Published in: The proceedings of the 25th Conference on *Uncertainty in Artificial Intelligence* (UAI*09)
Montreal QC Canada, 18-21 June 2009. http://www.cs.mcgill.ca/ uai2009/

Interpretation and Generalization of Score Matching

Siwei Lyu

Computer Science Department
University at Albany, SUNY
Albany, NY 12222, USA
lsw@cs.albany.edu

Abstract

SM is a recently developed parameter learning method that is particularly eff ective to complicated high dimensional density models with intractable partition functions. In this paper, we study two issues that have not been completely resolved for SM. 1, we provide a formal link between ML and SM. Our analysis shows that SM finds model parameters that are more robust with noisy training data. 2, we develop a generalization of SM. Based on this generalization, we further demonstrate an extension of SM to models of discrete data.

2.1 Score matching

Score matching is a different parameter learning methodology recently proposed in [Hyv05] . It cleverly obviates computing the partition function by using an alternative divergence metric of density functions, the Fisher divergence:

$$D_F(p\|q_\theta) = \int_{\vec{x}} p(\vec{x}) \left| \frac{\nabla_{\vec{x}} p(\vec{x})}{p(\vec{x})} - \frac{\nabla_{\vec{x}} q_\theta(\vec{x})}{q_\theta(\vec{x})} \right|^2 d\vec{x}, \quad (2)$$

The learning procedure that finds θ to minimize $D_F(p\|q_\theta)$ is given the name score matching due to

that $\frac{\nabla_{\vec{x}} p(\vec{x})}{p(\vec{x})}$ is known as the (Fisher) score function Just as the KL divergence is induced from the Shannon differential entropy, so does the Fisher divergence from the Fisher information [1],

$$J(p) = \int_{\vec{x}} p(\vec{x}) \left| \frac{\nabla p(\vec{x})}{p(\vec{x})} \right|^2 d\vec{x} = \int_{\vec{x}} p(\vec{x}) \left|\nabla \log p(\vec{x})\right|^2 d\vec{x}. \quad (3)$$

Other similarities to the KL divergence include that the Fisher divergence is non-negative and is zero if and only if $p(\vec{x}) = q_\theta(\vec{x})$ $(a.e.)$, yet it is not symmetric and does not form a distance metric.

It is not hard to see that in score matching, there is no need to use the partition function, in other words , it can work directly with un-normalized models. This is because in the score functions, $\frac{\nabla_{\vec{x}} p(\vec{x})}{p(\vec{x})}$ and $\frac{\nabla_{\vec{x}} q_\theta(\vec{x})}{q_\theta(\vec{x})}$, the partition functions appear in both the denominator and the numerator, which cancel out and thus have no effect to the Fisher divergence. Furthermore, as shown in [Hyv05] , the squared distance of the model score function from the data score function, as measured by the Fisher divergence, can be computed as a simple expectation of functions of the un-normalized model. To better see this, first expand (2) as:

$$\int_{\vec{x}} p\left|\nabla_{\vec{x}} \log p\right|^2 + \int_{\vec{x}} p\left|\nabla_{\vec{x}} \log q_\theta\right|^2 - 2\int_{\vec{x}} \nabla_{\vec{x}} p^T \frac{\nabla_{\vec{x}} q_\theta}{q_\theta}. \quad (4)$$

Assume that both $p(\vec{x})$ and $q_\theta(\vec{x})$ are smooth and fast decaying, such that their logarithms have growth at most polynomial at infinity.

==>

$$D_F(p\|q_\theta) = \int_{\vec{x}} p \left(\left|\nabla_{\vec{x}} \log p\right|^2 + \left|\nabla_{\vec{x}} \log q_\theta\right|^2 + 2\triangle_{\vec{x}} \log q_\theta\right). \quad (6)$$

3 SM and ML

There is a striking similarity between the Fisher divergence and the KL divergence as in (1).

