SCORE-BASED VARIATIONAL INFERENCE FOR INVERSE PROBLEMS

Author
Haitong Lan
lanhaitong@bupt.edu.cn

September 9, 2024

1 Background

1.1 Bayesian estimation

When using Bayesian method to build a statistical model, the parameters of the model need to be optimized to minimize the difference between the model parameter $\hat{\theta}$and the real parameter θ(that is, to minimize the risk function). To evaluate the difference, a metric $d()$needs to be selected:

$$\hat{\theta}(x) = \underset{\hat{\theta}}{\arg\min} \mathbb{E}_{x,\theta}[d(\hat{\theta}(x), \theta)] \tag{1}$$

One choice of metric is MSE, in which case the risk function can be defined as

$$R(\theta, \hat{\theta}) = \int (\theta - \hat{\theta})^2 p(\theta \mid x) d\theta \tag{2}$$

Let$\nabla_{\hat{\theta}} R(\theta, \hat{\theta}) = 0$，there is

$$\hat{\theta}^* = \frac{\int \theta p(\theta \mid x) d\theta}{\int p(\theta \mid x) d\theta} = \int \theta p(\theta \mid x) d\theta = E[\theta \mid x] \tag{3}$$

That is, the conditional mean is the optimal estimate of the model in the meaning of MMSE.

1.2 Variational inference

Getting the analytic form of the posterior distribution is intractable. Variational inference(VI) is a method of using the distribution $q(z)$**to approximate the posterior distribution** $p(z|x)$. In the framework of variational Bayes, the measure of approximation is KL divergence.

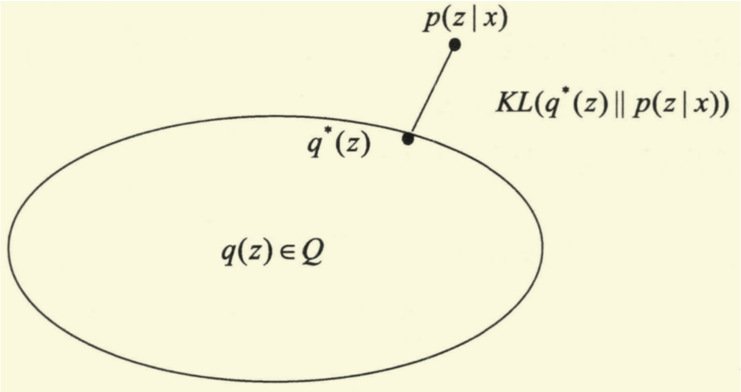


Figure 1: VI

Starting from the KL divergence, simplify:

$$\begin{aligned} KL(q(z)\|p(z\mid x)) &= -\int_z q(z)\log\frac{p(z\mid x)}{q(z)}dz \\ &= \int_z q(z)\log q(z)dz - \int_z q(z)\log p(z\mid x)dz. \\ &= E_q\left[\log q(z)\right] - E_q\left[\log p(z\mid x)\right] \\ &= E_q\left[\log q(z)\right] - E_q\left[\log p(z,x)\right] + \log p(x) \\ &= E_q\left[\log\frac{q(z)}{p(z,x)}\right] + \log p(x) \\ &= -KL\left[q(z) \mid\mid p(z,x)\right] + \log p(x) \\ &:= ELBO\left[q(z)\right] + \log p(x) \end{aligned} \tag{4}$$

Minimizing the KL divergence of $q(z)$and $p(z|x)$is equivalent to maximizing ELBO(Evidence Lower Bound), There are many kinds of derivation of ELBO. Another derivation in this paper is as follows:

$$\begin{aligned} \log p(\boldsymbol{y}) &= \log\int\frac{q_\phi\left(\boldsymbol{x}_0\mid\boldsymbol{y}\right)}{q_\phi\left(\boldsymbol{x}_0\mid\boldsymbol{y}\right)}p\left(\boldsymbol{y}\mid\boldsymbol{x}_0\right)p\left(\boldsymbol{x}_0\right)d\boldsymbol{x}_0 \\ &\geq \int q_\phi\left(\boldsymbol{x}_0\mid\boldsymbol{y}\right)\log\frac{p\left(\boldsymbol{y},\boldsymbol{x}_0\right)}{q_\phi\left(\boldsymbol{x}_0\mid\boldsymbol{y}\right)}d\boldsymbol{x}_0 = -\mathcal{F}_\phi(\boldsymbol{y}) \\ &= -\mathrm{KL}\left(q_\phi\left(\boldsymbol{x}_0\mid\boldsymbol{y}\right)\|p\left(\boldsymbol{x}_0\mid\boldsymbol{y}\right)\right) + \log p(\boldsymbol{y}) \end{aligned} \tag{5}$$

1.3 Inverse problem and diffusion model

An inverse problem is defined as the estimation of an unknown state or a latent $\boldsymbol{x}_0 \in \mathbb{R}^{N\times 1}$ from measurement $\boldsymbol{y} \in \mathbb{R}^{M\times 1}$. Specifically, the measurement process can be described by a measurement operator $\mathcal{A}: \mathbb{R}^{N\times 1} \rightarrow \mathbb{R}^{M\times 1}$, and the final output is a noisy version of the measurement:

$$\boldsymbol{y} = \mathcal{A}\left(\boldsymbol{x}_0\right) + \boldsymbol{w}_0 \tag{6}$$

This paper assumes that $\mathcal{A}$ is known. The inverse problem can be formulated as a **Bayesian estimation problem**. The posterior distribution of $\boldsymbol{x}_0$ is given by $p\left(\boldsymbol{x}_0\mid\boldsymbol{y}\right) = \frac{p(\boldsymbol{y}|\boldsymbol{x}_0)p(\boldsymbol{x}_0)}{p(\boldsymbol{y})}$, where $p\left(\boldsymbol{x}_0\right)$ is the prior distribution of $\boldsymbol{x}_0$ and $p\left(\boldsymbol{y}\mid\boldsymbol{x}_0\right)$ is the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{x}_0$. The **MMSE estimator** can be employed for the estimation of $\boldsymbol{x}_0$. However, the posterior $p\left(\boldsymbol{x}_0\mid\boldsymbol{y}\right)$ is intractable in general since the prior $p\left(\boldsymbol{x}_0\right)$ and the likelihood $p\left(\boldsymbol{y}\mid\boldsymbol{x}_0\right)$ may be very complicated in real applications.

