Temporal Difference Flows

Jesse Farebrother, Matteo Pirotta, Andrea Tirinzoni, Rémi Munos, Alessandro Lazaric, Ahmed Touati

Meta, Mila, McGill University

ICML 2025

Junhyeong Hong

learning agent

2025.10.30

Content

Preliminary

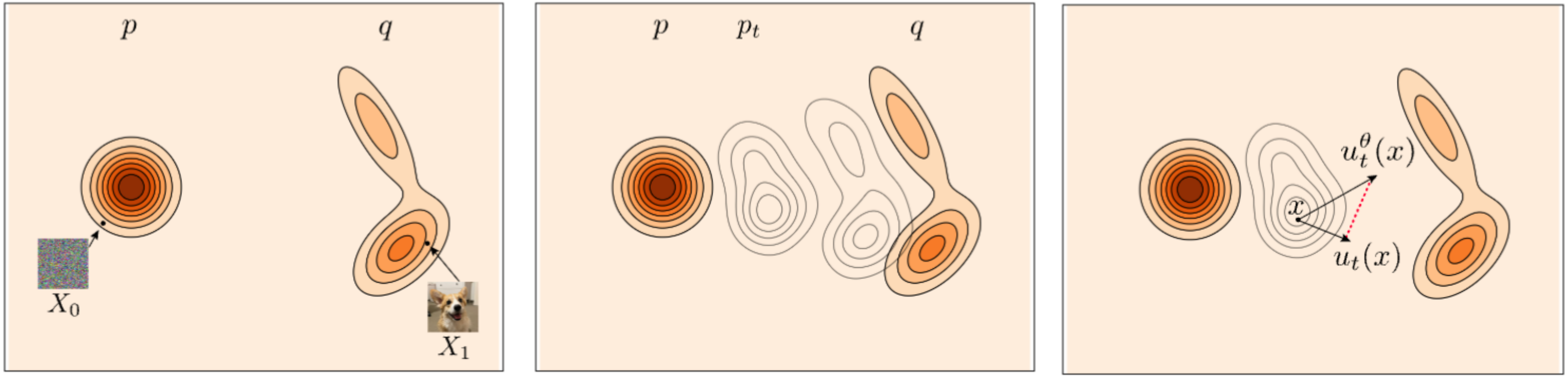
Flow matching

vector field : 다음과 같은 함수 $\mathbb{R}^n \rightarrow \mathbb{R}^n$

flow matching : source 분포를 target 분포로 서서히 변화하게 하는 vector field를 찾는게 목적

p : source distribution(정규분포), q : target distribution(data 분포), p_t : probability path(t시점의 분포)

$$L_{FM} = E_{t,p_t}[\|v_t(x;\theta) - u_t(x)\|^2]$$



Flow matching

diffeomorphism : 역함수 존재, 원함수 역함수 미분 가능

vector field u는 diffeomorphic map(미분 동형 사상)을 만든다. (diffeomporphism이면 change of variable 가능)

$$\frac{d\psi_t(x_0)}{dt} = u_t(\psi_t(x_0)), \quad x_t := \psi_t(x_0) \sim p_t \text{ for } x_0 \sim p_0$$

change of variable : mapping은 확률 총량을 변화시키지 않음. 1km와 1mile과의 관계

$|p_x(x)dV_x| = |p_y(y)dV_y|$, where $y = f(x)$. f : invertible and differentiable

then, $\frac{dV_x}{dV_y} = \det J_{f^{-1}}$

$$p_t(x) = p_0(\psi_t^{-1}(x))|\det[\frac{\partial \psi_t^{-1}}{\partial x}(x)]|$$

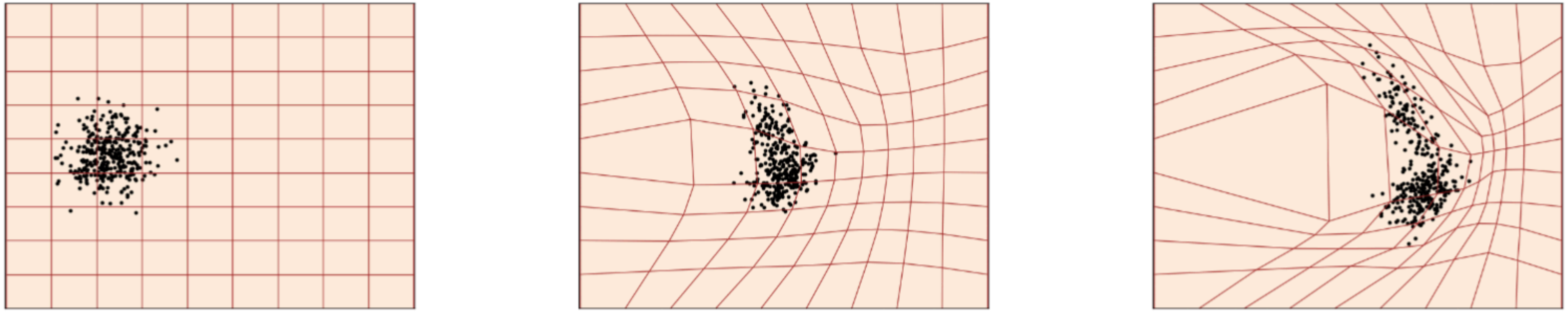
Preliminary

Flow matching

flow matching에서의 vector field는 모두 continuity equation을 만족함. diffusion에서는 fokker planck equation에 의해 기술됨

Continuity equation (PDE) : $\frac{\partial p_t}{\partial t} = -\nabla(p_t v_t)$

Continuity equation (ODE) : $\frac{dp_t}{dt} = -p_t \cdot \nabla v_t$, (change of variable하고 같은 의미)

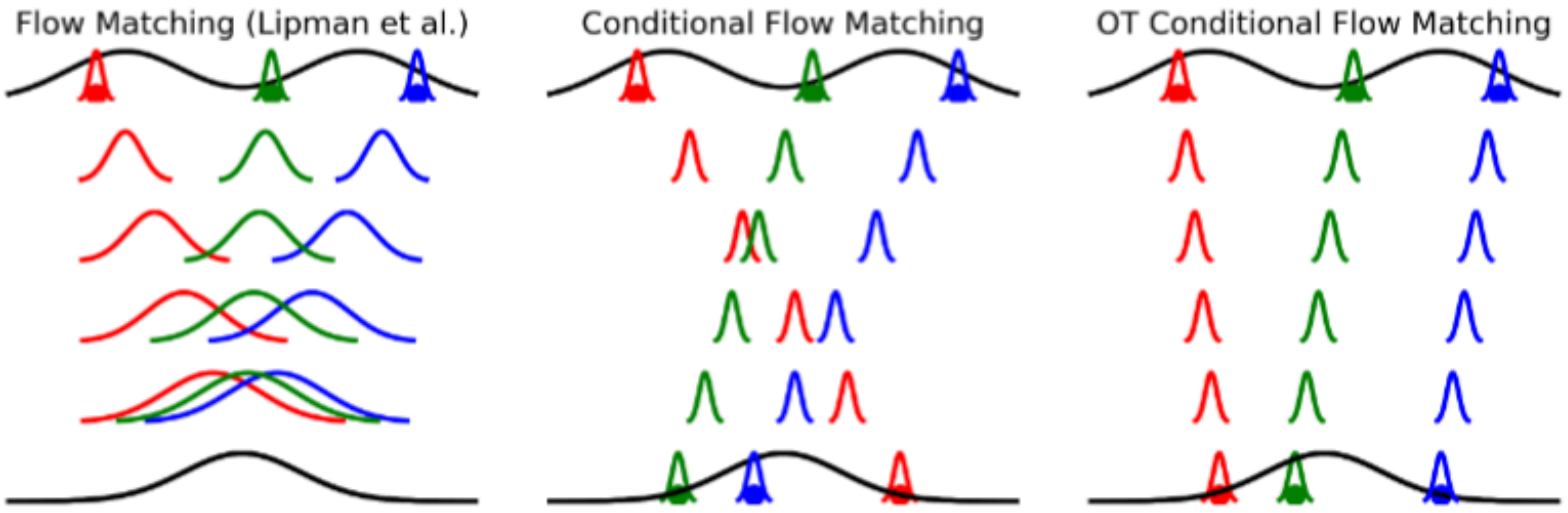


Flow matching

Condition flow matching : CFM loss를 줄이는건 FM을 최적화 하는 것과 같다. (Flow matching 3.1 과정을 거치면)