(2) ==>

$$D_F(p\|q_\theta) = \int_{\vec{x}} p(\vec{x}) \left|\nabla_{\vec{x}} \log \frac{p(\vec{x})}{q_\theta(\vec{x})}\right|^2 d\vec{x},$$

their difference lie in that instead of using the likelihood ratio, the Fisher divergence computes the l_2 norm of the gradient of the LR.

a deeper relation between them, and hence between SM and ML:

Theorem 1. *Let $\boxed{\vec{y} = \vec{x} + \sqrt{t}\vec{w}}$, for $t \geq 0$ and $\vec{w}$ a zero-mean white Gaussian vector. Denote $\tilde{p}_t(\vec{y})$ and $\tilde{q}_t(\vec{y})$ as the densities of $\vec{y}$ when $\vec{x}$ has distribution $p(\vec{x})$ or $q(\vec{x})$, respectively. Assume that $\tilde{p}_t(\vec{y})$ and $\tilde{q}_t(\vec{y})$ are smooth and fast decaying, such that their logarithms has growth at most polynomial at infinity. We have*

$$\boxed{\frac{d}{dt} D_{KL}(\tilde{p}_t(\vec{y})\|\tilde{q}_t(\vec{y})) = -\frac{1}{2} D_F(\tilde{p}_t(\vec{y})\|\tilde{q}_t(\vec{y})).} \quad (7)$$

As $\tilde{p}_0(\vec{y}) = p(\vec{x})$ and $\tilde{q}_0(\vec{y}) = q(\vec{x})$, we further have

$$\boxed{\frac{d}{dt} D_{KL}(\tilde{p}_t(\vec{y})\|\tilde{q}_t(\vec{y}))\Big|_{t=0} = -\frac{1}{2} D_F(p(\vec{x})\|q(\vec{x})).}$$

Lemma 1. *For any positive valued function $f(\vec{x})$ whose gradient $\nabla_{\vec{x}}$ and Laplacian $\triangle_{\vec{x}}$ are well defined,*

$$\frac{\triangle_{\vec{x}} f(\vec{x})}{f(\vec{x})} = \triangle_{\vec{x}} \log f(\vec{x}) + |\nabla_{\vec{x}} \log f(\vec{x})|^2. \quad (8)$$

Lemma 2. [Heat kernel] *For density $\tilde{p}_t(\vec{y})$ as defined in Theorem 1,*

$$\boxed{\frac{d}{dt}\tilde{p}_t(\vec{y}) = \frac{1}{2}\triangle_{\vec{y}}\tilde{p}_t(\vec{y}).} \quad (9)$$

Proof. [Theorem 1] For conciseness in notation, we drop references to variables $\vec{x}$ and $\vec{y}$ in the integration, the density functions, and the operators whenever this does not lead to ambiguity.

1, with Lemma 1,

$$\begin{aligned} D_F(\tilde{p}\|\tilde{q}) &= \int \tilde{p}\left(|\nabla \log \tilde{p}|^2 + |\nabla \log \tilde{q}|^2 + 2\triangle \log \tilde{q}\right) \\ &= \int \tilde{p}\left(|\nabla \log \tilde{p}|^2 + \frac{\triangle \tilde{q}}{\tilde{q}} + \triangle \log \tilde{q}\right). \quad (10) \end{aligned}$$

2, expanding the lhs of Eq.(7), we have:

$$\frac{d}{dt}D_{KL}(\tilde{p}\|\tilde{q}) = \int \frac{d}{dt}\tilde{p}\log\frac{\tilde{p}}{\tilde{q}} + \underbrace{\int \tilde{p}\frac{d}{dt}\log\tilde{p}}_{=0} - \int \tilde{p}\frac{d}{dt}\log\tilde{q}.$$

We can eliminate the second term by exchanging integration and differentiation of t:

$$\int \tilde{p}\frac{d}{dt}\log\tilde{p} = \int \tilde{p}\frac{\frac{d\tilde{p}}{dt}}{\tilde{p}} = \int \frac{d}{dt}\tilde{p} = \frac{d}{dt}\int \tilde{p} = 0.$$

