Bless of Gaussian Noise

$$\nabla \log \rho_t(x) = \frac{t v_t(x) - x}{1 - t}$$

$$\mathrm{KL}(\rho_1 \mid\mid \rho_1') = \int_0^1 \frac{t}{1-t} \mathbb{E}\left[\left\| v_t(X_t) - v_t'(X_t) \right\|^2 \right] \mathrm{d}t.$$

Bless of Gaussian Noise

- When noise X_0 is Gaussian, and the coupling (X_0, X_1) is independent, the RF velocity field is related to the score function.

Velocity field: $v_t(x) = \mathbb{E}[X_1 - X_0 \mid X_t = x]$,

Density: $\rho_t(x) =$ density of X_t (and Z_t),

Score function: $\nabla \log \rho_t(x)$.

Tweedie's Formula: Let ρ_t be the density of $X_t = tX_1 + (1-t)X_0$. Assume $X_0 \sim \mathtt{Normal}(0, I)$ and $X_0 \perp\!\!\!\perp X_1$, we have

$$\nabla \log \rho_t(x) = \frac{tv_t(x) - x}{1-t}$$

- *Score-based generative models of [SSDK+20]*
- Has important implications and applications:
  - Likelihood estimation.
  - Training-free conversion to SDE.
  - Distillation and control.

Score Function of Interpolated Variable

Tweedie (General Case): The density ρ_t of $X_t = (1-t)X_0 + tX_1$ satisfies

$$\nabla \log \rho_t(x) = \frac{1}{1-t}\,\mathbb{E}\left[\nabla_{X_0} \log \rho_{X_0|X_1}(X_0 \mid X_1) \,\middle|\, X_t = x\right].$$

This holds for any (X_0, X_1), whenever relevant densities exist & smooth.

Tweedie (Gaussian): When $X_0|X_1 \sim \texttt{Normal}(0, I)$, we have

$$\begin{aligned}
\nabla_x \log \rho_t(x) &= \mathbb{E}\left[\frac{-X_0}{1-t} \mid X_t = x\right] && //\nabla \log \rho_{X_0|X_1}(x_0|x_1) = -x_0 \\
&= \mathbb{E}\left[\frac{t(X_1 - X_0) - X_t}{1-t} \mid X_t = x\right] && //X_t = tX_1 + (1-t)X_0 \\
&= \frac{t \cdot v_t(x) - x}{1-t}.
\end{aligned}$$

Proof.

Proof of Tweedie's Formula (General Case) The density of $X_t = (1-t)X_0 + tX_1$ can be written as

$$\rho_t(x) = \mathbb{E}_{X_1}\left[\rho_{X_0|X_1}\left(\frac{x - tX_1}{1-t} \,\middle|\, X_1\right) \cdot \frac{1}{1-t}\right].$$

Taking the log and differentiating gives:

$$\nabla_x \log \rho_t(x) = \frac{\mathbb{E}_{X_1}\left[\frac{1}{1-t}\,\nabla\rho_{X_0|X_1}\left(\frac{x-tX_1}{1-t} \,\middle|\, X_1\right)\right]}{\mathbb{E}_{X_1}\left[\rho_{X_0|X_1}\left(\frac{x-tX_1}{1-t} \,\middle|\, X_1\right)\right]}$$
$$= \mathbb{E}\left[\nabla_{X_0} \log \rho_{X_0|X_1}(X_0|X_1) \mid X_t = x\right].$$

□

KL Divergence of Marginals

For any two stochastic processes

- $\{X_t\}$ with marginal ρ_t, RF velocity $v_t(x) = \mathbb{E}\left[\dot{X}_t \mid X_t = x\right]$.
- $\{X'_t\}$ with marginal ρ_t, RF velocity $v'_t(x) = \mathbb{E}\left[\dot{X}'_t \mid X'_t = x\right]$.

