# Comparing Implicit and Denoising Score-Matching Objectives

**Artem Artemev**
MediaTek Research
Cambridge, CB23 6DW, UK
artem.artemev@mtkresearch.com

**Ayan Das**
MediaTek Research
Cambridge, CB23 6DW, UK
ayan.das@mtkresearch.com

**Farhang Nabiei**
MediaTek Research
Cambridge, CB23 6DW, UK
farhang.nabiei@mtkresearch.com

**Alberto Bernacchia**
MediaTek Research
Cambridge, CB23 6DW, UK
alberto.bernacchia@mtkresearch.com

## Abstract

Score estimation has led to several state-of-the-art generative models, particularly in computer vision. Compared to maximum likelihood, one of the key advantages of score estimation is that it does not require the calculation of a normalization factor. However, explicit score matching necessitates knowledge of the true score of the data distribution, which is typically unavailable. To address this challenge, various approaches have been proposed to approximate the score-matching loss. The two main approaches are implicit score matching (ISM) and denoising score matching (DSM), which differ in their bias, making direct comparison difficult. In this work, we expand the ISM and DSM losses to remove the constant bias between them. While it is known that they are asymptotically equivalent, we show empirically that, in finite data regimes, differences in variance make DSM loss sensitive to the noise scale. ISM does not require noised data to learn and is more robust than DSM when learning from noised data, particularly when the noise scale is relatively small.

## 1 Introduction

Parameter estimation of probabilistic models is usually performed by *maximum likelihood* estimation. Given a probabilistic model $p_{\boldsymbol{\theta}}(\mathbf{x})$ parametrized by $\theta$ and the true data distribution $q(\mathbf{x})$, we seek to minimize the following loss:

$$\mathcal{L}(\boldsymbol{\theta}) = -\mathbb{E}_{q(\mathbf{x})} \log\left(p_{\boldsymbol{\theta}}(\mathbf{x})\right) \tag{1}$$

Estimating this loss is relatively straightforward, with a sample of data points $\mathbf{x} \sim q(\mathbf{x})$, when the probabilistic model has a simple form, such as Gaussian or Bernoulli (for regression and classification). However, the normalization factors of more complex probabilistic models are intractable, making maximum likelihood estimation usually infeasible, with some exceptions, such as Normalizing Flow models [1, 6]).

An alternative to maximum likelihood is *score matching*, which involves minimizing the following loss function (also called *Fisher divergence* or *relative Fisher information*):

$$\mathcal{L}_{\text{ESM}}(\boldsymbol{\theta}) = \mathbb{E}_{q(\mathbf{x})} \frac{1}{2} \left|\nabla_{\mathbf{x}} \log\left(p_{\boldsymbol{\theta}}(\mathbf{x})\right) - \nabla_{\mathbf{x}} \log\left(q(\mathbf{x})\right)\right|^2 \tag{2}$$

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

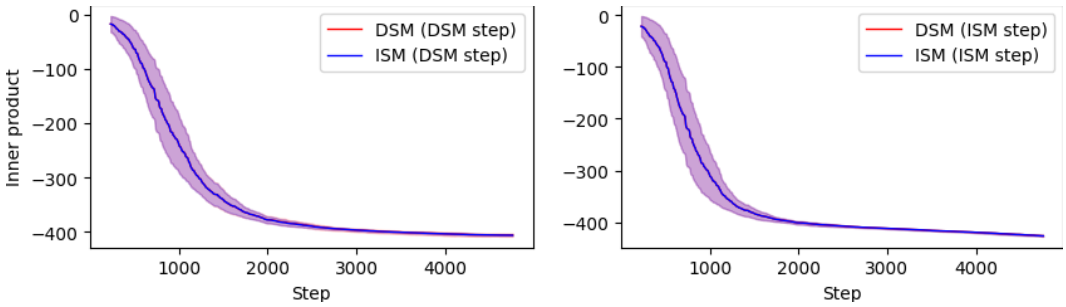


Figure 1: The *fitting* term (inner product) of ISM and DSM evaluated at every training step, where the model is updated with (left) DSM gradients and (right) ISM gradients.

