ICML International Conference On Machine Learning

code available

Efficiently Access Diffusion Fisher: Within the Outer Product Span Space

Fangyikang Wang[1*], Hubery Yin[2*], Shaobin Zhuang[3], Huminhao Zhu[1], Yinan Li[1], Lei Qian[1], Chao Zhang[1†], Hanbin Zhao[1], Hui Qian[1], Chen Li[2]

[1] Zhejiang University, [2] WeChat, Tencent Inc., [3] Shanghai Jiao Tong University

CONTENTS

1. Background
2. Diffusion Fisher
3. DF Trace Matching
4. DF Endpoint Approximation
5. DF Optimal Transport

PART ONE

Background

Background: Fisher Information in Diffusion Models

The diffusion Fisher (DF) in DMs, defined as the negative Hessian of the diffused distributions' log density:

$$\boldsymbol{F}_t(\boldsymbol{x}_t, t) := -\frac{\partial^2}{\partial \boldsymbol{x}_t^2} \log q_t\left(\boldsymbol{x}_t, t\right)$$

Current practices typically approximate the diffusion Fisher by applying auto-differentiation to the learned score network:

$$\begin{aligned}\boldsymbol{F}_t(\boldsymbol{x}_t, t) &= -\frac{\partial}{\partial \boldsymbol{x}_t}\left(\frac{\partial}{\partial \boldsymbol{x}_t} \log p\left(x_t, t\right)\right) \\ &\approx -\frac{\partial}{\partial \boldsymbol{x}_t}\left(-\frac{\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t)}{\sigma_t}\right) = \frac{1}{\sigma_t} \frac{\partial \boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t)}{\partial \boldsymbol{x}_t}\end{aligned} \tag{8}$$

Straightforward, but lacks accuracy guarantee, and is time-consuming

PART TWO

Diffusion Fisher

DF: Within the Outer Product Span Space

Data distribution under the Dirac assumption

$$\text{(Dirac Setting)} \qquad q(\boldsymbol{x}, t)|_{t=0} = \frac{1}{N}\sum_{i=0}^{N} \delta(\boldsymbol{x} - \boldsymbol{y}_i),$$

DF resides within a space spanned by the outer products of score and initial data.

Proposition 1. Defines $v_i(\boldsymbol{x}_t, t)$ as $\exp\left(-\frac{|\boldsymbol{x}_t - \alpha_t \boldsymbol{y}_i|^2}{2\sigma_t^2}\right) \in \mathbb{R}$ and $w_i(\boldsymbol{x}_t, t)$ as $\frac{v_i(\boldsymbol{x}_t, t)}{\sum_j v_j(\boldsymbol{x}_t, t)} \in \mathbb{R}$. If q_0 takes the form as in equation equation 10, the diffusion Fisher matrix of the diffused distribution q_t for $t \in (0, 1]$ can be analytically formulated as follows:

$$\boldsymbol{F}_t(\boldsymbol{x}_t, t) = \frac{1}{\sigma_t^2}\boldsymbol{I} - \frac{\alpha_t^2}{\sigma_t^4}\left[\sum_i w_i \boldsymbol{y}_i \boldsymbol{y}_i^\top - \left(\sum_i w_i \boldsymbol{y}_i\right)\left(\sum_i w_i \boldsymbol{y}_i\right)^\top\right] \qquad (11)$$

where we have simplified $w_i(\boldsymbol{x}_t, t)$ to w_i, as it does not lead to any confusion.

DF: Within the Outer Product Span Space

Data distribution under the general assumption

$$\text{(General Setting)} \qquad q_0 \in \mathcal{P}_2(\mathbb{R}^d),$$

DF resides within a space spanned by an infinite outer product basis of score and initial data.