$$L_{FM} = E_{t,p_t}[\|v_t(x;\theta) - u_t(x)\|^2]$$

$$L_{CFM} = E_{t,p_t(x|x_1),q(x_1)}[\|v_t(x;\theta) - u_t(x_t|x_1)\|^2]$$



CFM I-CFM OT - CFM

Bellman expectation equation

The state-value function can again be decomposed into immediate reward plus discounted value of successor state,

$$v_\pi(s) = \mathbb{E}_\pi\left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s\right]$$

The action-value function can similarly be decomposed,

$$q_\pi(s, a) = \mathbb{E}_\pi\left[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) \mid S_t = s, A_t = a\right]$$

Bellman operator

$$T^\pi : \mathbb{R}^s \to \mathbb{R}^s \quad T^\pi V = \sum_a \pi(a\,|\,s)\left\{ r(s,a) + \gamma \sum_{s'} P(s'\,|\,s,a)V(s') \right\} = r_\pi + \gamma P_\pi V,$$

value function의 정의에 의해 $T^\pi V^\pi = V^\pi$

bellman operator는 수축 사상

$$\|T^\pi U - T^\pi V\|_p = \gamma P_\pi \|U - V\|_p \le \gamma \|U - V\|_p \quad ,p \ge 1$$

banach's fixed point theorem에 의해 $\lim_{n\to\infty} T^n_\pi V = V^\pi$

Effective horizon

강화학습의 목적식은 $G_0 = \sum_{t=0} \gamma^t r_t$ 의 기댓값을 최대로 하는것이다. 만약 $r \leq r_{max}$ 일때

infinite horizon의 경우에서 return의 최댓값은 $\max G_0 = \sum_{t=0} \gamma^t r_{max} = \frac{r_{max}}{1-\gamma}$ 이고

finite horizion에서 시간이 $\frac{1}{1-\gamma}$ 인 경우의 return의 최댓값은 $\sum_{t=1}^{\frac{1}{1-\gamma}} r_{max}$ 이므로 같은 값을 갖는다.

따라서 감쇠율이 gamma인 경우 effective horizon은 $\frac{1}{1-\gamma}$이다.

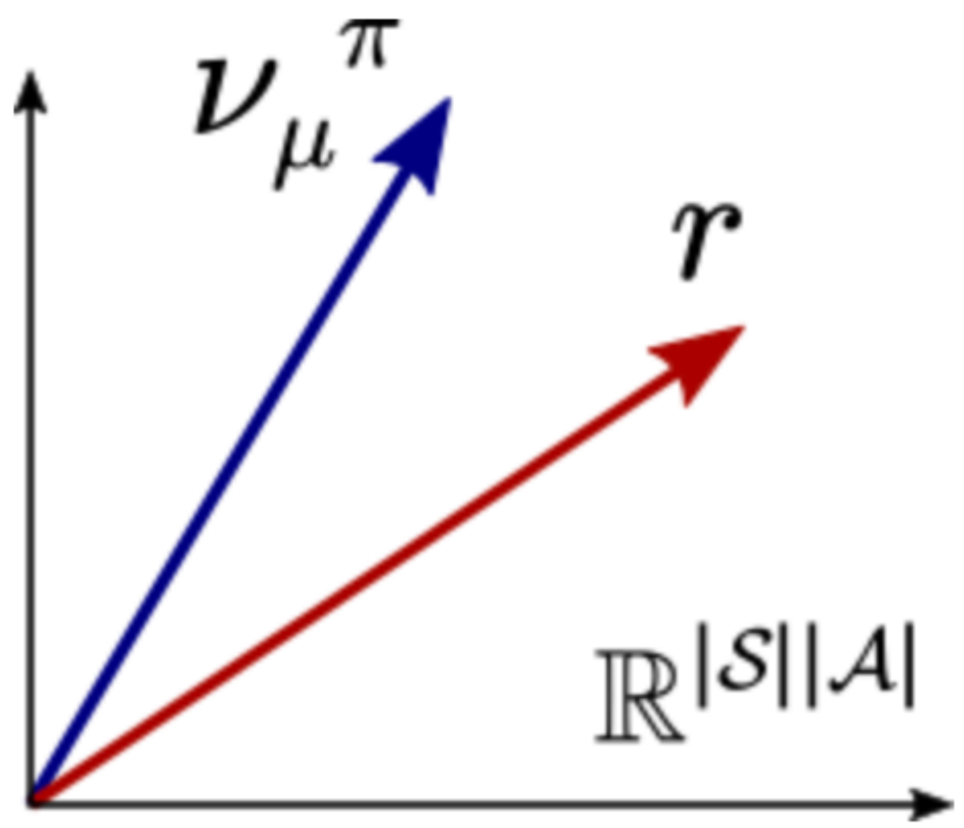
occupancy measure

occupancy measure는 policy를 따를 경우 얻어지는 s,a pair의 분포를 가중합한 분포이고 policy에 일대일 대응이다.

$$\rho_\pi(s,a) = (1-\gamma)\pi(a\,|\,s)\sum_{t=0}^{\infty}\gamma^t P(s_t = s\,|\,\pi)$$

reward가 state와 action의 함수일때 강화학습의 목적식은 occupancy measure와 reward의 내적으로 나타낼 수 있다.

$$J(\pi) = (1-\gamma)^{-1}\mathbb{E}_{s,a,\sim\rho_\pi}[r(s,a)] = (1-\gamma)^{-1}\sum_{s,a}\rho_\pi(s,a)r(s,a) = (1-\gamma)^{-1} < \rho_\pi, r >$$



successor measure

successor measure는 s,a에서 시작해 현재의 policy를 따를 경우 얻어지는 future state의 분포를 가중합한 분포이다.

$$m^\pi(s'|s,a) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t P(s_{t+1}=s'|s_0=s,a_0=a,\pi)$$

이를 통해 state action value function을 나타낼 수 있다.

$$\begin{aligned}
Q^\pi(s,a) &= \frac{1}{1-\gamma}\mathbb{E}_{s'\sim m(\cdot|s,a)}[r(s')] \\
&= \sum_{s'\in S} m(s'|s,a)r(s') \\
&= \sum_{s'\in S}\sum_{t=0}^{\infty}\gamma^t P(s_{t+1}=s'|s_0=s,a_0=a,\pi)r(s') \\
&= \sum_{t=0}^{\infty}\gamma^t\mathbb{E}[r_{t+1}|s,a] = \mathbb{E}[G_0|s,a]
\end{aligned}$$