In this paper, both VE-SDE and VP-SDE are mathematically derived to solve the inverse problem. Since both are **score based models**, the derivation processes are consistent. This note take **VE-SDE** as an example,the formula of VE-SDE's diffusion process and sampling process is as follows:

Diffusion:

$$\mathbf{x}_i = \mathbf{x}_{i-1} + \sqrt{\sigma_i^2 - \sigma_{i-1}^2}\mathbf{z}_{i-1}, \quad i = 1,\cdots,N \tag{7}$$

Sampling:

$$\mathbf{x}_{i+1}=\mathbf{x}_i+\tau\nabla_{\mathbf{x}}\log p\left(\mathbf{x}_i\right)+\sqrt{2\tau}\mathbf{z}_i \tag{8}$$

2 Diffusion Process and Posterior Estimation

In this chapter, some properties are used to obtain the posterior distribution, and the theoretical support of RMP framework is given.

2.1 Calculation of a posterior distribution

The measurement $\boldsymbol{y}$ and diffusion states $\{\boldsymbol{x}_k\}_{k=0}^{T}$ forms a new **Markov chain** $\boldsymbol{y}\rightarrow\boldsymbol{x}_0\rightarrow\boldsymbol{x}_1\cdots\rightarrow\boldsymbol{x}_T$ and the reverse conditional $p_k\left(\boldsymbol{x}_k\mid\boldsymbol{x}_{k+1},\boldsymbol{y}\right)$ is given by $p_k\left(\boldsymbol{x}_k\mid\boldsymbol{x}_{k+1},\boldsymbol{y}\right)=\frac{p(\boldsymbol{x}_{k+1}|\boldsymbol{x}_k)p(\boldsymbol{x}_k|\boldsymbol{y})}{p(\boldsymbol{x}_{k+1}|\boldsymbol{y})},\forall k=0,\cdots T-1$. In this part, we focus on the property of reverse conditional $p_k\left(\boldsymbol{x}_k\mid\boldsymbol{x}_{k+1},\boldsymbol{y}\right)$.

Proposition 2.1. $p_k\left(\boldsymbol{x}_k\mid\boldsymbol{x}_{k+1},\boldsymbol{y}\right),\forall k=0\cdots T-1$, is Gaussian when $\Delta t\rightarrow 0$. For VE and VP diffusion, the mean and covariance of $p_k\left(\boldsymbol{x}_k\mid\boldsymbol{x}_{k+1},\boldsymbol{y}\right)$ are tractable with mean given by

$$\boldsymbol{\mu}_k\left(\boldsymbol{x}_{k+1},\boldsymbol{y}\right)=\boldsymbol{V}_{k,1}\boldsymbol{x}_{k+1}+\boldsymbol{V}_{k,2}\mathbb{E}_{p(\boldsymbol{x}_0|\boldsymbol{y})}\left[\boldsymbol{x}_0\right]$$

where $\boldsymbol{V}_{k,1}=\left(\sigma_k^2\boldsymbol{I}+\boldsymbol{C}_{\boldsymbol{x}_0}\right)\left(\sigma_{k+1}^2\boldsymbol{I}+\boldsymbol{C}_{\boldsymbol{x}_0}\right)^{-1}$ and $\boldsymbol{V}_{k,2}=\left(\sigma_{k+1}^2-\sigma_k^2\right)\left(\sigma_{k+1}^2\boldsymbol{I}+\boldsymbol{C}_{\boldsymbol{x}_0}\right)^{-1}$ for VE diffusion. $\mathbb{E}_{p(\boldsymbol{x}_0|\boldsymbol{y})}\left[\boldsymbol{x}_0\right]$ and $\boldsymbol{C}_{\boldsymbol{x}_0}$ are the mean and covariance of $p\left(\boldsymbol{x}_0\mid\boldsymbol{y}\right)$ respectively. The covariance of $p_k\left(\boldsymbol{x}_k\mid\boldsymbol{x}_{k+1},\boldsymbol{y}\right)$ is given by

$$\boldsymbol{C}_{k,VE}=\left(\sigma_{k+1}^2-\sigma_k^2\right)\left(\sigma_k^2\boldsymbol{I}+\boldsymbol{C}_{\boldsymbol{x}_0}\right)\left(\sigma_{k+1}^2\boldsymbol{I}+\boldsymbol{C}_{\boldsymbol{x}_0}\right)^{-1}$$

By Bayes' formula $p_k\left(\boldsymbol{x}_k\mid\boldsymbol{x}_{k+1},\boldsymbol{y}\right)=\frac{p(\boldsymbol{x}_{k+1}|\boldsymbol{x}_k)p(\boldsymbol{x}_k|\boldsymbol{y})}{p(\boldsymbol{x}_{k+1}|\boldsymbol{y})}$can be found a posteriori distribution, according to the diffusion process and the sampling process, We know $p\left(\boldsymbol{x}_{k+1}\mid\boldsymbol{x}_k\right)$and $p_k\left(\boldsymbol{x}_k\mid\boldsymbol{x}_{k+1},\boldsymbol{y}\right)$is a Gaussian distribution, but the posterior probability obtained by Reverse SDE has differential and score. It is difficult for us to observe some of its properties intuitively, so we intend to use the condition that it is a Gaussian distribution to derive a posteriori distribution that is more conducive to analysis.

Note that given the premise of $x_{k+1}$and y, the evidence (denominator) of Bayes' formula is constant, so the **prior distribution** $p\left(\boldsymbol{x}_k\mid\boldsymbol{y}\right)$**must also be Gaussian**, If we can obtain the Gaussian kernel of the prior distribution, we may be able to obtain a posteriori distribution form that is easy to analyze. The following proof is done according to the steps of this analysis, and the proof process is very similar to the way DDPM obtains the posteriori distribution.

Proof. For VE diffusion models, the mean of $p\left(\boldsymbol{x}_k \mid \boldsymbol{y}\right)$ is given by

$$\begin{aligned}
\mathbb{E}_{p\left(\boldsymbol{x}_k \mid \boldsymbol{y}\right)}\left[\boldsymbol{x}_k\right] &= \int \boldsymbol{x}_k p\left(\boldsymbol{x}_k \mid \boldsymbol{y}\right) d\boldsymbol{x}_k \\
&= \int \boldsymbol{x}_k \int p\left(\boldsymbol{x}_k \mid \boldsymbol{x}_0\right) p\left(\boldsymbol{x}_0 \mid \boldsymbol{y}\right) d\boldsymbol{x}_0 d\boldsymbol{x}_k \\
&= \iint \boldsymbol{x}_k p\left(\boldsymbol{x}_k \mid \boldsymbol{x}_0\right) d\boldsymbol{x}_k p\left(\boldsymbol{x}_0 \mid \boldsymbol{y}\right) d\boldsymbol{x}_0 \\
&= \int \boldsymbol{x}_0 p\left(\boldsymbol{x}_0 \mid \boldsymbol{y}\right) d\boldsymbol{x}_0 = \mathbb{E}_{p\left(\boldsymbol{x}_0 \mid \boldsymbol{y}\right)}\left[\boldsymbol{x}_0\right]
\end{aligned}$$

