|x}^2 + \left\langle (\partial_{x_i} E(x, z; \theta))^2 \right\rangle_{z|x} - \left\langle \partial_{x_i}^2 E(x, z; \theta) \right\rangle_{z|x} . \tag{7}$$

Here, the first sum is over the observations $x \in \{x^{(n)}\}_{n=1\ldots N}$ and ∂_{x_i} denotes the partial derivative with respect to the i^{th} dimension of x.

3.2 Score matching in jointly exponential family models

The RBM models analyzed by [2] are "singly intractable" in that the latent variables are conditionally independent given the observations and the posterior distribution is easily computed. This structure allows the SM objective function (Eq. 7) to be written in closed form and thus optimized directly with respect to the model parameters.

However, for continuous latent variables without conditional independence both the posterior distribution and the expectations in the SM objective are typically intractable. In these cases the SM objective function is non-trivial to optimize since the posterior expectations appearing in $J^{LV}(\theta)$ (Eq. 7) depend on the intractable normalizer of $p(z|x)$.

The main contribution of this paper is to show that it is nonetheless possible to compute the gradients of the SM objective for doubly intractable jointly exponential family distributions.

The SM objective function for jointly exponential family models takes a form analogous to Eq. 7.

$$J^{ExpF}(\theta) = \frac{1}{N}\sum_x \sum_i -\frac{1}{2}\left\langle \theta^T \partial_{x_i} S(x,z)\right\rangle_{z|x}^2 + \left\langle \left(\theta^T \partial_{x_i} S(x,z)\right)^2\right\rangle_{z|x} + \left\langle \theta^T \partial_{x_i}^2 S(x,z)\right\rangle_{z|x} . \tag{8}$$

To compute the gradient of $J^{ExpF}(\theta)$ with respect to θ without knowing the normalizer of $p(z|x)$ or $p(x,z)$, we make use of the following exponential family property:

$$\nabla_\theta \log p(z|x) = \nabla_\theta \theta^T S(x,z) - \nabla_\theta A_{z|x}(\theta) = S(x,z) - \left\langle S(x,z)\right\rangle_{z|x} . \tag{9}$$

This relationship allows us to propagate the derivatives into the expectation integrals, arriving at an expression for the learning gradient in which the posterior $p(z|x)$ appears only in terms of its expectations:

$$\begin{aligned}\nabla_\theta J^{ExpF}(\theta) = \frac{1}{N}\sum_x \sum_i & -\theta^T \langle \partial_{x_i} S\rangle_{z|x} \left[\mathrm{Cov}_{z|x}(S, \partial_{x_i} S)\theta + \langle \partial_{x_i} S\rangle\right] \\ & + \langle (S - \langle S\rangle_{z|x})(\theta^T \partial_{x_i} S)^2\rangle_{z|x} + 2\langle \partial_{x_i} S \partial_{x_i} S^T\rangle_{z|x}\theta \\ & + \mathrm{Cov}_{z|x}(S, \partial_{x_i}^2 S)\theta + \langle \partial_{x_i}^2 S\rangle_{z|x} ,\end{aligned} \tag{10}$$

where we have omitted the arguments of the sufficient statistics vector $S(x,z)$ for compactness. These necessary expectations can be efficiently estimated using gradient-based Markov chain Monte Carlo samplers in the posterior – typically a simpler sampling problem than that of sampling from the marginal on the observed variables. Here we used the No-U-Turn sampler [3], a variant of Hamiltonian Monte Carlo.

4 Experiments

We ran preliminary experiments using synthetic data from a rectified latent Gaussian model (RLGM), with correlated Gaussian latent variables constrained to the positive quadrant and a Gaussian output distribution:

$$p(z) \propto \mathcal{N}(z|\mathbf{0}, \Sigma) \prod_l \Theta(z_l) \tag{11}$$

$$p(x|z) = \mathcal{N}(x|Wz, \sigma^2 I) . \tag{12}$$

where $\Theta(\cdot)$ is the Heaviside function.

In general, the normalizer for the joint RLGM cannot be computed analytically, making it unsuitable for learning by EM. On the other hand, the RLGM can be written in a jointly exponential family form with the following sufficient statistics: $S(x,z) = \mathbf{vec}\left[x^T x, xz^T, zz^T\right]$. Thus we can use the gradients of the SM objective to learn the parameters of the model: Σ, W, σ.

We chose both the latent space and the observation space to be 2-dimensional and used data generated from the model for learning. Figure 1 shows contours of the empirical densities for the true and learned model parameters for three different examples. Since the log-likelihood function is intractable for the RLGM, we evaluated the the quality of the learned parameters by the total variation distance between empirical data distributions:

$$\delta(P,Q) = \sup_x |P(x) - Q(x)| . \tag{13}$$

The distances were computed across pairs of data sets generated using the true and the learned parameters. We compared these values to distances between data sets both coming from the true model (Figure 2). Based on this metric, there was no significant difference between the learned and true densities.

5 Discussion

The introduction of score matching for exponential family latent variable models makes it possible to design and learn flexible latent variable models, directly incorporating domain-specific knowledge in

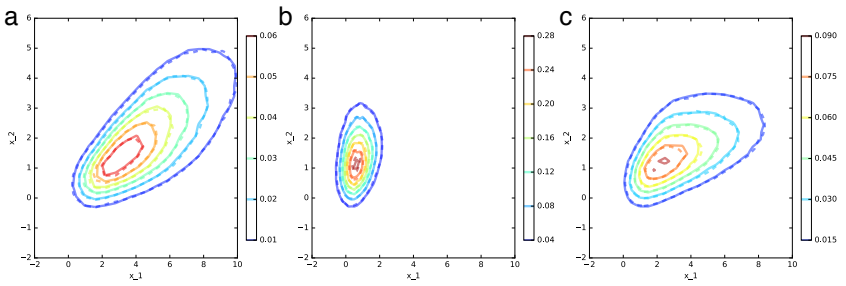

Figure 1: Contours of the data histograms generated using the true (solid lines) and the learned parameters (dashed lines).

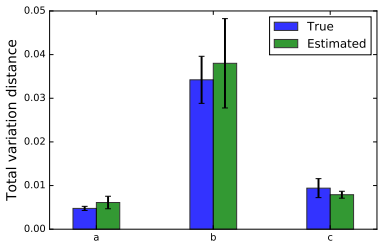

Figure 2: Total variation distance between pairs of data sets generated from the true model (blue), between pairs of data sets from the true and the learned model (green). The three sets of bars correspond to the examples shown in Fig.1.

the form of the sufficient statistics. SM is also well suited to learn models where the original model is normalized but the latent variables have a restricted domain, such as the RLGM examined in our experiments. In higher dimensions normalizing constants for these models often become analytically intractable.

As we have seen, SM can be applied to doubly intractable jointly exponential family models, and (based on consistency results shown for standard score-matching) should thus converge to the correct marginal distribution. No bias is introduced by resorting to fixed form approximations (e.g. factored, Gaussian). Furthermore, by contrast to fully stochastic learning (such as the Boltzmann machine learning rule, or contrastive divergence) Monte Carlo simulation is only required to estimate expectations under the posterior distribution, typically a more tractable problem.

Thus, we expect this new approach to be useful in many settings where the parameters of a doubly intractable model are to be learned.

References

[1] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(Apr):695–709, 2005.

[2] Kevin Swersky, David Buchman, Nando D Freitas, Benjamin M Marlin, et al. On autoencoders and score matching for energy based models. *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 1201–1208, 2011.

[3] Matthew D Hoffman and Andrew Gelman. The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. arxiv preprint arxiv: 1111.4246. 2011.