Preliminary

bellman operator of successor measure

$$m^\pi(\cdot \mid s, a) = T^\pi m^\pi$$
$$:= (1-\gamma)P(\cdot \mid s, a) + \gamma (P^\pi m^\pi)(\cdot \mid s, a)$$

$$(P^\pi m)(dx \mid s, a) = \int_{s'} m(dx \mid s', \pi(s'))P(ds' \mid s, a) = \mathbb{E}[m(dx \mid s', \pi(s'))]$$

$$m^\pi(s' \mid s, a) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t P(s_{t+1} = s' \mid s_0 = s, a_0 = a, \pi)$$
$$= (1-\gamma)P(s_1 = s' \mid s_0 = s, a_0 = a) + (1-\gamma)\gamma\sum_{t=1}^{\infty}\gamma^{t-1}P(s_{t+1} = s' \mid s_0 = s, a_0 = a, \pi)$$
$$= (1-\gamma)P(s_1 = s' \mid s_0 = s, a_0 = a) + (1-\gamma)\gamma\sum_{x \in S}P(s_1 = x \mid s_0 = s, a_0 = a)\sum_{t=1}^{\infty}\gamma^{t-1}P(s_{t+1} = s' \mid s_0 = s, a_0 = a, s_1 = x, a_1 = \pi(x), \pi)$$
$$= (1-\gamma)P(s_1 = s' \mid s_0 = s, a_0 = a) + (1-\gamma)\gamma\sum_{x \in S}P(s_1 = x \mid s_0 = s, a_0 = a)\sum_{t=1}^{\infty}\gamma^{t-1}P(s_{t+1} = s' \mid s_1 = x, a_1 = \pi(x), \pi)$$

markov property

$$= (1-\gamma)P(s_1 = s' \mid s_0 = s, a_0 = a) + (1-\gamma)\gamma\sum_{x \in S}P(s_1 = x \mid s_0 = s, a_0 = a)\sum_{t=0}^{\infty}\gamma^{t}P(s_{t+1} = s' \mid s_0 = x, a_0 = \pi(x), \pi)$$

time homogeneity

$$= (1-\gamma)P(s_1 = s' \mid s_0 = s, a_0 = a) + \gamma\mathbb{E}_{x \sim P(\cdot \mid s, a)}\left[(1-\gamma)\sum_{t=0}^{\infty}\gamma^t P(s_{t+1} = s' \mid s_0 = x, a_0 = \pi(x), \pi)\right]$$
$$= (1-\gamma)P(s_1 = s' \mid s_0 = s, a_0 = a) + \gamma\mathbb{E}_{x \sim P(\cdot \mid s, a)}\left[m^\pi(s' \mid x, \pi(x))\right]$$

Introduction

single step model

- 환경의 dynamics를 근사하는 모델
- 오차의 누적으로 인해 long horizon prediction이 어려움

$$P_\theta(s_{t+1} \mid s_t, a_t) \approx P(s_{t+1} \mid s_t, a_t)$$

gamma model

- 시간 평균을 가한 successor의 분포를 예측하는 모델
- effective horizon에 대해 평균적인 미래를 예측 가능함

$$m^\pi(s' \mid s, a) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t P(s_{t+1} = s' \mid s_0 = s, a_0 = a, \pi)$$

gamma model을 근사하는데 flow matching을 이용해보자

Method

Notation

probability measure over a set X : $\mathscr{P}(X)$

probability path : $m_t : S \times A \rightarrow \mathscr{P}(S)$ for $t \in [0,1]$

empirical distribution : ρ

source distribution : $m_0 = p_0$

target distribution : $m_1 = m^\pi$

flow : $\psi_t : S \times S \times A \rightarrow S$, velocity field :$v_t : S \times S \times A \rightarrow S$, parameterized velocity field : $\tilde{v}_t(\cdots; \theta)$

generated RV by velocity field $X_t := \psi_t(X_0 \,|\, S, A) \sim m_t(\,\cdot\,|\, S, A),$ where $X_0 \sim m_0$

conditional probability path : $p_{t|Z} : S \times Z \rightarrow \mathscr{P}(S)$, conditional velocity field : $u_{t|Z} : S \times Z \rightarrow S$

Flow matching

marginal velocity 는 marginal path를 만들고 conditional velocity는 conditional path를 만든다.

$$p_t(x) = \int p_t(x|x_1)q(x_1)dx_1, \qquad u_t(x) = \int u_t(x|x_1)\frac{p_t(x|x_1)q(x_1)}{p_t(x)}dx_1,$$

학습 과정

1. sampling time, data and noise : $t, x_0, x_1 \sim U(0,1), p_0, q$

2. compute flow and velocity at t : $\psi_t(x_0) = tx_1 + (1-t)x_0, u_t(x\,|\,x_1) = x_1 - x_0$

3. minimize cfm loss : $\min \|v_t(\psi_t(x_0);\theta) - (x_1 - x_0))\|^2$

MC-FM/CFM

MC-FM : 가장 이상적인 방법이지만 m_t를 알 수 없어서 할 수 없음

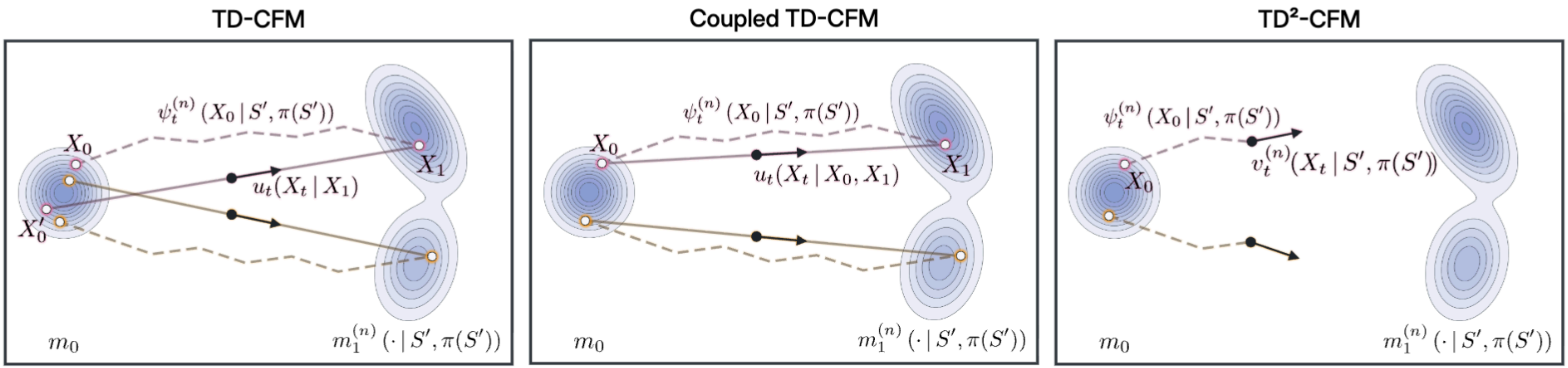
$$\frac{\mathrm{d}}{\mathrm{d}t}\psi_t(x \mid s,a) = v_t\big(\psi_t(x \mid s,a) \mid s,a\big),\ \psi_0(x \mid s,a) = x \iff \psi_t(x|s,a) = x + \int_0^t v_\tau\big(\psi_\tau(x|s,a) \mid s,a\big)\,\mathrm{d}\tau\,.$$

$$\ell_{\text{MC-FM}}(\theta) = \mathbb{E}_{\rho,t,X_t}\left[\left\|\tilde{v}_t(X_t \mid S,A;\theta) - v_t(X_t \mid S,A)\right\|^2\right],$$
$$\text{where } X_t \sim m_t(\cdot \mid S,A)\,.$$

MC-CFM : conditional flow matching 방법을 따라 유도되었지만 여전히 m^π라는 접근이 어려운 분포에 의존함

$$\ell_{\text{MC-CFM}}(\theta) = \mathbb{E}_{\rho,t,Z,X_t}\left[\left\|\tilde{v}_t(X_t \mid S,A;\theta) - u_{t|Z}(X_t \mid Z)\right\|^2\right],$$
$$\text{where } Z = X_1 \sim m^\pi(\cdot \mid S,A)\,, X_t \sim p_{t|Z}(\cdot \mid Z)\,.$$

TD-CFM



논문에서 MC-CFM을 대체 할 방법으로 총 세가지를 제시함

Method

TD-CFM

MC-CFM의 샘플링 분포 m^{π}를 아래와 같이 변경하여 transition data 만으로도 학습이 가능하다.