As a result, there are three remaining terms in computing $\frac{d}{dt}D_{KL}(\tilde{p}\|\tilde{q})$, which we can further substitute using Lemma 2:

$$\begin{aligned} \frac{d}{dt}D_{KL}(\tilde{p}\|\tilde{q}) &= \int \frac{d}{dt}\tilde{p}\log\tilde{p} - \int \frac{d}{dt}\tilde{p}\log\tilde{q} - \int \tilde{p}\frac{d}{dt}\log\tilde{q} \\ &= \frac{1}{2}\left(\int \triangle\tilde{p}\log\tilde{p} - \int \triangle\tilde{p}\log\tilde{q} - \int \tilde{p}\frac{\triangle\tilde{q}}{\tilde{q}}\right). \quad (11) \end{aligned}$$

the first term in (11) =

$$\int \triangle\tilde{p}\log\tilde{p} = \sum_{i=1}^{d} \frac{\partial\tilde{p}}{\partial y_i}\log\tilde{p}(\vec{y})\Big|_{y_i=-\infty}^{y_i=\infty} - \int \nabla\tilde{p}^T\nabla\log\tilde{p}.$$

The limits in the first term = 0 given the smoothness and fast decay properties of $\tilde{p}(\vec{y})$. The remaining term =

$$\int \nabla\tilde{p}^T\nabla\log\tilde{p} = \int \tilde{p}\frac{(\nabla\tilde{p})^T}{\tilde{p}}\nabla\log\tilde{p} = \int \tilde{p}|\nabla\log\tilde{p}|^2.$$

The second term in (11) =

$$\int \triangle\tilde{p}\log\tilde{q} = \sum_{i=1}^{d} \frac{\partial\tilde{p}}{\partial y_i}\log\tilde{q}\Big|_{y_i=-\infty}^{y_i=\infty} - \int \nabla\tilde{p}^T\nabla\log\tilde{q}.$$

~~Applying integration by parts again to $\int \nabla\tilde{p}^T\nabla\log\tilde{q}$, we have~~

$$\int \nabla\tilde{p}^T\nabla\log\tilde{q} = \sum_{i=1}^{d} \tilde{p}\frac{\partial\log\tilde{q}}{\partial y_i}\Big|_{y_i=-\infty}^{y_i=\infty} - \int \tilde{p}\triangle\log\tilde{q}.$$

~~The limits at the boundary values are all zero due to the smoothness and fast decay properties of $\tilde{p}(\vec{y})$.~~ Now collecting all terms, we have $\int \triangle\tilde{p}\log\tilde{p} = -\int \tilde{p}|\nabla\log\tilde{p}|^2$ and $\int \triangle\tilde{p}\log\tilde{q} = \int \tilde{p}\triangle\log\tilde{q}$, Eq.(11) becomes

$$\frac{d}{dt}D_{KL}(\tilde{p}\|\tilde{q}) = -\frac{1}{2}\int \tilde{p}\left(|\nabla\log\tilde{p}|^2 + \triangle\log\tilde{q} + \frac{\triangle\tilde{q}}{\tilde{q}}\right).$$

Combining with (10). □

Theorem1 reveals some intriguing aspects of the relation between SM and ML by setting up a formal relation between the Fisher divergence and the KL divergence.

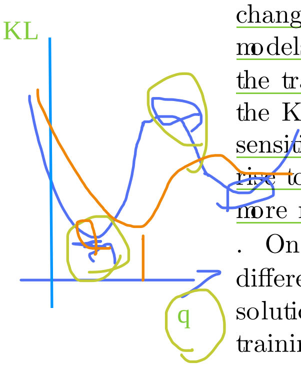
1. The effect of adding white Gaussian noise, $\sqrt{t}\vec{w}$, relates the density of $\vec{x}$ and $\vec{y}$ by

$$\tilde{p}_t(\vec{y}) = \int_{\vec{x}} \frac{1}{(2\pi t)^{d/2}} \exp\left(-\frac{|\vec{y}-\vec{x}|^2}{2t}\right) p(\vec{x}) d\vec{x},$$

   i.e., $\tilde{p}_t(\vec{y})$ is the convolution of $p(\vec{x})$ and a white Gaussian density N(0,t). It is known that this process forms a scale space [Lin94] over probability densities, which composes Gaussian smoothed density functions of different scale factor t. With large scale factors, small local structures in the density function are smoothed. So if parameter in q_θ to match p is sought in the scale space, it can put emphasis on large scale structures that survive the smoothing operation, and at the same time, spurious structures caused by the sampling effects of the training data are discounted. Indeed, this methodology has been adopted in clustering and non-parametric density estimation [LZX00] .