Score matching aims to find the gradient of the log-likelihood (the *score*) instead of the log-likelihood itself, as knowledge of the score is often sufficient for sampling from the probabilistic model [3, 9, 12]. However, the loss in Eq. 2 cannot be directly estimated because the score of the true distribution is unknown. In particular, the score function cannot be evaluated at sample data points.

Two approaches have been proposed to estimate the loss function in Eq. 2: 1) *implicit* score matching (ISM) [3] and 2) *denoising* score matching (DSM) [9, 11]. The former has led to several generative models (e.g. [4], [10]). However, ISM is known to be infeasible for large-scale problems, as it requires the trace of log-likelihood Hessian to compute the loss (Eq. 3). On the other hand, DSM underpins the diffusion models, which have been extensively studied and are considered state-of-the-art for generative computer vision tasks (e.g. [2], [7], [8], [9]).

Although ISM and DSM both approximate the score-matching loss (Eq. 2), comparing them is not straightforward due to their different biases. In this study, we eliminate the bias from the DSM loss and show that it becomes asymptotically equivalent to ISM loss on noisy data. However, the DSM loss and gradient variance is sensitive to the noise scale and can explode for low-scale noise. ISM can be used to learn from noisy or noiseless data, and it is more robust than DSM against the noise scale. Our results encourage further studies to improve the computational cost of ISM.

## 2 Methodology

### 2.1 Implicit score matching

For infinite data and under the following mild assumptions about $q(\mathbf{x})$ and $s(\mathbf{x}, \boldsymbol{\theta}) = \nabla_{\mathbf{x}} \log\left(p_{\boldsymbol{\theta}}(\mathbf{x})\right)$:

**Assumption 1.** *$q(\mathbf{x})$ and $s(\mathbf{x}, \boldsymbol{\theta})$ are differentiable, $\mathbb{E}_{q(\mathbf{x})}|\nabla_{\mathbf{x}} log(q(\mathbf{x}))|^2$ is finite, $\mathbb{E}_{q(\mathbf{x})}|s(\mathbf{x}, \boldsymbol{\theta})|^2$ is finite for any $\boldsymbol{\theta}$, and, importantly, $\lim_{|x|\to\infty} q(\mathbf{x}) s(\mathbf{x}, \boldsymbol{\theta}) = 0$.*

[3] showed that the score-matching loss in Eq. 2, $\mathcal{L}_{ESM}(\boldsymbol{\theta})$, is asymptotically equivalent to ISM loss defined in Eq. 3 plus some constant, $C_1$, that does not depend on parameters, $\boldsymbol{\theta}$.

$$\mathcal{L}_{\text{ESM}}(\boldsymbol{\theta}) = \mathcal{L}_{\text{ISM}}(\boldsymbol{\theta}) + C_1 = \mathbb{E}_{q(\mathbf{x})} \left\{ \text{Tr}\left[\nabla_{\mathbf{x}}^2 \log\left(p_{\boldsymbol{\theta}}(\mathbf{x})\right)\right] + \frac{1}{2} \left|s(\mathbf{x}, \boldsymbol{\theta})\right|^2 \right\} + C_1 \tag{3}$$

Unlike ESM, the ISM loss does not require a regression target depending on $q(\mathbf{x})$ and can be computed for a dataset of samples from an unknown distribution. The property in Eq. 3 can be derived from equality in Proposition 1:

**Proposition 1.** *For a distribution, $q(\mathbf{x})$, and its parameterized score, $s(\mathbf{x}, \boldsymbol{\theta})$, under the assumption 1, it can be shown that:*

$$\mathbb{E}_{q(\mathbf{x})} \left\langle s(\mathbf{x}, \boldsymbol{\theta}), \nabla_{\mathbf{x}} \log(q(\mathbf{x})) \right\rangle = -\mathbb{E}_{q(\mathbf{x})} \text{Tr}\left[\nabla_{\mathbf{x}}^2 \log\left(p_{\boldsymbol{\theta}}(\mathbf{x})\right)\right] \tag{4}$$

The left side of Eq. 4 represents the inner product term from the expansion of quadratic loss in Eq. 2. The right side of Eq. 4 is the trace of log-likelihood Hessian in ISM loss, Eq. 3. The proof of this proposition is provided in the appendix A.