$$m^{\pi}(s' \mid s, a) = (1-\gamma)P(s' \mid s, a) + \gamma \mathbb{E}_{x \sim P(\cdot \mid s,a)}[m^{\pi}(s' \mid x, \pi(x))]$$

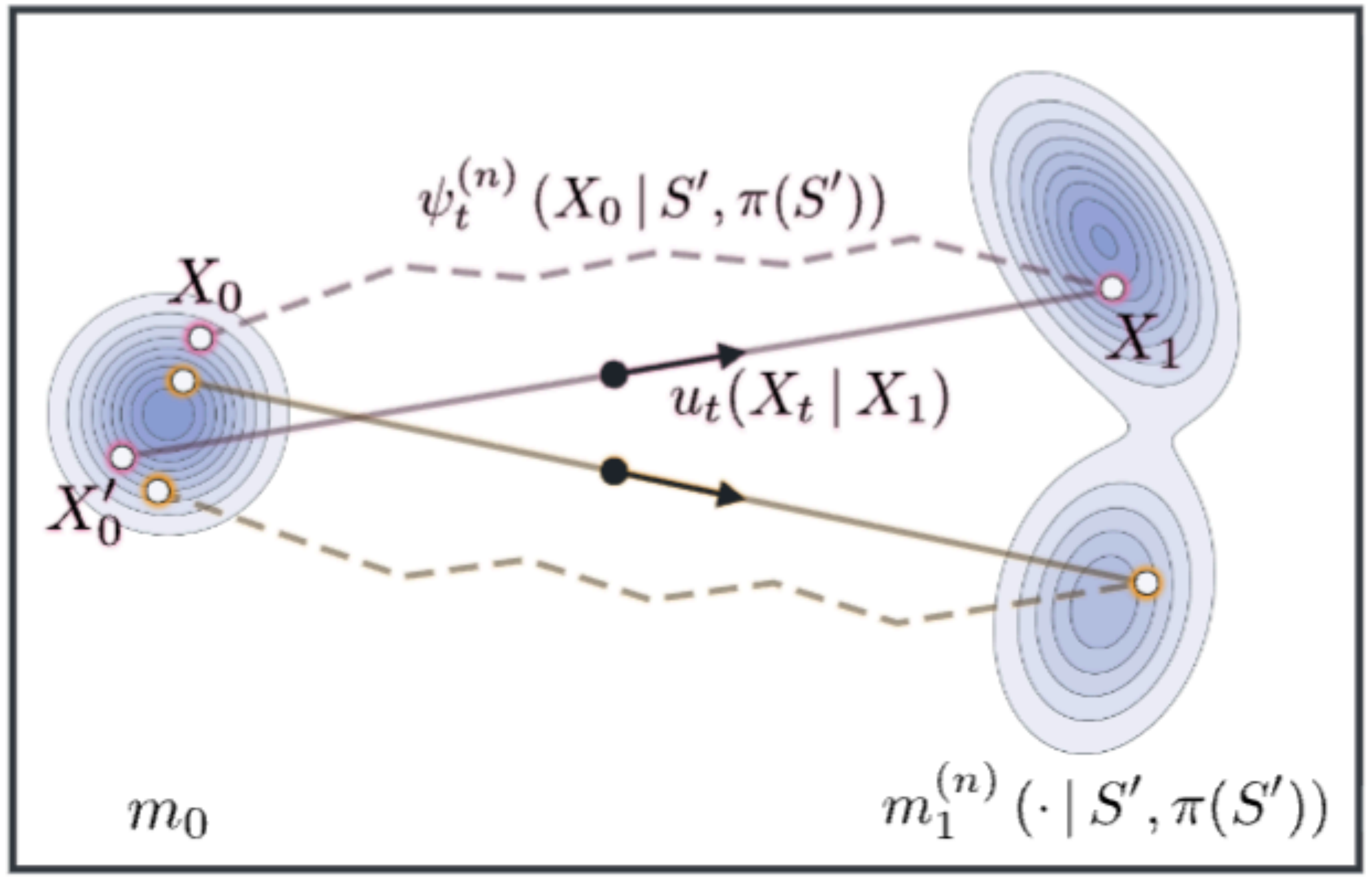
$$X_0 \sim p_0$$
$$Z = X_1 \sim (1-\gamma)\,\delta_{S'} + \gamma\,\delta_{\tilde{\psi}_1^{(n)}(X_0 \mid S', \pi(S'))}\,.$$

Coupled TD-CFM

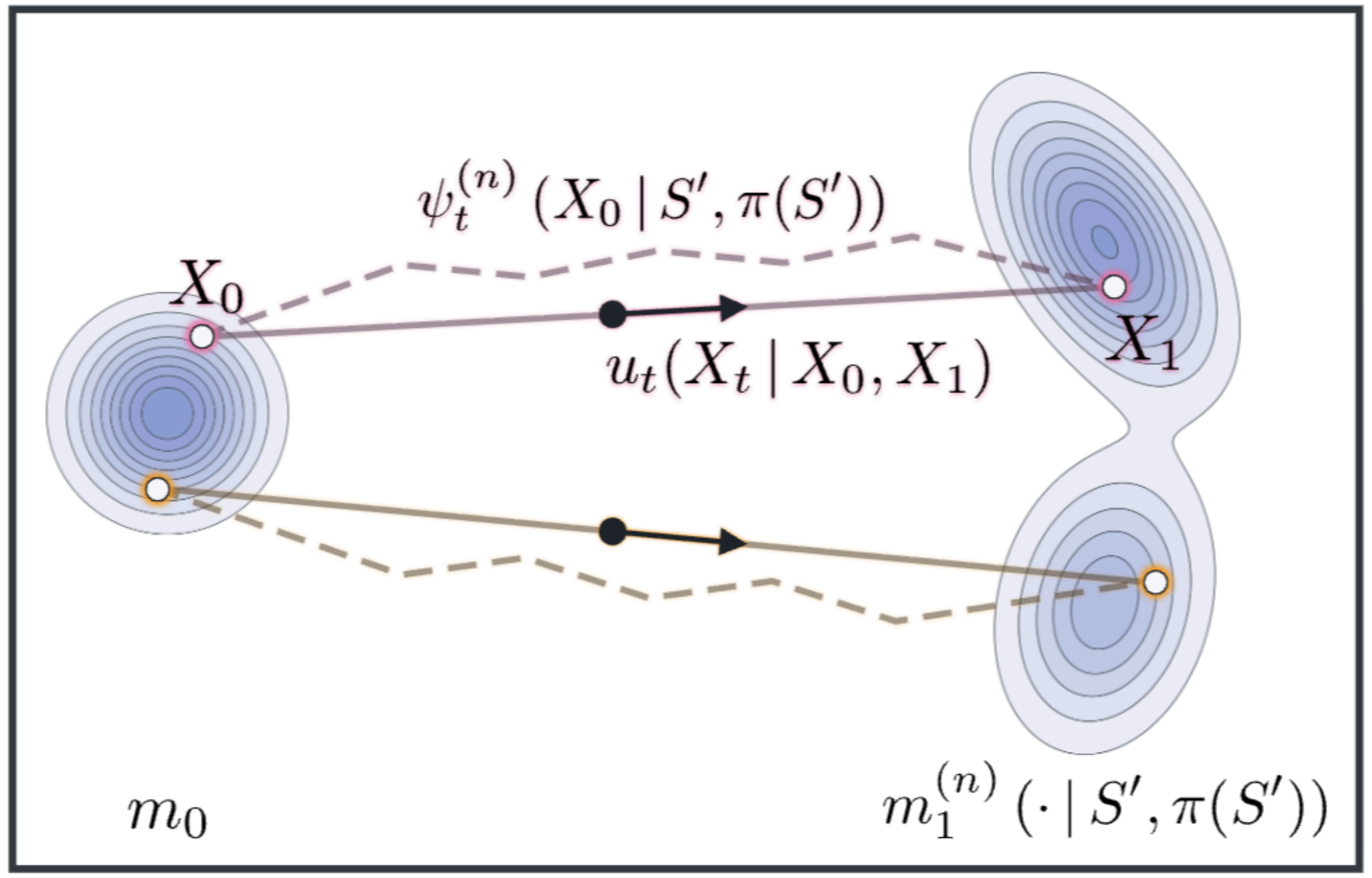
OT-CFM과 같이 coupling을 이용해 cross가 발생하는걸 억제한다.