We have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(\rho_t \mid\mid \rho'_t) = \mathbb{E}\left[(\nabla \log \rho_t(X_t) - \nabla \log \rho'_t(X_t))^\top (v_t(X_t) - v'_t(X_t))\right]$$

$$\frac{\mathrm{d}}{\mathrm{dt}}\mathtt{KL} = \mathbb{E}\left[\langle \mathtt{score\ difference}, \mathtt{velocity\ difference}\rangle\right].$$

KL Divergence of Marginals

For any two stochastic processes

- $\{X_t\}$ with marginal ρ_t, RF velocity $v_t(x) = \mathbb{E}\left[\dot{X}_t \mid X_t = x\right]$.
- $\{X'_t\}$ with marginal ρ_t, RF velocity $v'_t(x) = \mathbb{E}\left[\dot{X}'_t \mid X'_t = x\right]$.

We have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(\rho_t \mid\mid \rho'_t) = \mathbb{E}\left[(\nabla \log \rho_t(X_t) - \nabla \log \rho'_t(X_t))^\top (v_t(X_t) - v'_t(X_t))\right]$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\texttt{KL} = \mathbb{E}\left[\langle \texttt{score difference}, \texttt{velocity difference} \rangle\right].$$

- If $\nabla \log \rho_t(x) = \frac{t v_t(x) - x}{1-t}$ and $\nabla \log \rho'_t(x) = \frac{t v'_t(x) - x}{1-t}$, we can cancel out score function or velocity:

$$\nabla \log \rho_t(x) - \nabla \log \rho'_t(x) = \frac{t}{1-t}(v_t(x) - v'_t(x)).$$

KL Divergence of Marginals

For two interpolation processes from different data X_1 and X_1':

- $X_t = tX_1 + (1-t)X_0$ with $X_0 \perp\!\!\!\perp X_1$ and $X_0 \sim \texttt{Normal}(0, I)$.
- $X_t' = tX_1' + (1-t)X_0'$ with $X_0' \perp\!\!\!\perp X_1'$ and $X_0' \sim \texttt{Normal}(0, I)$.

We have

$$\mathrm{KL}(\rho_{X_1} \| \rho_{X_1'}) = \int_0^1 \frac{t}{1-t} \mathbb{E}\left[\left\| v_t(X_t) - v_t'(X_t) \right\|^2\right] \mathrm{d}t$$
$$= \int_0^1 \frac{1-t}{t} \mathbb{E}\left[\left\| \nabla \log \rho_t(X_t) - \nabla \log \rho_t'(X_t) \right\|^2\right] \mathrm{d}t.$$

- Connects KL divergence, RF loss, Fisher divergence:
  - Related: JKO Wasserstein gradient flows, De Bruijn's Identity, etc.
- With weight $w_t = \frac{t}{1-t}$, RF training = MLE.
- Applications: Sampling, exponential tilting with Gibbs variational principle.

Likelihood Evaluation

For a given data point $x^{\texttt{eval}}$, we can derive the following formula for log likelihood:

$$\log \rho_t(x^{\texttt{eval}}) = \int_0^1 \frac{t}{1-t}\mathbb{E}[\|\dot{X}_t^{\texttt{eval}}\|^2 - \underbrace{\|\dot{X}_t^{\texttt{eval}} - v_t(X_t^{\texttt{eval}})\|^2}_{\text{loss on data } x^{\texttt{eval}}}]\mathrm{d}t \quad (*)$$

where $X_t^{\texttt{eval}} = tx^{\texttt{eval}} + (1-t)X_0$, with $X_0 \sim \texttt{Normal}(0, I)$.

Likelihood Evaluation

For a given data point $x^{\mathtt{eval}}$, we can derive the following formula for log likelihood:

$$\log \rho_t(x^{\mathtt{eval}}) = \int_0^1 \frac{t}{1-t} \mathbb{E}[\|\dot{X}_t^{\mathtt{eval}}\|^2 - \underbrace{\|\dot{X}_t^{\mathtt{eval}} - v_t(X_t^{\mathtt{eval}})\|^2}_{\text{loss on data } x^{\mathtt{eval}}}]\mathrm{d}t \quad (*)$$

where $X_t^{\mathtt{eval}} = tx^{\mathtt{eval}} + (1-t)X_0$, with $X_0 \sim \mathtt{Normal}(0, I)$.