### 2.2 Denosing Score Matching

Denoising score matching is another objective for score matching that uses pairs of clean and corrupted samples, $(\mathbf{x}, \mathbf{z})$ [11]. For the joint distribution $q_\sigma(\mathbf{x}, \mathbf{z}) = q(\mathbf{x})q_\sigma(\mathbf{z}|\mathbf{x})$, DSM is defined as:

$$\mathcal{L}_{\text{DSM}}(\boldsymbol{\theta}) = \mathbb{E}_{q_\sigma(\mathbf{z},\mathbf{x})}\frac{1}{2}\left|s(\mathbf{z},\boldsymbol{\theta}) - \nabla_{\mathbf{z}}\log\left(q_\sigma(\mathbf{z}|\mathbf{x})\right)\right|^2 \tag{5}$$

DSM uses the score at the corrupted point to move it toward the clean sample. Selecting a Gaussian density for corruption, $q_\sigma(\mathbf{z}|\mathbf{x}) = \mathcal{N}(z; x, \sigma^2 I)$, we can write the second term of DSM loss in Eq. 5 as:

$$\frac{\partial \log q_\sigma(\mathbf{z}|\mathbf{x})}{\partial z} = -\frac{\mathbf{z}-\mathbf{x}}{\sigma^2} = -\frac{\epsilon}{\sigma} \tag{6}$$

where $\epsilon \sim \mathcal{N}(\epsilon, 0, I)$ is standard Gaussian noise. Thus, the DSM loss in Eq. 5 can be expanded as

$$\mathcal{L}_{\text{DSM}}(\boldsymbol{\theta}) = \mathbb{E}_{q_\sigma(\mathbf{z})}\frac{1}{2}|s(\mathbf{z},\boldsymbol{\theta})|^2 + \mathbb{E}_{q(\mathbf{x})}\mathbb{E}_{\mathcal{N}(\epsilon;0,I)}\langle s(\mathbf{z},\boldsymbol{\theta}), \frac{\epsilon}{\sigma}\rangle + \mathbb{E}_{\mathcal{N}(\epsilon,0,I)}\frac{|\epsilon|^2}{2\sigma^2} \tag{7}$$

### 2.3 Comparing ISM and DSM for noisy data

Similar to DSM, we can rewrite the ISM loss in Eq. 3 over the distribution of corrupted samples $\mathbf{z} \sim q_\sigma(\mathbf{z})$ :

$$\mathcal{L}_{\text{ISM}}(\boldsymbol{\theta}) = \underbrace{\mathbb{E}_{q(\mathbf{z})}\frac{1}{2}|s(\mathbf{z},\boldsymbol{\theta})|^2}_{\text{Norm}} + \underbrace{\mathbb{E}_{q(\mathbf{z})}\sum_i \frac{\partial s(\mathbf{z},\boldsymbol{\theta})_i}{\partial \mathbf{z}_i}}_{\text{Fitting}} \tag{8}$$

The first term in the ISM loss penalizes the score *norm*, while the second term is responsible for fitting the score to the noisy data [4]. Similarly, we remove the last term from expanded DSM loss in eq. 7 to obtain the modified DSM objective, $\mathcal{L}_{\text{DSM-S}}(\boldsymbol{\theta})$ (DSM-S) 9.