$$X_0 \sim p_0$$
$$X_1 \sim (1-\gamma)\,\delta_{S'} + \gamma\,\delta_{\tilde{\psi}_1^{(n)}(X_0 \mid S', \pi(S'))}$$
$$Z = (X_0, X_1)\,.$$

TD-CFM



Coupled TD-CFM



TD^2-CFM

vector field를 bellman operator처럼 쪼개어 학습하자

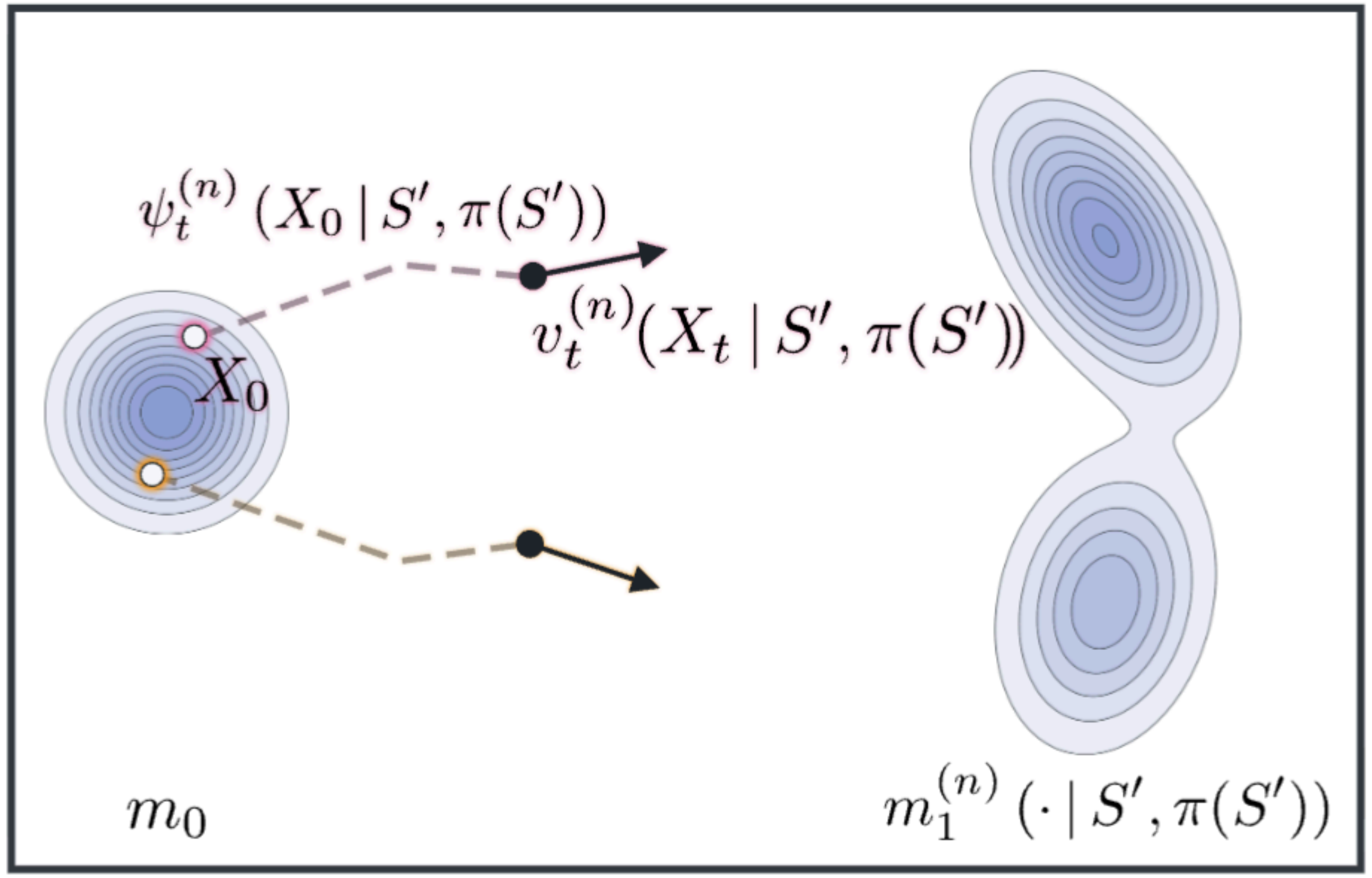
두개의 벡터장을 학습해 TD^2 CFM이다.

Lemma 1. *Let $\vec{p}_t$ be a probability path for P generated by vector field $\vec{v}_t$ and $\widehat{p}_t^{(n)}$ be a probability path for $P^\pi m_1^{(n)}$ generated by $\widehat{v}_t^{(n)}$ such that $\vec{p}_0 = \widehat{p}_0^{(n)} = m_0$. For any $t \in [0,1]$ and (s,a) let*[1]

$$v_t^{(n+1)}(\cdot \mid s,a) = \underset{v:\mathbb{R}^d \to \mathbb{R}^d}{\arg\min}\,(1-\gamma)\mathbb{E}_{\vec{X}_t \sim \vec{p}_t(\cdot\mid s,a)}\left[\left\|v(\vec{X}_t) - \vec{v}_t(\vec{X}_t \mid s,a)\right\|^2\right] + \gamma\mathbb{E}_{\widehat{X}_t \sim \widehat{p}_t^{(n)}(\cdot\mid s,a)}\left[\left\|v(\widehat{X}_t) - \widehat{v}_t^{(n)}(\widehat{X}_t \mid s,a)\right\|^2\right].$$

Then $v_t^{(n+1)}$ induces a probability path $m_t^{(n+1)}$ such that $m_0^{(n+1)} = m_0$ and $m_1^{(n+1)} = \mathcal{T}^\pi m_1^{(n)}$.

TD²-CFM



lemma 1 : bellman operator 처럼 쪼개어 학습하면 $m_1^{n+1} = T^\pi m_1^n$, 이때 $m_t^{(n+1)}(x|s,a) = (1-\gamma)\vec{p}_t(x|s,a) + \gamma\widehat{p}_t^{(n)}(x|s,a)$

$$\vec{v}_t(x \mid s,a) = \int \vec{u}_{t|1}(x \mid x_1)\frac{\vec{p}_{t|1}(x \mid x_1)P(\mathrm{d}x_1 \mid s,a)}{\vec{p}_t(x \mid s,a)},\quad \widehat{v}_t^{(n)}(x \mid s,a) = \int v_t^{(n)}(x \mid s',a')\frac{m_t^{(n)}(x \mid s',a')P(\mathrm{d}s' \mid s,a)}{\widehat{p}_t^{(n)}(x \mid s,a)},$$

$$\vec{p}_t(x|s,a) = \int \vec{p}_{t|1}(x|s')P(\mathrm{d}s'|s,a).\qquad \widehat{p}_t^{(n)}(x \mid s,a) = \int m_t^{(n)}(x \mid s',a')P(\mathrm{d}s' \mid s,a),$$