Another more common formula: Simultaneous change of variable:

$$\log \rho_t(x^{\mathtt{eval}}) = \log \rho_0(z_0^{\mathtt{eval}}) - \int_0^1 \nabla \cdot v_t(z_t^{\mathtt{eval}})\mathrm{d}t,$$

where $\{z_t^{\mathtt{eval}}\}$ is the solution of $\dot{z}_t^{\mathtt{eval}} = v_t(z_t^{\mathtt{eval}})$ with $z_1^{\mathtt{eval}} = x^{\mathtt{eval}}$.

- Eq. (*) offers a faster computation (avoiding ODE solving), but its accuracy relies on how well v_t is learned.

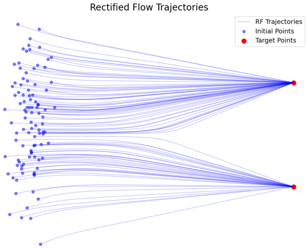
But let us face the issue of singularity...

- The affine relation of v_t and $\nabla \log \rho_t$:

$$\underbrace{\nabla \log \rho_t(x) = \frac{t v_t(x) - x}{1-t}}_{\text{singular at } t=1} \qquad \underbrace{v_t(x) = \frac{(1-t)\nabla \log \rho_t(x) + x}{t}}_{\text{singular at } t=0}.$$

Cause singularity in the formula of KL divergence and log-likelihood.

This is expected, because the ODE can overfit to the delta measure of training data, yielding no finite densities.



Stable Parameterization

Let us exam the boundary conditions carefully [HLX+25]:

- At $t = 1$, $v_1(x) = \mathbb{E}[X_1 - X_0 \mid X_1 = x] = x$.
- At $t = 0$, $\nabla \log \rho_0(x) = -x$, for $\rho_0 \sim \texttt{Normal}(0, I)$.

Hence, the relation is rewrite into a symmetric form:

$$\underbrace{\frac{v_1(x) - v_t(x)}{1 - t}}_{\text{slope of } v_t} = -\underbrace{\frac{\nabla \log \rho_t(x) - \nabla \log \rho_0(x)}{t}}_{\text{slope of } \nabla \log \rho_t}.$$

Stable Parameterization

Let us exam the boundary conditions carefully [HLX⁺25]:

- At $t = 1$, $v_1(x) = \mathbb{E}[X_1 - X_0 \mid X_1 = x] = x$.
- At $t = 0$, $\nabla \log \rho_0(x) = -x$, for $\rho_0 \sim \texttt{Normal}(0, I)$.

Hence, the relation is rewrite into a symmetric form:

$$m_t(x) := \underbrace{\frac{v_1(x) - v_t(x)}{1 - t}}_{\text{slope of } v_t} = - \underbrace{\frac{\nabla \log \rho_t(x) - \nabla \log \rho_0(x)}{t}}_{\text{slope of } \nabla \log \rho_t}.$$

All singularities are eliminated when parameterized by m_t:

$$v_t(x) = x + (t - 1) m_t(x),$$

$$\nabla \log \rho_t(x) = -t m_t(x) - x$$

$$\mathrm{KL}(\rho_1 \parallel \rho_1') = \int t(1 - t) \mathbb{E}\left[\left\| m_t - m_t' \right\|^2\right] \mathrm{d}t.$$

On the boundaries, $m_t(x)$ models the time-derivatives of v_t and $\nabla \log \rho_t$:

$$m_t(x) := \underbrace{\frac{v_1(x) - v_t(x)}{1-t}}_{\text{slope of } v_t} = -\underbrace{\frac{\nabla \log \rho_t(x) - \rho_0(x)}{t}}_{\text{slope of } \nabla \log \rho_t}.$$

Taking limit at $t = 1$ and $t = 0$:

$$m_1(x) = \partial_t v_t(x)\big|_{t=1}, \qquad m_0(x) = -\partial_t \nabla \log \rho_t(x)\big|_{t=0}.$$