$$\mathcal{L}_{\text{DSM-S}}(\boldsymbol{\theta}) = \underbrace{\mathbb{E}_{q_\sigma(\mathbf{z})}\frac{1}{2}|s(\mathbf{z},\boldsymbol{\theta})|^2}_{\text{Norm}} + \underbrace{\mathbb{E}_{q(\mathbf{x})}\mathbb{E}_{\mathcal{N}(\epsilon;0,I)}\langle s(\mathbf{z},\boldsymbol{\theta}), \frac{\epsilon}{\sigma}\rangle}_{\text{Fitting}} \tag{9}$$

The eliminated term, $C_{DSM} = \mathbb{E}_{\mathcal{N}(\epsilon,0,I)}(|\epsilon|^2/2\sigma^2)$, does not depend on the parameters, $\boldsymbol{\theta}$, and it is not expected to affect training dynamics. However, it might still introduce variance to the DSM loss due to sampling error. Thus, we remove this term in our experiments. The score *norm* regularization terms in ISM and DSM-S (Eq. 8 & 9) are equivalent. The difference lies in the *fitting* term. Using proposition 4, we can show that the *fitting* term in ISM and DSM are asymptotically equivalent for infinite samples.

**Proposition 2.** *For $\mathbf{z} \sim q_\sigma(\mathbf{z})$, where $q_\sigma(\mathbf{z}) = q_\sigma(\mathbf{z}|\mathbf{x})q(\mathbf{x}) = \mathcal{N}(\mathbf{z}; \mathbf{x}, \sigma^2 I)q(\mathbf{x})$, if $q_\sigma(\mathbf{z})$ and its parameterized score $s(\mathbf{z}, \boldsymbol{\theta})$ satisfy the assumptions 1, we have:*

$$\mathbb{E}_{q(\mathbf{z})}\sum_i \frac{\partial s(\mathbf{z},\boldsymbol{\theta})_i}{\partial \mathbf{z}_i} = \mathbb{E}_{q(\mathbf{x})}\mathbb{E}_{\mathcal{N}(\epsilon,0,I)}\langle s(\mathbf{z},\boldsymbol{\theta}), \frac{\epsilon}{\sigma}\rangle \tag{10}$$

*where $\epsilon \sim \mathcal{N}(\epsilon; 0, I)$.*

We derive the proposition in appendix B. Although the fitting terms in both ISM and DSM are expected to be equal for infinite sample, their variances are not the same. Thus, the sampling error arising from the finite data and mini-batch size can result in significantly different fitting behavior for DSM and ISM. Notably, the variance of the *fitting* term in ISM is zero for Gaussian data, i.e. $\tilde{\mathbf{z}} \sim \mathcal{N}(\tilde{\mathbf{z}}; \mu, \Sigma)$, as the Hessian of the log-likelihood is constant (i.e. $-\Sigma^{-1}$). On the other hand, the DSM variance for Gaussian data is non-zero and depends on the sample variance. Therefore, ISM has better stability and robustness to the noise for Gaussian data than DSM. However, their learning behavior for non-Gaussian data is not clear. In this study, we investigate this with simple experiments.