2. the Fisher divergence between two densities are non-negative ==> the KL divergence between two densities never increases as the scale factor increases (or equivalently, the SNR decreases). This is easy to understand, as the stronger noise is added, different signal sources get closer to the distribution of the noise and become more similar.

3. While ML aims to minimize the KL divergence directly, according to Theorem1, SM seeks to eliminate its derivative in the scale space at $t = 0$. In other words , SM looks for *stability*, where the optimal parameter θ leads to least changes in the KL divergence between the two models when a small amount of noise is present in the training data, while ML pursues *extremity* of the KL divergence. It is known that MLE is sensitive to noisy training data, which may give rise to many false extreme values, yet SM may be more robust to small perturbation in training data . On the other hand, due to this fundamental difference, SM and ML can lead to very different solutions for the same parametric density family and training data, the only known exception of which is when q_θ is Gaussian [Hyv05] .



4. There have been other interpretations of SM based on data corrupted by additive Gaussian noise. In [Hyv08] ,it was shown that SM is an approximation of the optimal parameter estimation when using the model as a prior in the inference of noise-free signal, and as the noise goes to infinitesimally small. In [RS07] ,SM was interpreted as searching parameters of q_θ so that when a Bayes LSE is constructed based on it, the overall MSE with the optimal estimator based on $p(\vec{x})$ is minimal (see Section 4). However , neither of these provide a direct relation between SM and ML.

5. Finally, as a special case of (7), when $\tilde{q}_t$ is set to a uniform distribution over the support of $\tilde{p}_t$ so that $\frac{d}{dt}D_{KL}(\tilde{p}\|\tilde{q}) = \frac{d}{dt}H(\tilde{p}_t)$ and $D_F(\tilde{p}\|\tilde{q}) = J(\tilde{p}_t)$, where $H(p)$ is the (Shannon) differential entropy and $J(\tilde{p}_t)$ is the Fisher information (3), we have

$$\frac{d}{dt}H(\tilde{p}_t) = \frac{1}{2}J(\tilde{p}_t).$$

   de Bruijn's identity: reveals a remarkable geometric relation between the differential entropy and the Fisher information: the former is related to the volume of the typical set of $\tilde{p}_t$, the latter is related to its surface area [CT06] .

4 Generalized SM

The SM learning can be generalized to a more flexible parametric learning methodology . Starting with the definition of the Fisher divergence (2), the main idea is to replace the gradient , which is a linear operator (functional) on density functions, with a general linear operator $\mathcal{L}$:

$$\boxed{D_{\mathcal{L}}(p\|q_\theta) = \int_{\vec{x}} p \left|\frac{\mathcal{L}p(\vec{x})}{p(\vec{x})} - \frac{\mathcal{L}q_\theta(\vec{x})}{q_\theta(\vec{x})}\right|^2 d\vec{x}.} \quad (12)$$

If $\vec{x}$ has discrete components, integration is substituted with summations. We term $D_{\mathcal{L}}$ the *generalized Fisher divergence*, and $\frac{\mathcal{L}p(\vec{x})}{p(\vec{x})}$ the *generalized score function*. Correspondingly, parametric learning using $D_{\mathcal{L}}$ is called the *generalized SM*. It is easy to see that $D_{\mathcal{L}} \geq 0$, and $=0$ when the two densities equal a.e.

The generalized Fisher divergence keeps several important computational advantages of the original Fisher divergence: 1, as an linear operator does not affect the normalizing partition function, it is canceled out in the generalized score function, and hence has no effect in the subsequent computation. 2, the generalized Fisher divergence can also be transformed to a form as an expectation of functions of the unnormalized model.