Proposition 3. Let us define $v(\boldsymbol{x}_t, t, \boldsymbol{y})$ as $\exp\left(-\frac{|\boldsymbol{x}_t - \alpha_t \boldsymbol{y}|^2}{2\sigma_t^2}\right) \in \mathbb{R}$ and $w(\boldsymbol{x}_t, t, \boldsymbol{y})$ as $\frac{v(\boldsymbol{x}_t, t, \boldsymbol{y})}{\int_{\mathbb{R}^d} v(\boldsymbol{x}_t, t, \boldsymbol{y}) \mathrm{d}q_0(\boldsymbol{y})} \in \mathbb{R}$. If q_0 takes the form as in equation 12, the diffusion Fisher matrix of the diffused distribution q_t for $t \in (0, 1]$ can be analytically formulated as follows:

$$\boldsymbol{F}_t(\boldsymbol{x}_t, t) = \frac{1}{\sigma_t^2}\boldsymbol{I} - \frac{\alpha_t^2}{\sigma_t^4}\left[\int w(\boldsymbol{y})\boldsymbol{y}\boldsymbol{y}^\top \mathrm{d}q_0 - \left(\int w(\boldsymbol{y})\boldsymbol{y}\mathrm{d}q_0\right)\left(\int w(\boldsymbol{y})\boldsymbol{y}\mathrm{d}q_0\right)^\top\right] \tag{13}$$

where we simply write $w(\boldsymbol{x}_t, t, \boldsymbol{y})$ as $w(\boldsymbol{y})$, as long as it does not lead to any confusion.

PART THREE

DF Trace Matching

DF Trace Matching

The log-likelihood of DM can be computed through:

$$\frac{\partial \log q_t(\boldsymbol{x}_t, t)}{\partial t} = -\operatorname{tr}\left(\frac{\partial}{\partial \boldsymbol{x}_t}\left(f(t)\boldsymbol{x}_t - \frac{1}{2}g^2(t)\partial_{\boldsymbol{x}_t}\log q_t(\boldsymbol{x}_t, t)\right)\right)$$
$$= -\operatorname{tr}\left(\left(f(t)\boldsymbol{I} - \frac{1}{2}g^2(t)\frac{\partial^2}{\partial {\boldsymbol{x}_t}^2}\log q_t(\boldsymbol{x}_t, t)\right)\right)$$
$$= -f(t)d - \frac{g^2(t)}{2}\operatorname{tr}\left(\boldsymbol{F}_t(\boldsymbol{x}_t, t)\right) \qquad (14)$$

We need to access the trace of diffusion Fisher!

Current VJP-based method:

$$\operatorname{tr}\left(\boldsymbol{F}_t(\boldsymbol{x}_t, t)\right) \approx \frac{1}{\sigma_t}\sum_{i=1}^{d}\frac{\partial\left[\left\langle \boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t) \middle| \boldsymbol{e}^{(i)}\right\rangle\right]}{\partial \boldsymbol{x}_t}.$$

A time complexity of $\mathcal{O}(d^2)$

DF Trace Matching

Our approach for accessing the trace of diffusion Fisher:

Proposition 5. In the same context as Proposition 1, the trace of the diffusion Fisher matrix for the diffused distribution q_t, where $t \in (0, 1]$, is given by:

$$\operatorname{tr}(\boldsymbol{F}_t(\boldsymbol{x}_t, t)) = \frac{d}{\sigma_t^2} - \frac{\alpha_t^2}{\sigma_t^4}\left[\sum_i w_i \|\boldsymbol{y}_i\|^2 - \left\|\sum_i w_i \boldsymbol{y}_i\right\|^2\right] \tag{16}$$

Learned via a trace network

Algorithm 1 Training of DF-TM Network

1: **Input**: data space dimension d, initial network $\boldsymbol{t}_\theta(\cdot, \cdot) : \mathbb{R}^d \times \mathbb{R} \mapsto \mathbb{R}$, noise schedule $\{\alpha_t\}$ and $\{\sigma_t\}$.
2: **repeat**
3: $\quad \boldsymbol{x}_0 \sim q_0(\boldsymbol{x}_0)$
4: $\quad t \sim \text{Uniform}(\{1, \ldots, T\})$
5: $\quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
6: $\quad \boldsymbol{x}_t = \alpha_t \boldsymbol{x}_0 + \sigma_t \boldsymbol{\varepsilon}$
7: $\quad$ Take gradient descent on $\nabla_\theta \left| \boldsymbol{t}_\theta(\boldsymbol{x}_t, t) - \frac{\|\boldsymbol{x}_0\|^2}{d} \right|^2$
8: **until** converged
9: **Output**: $\boldsymbol{t}_\theta(\cdot, \cdot)$

Approximated via learned score

Proposition 2. Given the diffusion training loss in equation 4, and if q_0 conforms to the form presented in equation 10, then the optimal $\bar{\boldsymbol{y}}_\theta(\boldsymbol{x}_t, t)$ can accurately estimate $\sum_i w_i \boldsymbol{y}_i$.