and its covariance matrix is given by

$$\begin{aligned}
&\operatorname{Cov}_{p\left(\boldsymbol{x}_k \mid \boldsymbol{y}\right)}\left[\boldsymbol{x}_k\right] \\
=& \int \boldsymbol{x}_k \boldsymbol{x}_k^T p\left(\boldsymbol{x}_k \mid \boldsymbol{y}\right) d\boldsymbol{x}_k - \mathbb{E}_{p\left(\boldsymbol{x}_0 \mid \boldsymbol{y}\right)}\left[\boldsymbol{x}_0\right] \mathbb{E}_{p\left(\boldsymbol{x}_0 \mid \boldsymbol{y}\right)}\left[\boldsymbol{x}_0\right]^T \\
=& \int \boldsymbol{x}_k \boldsymbol{x}_k^T \int p\left(\boldsymbol{x}_k \mid \boldsymbol{x}_0\right) p\left(\boldsymbol{x}_0 \mid \boldsymbol{y}\right) d\boldsymbol{x}_0 d\boldsymbol{x}_k - \mathbb{E}_{p\left(\boldsymbol{x}_0 \mid \boldsymbol{y}\right)}\left[\boldsymbol{x}_0\right] \mathbb{E}_{p\left(\boldsymbol{x}_0 \mid \boldsymbol{y}\right)}\left[\boldsymbol{x}_0\right]^T \\
=& \iint \boldsymbol{x}_k \boldsymbol{x}_k^T p\left(\boldsymbol{x}_k \mid \boldsymbol{x}_0\right) d\boldsymbol{x}_k p\left(\boldsymbol{x}_0 \mid \boldsymbol{y}\right) d\boldsymbol{x}_0 - \mathbb{E}_{p\left(\boldsymbol{x}_0 \mid \boldsymbol{y}\right)}\left[\boldsymbol{x}_0\right] \mathbb{E}_{p\left(\boldsymbol{x}_0 \mid \boldsymbol{y}\right)}\left[\boldsymbol{x}_0\right]^T \\
=& \int \left(\boldsymbol{x}_0 \boldsymbol{x}_0^T + \sigma_k^2 \boldsymbol{I}\right) p\left(\boldsymbol{x}_0 \mid \boldsymbol{y}\right) d\boldsymbol{x}_0 - \mathbb{E}_{p\left(\boldsymbol{x}_0 \mid \boldsymbol{y}\right)}\left[\boldsymbol{x}_0\right] \mathbb{E}_{p\left(\boldsymbol{x}_0 \mid \boldsymbol{y}\right)}\left[\boldsymbol{x}_0\right]^T \\
=& \sigma_k^2 \boldsymbol{I} + \operatorname{Cov}_{p\left(\boldsymbol{x}_0 \mid \boldsymbol{y}\right)}\left[\boldsymbol{x}_0\right] + \mathbb{E}_{p\left(\boldsymbol{x}_0 \mid \boldsymbol{y}\right)}\left[\boldsymbol{x}_0\right] \mathbb{E}_{p\left(\boldsymbol{x}_0 \mid \boldsymbol{y}\right)}\left[\boldsymbol{x}_0\right]^T - \mathbb{E}_{p\left(\boldsymbol{x}_0 \mid \boldsymbol{y}\right)}\left[\boldsymbol{x}_0\right] \mathbb{E}_{p\left(\boldsymbol{x}_0 \mid \boldsymbol{y}\right)}\left[\boldsymbol{x}_0\right]^T \\
=& \sigma_k^2 \boldsymbol{I} + \operatorname{Cov}_{p\left(\boldsymbol{x}_0 \mid \boldsymbol{y}\right)}\left[\boldsymbol{x}_0\right] = \sigma_k^2 \boldsymbol{I} + \boldsymbol{C}_{\boldsymbol{x}_0}
\end{aligned}$$

Since $p_k\left(\boldsymbol{x}_k \mid \boldsymbol{x}_{k+1}, \boldsymbol{y}\right)$ and $p\left(\boldsymbol{x}_{k+1} \mid \boldsymbol{x}_k\right) = \mathcal{N}\left(\boldsymbol{x}_{k+1}; \boldsymbol{x}_k, \left(\sigma_{k+1}^2 - \sigma_k^2\right) \boldsymbol{I}\right)$ are both Gaussian, then $p\left(\boldsymbol{x}_k \mid \boldsymbol{y}\right)$ is also Gaussian when $\boldsymbol{x}_{k+1}$ and $\boldsymbol{y}$ are given, i.e., $p\left(\boldsymbol{x}_k \mid \boldsymbol{y}\right) = \mathcal{N}\left(\boldsymbol{x}_k; \mathbb{E}_{p\left(\boldsymbol{x}_0 \mid \boldsymbol{y}\right)}\left[\boldsymbol{x}_0\right], \left(\sigma_k^2 \boldsymbol{I} + \ \boldsymbol{C}_{\boldsymbol{x}_0}\ \right)\right)$. For two multivariate Gaussian distribution $G_1(\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{x}, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\right)$ and $G_2(\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{x}, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2\right)$, the product $G_1(\boldsymbol{x}) G_2(\boldsymbol{x})$ is also Gaussian with mean and covariance given by

$$\begin{aligned}
&\boldsymbol{\mu}_3 = \boldsymbol{\Sigma}_2\left(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2\right)^{-1} \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_1\left(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2\right)^{-1} \boldsymbol{\mu}_2 \\
&\boldsymbol{\Sigma}_3 = \boldsymbol{\Sigma}_1\left(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2\right)^{-1} \boldsymbol{\Sigma}_2
\end{aligned}$$

Thus, the mean and covariance of $p_k\left(\boldsymbol{x}_k \mid \boldsymbol{x}_{k+1}, \boldsymbol{y}\right)$ can be calculated and are given respectively by:

$$\begin{aligned}
\boldsymbol{\mu}_k &= \left(\sigma_k^2 \boldsymbol{I} + \boldsymbol{C}_{\boldsymbol{x}_0}\right)\left(\sigma_{k+1}^2 \boldsymbol{I} + \boldsymbol{C}_{\boldsymbol{x}_0}\right)^{-1} \boldsymbol{x}_{k+1} + \left(\sigma_{k+1}^2 - \sigma_k^2\right)\left(\sigma_{k+1}^2 \boldsymbol{I} + \boldsymbol{C}_{\boldsymbol{x}_0}\right)^{-1} \mathbb{E}_{p\left(\boldsymbol{x}_0 \mid \boldsymbol{y}\right)}\left[\boldsymbol{x}_0\right] \\
&= \boldsymbol{V}_{k, p_1} \boldsymbol{x}_{k+1} + \boldsymbol{V}_{k, p_2} \mathbb{E}_{p\left(\boldsymbol{x}_0 \mid \boldsymbol{y}\right)}\left[\boldsymbol{x}_0\right] \\
\boldsymbol{C}_k &= \left(\sigma_{k+1}^2 - \sigma_k^2\right)\left(\sigma_k^2 \boldsymbol{I} + \boldsymbol{C}_{\boldsymbol{x}_0}\right)\left(\sigma_{k+1}^2 \boldsymbol{I} + \boldsymbol{C}_{\boldsymbol{x}_0}\right)^{-1}
\end{aligned}$$