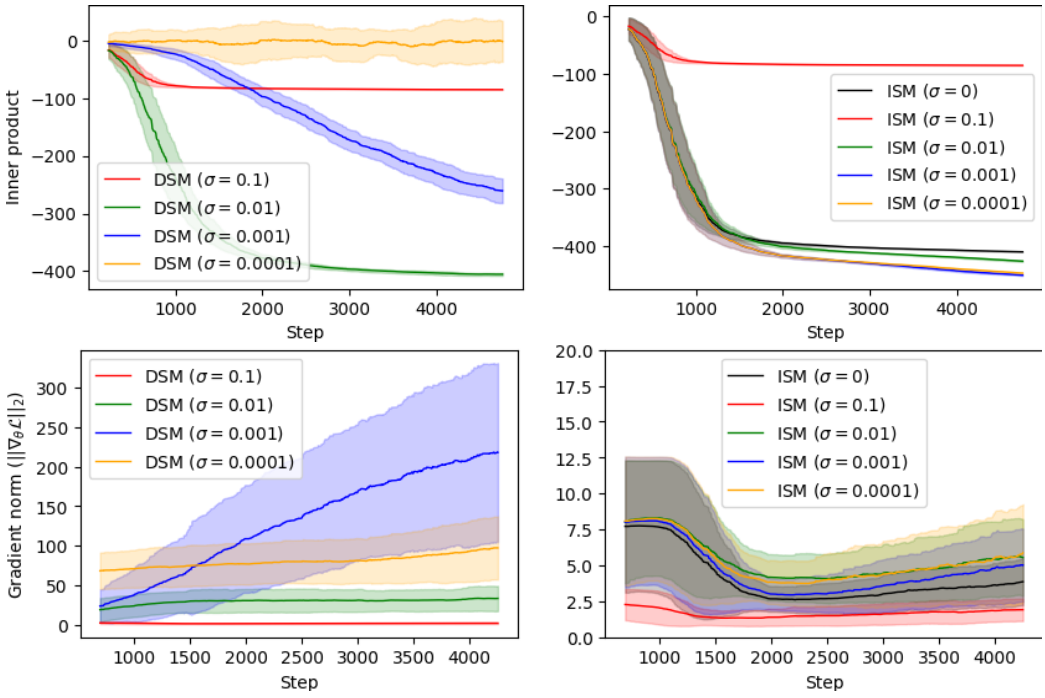


Figure 2: Top: *fitting* term (inner product) of the loss for training with DSM-S (left) and ISM (right) losses in Eq. 8, 9. Both objectives are trained on the data corrupted by Gaussian noise for a range of noise variance. However, ISM is also trained without any data corruption. Bottom: The gradient norm of the loss for each experiment on the top row.

## 3 Results & Discussion

**Experimental setup & dataset** We use the half-moon synthetic dataset [5] to compare ISM and DSM on non-Gaussian data distribution. The corruption kernel is a Gaussian centered around samples, and the noise scale is adjusted by choosing the standard deviation $\sigma = 0.01$. AdamW with a learning rate of $10^{-3}$ is used to optimize ISM and DSM-S losses (Eq. 8 & Eq. 9) with their respective gradients. We use $N = 1000$ data samples (in full batch) and 500 noise samples to approximate the $\mathbb{E}_{\mathcal{N}(\epsilon;0,I)}$ in Eq. 7. The model is a 3-layer multi-layer perception (MLP) with 16, 32, and 2 neurons and *tanh* activations.

**Fitting the data with ISM & DSM** To compare ISM and DSM-S, we train two separate models. One model is trained with ISM loss, but we also calculate the DSM-S loss in every training step (Fig. 1 left). The second model is trained using DSM-S loss gradients while also recording the ISM loss for every training step (Fig. 1 left). We show the variation of the *fitting* term for both losses (Eq. 8 & Eq. 9) during the training in Fig. 1. The mean of *fitting* term is the same for ISM and DSM in both cases and every training step. However, we show in the next section that results may differ significantly when using smaller values of the noise $\sigma$.

**Sensitivity of DSM to the noise scale $\sigma$** The above experiments demonstrate the convergence of ISM and DSM-S losses for a fixed noise scale, i.e. $\sigma = 10^{-2}$. Next, we exponentially vary the noise scale by setting the standard deviation of Gaussian corruption to $\sigma \in \{0.0001, 0.001, 0.01, 0.1, 0\}$. Note that ISM loss can be trained without noise (i.e. $\sigma = 0$). When trained on the noisy data, the ISM treats the noise as part of the data and does not explicitly depend on the corruption scheme. Fig. 2 shows the resilience of ISM training dynamic to the varying noise scale. On the other hand, DSM explicitly relies on the added noise (Eq. 6). Thus, noiseless training is not feasible (i.e. $\sigma \neq 0$ in eq. 9). Fig. 2 shows that for low noise scale, $\sigma = 10^{-3}$, the DSM loss gradient and its variance is increasing rapidly. In the case of $\sigma = 10^{-4}$, the loss barely decreases and the score is not fitted to the data. This suggests a large enough noise scale is required for DSM to be feasible. However, when the noise scale is too large, i.e. $\sigma = 10^{-1}$, the learned score is far away from the score of the data. Our

results indicate that an intermediate noise scale is required to ensure stability and convergence of DSM loss.