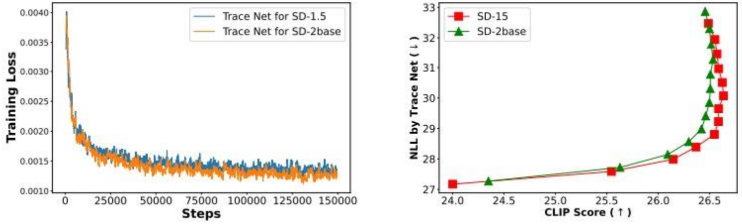
Proposition 6. $\forall (x_t, t) \in \mathbb{R}^d \times \mathbb{R}_{\geq 0}$, the optimal $t_\theta(\boldsymbol{x}_t, t)$s trained by the objective in Algorithm 1 are equal to $\frac{1}{d} \sum_i w_i(\boldsymbol{x}_t, t) \|\boldsymbol{y}_i\|^2$.

DF Trace Matching

Our DF-TM method:

$$\operatorname{tr}\left(\boldsymbol{F}_t(\boldsymbol{x}_t,t)\right) \approx \frac{d}{\sigma_t^2} - \frac{\alpha_t^2}{\sigma_t^4}\left(d * t_\theta(\boldsymbol{x}_t,t) - \left\|\boldsymbol{y}_\theta(\boldsymbol{x}_t,t)\right\|^2\right) \tag{17}$$

Training stability of our DF-TM method:



(a) (b)

Figure 1: (a) The training loss of DF-TM for SD-1.5 and SD-2base. It demonstrates commendable convergence behavior. (b) The trade-off curve of NLL and Clip score of SD-1.5 and SD-2base across various guidance scales in [1.5, 2.5, ..., 12.5, 13.5]

Theoretical analysis of our DF-TM method:

Proposition 7. Assume the approximation error on $t_\theta(\boldsymbol{x}_t,t)$ is δ_1 and on $\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t,t)$ is δ_2, then the approximation error of the approximated Fisher trace in equation 17 is at most $\frac{\alpha_t^2}{\sigma_t^4}\delta_1 + \frac{1}{\sigma_t^2}\delta_2^2$.

DF Trace Matching

Experiments on our DF-TM method:

Methods	The relative error of NLL evaluation					
	t = 1.0	t = 0.8	t = 0.6	t = 0.4	t = 0.2	t = 0.0
VJP (eq. 15)	6.68%	5.79%	10.46%	20.13%	51.14%	70.95%
DF-TM (**Ours**)	**3.41%**	**4.56%**	**4.13%**	**4.28%**	**5.33%**	**5.81%**

Table 2: Comparison of the VJP method and our DF-TM in terms of the diffusion Fisher trace evaluation error across different timesteps. The error is evaluated on the 2-D chessboard data with the VE schedule.