□

Comparing the posterior distribution obtained by DDPM, we can find that the two are very similar:

$$\begin{aligned} & q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right)=\mathcal{N}\left(\mathbf{x}_{t-1} ; \tilde{\boldsymbol{\mu}}_t\left(\mathbf{x}_t, \mathbf{x}_0\right), \tilde{\beta}_t \mathbf{I}\right), \\ & \text{where} \quad \tilde{\boldsymbol{\mu}}_t\left(\mathbf{x}_t, \mathbf{x}_0\right):=\frac{\sqrt{\bar{\alpha}_{t-1}} \beta_t}{1-\bar{\alpha}_t} \mathbf{x}_0+\frac{\sqrt{\alpha_t}\left(1-\bar{\alpha}_{t-1}\right)}{1-\bar{\alpha}_t} \mathbf{x}_t \quad \text{and} \quad \tilde{\beta}_t:=\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \beta_t \end{aligned} \tag{9}$$

The difference is that this paper focuses on solving the inverse problem. There is observation data y, and $x_0$is also put into the hidden space variable, which is different from the assumption of DDPM. Note that $\mathbb{E}_{p(\boldsymbol{x}_k|\boldsymbol{y})}\left[\boldsymbol{x}_k\right]$is finally determined independent of k, This is also consistent with the diffusion process of VE-SDE, for VP-SDE, because there is a drift term, For different k,$\mathbb{E}_{p(\boldsymbol{x}_k|\boldsymbol{y})}\left[\boldsymbol{x}_k\right]$ must be different.

2.2 RMP framework and convergence proof

Define 2.2. The reverse mean propagation chain of a diffusion process is defined as

$$\boldsymbol{\mu}_T \rightarrow \boldsymbol{\mu}_{T-1}\left(\boldsymbol{x}_T=\boldsymbol{\mu}_T, \boldsymbol{y}\right) \rightarrow \cdots \rightarrow \boldsymbol{\mu}_1\left(\boldsymbol{x}_2=\boldsymbol{\mu}_2, \boldsymbol{y}\right) \rightarrow \boldsymbol{\mu}_0\left(\boldsymbol{x}_1=\boldsymbol{\mu}_1, \boldsymbol{y}\right)$$

where $\boldsymbol{\mu}_k\left(\boldsymbol{x}_{k+1}=\boldsymbol{\mu}_{k+1}, \boldsymbol{y}\right)$ is the mean of $p_k\left(\boldsymbol{x}_k \mid \boldsymbol{x}_{k+1}=\boldsymbol{\mu}_{k+1}, \boldsymbol{y}\right), \forall k = 0, \cdots, T-1$, and $\boldsymbol{\mu}_T$ and $\boldsymbol{\mu}_{\mathbf{0}}$ are the initial point and end point of the reverse chain respectively.

Note that the random variable in the definition $\boldsymbol{x}_k=\boldsymbol{\mu}_k$, that is, the article makes an assumption that all **random variables take the mean**, so the propagation process will eventually output a mean. **RMP is a deterministic algorithm**. We can also see from the later theorem that only by making the random variable equal to its mean can we guarantee the optimal estimation of the RMP to the final output in the sense of MMSE.

Theorem 2.3. For VE diffusion, when $\Delta t \rightarrow 0$, the end point of the reverse chain, i.e., $\boldsymbol{\mu}_0$ is given by

$$\boldsymbol{\mu}_0=\left(\sigma_0^2 \boldsymbol{I}+\boldsymbol{C}_{\boldsymbol{x}_0}\right)\left(\sigma_T^2 \boldsymbol{I}+\boldsymbol{C}_{\boldsymbol{x}_0}\right)^{-1} \boldsymbol{\mu}_T+\left(\sigma_T^2-\sigma_0^2\right)\left(\sigma_T^2 \boldsymbol{I}+\boldsymbol{C}_{\boldsymbol{x}_0}\right)^{-1} \mathbb{E}_{p(\boldsymbol{x}_0|\boldsymbol{y})}\left[\boldsymbol{x}_0\right]$$

and $\boldsymbol{\mu}_0 \rightarrow \mathbb{E}_{p(\boldsymbol{x}_0|\boldsymbol{u})}\left[\boldsymbol{x}_0\right]$ as $\sigma_T \rightarrow \infty$.

Now we want to get the analytic form of the final output of RMP $\boldsymbol{\mu}_0$, according to proposition 2.1, We have obtained $\boldsymbol{\mu}_k$, let $k=0$, and consider the hypothesis of $\boldsymbol{x}_k=\boldsymbol{\mu}_k$, we can iteratively obtain the final output mean of RMP.

Proof. For VE diffusion model, if we set $\boldsymbol{x}_k=\boldsymbol{\mu}_k\left(\boldsymbol{x}_{k+1}=\boldsymbol{\mu}_{k+1}, \boldsymbol{y}\right), \forall k=0: T-1$, then from Proposition 2.1, we have

$$\begin{aligned}
&\boldsymbol{\mu}_0\left(\boldsymbol{x}_1, \boldsymbol{y}\right) \\
&= \boldsymbol{V}_{0,p_1}\boldsymbol{x}_1 + \boldsymbol{V}_{0,p_2}\mathbb{E}_{p(\boldsymbol{x}_0|\boldsymbol{y})}\left[\boldsymbol{x}_0\right] \\
&= \boldsymbol{V}_{0,p_1}\left(\boldsymbol{V}_{1,p_1}\boldsymbol{x}_2 + \boldsymbol{V}_{1,p_2}\mathbb{E}_{p(\boldsymbol{x}_0|\boldsymbol{y})}\left[\boldsymbol{x}_0\right]\right) + \boldsymbol{V}_{0,p_2}\mathbb{E}_{p(\boldsymbol{x}_0|\boldsymbol{y})}\left[\boldsymbol{x}_0\right] \\
&= V_{0,p_1}V_{1,p_1}\boldsymbol{x}_2 + \boldsymbol{V}_{0,p_1}\boldsymbol{V}_{1,p_2}\mathbb{E}_{p(\boldsymbol{x}_0|\boldsymbol{y})}\left[\boldsymbol{x}_0\right] + \boldsymbol{V}_{0,p_2}\mathbb{E}_{p(\boldsymbol{x}_0|\boldsymbol{y})}\left[\boldsymbol{x}_0\right] \\
&= \cdots \\
&= \prod_{i=0}^{T-1}\boldsymbol{V}_{i,p_1}\boldsymbol{\mu}_T + \left(\boldsymbol{V}_{0,p_1}\left(\boldsymbol{V}_{1,p_1}\left(\cdots\boldsymbol{V}_{T-2,p_1}\left(\boldsymbol{V}_{T-1,p_2}\right) + \boldsymbol{V}_{T-2,p_2}\right) + \boldsymbol{V}_{1,p_2}\right) + \boldsymbol{V}_{0,p_2}\right)\mathbb{E}_{p(\boldsymbol{x}_0|\boldsymbol{y})}\left[\boldsymbol{x}_0\right].
\end{aligned}$$

