Generative Diffusion Models (18): Score Matching = Conditional Score Matching

Su Jianlin

February 28, 2023

In previous introductions, we have frequently mentioned "Score Matching" and "Conditional Score Matching." These are concepts that often appear in diffusion models, energy-based models, and similar frameworks. In particular, many articles directly state that the training objective of diffusion models is "Score Matching," but in fact, the training objective of current mainstream diffusion models such as DDPM is actually "Conditional Score Matching."

So, what exactly is the relationship between "Score Matching" and "Conditional Score Matching"? Are they equivalent? This article discusses this issue in detail.

1 Score Matching

First, Score Matching refers to the training objective:

$$\mathbb{E}_{\boldsymbol{x}_t\sim p_t(\boldsymbol{x}_t)}\left[\left\|\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t) - \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\right\|^2\right] \tag{1}$$

where $\boldsymbol{\theta}$ represents the training parameters. Clearly, Score Matching aims to learn a model $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$ to approximate $\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t)$. Here, $\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t)$ is what we call the "score."

In the context of diffusion models, $p_t(\boldsymbol{x}_t)$ is given by:

$$p_t(\boldsymbol{x}_t) = \int p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)p_0(\boldsymbol{x}_0)d\boldsymbol{x}_0 = \mathbb{E}_{\boldsymbol{x}_0\sim p_0(\boldsymbol{x}_0)}\left[p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)\right] \tag{2}$$

where $p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)$ is generally a simple distribution with a known analytical probability density (such as a conditional normal distribution), and $p_0(\boldsymbol{x}_0)$ is also a given distribution, typically representing the training data. This means we can only sample from $p_0(\boldsymbol{x}_0)$ but do not know its specific analytical expression.

According to Equation (2), we can derive:

$$\begin{aligned}
\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t) &= \frac{\nabla_{\boldsymbol{x}_t} p_t(\boldsymbol{x}_t)}{p_t(\boldsymbol{x}_t)} \\
&= \frac{\int p_0(\boldsymbol{x}_0)\nabla_{\boldsymbol{x}_t} p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)d\boldsymbol{x}_0}{p_t(\boldsymbol{x}_t)} \\
&= \frac{\int p_0(\boldsymbol{x}_0)\nabla_{\boldsymbol{x}_t} p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)d\boldsymbol{x}_0}{\int p_0(\boldsymbol{x}_0)p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)d\boldsymbol{x}_0} \\
&= \frac{\mathbb{E}_{\boldsymbol{x}_0\sim p_0(\boldsymbol{x}_0)}\left[\nabla_{\boldsymbol{x}_t} p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)\right]}{\mathbb{E}_{\boldsymbol{x}_0\sim p_0(\boldsymbol{x}_0)}\left[p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)\right]}
\end{aligned} \tag{3}$$

Based on our assumptions, both $\nabla_{\boldsymbol{x}_t} p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)$ and $p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)$ have known analytical forms. Therefore, in theory, we could estimate $\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t)$ by sampling $\boldsymbol{x}_0$. However, since this involves the division of two expectations, it is a biased estimate (refer to "Briefly on Unbiased and Biased Estimation"). Consequently, a sufficiently large number of points must be sampled to make an accurate estimate. Thus, if Equation (1) is used directly as the training objective, a large batch size is required to achieve good results.

2 Conditional Score

In practice, the training objective used by general diffusion models is "Conditional Score Matching":

$$\mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{x}_t\sim p_0(\boldsymbol{x}_0)p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\left\|\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t|\boldsymbol{x}_0) - \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)\right\|^2\right] \tag{4}$$

By assumption, $\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)$ has a known analytical form, so the above objective is directly usable by sampling pairs of $(\boldsymbol{x}_0,\boldsymbol{x}_t)$ for estimation. Notably, this is an unbiased estimate, which means it does not particularly rely on a large batch size, making it a more practical training objective.

To analyze the relationship between "Score Matching" and "Conditional Score Matching," we also need another identity for $\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t)$:

$$\begin{aligned}
\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t) =& \frac{\nabla_{\boldsymbol{x}_t} p_t(\boldsymbol{x}_t)}{p_t(\boldsymbol{x}_t)} \\
=& \frac{\int p_0(\boldsymbol{x}_0)\nabla_{\boldsymbol{x}_t} p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)d\boldsymbol{x}_0}{p_t(\boldsymbol{x}_t)} \\
=& \frac{\int p_0(\boldsymbol{x}_0)p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)d\boldsymbol{x}_0}{p_t(\boldsymbol{x}_t)} \\
=& \int p_t(\boldsymbol{x}_0|\boldsymbol{x}_t)\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)d\boldsymbol{x}_0 \\
=& \mathbb{E}_{\boldsymbol{x}_0\sim p_t(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)\right]
\end{aligned} \tag{5}$$

3 Inequality Relationship

First, we can quickly prove the first result between the two: Conditional Score Matching is an upper bound of Score Matching. This implies that minimizing Conditional Score Matching, to some extent, also minimizes Score Matching.

