My Math Paper

Your Name

February 14, 2024

Abstract

1 Some stuff

Stuff and more stuff aa

2 Validity of the loss

This is a very short proof of why what we discussed yesterday is correct. I will lay it down in full generality and we can summarize it to include in the paper if we want later.

We assume that there are $\mathbf{x} \in \mathbb{R}^{d_x}$ and $\mathbf{y} \in \mathbb{R}^{d_y}$, and that $p^*(\mathbf{x}, \mathbf{y})$ is the true distribution of the data.

2.1 Short review of denoising score-matching

The references on this are Vincent (2010); Song and Ermon (2020). I will give a very short review of the denoising score-matching loss.

First, assume that we have only samples of the $\{\mathbf{y}_i\}_{i=1}^n$. and our goal is to estimate the score function $\nabla_{\mathbf{y}} \log p^*(\mathbf{y})$. For the moment let's assume that to do this we have some class of functions

$$\mathcal{S} = \{\mathbf{s} : \mathbb{R}^{d_y} \to \mathbb{R}^{d_y}\}$$

and we simply want to find some good $\mathbf{s} \in \mathcal{S}$ that approximates the true score function.

A resonable objective is to minimize the quanity

$$l(\mathbf{s}, p) = \mathbb{E}_{p(\mathbf{y})}\left[||\nabla_{\mathbf{y}} \log p(\mathbf{y}) - \mathbf{s}(\mathbf{y})||^2\right] \tag{1}$$

where the p is just a stand in for some distribution. For example if we want to get the score of the true distribution we can minimize $\mathbf{s} = \arg\min_{\mathbf{s} \in \mathcal{S}} l(\mathbf{s}, p^*)$ where we have plugged in the true distribution p^*.

This is reasonable because the minimum of this quanity is neccesarily zero if and only if $\mathbf{s}(\mathbf{y}) = \nabla_{\mathbf{y}} \log p(\mathbf{y})$ almost surely. However, we can't use this objective since we don't have access to the true distribution of the data. As a consequence we use denoising score-matching (DSM) which was proposed by Vincent (2010).
The idea of DSM is that we have noise kernel $q(\tilde{\mathbf{y}}|\mathbf{y})$ which we can use to estimate the score of the distribution $q_p(\tilde{\mathbf{y}}) = \int q(\tilde{\mathbf{y}}|\mathbf{y})p(\mathbf{y})d\mathbf{y}$. Using this is better because one can show that minimizing $l(\mathbf{s}, q_p)$ i.e eq. (1) is equivalent to minimizing $l'(\mathbf{s}, q_p)$ where

$$l'(\mathbf{s}, q_p) = \mathbb{E}_{q(\tilde{\mathbf{y}}|\mathbf{y})p(\mathbf{y})} \left[||\nabla_{\mathbf{y}} \log q(\tilde{\mathbf{y}}|\mathbf{y}) - \mathbf{s}(\tilde{\mathbf{y}})||^2 \right] \tag{2}$$

The objective is the same in the sense that the value $\mathbf{s}$ that minimizes $l'(\mathbf{s}, q_p)$ has the property that $\mathbf{s}(\mathbf{y}) = \nabla_{\mathbf{y}} \log q_p(\mathbf{y})$. Of course, if we set $p = p^*$ and if the perturbation is small enough, then $q_p(\tilde{\mathbf{y}}) \approx p^*(\tilde{\mathbf{y}})$. And so having the score function for q_p is a good approximation for the score function of p^*.

2.2 How this connects with SDEs

I will ommit the details the setup for the SDEs (take a look at the paper for the details). Hence I'll describe this in a bit more generality.
Now, assume that we don't want to estimate the score for a single distribution $q_p(\mathbf{y})$, but rather I want to estimate the score for a family of distributions

$$Q(p) = \left\{ q_{p,t}(\tilde{\mathbf{y}}) | q_{p,t}(\tilde{\mathbf{y}}) = \int q_t(\tilde{\mathbf{y}}|\mathbf{y})p(\mathbf{y})d\mathbf{y}, t \in \mathcal{T} \right\}$$

where $t \in \mathcal{T}$ is a parameter.[1] In other words we have a family of distributions with different perturbation kernels.
Then the reasonable thing to do is to define a new family

$$\mathcal{S} = \{\mathbf{s} : \mathbb{R}^{d_y} \times \mathbb{R} \to \mathbb{R}^{d_y}\}$$