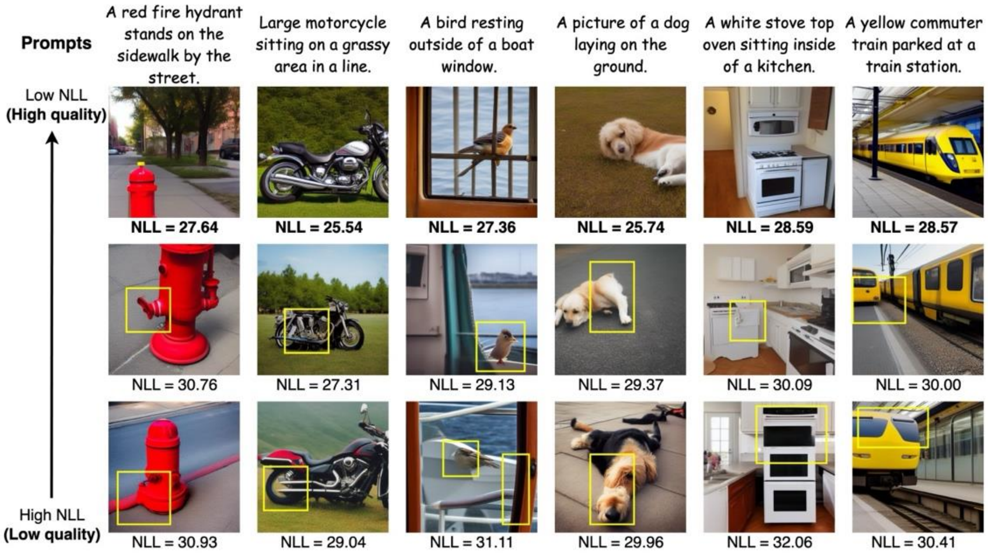


Figure 2: Our DF-TM method facilitates the effective evaluation of the NLL of generated samples with varying seeds. It can be demonstrated that a lower NLL signifies a region of higher possibility, thereby consistently indicating superior image quality.

PART FOUR

DF Endpoint Approximation

DF Endpoint Approximation

When doing adjoint ODE, we need to access the matrix multiplication of the diffusion Fisher:

Consider optimizing a scalar-valued loss function $\mathcal{L}(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}$, which takes $\boldsymbol{x}_0$ in the data space as input. Adjoint guidance is implemented by applying gradient descent on $\boldsymbol{x}_t$ in the direction of $\frac{\partial \mathcal{L}(\boldsymbol{x}_0(\boldsymbol{x}_t))}{\partial \boldsymbol{x}_t}$. The essence of adjoint guidance is to use the gradient at $t = 0$ and follow the adjoint ODE (Pollini et al., 2018; Chen et al., 2018) to compute $\boldsymbol{\lambda}_t := \frac{\partial \mathcal{L}(\boldsymbol{x}_0(\boldsymbol{x}_t))}{\partial \boldsymbol{x}_t}$ for $t > 0$.

$$\frac{\mathrm{d}\boldsymbol{\lambda}_t}{\mathrm{d}t} = -\boldsymbol{\lambda}_t^\top \frac{\partial \boldsymbol{h}_\theta(\boldsymbol{x}_t, t)}{\partial \boldsymbol{x}_t}, \quad \boldsymbol{\lambda}_0 = \frac{\partial \mathcal{L}(\boldsymbol{x}_0)}{\partial \boldsymbol{x}_0} \tag{18}$$

The current method mainly uses the VJP-based method, which needs time-consuming auto-differentiation.

$$\boldsymbol{F}(\boldsymbol{x}_t, t)^\top \boldsymbol{\lambda}_t \approx \frac{1}{\sigma_t} \frac{\partial \boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t)}{\partial \boldsymbol{x}_t}^\top \boldsymbol{\lambda}_t$$

$$\approx \frac{1}{\sigma_t} \frac{\partial [\langle \boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t) | \boldsymbol{\lambda}_t \rangle]}{\partial \boldsymbol{x}_t}$$

DF Endpoint Approximation

Our DF-EA method:

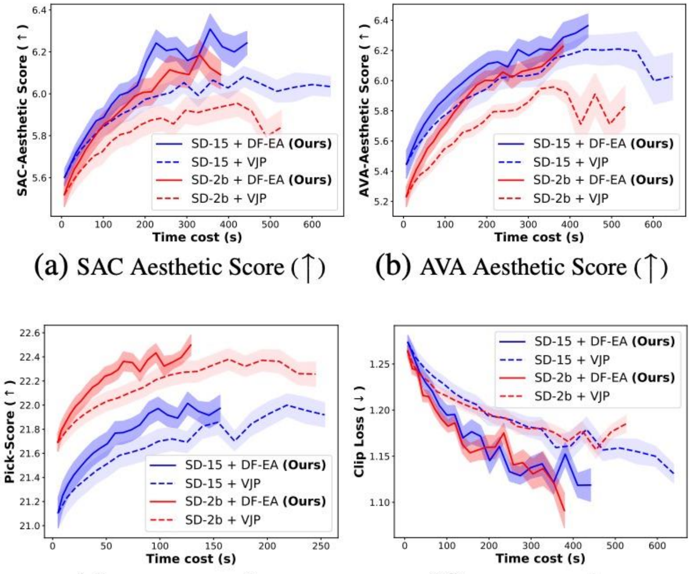
$$
\begin{aligned}
&\boldsymbol{F}(\boldsymbol{x}_t, t)^\top \boldsymbol{\lambda}_t \\
\approx &\left(\frac{1}{\sigma_t^2}\boldsymbol{I} - \frac{\alpha_t^2}{\sigma_t^4}\left(\sum_i w_i \boldsymbol{y}_i \boldsymbol{y}_i^\top - \bar{\boldsymbol{y}}_\theta(\boldsymbol{x}_t, t)\bar{\boldsymbol{y}}_\theta(\boldsymbol{x}_t, t)^\top\right)\right)^\top \boldsymbol{\lambda}_t \\
\approx &\left(\frac{1}{\sigma_t^2}\boldsymbol{I} - \frac{\alpha_t^2}{\sigma_t^4}\left(\boldsymbol{x}_0 \boldsymbol{x}_0^\top - \bar{\boldsymbol{y}}_\theta(\boldsymbol{x}_t, t)\bar{\boldsymbol{y}}_\theta(\boldsymbol{x}_t, t)^\top\right)\right)^\top \boldsymbol{\lambda}_t \\
= &\frac{1}{\sigma_t^2}\boldsymbol{\lambda}_t - \frac{\alpha_t^2}{\sigma_t^4}\left\langle \boldsymbol{x}_0, \boldsymbol{\lambda}_t \right\rangle \boldsymbol{x}_0 + \frac{\alpha_t^2}{\sigma_t^4}\left\langle \bar{\boldsymbol{y}}_\theta(\boldsymbol{x}_t, t), \boldsymbol{\lambda}_t \right\rangle \bar{\boldsymbol{y}}_\theta(\boldsymbol{x}_t, t)
\end{aligned}
\tag{20}
$$

Theoretical approximation error bound:

Proposition 8. Assume that the approximation error on $\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t)$ is δ_2, the approximation error of the DF-EA linear operator, as referenced in 20, is at most $\frac{\alpha_t^2}{\sigma_t^3}\left(2\mathcal{D}_y^2 + \sqrt{d}\delta_2\right)$ when measured in terms of the Hilbert–Schmidt norm.

DF Endpoint Approximation

Experiments on our DF-EA method:



(a) SAC Aesthetic Score (↑) (b) AVA Aesthetic Score (↑)

(c) Pick-Score (↑) (d) Clip Loss (↓)