The proof is not difficult, as we already demonstrated in "Generative Diffusion Models (16): W-Distance $\leq$ Score Matching":

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}_t\sim p_t(\boldsymbol{x}_t)}\left[\left\|\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t) - \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)\right\|^2\right] \\
=&\mathbb{E}_{\boldsymbol{x}_t\sim p_t(\boldsymbol{x}_t)}\left[\left\|\mathbb{E}_{\boldsymbol{x}_0\sim p_t(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)\right] - \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)\right\|^2\right] \\
\leq&\mathbb{E}_{\boldsymbol{x}_t\sim p_t(\boldsymbol{x}_t)}\mathbb{E}_{\boldsymbol{x}_0\sim p_t(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[\left\|\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t|\boldsymbol{x}_0) - \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)\right\|^2\right] \\
=&\mathbb{E}_{\boldsymbol{x}_0\sim p_0(\boldsymbol{x}_0),\boldsymbol{x}_t\sim p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\left\|\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t|\boldsymbol{x}_0) - \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)\right\|^2\right]
\end{aligned} \tag{6}$$

The first equality follows from identity (5), the second inequality follows from the generalization of the mean square inequality or Jensen's inequality, and the third equality follows from Bayes' rule.

4 Equivalence Relationship

A few days ago, during a discussion about Score Matching in a WeChat group, a member pointed out: The difference between Conditional Score Matching and Score Matching is a constant independent of optimization, so the two are actually completely equivalent! When I first heard this conclusion, I was quite surprised—that they are equivalent, not just in an upper-bound relationship. Furthermore, after I tried to prove it, I found the proof process to be quite simple!

First, regarding Score Matching, we have:

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{x}_t\sim p_t(\boldsymbol{x}_t)}\left[\|\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t) - \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)\|^2\right]\\ =&\mathbb{E}_{\boldsymbol{x}_t\sim p_t(\boldsymbol{x}_t)}\left[\|\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t)\|^2 + \|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)\|^2 - 2\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)\cdot\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t)\right]\end{aligned} \tag{7}$$

Then, regarding Conditional Score Matching, we have:

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{x}_t\sim p_0(\boldsymbol{x}_0)p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\|\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t|\boldsymbol{x}_0) - \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)\|^2\right]\\ =&\mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{x}_t\sim p_0(\boldsymbol{x}_0)p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\|\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)\|^2 + \|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)\|^2 - 2\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)\cdot\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)\right]\\ =&\mathbb{E}_{\boldsymbol{x}_t\sim p_t(\boldsymbol{x}_t),\boldsymbol{x}_0\sim p_t(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[\|\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)\|^2 + \|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)\|^2 - 2\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)\cdot\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)\right]\\ =&\mathbb{E}_{\boldsymbol{x}_t\sim p_t(\boldsymbol{x}_t)}\left[\begin{aligned}&\mathbb{E}_{\boldsymbol{x}_0\sim p_t(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[\|\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)\|^2\right] + \|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)\|^2\\ &- 2\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)\cdot\mathbb{E}_{\boldsymbol{x}_0\sim p_t(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)\right]\end{aligned}\right]\\ =&\mathbb{E}_{\boldsymbol{x}_t\sim p_t(\boldsymbol{x}_t)}\left[\mathbb{E}_{\boldsymbol{x}_0\sim p_t(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[\|\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)\|^2\right] + \|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)\|^2 - 2\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)\cdot\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t)\right]\end{aligned} \tag{8}$$

Taking the difference between the two, we find the result is:

$$\mathbb{E}_{\boldsymbol{x}_t\sim p_t(\boldsymbol{x}_t)}\left[\mathbb{E}_{\boldsymbol{x}_0\sim p_t(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[\|\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t|\boldsymbol{x}_0)\|^2\right] - \|\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t)\|^2\right] \tag{9}$$

It is independent of the parameter $\boldsymbol{\theta}$. So minimizing the Score Matching objective is theoretically equivalent to minimizing the Conditional Score Matching objective. According to the group member, this result first appeared in the article "A Connection Between Score Matching and Denoising Autoencoders".

Since the two are theoretically equivalent, does this mean our previous statement that "Score Matching" requires a larger batch size than "Conditional Score Matching" is incorrect? Not necessarily. If we still estimate $\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t)$ directly from Equation (3) and then perform Score Matching, the result is indeed still biased and depends on a large batch size. However, when we expand and further simplify the objective (1), we gradually transform the biased estimate into an unbiased one, which then does not rely heavily on the batch size. In other words, although the two objectives are theoretically equivalent, from the perspective of statistics, they belong to different types of estimators; their equivalence is an exact equivalence only when the number of samples tends to infinity.

5 Summary

This article mainly analyzes the connection between the two training objectives: "Score Matching" and "Conditional Score Matching."