## 4 Conclusion and Perspective

ISM and DSM are two approaches for score matching, each with their derivative methods. In this study we recount the theory behind these two methods to highlight their similarities and differences. Then, we experimentally compare ISM and DSM data-fitting behaviors for a simple non-Gaussian dataset. Unlike DSM, ISM does not require noise to learn the data distribution. ISM is stable and robust against changing noise variance, while DSM requires nuanced adjustment of the noise scale. Nevertheless, ISM has often been overlooked due to the cost of computing the Hessian (Eq. 3). Our study emphasizes the advantages of ISM and motivates further efforts to improve its computational costs.

## References


[1] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=HkpbnH9lx`.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[3] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

[4] Diederik P. Kingma and Yann LeCun. Regularized estimation of image statistics by score matching. In *Neural Information Processing Systems*, 2010. URL `https://api.semanticscholar.org/CorpusID:8821883`.

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[6] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, 2015.

[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[8] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL `https://proceedings.mlr.press/v37/sohl-dickstein15.html`.

[9] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[10] Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. Sliced score matching: A scalable approach to density and score estimation. In *Conference on Uncertainty in Artificial Intelligence*, 2019. URL `https://api.semanticscholar.org/CorpusID:158047026`.

[11] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. doi: 10.1162/NECO_a_00142.

[12] Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, page 681–688, 2011.

## A Proposition 1. derivation

**Proposition A.1.** *For a distribution, $q(\mathbf{x})$, and its parameterized score, $s(\mathbf{x}, \boldsymbol{\theta})$, under the assumption 1, it can be shown that:*

$$\mathbb{E}_{q(\mathbf{x})}\big\langle s(\mathbf{x}, \boldsymbol{\theta}), \nabla_{\mathbf{x}} \log(q(\mathbf{x}))\big\rangle = \mathbb{E}_{q(\mathbf{x})} \mathrm{Tr}\left[\nabla^2_{\mathbf{x}} \log\left(p_{\boldsymbol{\theta}}(\mathbf{x})\right)\right] \tag{11}$$

*Proof.* We use partial derivative notation and start from the left side of eq. A.1:

$$\begin{aligned}
\mathbb{E}_{q(\mathbf{z})}\big\langle s(\mathbf{z}, \boldsymbol{\theta}), \frac{\partial log q(\mathbf{z})}{\partial \mathbf{z}}\big\rangle &= \int_{\mathbf{z}} s(\mathbf{z}, \boldsymbol{\theta})^T \frac{\partial \log q(\mathbf{z})}{\partial \mathbf{z}} q(\mathbf{z}) d\mathbf{z} \\
&= \int_{\mathbf{z}} s(\mathbf{z}, \boldsymbol{\theta})^T \frac{\partial q(\mathbf{z})}{\partial \mathbf{z}} \frac{q(\mathbf{z})}{q(\mathbf{z})} d\mathbf{z} \\
&= \int_{\mathbf{z}} s(\mathbf{z}, \boldsymbol{\theta})^T \frac{\partial q(\mathbf{z})}{\partial \mathbf{z}} d\mathbf{z}
\end{aligned}$$

Now, using integration by parts and the assumption $\lim_{|x| \to \infty} q(\mathbf{x}) s(\mathbf{x}, \boldsymbol{\theta}) = 0$, we can simplify the integral.