For the first part, the coefficient of $\boldsymbol{\mu}_T$ equals

$$\prod_{i=0}^{T-1}\boldsymbol{V}_{i,p_1} = \prod_{i=0}^{T-1}\left(\sigma_i^2\boldsymbol{I} + \boldsymbol{C}_{\boldsymbol{x}_0}\right)\left(\sigma_{i+1}^2\boldsymbol{I} + \boldsymbol{C}_{\boldsymbol{x}_0}\right)^{-1} = \left(\sigma_0^2\boldsymbol{I} + \boldsymbol{C}_{\boldsymbol{x}_0}\right)\left(\sigma_T^2\boldsymbol{I} + \boldsymbol{C}_{\boldsymbol{x}_0}\right)^{-1}$$

For the second part, $\forall k = 1 : T-1$, we have

$$\begin{aligned}
\boldsymbol{V}_{k-1,p_1}\boldsymbol{V}_{k,p_2} + \boldsymbol{V}_{k-1,p_2} &= \left(\sigma_{k-1}^2\boldsymbol{I} + \boldsymbol{C}_{\boldsymbol{x}_0}\right)\left(\sigma_k^2\boldsymbol{I} + \boldsymbol{C}_{\boldsymbol{x}_0}\right)^{-1}\left(\sigma_{k+1}^2 - \sigma_k^2\right)\left(\sigma_{k+1}^2\boldsymbol{I} + \boldsymbol{C}_{\boldsymbol{x}_0}\right)^{-1} \\
&\quad + \left(\sigma_k^2 - \sigma_{k-1}^2\right)\left(\sigma_k^2\boldsymbol{I} + \boldsymbol{C}_{\boldsymbol{x}_0}\right)^{-1} \\
&= \left(\sigma_{k+1}^2 - \sigma_{k-1}^2\right)\left(\sigma_{k+1}^2\boldsymbol{I} + \boldsymbol{C}_{\boldsymbol{x}_0}\right)^{-1}
\end{aligned}$$

Similarly, we get

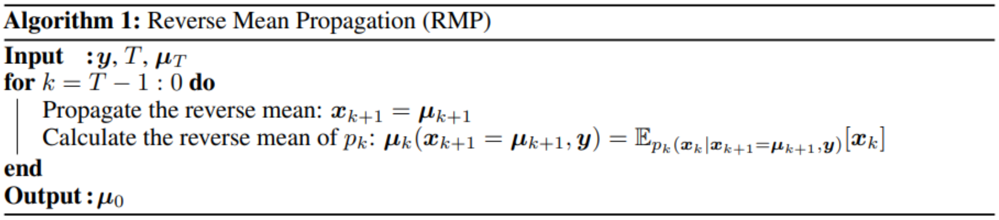
$$\left(\boldsymbol{V}_{0,p_1}\left(\boldsymbol{V}_{1,p_1}\left(\cdots\boldsymbol{V}_{T-2,p_1}\left(\boldsymbol{V}_{T-1,p_2}\right) + \boldsymbol{V}_{T-2,p_2}\right) + \boldsymbol{V}_{1,p_2}\right) + \boldsymbol{V}_{0,p_2}\right) = \left(\sigma_T^2 - \sigma_0^2\right)\left(\sigma_T^2\boldsymbol{I} + \boldsymbol{C}_{\boldsymbol{x}_0}\right)^{-1}$$

Thus, we have

$$\boldsymbol{\mu}_0 = \left(\sigma_0^2\boldsymbol{I} + \boldsymbol{C}_{\boldsymbol{x}_0}\right)\left(\sigma_T^2\boldsymbol{I} + \boldsymbol{C}_{\boldsymbol{x}_0}\right)^{-1}\boldsymbol{\mu}_T + \left(\sigma_T^2 - \sigma_0^2\right)\left(\sigma_T^2\boldsymbol{I} + \boldsymbol{C}_{\boldsymbol{x}_0}\right)^{-1}\mathbb{E}_{p(\boldsymbol{x}_0|\boldsymbol{y})}\left[\boldsymbol{x}_0\right]$$

When $\sigma_T^2 \to \infty, \boldsymbol{\mu}_0\left(\boldsymbol{x}_1, \boldsymbol{y}\right) \to \mathbb{E}_{p(\boldsymbol{x}_0|\boldsymbol{y})}\left[\boldsymbol{x}_0\right]$, which is the posterior mean. □

Using definition 2.2, this paper establishes a RMP framework for estimating posterior statistics, which only needs to propagate the implicit space posterior distribution mean in the way of RMP step by step. This framework can ensure that the final output mean is the optimal estimation in the sense of MMSE when **the number of propagation steps is large enough**. The pseudo-code of this framework algorithm is as follows:



```
Algorithm 1: Reverse Mean Propagation (RMP)
Input   : $\boldsymbol{y}, T, \boldsymbol{\mu}_T$
for $k = T-1 : 0$ do
    Propagate the reverse mean: $\boldsymbol{x}_{k+1} = \boldsymbol{\mu}_{k+1}$
    Calculate the reverse mean of $p_k$: $\boldsymbol{\mu}_k(\boldsymbol{x}_{k+1} = \boldsymbol{\mu}_{k+1}, \boldsymbol{y}) = \mathbb{E}_{p_k(\boldsymbol{x}_k|\boldsymbol{x}_{k+1}=\boldsymbol{\mu}_{k+1},\boldsymbol{y})}[\boldsymbol{x}_k]$
end
Output : $\boldsymbol{\mu}_0$
```

Figure 2: RMP pseudocode

3 Score-Based Variational Inference

3.1 RMP as Variational Inference

Although the RMP framework can guarantee the optimal estimation in the meaning of MMSE, the algorithm cannot be implemented. Because $\boldsymbol{C}_{\boldsymbol{x}_0}$and $\mathbb{E}_{p(\boldsymbol{x}_0|\boldsymbol{y})}\left[\boldsymbol{x}_0\right]$ is unknown.DDPM has a similar problem, it uses VI to fit the posterior distribution, and its neural network learns the residual, which we already know is actually score. DDPM objective function before variational is

$$DDPM:\min D_{kL}\left(q\left(x_{1:T}\mid x_0\right)\|p_0\left(x_{1:T}\mid x_0\right)\right) \tag{10}$$