Definition 1. *Denote $\mathcal{F}^1$ and $\mathcal{F}^D$ as the space of all scalar-valued and D-variate functions for $\vec{x}$, respectively. $\mathcal{L} : \mathcal{F}^1 \mapsto \mathcal{F}^D$ is an linear operator. Further, assume that both $\mathcal{L}f(\vec{x})$ and $g(\vec{x})$ are square integrable, i.e., $\int_{\vec{x}} |\mathcal{L}f(\vec{x})|^2 d\vec{x} < \infty$ and $\int_{\vec{x}} |g(\vec{x})|^2 d\vec{x} < \infty$. The adjoint of $\mathcal{L}$, $\mathcal{L}^+ : \mathcal{F}^D \mapsto \mathcal{F}^1$, is a linear operator satisfying that $\forall f \in \mathcal{F}^1$ and $g \in \mathcal{F}^D$,*

$$\int_{\vec{x}} (\mathcal{L}f(\vec{x}))^T g(\vec{x}) d\vec{x} = \int_{\vec{x}} f(\vec{x})(\mathcal{L}^+ g(\vec{x})) d\vec{x}.$$

==>

Fact.

$$\begin{aligned} D_{\mathcal{L}}(p \| q_\theta) &= \int_{\vec{x}} p \left[\left| \frac{\mathcal{L}p}{p} \right|^2 + \left| \frac{\mathcal{L}q_\theta}{q_\theta} \right|^2 - 2 \left(\frac{\mathcal{L}p}{p} \right)^T \left(\frac{\mathcal{L}q_\theta}{q_\theta} \right) \right]. \\ &= \int_{\vec{x}} p \left[\left| \frac{\mathcal{L}p}{p} \right|^2 + \boxed{\left| \frac{\mathcal{L}q_\theta}{q_\theta} \right|^2 - 2\mathcal{L}^+ \left(\frac{\mathcal{L}q_\theta}{q_\theta} \right)} \right]. \end{aligned} \tag{13}$$

Though in principle we can use any linear operator, for parameter learning, we need operators leading to score functions that do not "lose" information about the original density.

Definition 2. *A linear operator $\mathcal{L}$ is* complete *if for two densities $p(\vec{x}), q(\vec{x})$, $\frac{\mathcal{L}\ p(x)}{p} = \frac{\mathcal{L}\ q(x)}{q}$ (a.e.) $\Rightarrow p(\vec{x}) = q(\vec{x})$ (a.e.). Otherwise, it is* incomplete.

Gradient. If we choose $\mathcal{L}$ to be the gradient operator, ∇, $D_{\mathcal{L}}$ reduces to the original Fisher divergence, and the corresponding GSM becomes the original SM.

Marginalization. Another choice for $\mathcal{L}$ is what we call the marginalization operator, $\mathcal{M} : \mathcal{F}^1 \mapsto \mathcal{F}^d$,

$$\mathcal{M}f(\vec{x}) := \begin{pmatrix} \vdots \\ \mathcal{M}_i f(\vec{x}) \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ \int_{x_i} f(\vec{x}) dx_i \\ \vdots \end{pmatrix}, \tag{14}$$

for any $f \in \mathcal{F}^1$. Integration is to be replaced with summation when $\vec{x}$ has discrete components. If $f(\vec{x}) = p(\vec{x})$ is a probability density over $\vec{x}$, $\mathcal{M}_i p(\vec{x})$ is the marginal density of $\vec{x}^{\setminus i}$ induced from $p(\vec{x})$, where $\vec{x}^{\setminus i}$ denotes the vector formed by dropping x_i from $\vec{x}$ ($\mathcal{M}$ as the marginalization operator). ==>

$$\frac{\mathcal{M}_i p(\vec{x})}{p(\vec{x})} = \frac{p(\vec{x}^{\setminus i})}{p(\vec{x})} = \frac{1}{p(x_i | \vec{x}^{\setminus i})}, \tag{15}$$

each component of $\mathcal{M}p(\vec{x})$ is the reciprocal of conditional density $p(x_i | \vec{x}^{\setminus i})$ induced from $p(\vec{x})$. Wlog, we assume here that $p(x_i | \vec{x}^{\setminus i}) \neq 0$.
Such conditional densities are known as the singleton conditionals. Therefore, minimizing the generalized Fisher divergence between $p(\vec{x})$ and $q(\vec{x})$ under the marginalization operator <==> match their corresponding singletons under the induce divergence.