And to minimize the quantity

$$l_{\text{family}}(\mathbf{s}, p) = \int_{\mathcal{T}} \mathbb{E}_{q_t(\tilde{\mathbf{y}}|\mathbf{y})p(\mathbf{y})} \left[||\nabla_{\mathbf{y}} \log q_t(\tilde{\mathbf{y}}|\mathbf{y}) - \mathbf{s}(\tilde{\mathbf{y}}, t)||^2 \right] dt \tag{3}$$

Where the integral is taken with respect to the domain of t. Now, notice that, from the explanation above (and as it was shown in Song and Ermon (2020); Vincent (2010)) we know that for any given t the value of $\mathbf{s}(\tilde{\mathbf{y}}, t)$ that minimizes the inner expectation is $\nabla_{\mathbf{y}} \log q_{p,t}(\tilde{\mathbf{y}})$, then it must be the case that defining $\mathbf{s}(\tilde{\mathbf{y}}, t) = \nabla_{\mathbf{y}} \log q_{p,t}(\tilde{\mathbf{y}})$ for all $\tilde{\mathbf{y}}$ and t will minimize the quantity eq. (4). In which case it will be zero.

[1] I am also being a little bit loose with my tilde notation but I hope it is clear

Now, integrals are expectations and we will need to optimize this and weigh some values of t more than others so in reality what we will do is minimize

$$l_{\text{family}}(\mathbf{s}, p) = \mathbb{E}_{p(t)}\mathbb{E}_{q_t(\tilde{\mathbf{y}}|\mathbf{y})p(\mathbf{y})}\left[||\nabla_{\mathbf{y}} \log q_t(\tilde{\mathbf{y}}|\mathbf{y}) - \mathbf{s}(\tilde{\mathbf{y}}, t)||^2\right] \tag{4}$$

But the exact same reasoning applies. To summarize, in this section we have shown that if we have a family of distributions $Q(p)$ and we want to estimate the score for all of them simultaneously then defining a family of score functions $\mathbf{s}(\tilde{\mathbf{y}}, t) = \nabla_{\mathbf{y}} \log q_{p,t}(\tilde{\mathbf{y}})$ and choosing the one that minimizes eq. (4) will give us the score for all of them. In particular, it should be clear that the score function is a minima for such an objective and (with a little bit more formalism) one could prove it is unique almost everywhere (this follows from the fact that integral is zero).

An even larger family of distributions

Ok so, thus far we have specified how to handle a family of distributions where the kernel varies. This is the setting used by song. However, we can also consider a family of distributions where the density varies as well. This is the setting we are in and the one that we care about.
Even though all of the above is rigurous and true I will be a little bit more formal here to avoid any confusion.
Now, assume that we are now talking about a single p. In particular, we are talking about the $p^*(\mathbf{y}|\mathbf{x})$ so we can ommit the dependence on p^* in subscript (but it should be the same as in the previous section).
First, let's define a family of distributions Q as the set of all distributions

$$Q = \left\{q_{t,x}(\tilde{\mathbf{y}}) | q_{t,x}(\tilde{\mathbf{y}}) = \int q_t(\tilde{\mathbf{y}}|\mathbf{y})p^*(\mathbf{y}|\mathbf{x})d\mathbf{y}, t \in \mathcal{T}, x \in \mathcal{X} \subseteq \mathbb{R}^{d_x}\right\}$$

Where we assume that $\mathcal{X}$ is the subset of $\mathbb{R}^{d_x}$ where the data is supported, $q_t(\tilde{\mathbf{y}}|\mathbf{y})$ are well defined perturbation kernels and $p^*(\mathbf{y}|\mathbf{x})$ is the conditional density of the data.
Now define

$$\mathcal{S}_{t,x} = \{\mathbf{s} : \mathbb{R}^{d_y} \times \mathbb{R}^{d_x} \times \mathbb{R} \to \mathbb{R}^{d_y}\}$$