$$\begin{aligned}
\mathbb{E}_{q(\mathbf{z})}\big\langle s(\mathbf{z}, \boldsymbol{\theta}), \frac{\partial \log q(\mathbf{z})}{\partial \mathbf{z}}\big\rangle &= \int_{\mathbf{z}} s(\mathbf{z}, \boldsymbol{\theta})^T \frac{\partial \log q(\mathbf{z})}{\partial \mathbf{z}} q(\mathbf{z}) d\mathbf{z} \\
&= -\sum_i \int_z \frac{\partial s(\mathbf{z}, \boldsymbol{\theta})_i}{\partial \mathbf{z}_i} q(\mathbf{z}) d\mathbf{z} \\
&= -\mathbb{E}_{q(\mathbf{z})} \sum_i \frac{\partial s(\mathbf{z}, \boldsymbol{\theta})_i}{\partial \mathbf{z}_i} \\
&= -\mathbb{E}_{q(\mathbf{z})} \mathrm{Tr}\left[\nabla^2_{\mathbf{z}} \log p_{\boldsymbol{\theta}}(\mathbf{z})\right]
\end{aligned}$$

which proves the proposition.

□

## B Proposition 2. derivation

**Proposition B.2.** *For $\mathbf{z} \sim q_\sigma(\mathbf{z})$, where $q_\sigma(\mathbf{z}) = q_\sigma(\mathbf{z}|\mathbf{x}) q(\mathbf{x}) = \mathcal{N}(\mathbf{z}; \mathbf{x}, \sigma^2 I) q(\mathbf{x})$, if $q_\sigma(\mathbf{z})$ and its parameterized score $s(\mathbf{z}, \boldsymbol{\theta})$ satisfy the assumptions 1, we have:*

$$\mathbb{E}_{q(\mathbf{z})} \sum_i \frac{\partial s(\mathbf{z}, \boldsymbol{\theta})_i}{\partial \mathbf{z}_i} = \mathbb{E}_{q(\mathbf{x})} \mathbb{E}_{\mathcal{N}(\epsilon, 0, I)} \langle s(\mathbf{z}, \boldsymbol{\theta}), \frac{\epsilon}{\sigma} \rangle \tag{12}$$

*where $\epsilon \sim \mathcal{N}(\epsilon; 0, I)$.*

Instead of directly proving proposition B.2, we derive the following equality:

$$\mathbb{E}_{q(\mathbf{z})}\big\langle s(\mathbf{z}, \boldsymbol{\theta}), \frac{\partial \log q(\mathbf{z})}{\partial \mathbf{z}}\big\rangle = \mathbb{E}_{q(\mathbf{x})} \mathbb{E}_{\mathcal{N}(\epsilon, 0, I)} \langle s(\mathbf{z}, \boldsymbol{\theta}), \frac{\epsilon}{\sigma} \rangle \tag{13}$$

*Proof.*

$$\begin{aligned}
\mathbb{E}_{q(\mathbf{z})}\big\langle s(\mathbf{z},\boldsymbol{\theta}), \frac{\partial \log q(\mathbf{z})}{\partial \mathbf{z}}\big\rangle &= \int_{\mathbf{z}} s(\mathbf{z},\boldsymbol{\theta})^T \frac{\partial \log q(\mathbf{z})}{\partial \mathbf{z}} q(\mathbf{z}) d\mathbf{z} \\
&= \int_{\mathbf{z}} s(\mathbf{z},\boldsymbol{\theta})^T \frac{\partial q(\mathbf{z})}{\partial \mathbf{z}} \frac{q(\mathbf{z})}{q(\mathbf{z})} d\mathbf{z} \\
&= \int_{\mathbf{z}} s(\mathbf{z},\boldsymbol{\theta})^T \frac{\partial q(\mathbf{z})}{\partial \mathbf{z}} d\mathbf{z} \\
&= \int_{\mathbf{z}} s(\mathbf{z},\boldsymbol{\theta})^T \frac{\partial}{\partial \mathbf{z}} \big[\int_{x} q(\mathbf{z}|\mathbf{x}) q(\mathbf{x}) d\mathbf{x}\big] d\mathbf{z} \\
&= \int_{\mathbf{z}} \int_{\mathbf{x}} s(\mathbf{z},\boldsymbol{\theta})^T \frac{\partial q(\mathbf{z}|\mathbf{x})}{\partial \mathbf{z}} q(\mathbf{x}) d\mathbf{x} d\mathbf{z} \\
&= \int_{\mathbf{z}} \int_{\mathbf{x}} s(\mathbf{z},\boldsymbol{\theta})^T \frac{\partial \log(q(\mathbf{z}|\mathbf{x}))}{\partial \mathbf{z}} q(\mathbf{z}|\mathbf{x}) q(\mathbf{x}) d\mathbf{x} d\mathbf{z} \\
&= \mathbb{E}_{q(\mathbf{x})} \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \big\langle s(\mathbf{z},\boldsymbol{\theta}), \frac{\partial \log(q(\mathbf{z}|\mathbf{x}))}{\partial \mathbf{z}} \big\rangle
\end{aligned}$$