Fact.
The marginalization operator is complete because of

Lemma 3 (Brook's Lemma [Bro64]) *The joint density of rvs $(x_1, \cdots, x_d)$ are completely determined by the ensemble of singleton conditional densities, $p(x_i | \vec{x}^{\setminus i}), \forall i$.*

Lemma 4. *For $g(\vec{x}) \in \mathcal{F}^d$, and denote $(g(\vec{x}))_i = g_i(\vec{x})$, and $f(\vec{x}) \in \mathcal{F}^1$,*==>

$$\int_{\vec{x}} \mathcal{M}(f(\vec{x}))^T g(\vec{x}) d\vec{x} = \int_{\vec{x}} f(\vec{x}) \sum_{i=1}^d \mathcal{M}_i g_i(\vec{x}) d\vec{x},$$

in other words, $\mathcal{M}^+ = \sum_{i=1}^d \mathcal{M}_i$.

Posterior mean. In [RS07] ,the generalized score function, $\mathcal{L}p/p$, was given a very different statistical interpretation . Assume p is the density for variable $\vec{y}$, which depends on a latent variable $\vec{x}$. The mean for the posterior distribution , $p_{X|Y}(\vec{x}|\vec{y})$, defined as $E(\vec{x}|\vec{y}) = \int_{\vec{x}} \vec{x} p_{X|Y}(\vec{x}|\vec{y}) d\vec{x}$, is shown to take the form $\mathcal{L}p(\vec{y})/p(\vec{y})$, where $\mathcal{L}$ is determined from the conditional density $p_{Y|X}(\vec{y}|\vec{x})$. As $E(\vec{x}|\vec{y})$ is the optimal estimator of $\vec{x}$ given $\vec{y}$ that minimizes the MSEs , optimizing the resulting generalized Fisher divergence == to find the optimal density q such that when used as a model for $\vec{y}$, it achieves the best performance in the inference of $\vec{x}$. Especially , when $\vec{y}$ is obtained by adding noise of some known density to $\vec{x}$, $\mathcal{L}$ and its adjoint may have simple closeform solutions (e.g., the additive Gaussian noise case corresponds to the Fisher divergence and original SM). However , not all complete linear operators suitable for $\mathcal{L}$ affords such an interpretation , such as the marginalization operator $\mathcal{M}$. On the other hand, it is hard to check the completeness of operators originated from the posterior means in general.

5 GSM for Discrete Data

There are two important restrictions in the original SM method , being that $\vec{x}$ has to be continuous-valued and the densities have to be differentiable in the space of $\mathcal{R}^d$. Due to these restrictions , one cannot directly apply SM to discrete data as $\nabla \log p(\vec{x})$ is not well defined in such cases . In this section , we show that using the marginalization operator $\mathcal{M}$ (Eq .(14)) with the GSM leads to a natural extension of SM to discrete data .

Consider discrete vectors $\vec{x} \in \{c_1, \cdots, c_m\}^d$ with density $p(\vec{x})$. Correspondingly, the integration in Eq.(12) is replaced with summation . As in the continuous case, learning is to find the optimal parameter for $q_\theta(\vec{x})$ that minimizes its generalized Fisher divergence with $p(\vec{x})$, which is

$$D_{\mathcal{M}}(p\|q_\theta) = \sum_{\vec{x}} p(\vec{x}) \sum_{i=1}^{d} \left(\frac{\mathcal{M}_i p(\vec{x})}{p(\vec{x})} - \frac{\mathcal{M}_i q_\theta(\vec{x})}{q_\theta(\vec{x})} \right)^2 .$$