In words, this is the set of all functions that take an $\mathbf{x}$ and t and return a score function for $\mathbf{y}$.
Then we can define the loss

$$\mathcal{L}(\mathbf{s}) = \mathbb{E}_{p^*(\mathbf{x})}\mathbb{E}_{p(t)}\mathbb{E}_{p^*(\mathbf{y}|\mathbf{x})q_t(\tilde{\mathbf{y}}|\mathbf{y})}\left[||\nabla_{\mathbf{y}} \log q_t(\tilde{\mathbf{y}}|\mathbf{y}) - \mathbf{s}(\tilde{\mathbf{y}}, \mathbf{x}, t)||^2\right] \tag{5}$$

Here is the main result of this section:

Theorem 1. *The value* $\mathbf{s} \in \mathcal{S}_{t,x}$ *that minimizes* $\mathcal{L}(\mathbf{s})$ *is such that*

$$\mathbf{s}(\tilde{\mathbf{y}}, \mathbf{x}, t) = \nabla_{\mathbf{y}} \log q_{t,x}(\tilde{\mathbf{y}})$$

almost everywhere.

Remark 1. *This means that if we get a good estimate of* $\mathbf{s}(\tilde{\mathbf{y}}, \mathbf{x}, t)$ *then by fixing* $\mathbf{x}$ *we can get a good estimate of the score function for the conditional distribution* $q_{t,x}(\tilde{\mathbf{y}})$*. Looking at the definition of* $q_{t,x}(\tilde{\mathbf{y}})$ *we can see that with* x *fixed this reduces to the problem in the previous section. Hence everything works out and we can be confident that what we are doing is legit. Importantly it also means that by fixing* $\mathbf{x}$ *we can then use the techniques in song to sample form* $p^*(\mathbf{y}|\mathbf{x})$*.*

Proof. The proof is basically the same as in the previous section. Fix $\mathbf{x}$ and t and notice that the two inner expectations reduce to the loss in eq. (2). Specifically, by the results in section 2.1 we know that

$$\mathbb{E}_{q_t(\tilde{\mathbf{y}}|\mathbf{y})p^*(\mathbf{y}|\mathbf{x})}\left[||\nabla_{\mathbf{y}} \log q_t(\tilde{\mathbf{y}}|\mathbf{y}) - \mathbf{s}(\tilde{\mathbf{y}}, \mathbf{x}, t)||^2\right]$$

Is minimized by any $s \in \mathcal{S}$ such that $\mathbf{s}(\tilde{\mathbf{y}}, x, t) = \nabla_{\mathbf{y}} \log q_{t,x}(\tilde{\mathbf{y}})$ in which case it is zero. Now, we know that eq. (5) is greater than or equal to zero. Hence, the minimum is achieved.
Thus we have shown that $s(\tilde{\mathbf{y}}, \mathbf{x}, t) = \nabla_{\mathbf{y}} \log q_{t,x}(\tilde{\mathbf{y}})$ is a minimum of the loss $\mathcal{L}$. Now, clearly for any $\mathbf{s}$ we have that $\mathcal{L} \geq 0$ and that the inner expectation is always greater than or equal to zero. Hence we know that if the minimum is achieved then it must be the case that

$$0 = \mathbb{E}_{q_t(\tilde{\mathbf{y}}|\mathbf{y})p^*(\mathbf{y}|\mathbf{x})}\left[||\nabla_{\mathbf{y}} \log q_t(\tilde{\mathbf{y}}|\mathbf{y}) - \mathbf{s}(\tilde{\mathbf{y}}, \mathbf{x}, t)||^2\right]$$

almost everywhere (with respect to the measure $p^*(\mathbf{x})$ and $p(t)$)[2] However, we know then by the results in section 2.1 that this implies that $\mathbf{s}(\tilde{\mathbf{y}}, \mathbf{x}, t) = \nabla_{\mathbf{y}} \log q_{t,x}(\tilde{\mathbf{y}})$ almost everywhere with respect to $q^*(\tilde{\mathbf{y}}|\mathbf{y})p^*(\mathbf{y}|\mathbf{x})$. Which is the result we wanted to show.

□

References

Song, Y. and Ermon, S. (2020). Generative modeling by estimating gradients of the data distribution.

Vincent, P. (2010). A connection between score matching and denoising autoencoders.

[2] We are using the fact that if $f \geq 0$ and $\mathbb{E}[f] = 0$ then $f = 0$ almost everywhere. For any f that is measurable.