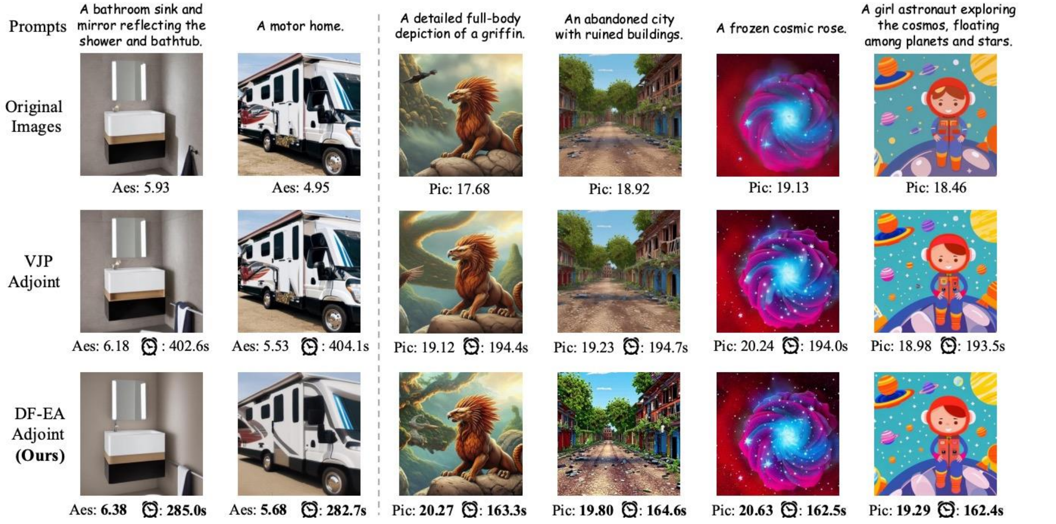


Figure 4: Visual comparison of DF-EA (**Ours**) and VJP in the adjoint improvement task on (left) SAC aesthetic score and (right) Pick-Score. DF-EA consistently generates images with better visual effects and reduced time expenditure.

PART FIVE

DF Optimal Transport

DF Optimal Transport

Numerical test for the OT property of PF-ODE map

Corollary 1. Denote the diffeomorphism deduced by the PF-ODE in equation 5 as follows

$$T_{s,t} : \mathbb{R}^n \longrightarrow \mathbb{R}^n; \boldsymbol{x}_s \longmapsto \boldsymbol{x}_t, \quad \forall t \geq s > 0. \tag{21}$$

The diffeomorphism $T_{s,T}$ is a Monge optimal transport map **if and only if** the normalized fundamental matrix for $\boldsymbol{B}(t) \equiv \boldsymbol{B}(t, \boldsymbol{x}_t)$ at s is s.p.d. for every PF-ODE chain that starts from a $\boldsymbol{x}_T \in \mathbb{R}^d$. where

$$\boldsymbol{B}(t, \boldsymbol{x}_t) = \left[f(t) - \frac{g^2(t)}{2\sigma_t^2}\right] \boldsymbol{I} + \frac{\alpha_t^2 g^2(t)}{2\sigma_t^4} \left[\sum_i w_i \boldsymbol{y}_i \boldsymbol{y}_i^\top - \left(\sum_i w_i \boldsymbol{y}_i\right)\left(\sum_i w_i \boldsymbol{y}_i\right)^\top\right]. \tag{22}$$

The definition of the normalized fundamental matrix is deferred to Appendix A.10.

Algorithm 2 Numerical OT test for PF-ODE map

1: **Input**: initial data $\{\boldsymbol{y}_i\}_{i=1}^N$, noise schedule $\{\alpha_t\}$ and $\{\sigma_t\}$, discretization steps M.
2: Initialize $\boldsymbol{A}_M = \boldsymbol{I}, \boldsymbol{x}_M \sim \mathcal{N}(0, \sigma_T \boldsymbol{I})$.
3: **for** $i = M, M-1, \cdots, 1$ **do**
4: $\quad \mathrm{d}t = t_{i-1} - t_i$.
5: $\quad$ Calculate $\boldsymbol{B}_i$ by equation 22.
6: $\quad \boldsymbol{A}_{i-1} = \boldsymbol{A}_i + \mathrm{d}t * \boldsymbol{A}_i^\top \boldsymbol{B}_i$ {solve fundamental matrix.}
7: $\quad \boldsymbol{x}_{i-1} = \text{PF-ODE Solver}(\boldsymbol{x}_i, i)$
8: **end for**
9: **Output**: $\boldsymbol{A}_0$. {The result fundamental matrix.}

DF Optimal Transport

Numerical OT verification results of common noise schedules:

Initial Data	Single-Gaussian		Affine		Non-affine	
Noise Schedule	Asym.	OT	Asym.	OT	Asym.	OT
VE (Song & Ermon, 2019)	0.00%	✓	0.00%	✓	25.28%	✗
VP (Ho et al., 2020)	0.00%	✓	0.00%	✓	23.36%	✗
sub-VP (Song et al., 2020)	0.00%	✓	0.00%	✓	13.84%	✗
EDM (Karras et al., 2022)	0.00%	✓	0.00%	✓	27.09%	✗

Table 3: Comparison of numerical OT verification results of four commonly used noise schedulers with different initial data.

THANKS!

Codes repository: https://github.com/zituitui/BELM