Substituting with Eq(15), and using dummy variable ξ_i where we need to marginalize over the i th component of $\vec{x}$ in the inner integral, we have

$$D_{\mathcal{M}}(p\|q_\theta) = \sum_{\vec{x}} p(\vec{x}) \sum_{i=1}^{d} \sum_{\xi_i} \left(p(\xi_i|\vec{x}^{\setminus i}) - q_\theta(\xi_i|\vec{x}^{\setminus i}) \right)^2 \tag{16}$$

On the other hand, with Eq.(13) , Lemma 4 ==>

$$\sum_{\vec{x}} p(\vec{x}) \sum_{i=1}^{d} \left[\left(\cancel{\frac{\mathcal{M}_i p}{p}} \right)^2 + \left(\frac{\mathcal{M}_i q_\theta}{q_\theta} \right)^2 - 2\mathcal{M}_i \left(\frac{\mathcal{M}_i q_\theta}{q_\theta} \right) \right].$$

Dropping the first term ==>

$$\begin{aligned} & \sum_{\vec{x}} p(\vec{x}) \sum_{i=1}^{d} \left[\left(\frac{\mathcal{M}_i q_\theta}{q_\theta} \right)^2 - 2\mathcal{M}_i \left(\frac{\mathcal{M}_i q_\theta}{q_\theta} \right) \right] \\ = & \sum_{\vec{x}} p(\vec{x}) \sum_{i=1}^{d} \sum_{\xi_i} \left[\frac{1}{q_\theta^2(\xi_i|\vec{x}^{\setminus i})} - \frac{2}{q_\theta(\xi_i|\vec{x}^{\setminus i})} \right] \\ = & \sum_{\vec{x}} p(\vec{x}) \sum_{i=1}^{d} \sum_{\xi_i} \frac{1 - 2q_\theta(\xi_i|\vec{x}^{\setminus i})}{q_\theta^2(\xi_i|\vec{x}^{\setminus i})}. \end{aligned}$$

==>

$$\begin{aligned} & \sum_{\vec{x}} p(\vec{x}) \sum_{i=1}^{d} \sum_{\xi_i} \left(\frac{1 - 2q_\theta(\xi_i|\vec{x}^{\setminus i}) + q_\theta^2(\xi_i|\vec{x}^{\setminus i})}{q_\theta^2(\xi_i|\vec{x}^{\setminus i})} - 1 \right) \\ = & \sum_{\vec{x}} p(\vec{x}) \sum_{i=1}^{d} \sum_{\xi_i} \frac{(1 - q_\theta(\xi_i|\vec{x}^{\setminus i}))^2}{q_\theta^2(\xi_i|\vec{x}^{\setminus i})} - md \\ = & \sum_{\vec{x}} p(\vec{x}) \sum_{i=1}^{d} \sum_{\xi_i} \left(\frac{q_\theta(\sim\xi_i|\vec{x}^{\setminus i})}{q_\theta(\xi_i|\vec{x}^{\setminus i})} \right)^2 - md, \end{aligned} \tag{17}$$

(1-q ^ {-1}) ^ 2

where we use $q_\theta(\sim \xi_i|\vec{x}^{\setminus i})$ to shorthand the conditional probability for the i$^{\text{th}}$ element of $\vec{x}$ not taking value ξ_i. As $\sum_{\xi_i} q_\theta(\xi_i|\vec{x}^{\setminus i}) = 1$, and $\sum_{\xi_i} \frac{(1-q_\theta(\xi_i|\vec{x}^{\setminus i}))^2}{q_\theta^2(\xi_i|\vec{x}^{\setminus i})}$ reaches its minimum when $q_\theta(\xi_i|\vec{x}^{\setminus i})$ approaches constant value, minimizing the generalized Fisher divergence has an overall effect of balancing the values of the singleton conditional densities.