In this paper, the author also uses VI to fit the posterior distribution, and the objective function of this paper is

$$RMP:\min D_{KL}(q\left(\boldsymbol{x}_{0:T}\mid \boldsymbol{y}\right)\|p\left(\boldsymbol{x}_{0:T}\mid \boldsymbol{y}\right)) \tag{11}$$

The objective function of RMP can be simplified to

$$RMP:\min \sum_{k=T-1}^{0} E_{q(\boldsymbol{x}_{k+1}|\boldsymbol{y})}\left[D_{KL}(q_k\left(\boldsymbol{x}_k\mid \boldsymbol{x}_{k+1},\boldsymbol{y}\right)||p_k\left(\boldsymbol{x}_k\mid \boldsymbol{x}_{k+1},\boldsymbol{y}\right))\right] \tag{12}$$

$q\left(\boldsymbol{x}_{k+1}\mid \boldsymbol{y}\right)$is used to approximate $p\left(\boldsymbol{x}_{k+1}\mid \boldsymbol{y}\right)$. In proposition 2.1 we have learned that $p\left(\boldsymbol{x}_{k+1}\mid \boldsymbol{y}\right)$is a Gaussian distribution, So here we can see $D_{KL}(q_k\left(\boldsymbol{x}_k\mid \boldsymbol{x}_{k+1},\boldsymbol{y}\right)||p_k\left(\boldsymbol{x}_k\mid \boldsymbol{x}_{k+1},\boldsymbol{y}\right))$ for Gaussian weighting, the author mentioned in the paper, Optimizing this objective function is equivalent to the optimization problem:

$$RMP:q_k^\star=\arg\min D_{KL}(q_k\left(\boldsymbol{x}_k\mid \boldsymbol{x}_{k+1},\boldsymbol{y}\right)||p_k\left(\boldsymbol{x}_k\mid \boldsymbol{x}_{k+1},\boldsymbol{y}\right)),\forall k=0,\cdots,T-1 \tag{13}$$

Compare the simplified objective function in DDPM:

$$DDPM:\min \sum_{t>0} \underbrace{D_{\mathrm{KL}}\left(q\left(\mathbf{x}_{t-1}\mid \mathbf{x}_t,\mathbf{x}_0\right)\|p_\theta\left(\mathbf{x}_{t-1}\mid \mathbf{x}_t\right)\right)}_{L_{t-1}} \tag{14}$$

DDPM is also optimized with equal weights in the training process, because VI is used for approximation, so DDPM is similar to the RMP-VI method in this paper.

Proposition 3.1. For a diffusion process with forward process , the KL divergence between variational $q\left(\boldsymbol{x}_{0:T} \mid \boldsymbol{y}\right)$ and joint posterior $p\left(\boldsymbol{x}_{0:T} \mid \boldsymbol{y}\right)$ equals

$$KL(q\|p) = \sum_{k=T-1}^{0} \int q\left(\boldsymbol{x}_{k+1} \mid \boldsymbol{y}\right) \int q_k\left(\boldsymbol{x}_k \mid \boldsymbol{x}_{k+1}, \boldsymbol{y}\right) \log \frac{q_k\left(\boldsymbol{x}_k \mid \boldsymbol{x}_{k+1}, \boldsymbol{y}\right)}{p_k\left(\boldsymbol{x}_k \mid \boldsymbol{x}_{k+1}, \boldsymbol{y}\right)} d\boldsymbol{x}_k d\boldsymbol{x}_{k+1}$$

and the minimization of $KL(q\|p)$ is equivalent to the minimization of

$$KL\left(q_k \| p_k\right) = \int q_k\left(\boldsymbol{x}_k \mid \boldsymbol{x}_{k+1}, \boldsymbol{y}\right) \log \frac{q_k\left(\boldsymbol{x}_k \mid \boldsymbol{x}_{k+1}, \boldsymbol{y}\right)}{p_k\left(\boldsymbol{x}_k \mid \boldsymbol{x}_{k+1}, \boldsymbol{y}\right)} d\boldsymbol{x}_k, \forall k = 0, \cdots, T-1$$

Proof. We have

$$\begin{aligned}
p\left(x_k \mid x_{k+1:T}, \boldsymbol{y}\right) &= \frac{p\left(x_{k:T}, \boldsymbol{y}\right)}{p\left(x_{k+1:T}, \boldsymbol{y}\right)} \\
&= \frac{\int p\left(x_0\right) p\left(\boldsymbol{y} \mid x_0\right) p\left(\boldsymbol{x}_{k:T} \mid x_0\right) dx_0}{\int p\left(x_0\right) p\left(\boldsymbol{y} \mid x_0\right) p\left(x_{k+1:T} \mid x_0\right) dx_0} \\
&= \frac{p\left(\boldsymbol{x}_{k+1:T} \mid \boldsymbol{x}_k\right) \int p\left(x_0\right) p\left(\boldsymbol{y} \mid x_0\right) p\left(x_k \mid x_0\right) dx_0}{p\left(x_{k+2:T} \mid x_{k+1}\right) \int p\left(x_0\right) p\left(\boldsymbol{y} \mid x_0\right) p\left(x_{k+1} \mid x_0\right) dx_0} \\
&= \frac{p\left(x_{k+1} \mid x_k\right) \int p\left(x_0\right) p\left(\boldsymbol{y} \mid x_0\right) p\left(x_k \mid x_0\right) dx_0}{\int p\left(x_0\right) p\left(\boldsymbol{y} \mid x_0\right) p\left(x_{k+1} \mid x_0\right) dx_0} \\
&= \frac{p\left(x_k, x_{k+1}, \boldsymbol{y}\right)}{p\left(x_{k+1}, \boldsymbol{y}\right)} = p_k\left(x_k \mid \boldsymbol{x}_{k+1}, \boldsymbol{y}\right)
\end{aligned} \tag{15}$$

We minimize the KL divergence between variational joint posterior $q\left(\boldsymbol{x}_{0:T} \mid \boldsymbol{y}\right) = q\left(\boldsymbol{x}_T \mid \boldsymbol{y}\right) \prod_{k=T-1}^{0} q_k\left(\boldsymbol{x}_k \mid x_{k+1}, \boldsymbol{y}\right)$ and joint posterior $p\left(\boldsymbol{x}_{0:T} \mid \boldsymbol{y}\right) = p\left(\boldsymbol{x}_T \mid \boldsymbol{y}\right) \prod_{k=T-1}^{0} p\left(\boldsymbol{x}_k \mid \boldsymbol{x}_{k+1:T}, \boldsymbol{y}\right)$ to obtain the optimal variational distribution q :