- The rectified flow model yields a finite final density iff $v_t(x)$ is differentiable w.r.t. t at time $t = 1$:

$$\nabla \log \rho_1(x) \text{ exists} \quad \Longleftrightarrow \quad m_1(x) = \partial_t v_t(x)\big|_{t=1} \text{ exists.}$$

Various Model Parameterizations

$$\begin{aligned}
\textbf{Velocity field}: &\quad v_t(x) = \mathbb{E}[X_1 - X_0 \mid X_t] \\
\textbf{Expected Noise}: &\quad \mu_{0,t}(x) = \mathbb{E}[X_0 \mid X_t] \\
\textbf{Expected Data}: &\quad \mu_{1,t}(x) = \mathbb{E}[X_1 \mid X_t] \\
\textbf{Score Function}: &\quad \nabla \log \rho_t(x) = -\frac{1}{1-t}\mathbb{E}[X_0 \mid X_t]
\end{aligned}$$

Plugging $X_t = tX_1 + (1-t)X_0$, they are related by

$$\underbrace{v_t(x) = \frac{\mu_{1,t} - x}{1-t} = \frac{x - \mu_{0,t}}{t}}_{\text{holds for any coupling } (X_0, X_1)} = \underbrace{\frac{x + (1-t)\nabla \log \rho_t(x)}{t}}_{\text{only for } X_0 \perp\!\!\!\perp X_1,\ \text{Gaussian } X_0}.$$

Various Model Parameterizations

$$\begin{aligned}
\textbf{Velocity field}: &\quad v_t(x) = \mathbb{E}[X_1 - X_0 \mid X_t] \\
\textbf{Expected Noise}: &\quad \mu_{0,t}(x) = \mathbb{E}[X_0 \mid X_t] \\
\textbf{Expected Data}: &\quad \mu_{1,t}(x) = \mathbb{E}[X_1 \mid X_t] \\
\textbf{Score Function}: &\quad \nabla \log \rho_t(x) = -\frac{1}{1-t}\mathbb{E}[X_0 \mid X_t]
\end{aligned}$$

Plugging $X_t = tX_1 + (1-t)X_0$, they are related by

$$v_t(x) = \frac{\mu_{1,t} - x}{1-t} = \frac{x - \mu_{0,t}}{t} = \frac{x + (1-t)\nabla \log \rho_t(x)}{t}.$$

Different prediction targets implicitly change the loss weights.

- Predicting velocity:

$$\mathbb{E}\left[w_t \left\|\dot{X}_t - v_t(X_t)\right\|^2\right]$$

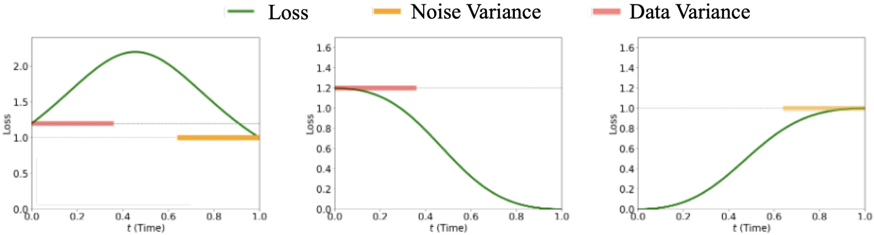
$$\Longleftrightarrow$$

- Predicting noise:

$$\mathbb{E}\left[\frac{w_t}{t^2}\left\|X_0 - \mu_{0,t}(X_t)\right\|^2\right]$$

Loss Comparisons

- Velocity prediction seems to sever a balanced baseline



Predicting Velocity | Predicting Data X_1 | Predicting Noise X_0

$\mathbb{E}\left[\|\dot{X}_t - v_t^*(X_t)\|^2\right]$ | $\mathbb{E}\left[\|X_1 - \mu_{1,t}(X_t)\|^2\right]$ | $\mathbb{E}\left[\|X_0 - \mu_{0,t}(X_t)\|^2\right]$

Figure: *Optimal losses when X_0, X_1 are independent Gaussian.*