5.1 Relation with Ratio Matching

We compare the aforementioned discrete extension of generalized score matching with another similar method, known as *ratio matching* [Hyv07a] . Originally, the ratio matching algorithm was described for binary data. Here we describe an extended version that can be applied to general discrete data types. First define a scalar function ϕ, as $\phi(u) = \frac{1}{1+u}$, for $u \in$

$\mathcal{R}^+$. In ratio matching [Hyv07a] , we find the optimal parameter θ that minimizes

$$\sum_{\vec{x}} p(\vec{x}) \sum_{i=1}^{d} \sum_{\xi_i} \left[\phi\left(\frac{p(\xi_i, \vec{x}^{\setminus i})}{p(\sim\xi_i, \vec{x}^{\setminus i})} \right) - \phi\left(\frac{q_\theta(\xi_i, \vec{x}^{\setminus i})}{q_\theta(\sim\xi_i, \vec{x}^{\setminus i})} \right) \right]^2 .$$

Xi ≠ ξi, Xj = xj

This is slightly different from that used in [Hyv07a] , as a result of merging identical terms and dropping irrelevant terms. Using the definition of ϕ, we have $\phi\left(\frac{p(\xi_i, \vec{x}^{\setminus i})}{p(\sim\xi_i, \vec{x}^{\setminus i})} \right) = p(\sim\xi_i | \vec{x}^{\setminus i})$. The ratio matching objective function is:

$$\begin{aligned} & \sum_{\vec{x}} p(\vec{x}) \sum_{i=1}^{d} \sum_{\xi_i} \left(p(\sim\xi_i | \vec{x}^{\setminus i}) - q_\theta(\sim\xi_i | \vec{x}^{\setminus i}) \right)^2 \\ = & \sum_{\vec{x}} p(\vec{x}) \sum_{i=1}^{d} \sum_{\xi_i} \left(p(\xi_i | \vec{x}^{\setminus i}) - q_\theta(\xi_i | \vec{x}^{\setminus i}) \right)^2 . \end{aligned}$$

As shown in [Hyv07a] ==>

$$\begin{aligned} & \sum_{\vec{x}} p(\vec{x}) \sum_{i=1}^{d} \sum_{\xi_i} \left(\phi\left(\frac{q_\theta(\xi_i, \vec{x}^{\setminus i})}{q_\theta(\sim\xi_i, \vec{x}^{\setminus i})} \right) \right)^2 \\ = & \sum_{\vec{x}} p(\vec{x}) \sum_{i=1}^{d} \sum_{\xi_i} \left(1 - q_\theta(\xi_i | \vec{x}^{\setminus i}) \right)^2 + \text{const.} \end{aligned}$$

Note that the minimum of $\sum_{\xi_i} \left(1 - q_\theta(\xi_i | \vec{x}^{\setminus i}) \right)^2$ is also reached when $q_\theta(\xi_i | \vec{x}^{\setminus i})$ is a constant, therefore at optimum, ratio matching and our extension agree with each other, and both of them are different from maximum pseudo-likelihood. On the other hand, note that the objective function in ratio matching is quite different from that in SM [Hyv07a] . □

6 Conclusion

In this paper, we show two new results regarding the recently developed parameter learning method known as score matching . First, we establish a formal link between ML and SM, by showing the relation between the corresponding divergence functions . Specifically, we show that the Fisher divergence is the derivative of the KL divergence in a scale space with regards to the scale factor . This suggests that SM searches for parameters that are stable with small noise perturbation in training data . Second, we provide a generalization of SM by employing general linear operators in the Fisher divergence , and demonstrate a specific instantiation of the GSM to discrete data to be a more natural extension of SM to discrete data .

There are several directions that we hope to further explore in the future . 1, by using other type of diffusion kernels, it may be possible to establish a similar relation between the ML and the GSM . 2, the GSM provides more flexibility in applying the score matching methodology to different parameter estimation problems . Especially, it will be of great interest to study appropriate complete linear operators for specific high dimensional data models such as MRFs. 3, we are currently working on applying the GSM learning to practical problems such as bioinformatics and image modeling . We hope the work presented in this paper may deepen our understanding on score matching and help to extend its applications in machine learning and related fields .

□