For $q(\mathbf{z}|\mathbf{x}) = q_\sigma(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \mathbf{x}, \sigma^2 I)$ and $\epsilon \sim \mathcal{N}(\epsilon; 0, I)$, we continue the proof as:

$$\begin{aligned}
\mathbb{E}_{q(\mathbf{z})}\big\langle s(\mathbf{z},\boldsymbol{\theta}), \frac{\partial \log q(\mathbf{z})}{\partial \mathbf{z}}\big\rangle &= \mathbb{E}_{q(\mathbf{x})} \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \big\langle s(\mathbf{z},\boldsymbol{\theta}), \frac{\partial \log(q(\mathbf{z}|\mathbf{x}))}{\partial \mathbf{z}} \big\rangle \\
&= -\mathbb{E}_{q(\mathbf{x})} \mathbb{E}_{\mathcal{N}(\mathbf{z};\mathbf{x},\sigma^2 I)} \big\langle s(\mathbf{z},\boldsymbol{\theta}), \frac{\mathbf{z}-\mathbf{x}}{\sigma^2} \big\rangle \\
&= -\mathbb{E}_{q(\mathbf{x})} \mathbb{E}_{\mathcal{N}(\epsilon;0,I)} \big\langle s(\mathbf{z},\boldsymbol{\theta}), \frac{\epsilon}{\sigma} \big\rangle
\end{aligned}$$

This proves equality 13. Equality 13 and proposition A.1 prove proposition B.2.

□

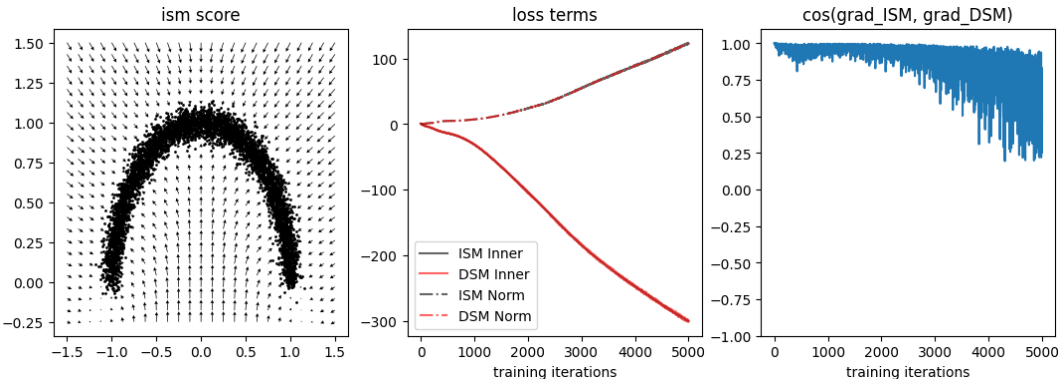


Figure 3: (left) learned score field, (center) *norm* and *fitting* terms of ISM and DSM, and (right) the gradient similarity between ISM and DSM loss for steps during training with ISM loss.