Note on Consistency models

(1) Diffusion models

$$dX_t = b(t, X_t)\,dt + \sigma(t)\,dB_t, \quad X_0 \sim P_{data}(\cdot) \qquad (1)$$

time reversal:

$$d\bar{X}_t = \Big(-b(T-t, \bar{X}_t) + \sigma^2(T-t)\underbrace{\nabla \log p(T-t, \bar{X}_t)}_{s(T-t,\, \bar{X}_t)}\Big)\,dt + \sigma(T-t)\,d\bar{B}_t,$$

$$\bar{X}_0 \sim P_{noise}(\cdot) \quad (\text{e.g. stationary dist} \ldots) \qquad (2)$$

A more convenient ODE sampling/solving by matching the Focker-Planck equation:

$$dX'_t = \Big[b(t, X'_t) - \frac{1}{2}\sigma^2(t)\nabla \log p(t, X'_t)\Big]\,dt$$

$$X'_0 \sim P_{data}(\cdot) \qquad (3)$$

time reversal:

$$d\bar{X}'_t = \Big[-b(T-t, \bar{X}'_t) + \frac{1}{2}\sigma^2(T-t)\underbrace{\nabla \log p(T-t, \bar{X}'_t)}_{s(T-t,\, \bar{X}'_t)}\Big]\,dt$$

$$\bar{X}'_0 \sim P_{noise}(\cdot) \qquad (4)$$

Disadvantage: slow sampling/generation

often requires 10~20 steps/iterations

(iii) Consistency models : allows one-step generation

To simplify the discussion (and for practical purposes),

consider $b \equiv 0$ and $\sigma(t) = \sqrt{2t}$ (VE)

(which can be easily extended to general settings).

ODE sampling :

$$d\overline{X}'_t = -t\, s(T-t, \overline{X}'_t)\, dt, \quad \overline{X}'_0 \sim p_{noise}(\cdot) \qquad (5)$$

Fact : For each t, $\overline{X}'_t \overset{d}{=} \overline{X}_t$ (marginal dist. matches).

$\boxed{\text{Consistency / ODE flow}}$:

If $\overline{X}'_0 \sim p(t, \cdot)$ and dynamics (5)

then $\overline{X}'_{t-s} \sim p(s, \cdot)$ (and $\overline{X}'_t \sim p_{data}(\cdot)$)

Idea : Find the ODE flow $F(s, x)$ to (5)

called the consistency fct.

So $Z \sim p_{noise}(\cdot)$

one step!

Sample $F(T, Z) \approx p_{data}(\cdot)$

Function approximation $f_\theta(s, x)$.

(a). Consistency distillation (CD):

Pretrained score $s_\varphi(t, x)$:

(i) sample $X_{t_{n+1}} \sim p_{data}(\cdot) \otimes \mathcal{N}(0, t_{n+1}^2 I)$

(ii) ODE approximation

$$X_{t_n} = X_{t_{n+1}} + (t_{n+1} - t_n)\, t_{n+1}\, s_\varphi(t_{n+1}, X_{t_{n+1}})$$

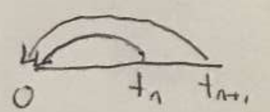
(iii) minimize e.g.

$$\boxed{\left| f_\theta(t_{n+1}, X_{t_{n+1}}) - f_\theta(t_n, X_{t_n}) \right|^2}$$

← in the paper

(Intuition: $f(t_{n+1}, X_{t_{n+1}}) = f(t_n, X_{t_n}) = f(0, X_0)$)

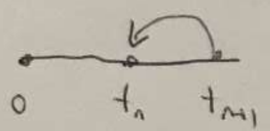
ODE flow



↖ boundary condition $f(0, x) = x$

A slight modification to minimize

$$\boxed{\left| f_\theta(t_{n+1} - t_n, X_{t_{n+1}}) - X_{t_n} \right|^2}$$



(b) Consistency training (CT)

Similar to consistency distillation except the way to "couple" $(X_{t_n}, X_{t_{n+1}})$.

(i) sample $(X_{t_n}, X_{t_{n+1}}) \sim X + (t_n, t_{n+1}) \mathcal{N}(0,1)$. (*)

(ii) minimize

$$| f_\theta(X_{t_n}, t_n) - f_\theta(X_{t_{n+1}}, t_{n+1}) |^2.$$

$\boxed{\text{Advantage}}$: no need to use "pretrained" $s_\varphi(t,x)$

$\boxed{\text{Bad thing}}$: Performance worse than CD!

Reason: the "consistency function" $f(t,x)$ is the (reversed) ODE flow

So CD couples $(X_{t_n}, X_{t_{n+1}})$ in the correct way

while CT couples $(X_{t_n}, X_{t_{n+1}})$ in the "wrong" way

↳ Q: How wrong?

↑

the marginal matches but the joint differs!

Another idea 1 : find $f(t,x)$ through PDE

We see that $f(t,x)$ is the (reversed) ODE flow and there is a connection between ODE flow and transport equation (characteristic).

$$d f(t, X_t) = 0$$

$$\frac{\partial f}{\partial t} dt + \frac{\partial f}{\partial x} \underbrace{dX_t}_{-t s_\varphi(t, X_t)\, dt} = 0$$

So f solves
$$\frac{\partial f}{\partial t} - t s_\varphi(t,x) \frac{\partial f}{\partial x} = 0 \qquad (6)$$
with boundary conditions.

$\boxed{\text{Idea}}$: Solve 6 using f_θ ?

Some one may have thought of this ?

Another idea 2 (ad-hoc):

From the SDE sampling 2, one can consider to find the "stochastic" flow $f(t, x, (B_\cdot))$

$$f_\theta = F_\theta(x,t) + G_\theta(x,t)\, N(0,1)$$

randomization