공통 : successor vector field 일반적인 CFM

TD-CFM & Coupled TD-CFM

$$x_0 \sim N(0,I), x_1 = \psi_1(x_0 \,|\, s', a', \bar{\theta})$$

$$x_t = tx_1 + (1-t)x_0$$

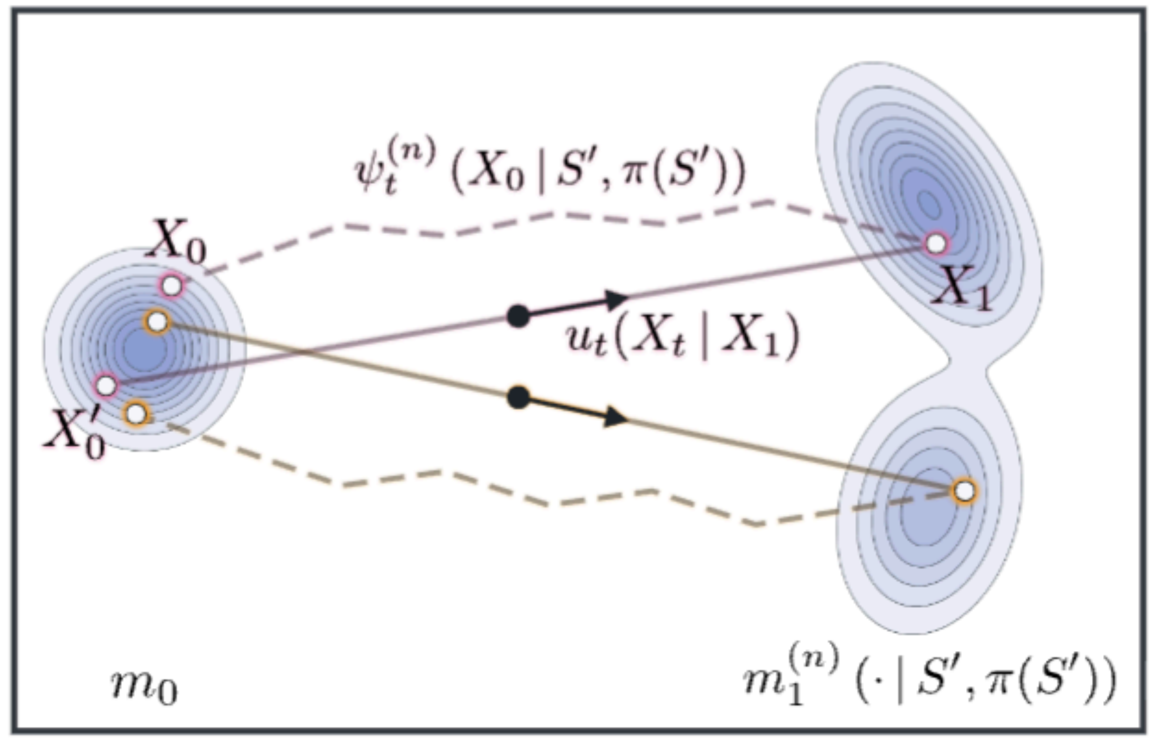
$$u_{t|1} = x_1 - x_0$$

TD^2-CFM

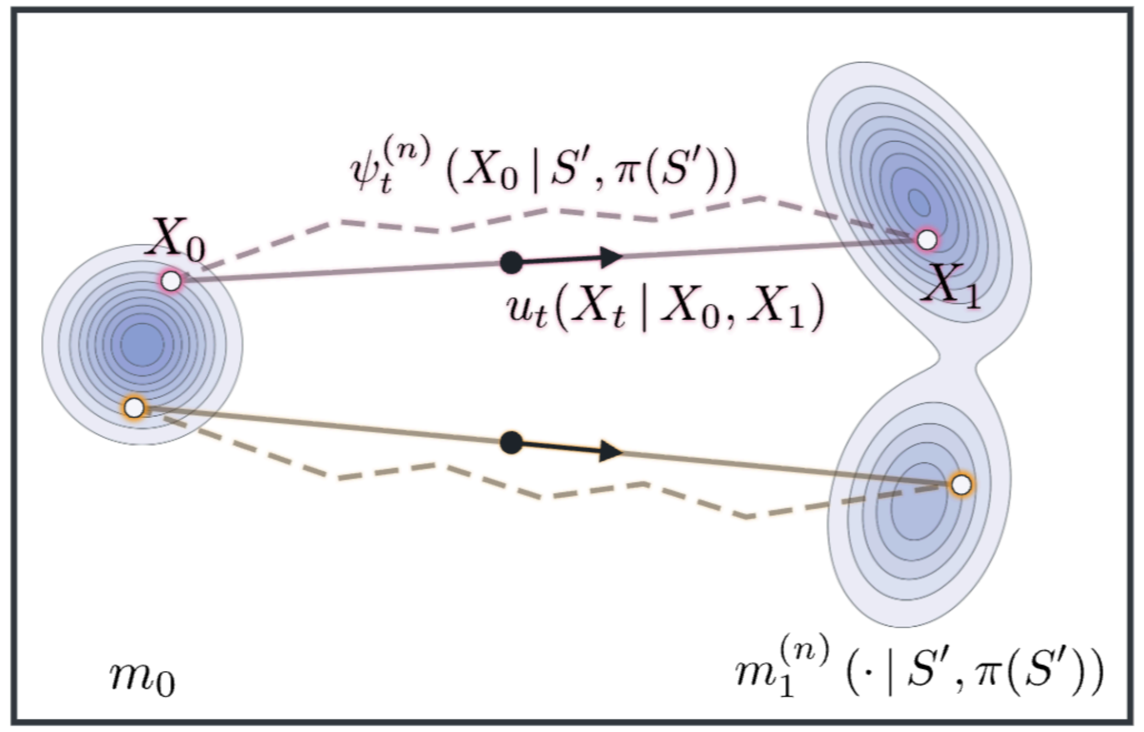
$$x_0 \sim N(0,I), x_t = \psi_t(x_0 \,|\, s', a', \bar{\theta})$$