$$\begin{aligned}
\mathcal{F} &= \int q\left(\boldsymbol{x}_{0:T} \mid \boldsymbol{y}\right) \log \frac{q\left(x_{0:T} \mid \boldsymbol{y}\right)}{p\left(x_{0:T} \mid \boldsymbol{y}\right)} dx_{0:T} \\
&= \sum_{k=T-1}^{0} \int q\left(x_{0:T} \mid \boldsymbol{y}\right) \log \frac{q_k\left(x_k \mid x_{k+1}, \boldsymbol{y}\right)}{p\left(x_k \mid x_{k+1:T}, \boldsymbol{y}\right)} dx_{0:T} \\
&= \sum_{k=T-1}^{0} \int q\left(x_{0:T} \mid \boldsymbol{y}\right) \log \frac{q_k\left(x_k \mid x_{k+1}, \boldsymbol{y}\right)}{p_k\left(\boldsymbol{x}_k \mid x_{k+1}, \boldsymbol{y}\right)} dx_{0:T} \\
&= \sum_{k=T-1}^{0} \int q\left(x_{k:k+1} \mid \boldsymbol{y}\right) \log \frac{q_k\left(x_k \mid x_{k+1}, \boldsymbol{y}\right)}{p_k\left(x_k \mid x_{k+1}, \boldsymbol{y}\right)} dx_k dx_{k+1} \\
&= \sum_{k=T-1}^{0} \int q\left(x_{k+1} \mid \boldsymbol{y}\right) \int q_k\left(x_k \mid x_{k+1}, \boldsymbol{y}\right) \log \frac{q_k\left(x_k \mid x_{k+1}, \boldsymbol{y}\right)}{p_k\left(x_k \mid x_{k+1}, \boldsymbol{y}\right)} dx_k dx_{k+1}
\end{aligned}$$

Thus, the minimization of $\mathcal{F}$ is equivalent to the minimization of

$$\mathcal{F}_k = \int q_k\left(\boldsymbol{x}_k \mid x_{k+1}, \boldsymbol{y}\right) \log \frac{q_k\left(\boldsymbol{x}_k \mid x_{k+1}, \boldsymbol{y}\right)}{p_k\left(\boldsymbol{x}_k \mid x_{k+1}, \boldsymbol{y}\right)} dx_k, \forall k = 0, 1, \cdots T-1$$

□

It is worth noting that eq.(15) reflects the **inverse process of a Markov process**, which is also a Markov process. In this way, we successfully transform the variational inference into an optimization problem, which can be solved using **any feasible optimization method**.

3.2 VI by Natural Gradient Dsecent

Here, the author adopts mini-batch stochastic gradient descent to deal with the optimization problem, and the final iterative formula is

$$\begin{aligned}
&\boldsymbol{\mu}_k \leftarrow \boldsymbol{\mu}_k + s_1 \Lambda_k^{-1} \frac{1}{L} \sum_{i=1}^{L} \nabla_{\boldsymbol{x}_k} \log p_k \left(\boldsymbol{x}_k \mid \boldsymbol{x}_{k+1}, \boldsymbol{y}\right) \Big|_{\boldsymbol{x}_k = \boldsymbol{x}_k^{(i)} \sim q_k} \\
&\Lambda_k \leftarrow \Lambda_k - s_2 \left(N \Lambda_k + \frac{1}{L} \sum_{i=1}^{L} \operatorname{Tr}\left(\nabla_{\boldsymbol{x}_k}^2 \log p_k \left(\boldsymbol{x}_k \mid \boldsymbol{x}_{k+1}, \boldsymbol{y}\right)\right) \Big|_{\boldsymbol{x}_k = \boldsymbol{x}_k^{(i)} \sim q_k} \right)
\end{aligned} \tag{16}$$

3.3 Score-Based Gradient Calculation

The title of this article is *Score-Based Variational Inference for Inverse Problems*, which does not only refer to the use of VP-SDE and VE-SDE to solve inverse problems. It also means that **score is used to simplify the computational complexity** of gradient descent in the variational inference optimization of RMP framework.

In gradient descent, The update of $\boldsymbol{\mu}_k$requires us to calculate the gradient $\nabla_{\boldsymbol{x}_k} \log p_k \left(\boldsymbol{x}_k \mid \boldsymbol{x}_{k+1}, \boldsymbol{y}\right)$. There are variables in this gradient that we do not know, and we need to use learning to calculate them.

From Bayes' rule, the score of reverse conditional $p\left(\boldsymbol{x}_k \mid \boldsymbol{x}_{k+1}, \boldsymbol{y}\right) = \frac{p(\boldsymbol{x}_{k+1} \mid \boldsymbol{x}_k) p(\boldsymbol{x}_k \mid \boldsymbol{y})}{p(\boldsymbol{x}_{k+1} \mid \boldsymbol{y})}$ is given by $\nabla_{\boldsymbol{x}_k} \log p_k \left(\boldsymbol{x}_k \mid \boldsymbol{x}_{k+1}, \boldsymbol{y}\right) = \nabla_{\boldsymbol{x}_k} \log p\left(\boldsymbol{x}_{k+1} \mid \boldsymbol{x}_k\right) + \nabla_{\boldsymbol{x}_k} \log p\left(\boldsymbol{y} \mid \boldsymbol{x}_k\right) + \nabla_{\boldsymbol{x}_k} \log p\left(\boldsymbol{x}_k\right)$, where $\nabla_{\boldsymbol{x}_k} \log p\left(\boldsymbol{x}_k\right)$ is the noisy **score function** which can be approximated by a well-trained score network $\boldsymbol{s}_{\boldsymbol{\theta}}\left(\boldsymbol{x}_k, \sigma_k\right)$ and $\nabla_{\boldsymbol{x}_k} \log p\left(\boldsymbol{x}_{k+1} \mid \boldsymbol{x}_k\right)$ can be calculated explicitly for both VE and VP diffusion models. For VE diffusion $\nabla_{\boldsymbol{x}_k} \log p\left(\boldsymbol{x}_{k+1} \mid \boldsymbol{x}_k\right) = \frac{\boldsymbol{x}_{k+1} - \boldsymbol{x}_k}{\sigma_{k+1}^2 - \sigma_k}$. However, the likelihood score, i.e., the gradient of logarithm conditional $\nabla_{\boldsymbol{x}_k} \log p\left(\boldsymbol{y} \mid \boldsymbol{x}_k\right)$ is hard to handle in general. In the paper of DPS, the authors proposed the following approximation that can be applied for general measurements:

Proposition 3.2.

$$\log p\left(\boldsymbol{y} \mid \boldsymbol{x}_k\right) \approx \log p\left(\boldsymbol{y} \mid \hat{\boldsymbol{x}}_0\left(\boldsymbol{x}_k\right)\right)$$

where $\hat{\boldsymbol{x}}_0\left(\boldsymbol{x}_k\right)$ is the MMSE estimate of $\boldsymbol{x}_0$.