$$u_{t|1} = v_t(x_t \,|\, s', a'; \theta)$$

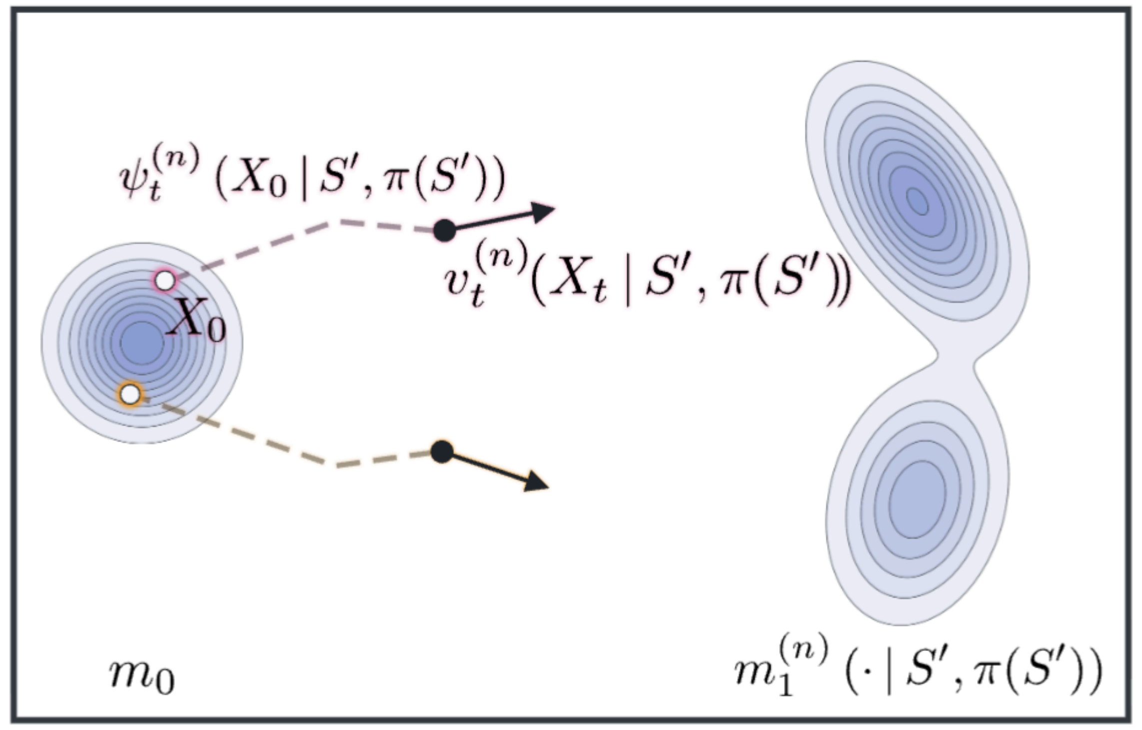
TD-CFM



Coupled TD-CFM



TD²-CFM



Method

TD^2-CFM

dynamics $P(s_{t+1} \mid s_t, a_t)$를 근사하는
velocity field를 학습

$$\vec{\ell}(\theta) = \mathbb{E}_{\rho,t,Z,\vec{X}_t}\left[\left\|\tilde{v}_t(\vec{X}_t \mid S, A; \theta) - \vec{u}_{t|Z}(\vec{X}_t \mid Z)\right\|^2\right],$$
$$\text{where } Z = X_1 \sim P(\cdot \mid S, A),\ \vec{X}_t \sim \vec{p}_{t|Z}(\cdot \mid Z),$$

현재 학습중인 NN으로 target
velocity sampling

$$\widehat{\ell}(\theta) = \mathbb{E}_{\rho,t,\widehat{X}_t}\left[\left\|\tilde{v}_t(\widehat{X}_t \mid S, A; \theta) - \tilde{v}_t^{(n)}(\widehat{X}_t \mid S', \pi(S'))\right\|^2\right],$$
$$\text{where } X_0 \sim p_0,\ S' \sim P(\cdot \mid S, A),\ \widehat{X}_t = \widetilde{\psi}_t^{(n)}(X_0 \mid S', \pi(S')),$$

$$\ell_{\text{TD}^2\text{-CFM}}(\theta) = (1-\gamma)\vec{\ell}(\theta) + \gamma\widehat{\ell}(\theta).$$

Algorithm 1 Template for TD-Flow algorithms

Inputs: offline dataset $\mathcal{D}$, policy π, batch size n, Polyak coefficient ζ, weight decay λ, randomly initialized weights θ, discount factor γ, learning rate η, one-step conditional path $\vec{p}_{t|1}$ and conditional vector-field $\vec{u}_{t|1}$, bootstrap path $\widehat{p}_t$ and vector-field $\widehat{v}_t$.
for $n = 1, \ldots$ **do**
  Sample mini-batch $\{(S_k, A_k, S'_k)\}_{k=1}^{K}$ from $\mathcal{D}$
  for $k = 1, \ldots, K$ **do**
    Sample $t_k \sim \mathcal{U}([0,1])$
    Sample $\vec{X}_k \sim \vec{p}_{t_k|1}(\cdot \mid S'_k)$
    $\vec{\ell}_k(\theta) = \left\|v_{t_k}(\vec{X}_k \mid S_k, A_k; \theta) - \vec{u}_{t_k|1}(\vec{X}_k \mid S'_k)\right\|^2$
    Sample $\widehat{X}_k \sim \widehat{p}_{t_k}(\cdot \mid S'_k, \pi(S'_k); \bar{\theta})$
    $\widehat{\ell}_k(\theta) = \left\|v_{t_k}(\widehat{X}_k \mid S_k, A_k; \theta) - \widehat{v}_{t_k}(\widehat{X}_k \mid S'_k, \pi(S'_k); \bar{\theta})\right\|^2$
  end for
  \# Compute loss
  $\ell(\theta) = \frac{1}{K}\sum_{k=1}^{K}(1-\gamma)\vec{\ell}_k(\theta) + \gamma\widehat{\ell}_k(\theta)$
  \# Perform gradient step
  $\theta \leftarrow \theta - \eta\nabla_\theta\left(\ell(\theta) + \lambda\|\theta\|^2\right)$
  \# Update parameters of target vector field
  $\bar{\theta} \leftarrow \zeta\bar{\theta} + (1-\zeta)\theta$
end for

Theoretical Analysis

Thm1.

Theorem 1. *For any $n \geq 1$, the probability paths generated by* TD-CFM, TD-CFM(C), *or* TD^2-CFM *satisfy*

$$m_t^{(n+1)}(x \mid s,a) = \left(\mathcal{B}_t^\pi m_t^{(n)}\right)(x \mid s,a), \quad \forall t \in [0,1]$$

where $\mathcal{B}_t^\pi m := (1-\gamma)P_t + \gamma P^\pi m$ *and* $P_t(x|s,a) := \int p_{t|1}(x \mid x_1)P(x_1|s,a)\mathrm{d}x_1$. *For any* $t \in [0,1]$, *the operator* $\mathcal{B}_t^\pi$ *is a* γ*-contraction in 1-Wasserstein distance, that is, for any couple of probability paths* p_t, q_t,

$$\sup_{s,a} W_1\left(\left(\mathcal{B}_t^\pi p_t\right)(\cdot \mid s,a), \left(\mathcal{B}_t^\pi q_t\right)(\cdot \mid s,a)\right) \leq \gamma \sup_{s,a} W_1\left(p_t(\cdot \mid s,a), q_t(\cdot \mid s,a)\right).$$

thm1 : 앞서 언급한 방법들은 모든 flow의 시점 t에서 probability path에 대해 gamma-contraction을 하는 학습이다.

Corollary1.

Corollary 1. *Let* $\{m_t^{(n)}\}_{n\geq 0}$ *be the sequence of probability paths produced by* TD-CFM, TD-CFM(C), *or* TD^2-CFM *starting from an arbitrary vector field* $v_t^{(0)}$. *Then,*

$$\lim_{n\to\infty} m_t^{(n)} = \overline{m}_t = \mathcal{B}_t\overline{m}_t,$$

where $\overline{m}_t$ *is the unique fixed point of* $\mathcal{B}_t$, *and* $\overline{m}_t = m_t^{MC}$, *where* $m_t^{MC}(\cdot \mid s,a) = \int p_{t|1}(\cdot \mid x_1)\, m^\pi(x_1 \mid s,a)$ *is the probability path of the Monte-Carlo approach in* (MC-CFM; 5).

Corollary : 앞서 언급한 bellman like mapping은 m_t^{MC} 를 fixed point로 갖는다. 따라서 모든 t에서 TD-CFM의 probability path는 MC-CFM의 probability path로 수렴한다.

Thm 2,3.