Proof.

$$
\begin{aligned}
& p\left(y \mid x_k\right) \\
= & \int p\left(y \mid x_0\right) p\left(x_0 \mid x_k\right) d x_0 \\
= & E_{p\left(x_0 \mid x_k\right)}\left[p\left(y \mid x_0\right)\right] \\
\approx & p\left(y \mid E\left[x_0 \mid x_k\right]\right) \\
= & p\left(y \mid \hat{x}_0\left(x_k\right)\right)
\end{aligned}
\tag{17}
$$

The approximation error of eq.(17) can be quantified with **Jenson's Gap** as given in the paper of DPS. □

For VE diffusion, according to the Tweedie formula (For Gaussian likelihood cases in Bayesian estimation):

$$\hat{\boldsymbol{x}}_0\left(\boldsymbol{x}_k\right)=\mathbb{E}_{p\left(\boldsymbol{x}_0 \mid \boldsymbol{x}_k\right)}\left[\boldsymbol{x}_0\right]=\boldsymbol{x}_k+\sigma_k^2 \nabla_{\boldsymbol{x}_k} \log p\left(\boldsymbol{x}_k\right)$$

As a conclusion, the gradient is calculated as

$$\nabla_{\boldsymbol{x}_k} \log p_k\left(\boldsymbol{x}_k \mid \boldsymbol{x}_{k+1}, \boldsymbol{y}\right) \approx \nabla_{\boldsymbol{x}_k} \log p\left(\boldsymbol{x}_{k+1} \mid \boldsymbol{x}_k\right)+\gamma_k \nabla_{\boldsymbol{x}_k} \log p\left(\boldsymbol{y} \mid \hat{\boldsymbol{x}}_0\left(\boldsymbol{x}_k\right)\right)+\boldsymbol{s}_{\boldsymbol{\theta}}\left(\boldsymbol{x}_k, \sigma_k\right)$$

where the parameter γ_k is added to balance the approximated likelihood score and prior score. The setting of parameter γ_k is key to algorithm's performance. We set $\gamma_k=\zeta \frac{\left\|\nabla_{\boldsymbol{x}_k} \log p\left(\boldsymbol{x}_{k+1} \mid \boldsymbol{x}_k\right)\right\|_2}{\left\| \log p\left(\boldsymbol{y} \mid \hat{\boldsymbol{x}}_0\left(\boldsymbol{x}_k\right)\right)\right\|_2}$ where ζ is tuned for different problems. The idea behind the strategy is that we should always **keep a balance between the data score and the likelihood score**.

3.4 Hessian calculation and Fixed Precision Update

The Hessian matrix $\nabla_{\boldsymbol{x}_k,}^2 \log p_k\left(\boldsymbol{x}_k \mid \boldsymbol{x}_{k+1}, \boldsymbol{y}\right)$ is difficult to acquire in general. Thus, we introduce an approximation that does not require the calculation of Hessian. Because of VI, the update of precision $\Lambda_k^{(i+1)}$ **converges to the precision** of $p_k\left(\boldsymbol{x}_k \mid \boldsymbol{x}_{k+1}, \boldsymbol{y}\right)$. Thus, we can **fix the update** of $\Lambda_k^{(i+1)}$ in Algorithm 1 to the precision of $p_k\left(\boldsymbol{x}_k \mid \boldsymbol{x}_{k+1}, \boldsymbol{y}\right)$ and only update $\boldsymbol{\mu}_k$ at each step. According to Proposition 1, for VE diffusion model, if we set $\boldsymbol{C}_{\boldsymbol{x}_0}=v_{\boldsymbol{x}_0} \boldsymbol{I}$, then the inverse of precision is given by

$$\left(\Lambda_k^{(i)}\right)^{-1}=\frac{\left(\sigma_k^2+v_{\boldsymbol{x}_0}\right)\left(\sigma_{k+1}^2-\sigma_k^2\right)}{\sigma_{k+1}^2+v_{\boldsymbol{x}_0}}$$

We cannot calculate $\left(\Lambda_k^{(i)}\right)^{-1}$ directly since $v_{\boldsymbol{x}_0}$ is unknown. However, for the cases that k is close to T, σ_k^2 is large enough comparing to $v_{\boldsymbol{x}_0}$. Thus, $\left(\Lambda_k^{(i)}\right)^{-1}$ can be approximated by $\frac{\sigma_k^2\left(\sigma_{k+1}^2-\sigma_k^2\right)}{\sigma_{k+1}^2}$. For the case that k is close to $0, \sigma_{k+1}^2, \sigma_k^2 \rightarrow 0$, we have $\left(\Lambda_k^{(i)}\right)^{-1} \approx \sigma_{k+1}^2-\sigma_k^2$.

With stochastic NGD based VI and score-based approximations of gradient and Hessian, we summarize a practical algorithm given in Algorithm2.

```
Algorithm 2: RMP with Score-based Stochastic NGD
Input   :$\boldsymbol{y}$, $s_1$, $T$, $T_{in}$, $T_s$, $\boldsymbol{x}_T$, $\boldsymbol{\mu}_{T-1}^0$
for $k = T - 1 : 0$ do
    For VE $\Lambda_k^{-1} = \frac{\sigma_k^2(\sigma_{k+1}^2 - \sigma_k^2)}{\sigma_{k+1}^2}$if $k > T_s$ else $\Lambda_k^{-1} = \sigma_{k+1}^2 - \sigma_k^2$ (for VP $\Lambda_k^{-1} = \beta_{k+1}$)
    for $i = 0 : T_{in} - 1$ do
        $\boldsymbol{\mu}_k^{(i+1)} = \boldsymbol{\mu}_k^{(i)} + s_1 \Lambda_k^{-1} \left(\nabla_{\boldsymbol{x}_k} \log p(\boldsymbol{x}_{k+1}|\boldsymbol{x}_k) + \gamma_k \nabla_{\boldsymbol{x}_k} \log p(\boldsymbol{y}|\hat{\boldsymbol{x}}_0(\boldsymbol{x}_k)) + \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_k, \sigma_k)\right)$
        where $\boldsymbol{x}_k \sim \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k^{(i)}, \Lambda_k^{-1}\boldsymbol{I})$, $\hat{\boldsymbol{x}}_0(\boldsymbol{x}_k) = \boldsymbol{x}_k + \sigma_k^2 \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_k, \sigma_k)$ for VE
        (for VP $\hat{\boldsymbol{x}}_0(\boldsymbol{x}_k) = \frac{1}{\sqrt{\bar{\alpha}_k}}(\boldsymbol{x}_k + (1 - \bar{\alpha}_k)\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_k, \sigma_k))$
    end
    $\boldsymbol{x}_k = \boldsymbol{\mu}_k^{(T_{in})}$ and $\boldsymbol{\mu}_{k-1}^{(0)} = \boldsymbol{\mu}_k^{(T_{in})}$
end
Output :$\boldsymbol{\mu}_0^{(T_{in})}$
```

Figure 3: RMP-VI pseudocode

4 Experiments