Theorem 2. *For any $n \geq 1$ and $t \in [0,1]$, assume that* $m_t^{(n)}(x \mid s,a) = \int p_{t|1}(x \mid x_1) m_1^{(n)}(x_1 \mid s,a) \mathrm{d}x_1$, *then*

$$\sigma^2_{\text{TD-CFM}} = \sigma^2_{\text{TD}^2\text{-CFM}} + \gamma^2 \mathbb{E}_\rho \left[\operatorname{Tr} \left(\operatorname{Cov}_{X_1|S,A,X_t} \left[\nabla_\theta v_t(X_t \mid S,A;\theta)^\top u_{t|1}(X_t \mid X_1) \right] \right) \right].$$

Theorem 3. *For any $n \geq 1$ and $t \in [0,1]$, assume that* $m_t^{(n)}(x \mid s,a) = \int p_{t|0,1}(x \mid x_0, x_1) m_{0,1}^{(n)}(x_0, x_1 \mid s,a) \mathrm{d}x_0 \mathrm{d}x_1$[3], *then we obtain*

$$\sigma^2_{\text{TD-CFM}(C)} = \sigma^2_{\text{TD}^2\text{-CFM}} + \gamma^2 \mathbb{E}_\rho \left[\operatorname{Tr} \left(\operatorname{Cov}_{Z|S,A,X_t} \left[\nabla_\theta v_t(X_t \mid S,A;\theta)^\top u_{t|Z}(X_t \mid Z) \right] \right) \right],$$

where $Z = (X_0, X_1)$. Furthermore, if we use straight conditional paths, i.e., $X_t = tX_1 + (1-t)X_0$, and the linear interpolant X_t does not intersect for any s, a, s', then $\sigma^2_{\text{TD-CFM}(C)} = \sigma^2_{\text{TD}^2\text{-CFM}}$.

Thm 2,3 : TD^2 CFM의 기울기는 다른 방법들 보다 더 작은 분산을 가져 더 안정적인 수렴을 한다.

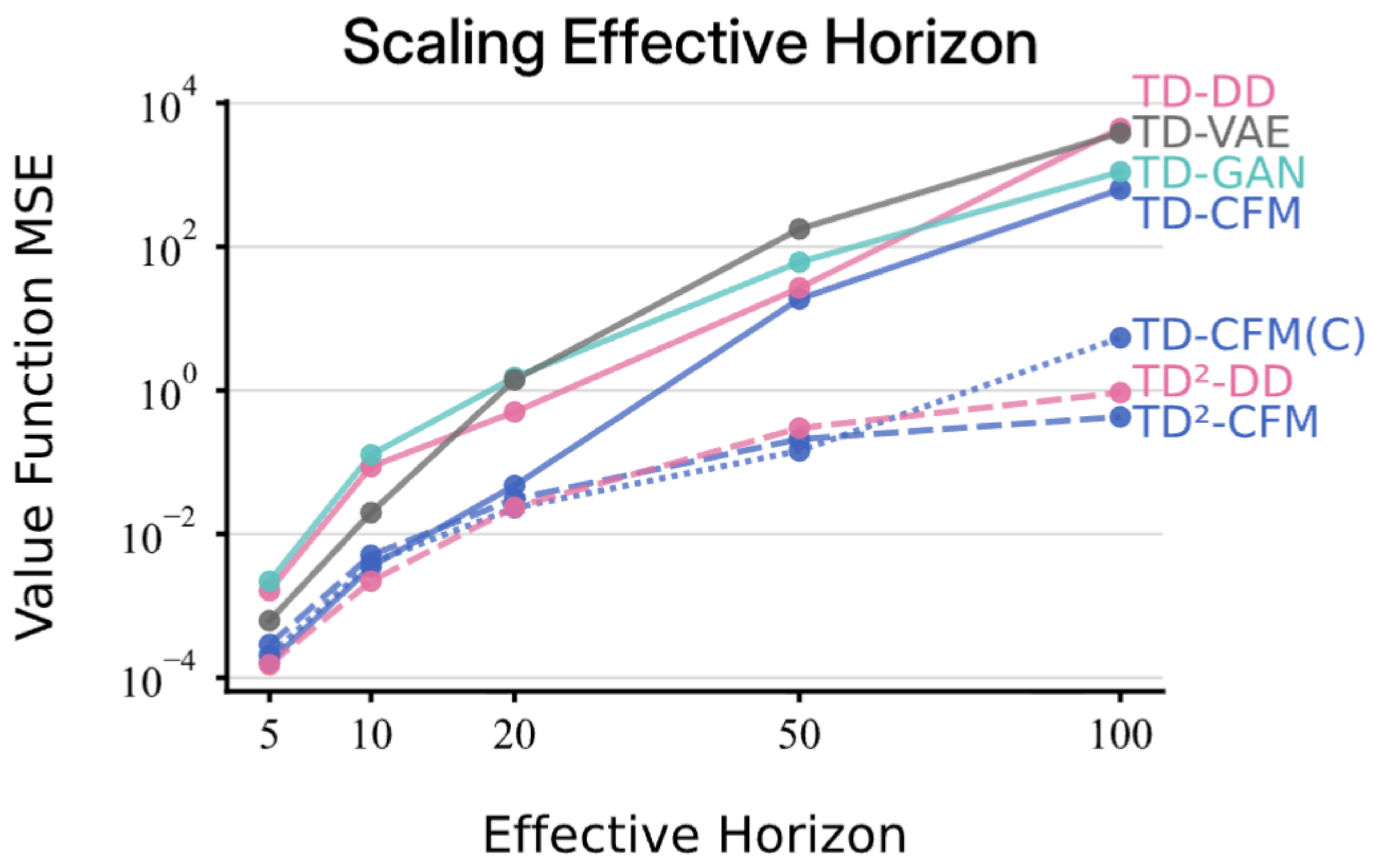
Experiments

ExoRL 데이터를 이용해 TD3로 학습

64개의 initial state로 부터 1000step rollout

각 trajectory 별로 2048개의 state를 복원 추출

이때 $t \sim$ Geometric$(1-\gamma)$ 시점의 state를 샘플링



	Method	EMD ↓	Norm NLL ↓	MSE(V) ↓
Cheetah	TD-DD	20.22 (0.26)	2.824 (0.195)	454.49 (131.97)
	TD^2-DD	14.14 (1.08)	0.806 (0.016)	189.15 (23.63)
	TD-CFM	12.26 (0.02)	0.886 (0.024)	228.77 (2.20)
	TD-CFM(c)	10.51 (0.06)	0.447 (0.020)	140.78 (18.72)
	TD^2-CFM	10.57 (0.07)	0.422 (0.014)	135.22 (19.79)
	GAN	23.97 (0.46)	—	2463.22 (628.05)
	VAE	83.77 (0.41)	—	1284.27 (37.62)
Pointmass	TD-DD	0.149 (0.001)	2.974 (0.100)	1245.20 (29.27)
	TD^2-DD	0.027 (0.001)	0.761 (0.082)	11.13 (3.09)
	TD-CFM	0.062 (0.003)	0.554 (0.033)	355.56 (82.83)
	TD-CFM(c)	0.022 (0.002)	−0.696 (0.094)	11.89 (3.16)
	TD^2-CFM	0.021 (0.000)	−0.843 (0.027)	8.74 (2.09)
	GAN	0.203 (0.037)	—	1257.26 (112.86)
	VAE	0.410 (0.036)	—	1821.89 (69.78)
Quadruped	TD-DD	28.33 (0.33)	1.908 (0.041)	1490.75 (444.49)
	TD^2-DD	22.64 (2.47)	0.861 (0.028)	159.03 (14.64)
	TD-CFM	15.73 (0.06)	1.056 (0.002)	525.06 (28.90)
	TD-CFM(c)	14.38 (0.03)	0.488 (0.003)	155.25 (5.58)
	TD^2-CFM	14.51 (0.05)	0.379 (0.011)	141.77 (3.10)
	GAN	36772.12 (13898.25)	—	2634.69 (798.38)
	VAE	60.27 (0.28)	—	1156.33 (36.52)
Walker	TD-DD	20.58 (0.24)	2.649 (0.137)	382.40 (458.63)
	TD^2-DD	12.09 (0.12)	0.537 (0.060)	39.04 (6.08)
	TD-CFM	13.53 (0.11)	0.713 (0.028)	225.27 (42.43)
	TD-CFM(c)	11.91 (0.02)	0.219 (0.016)	30.71 (3.44)
	TD^2-CFM	11.92 (0.10)	0.104 (0.001)	28.35 (6.10)
	GAN	24.51 (0.89)	—	3690.65 (1117.94)
	VAE	111.73 (2.53)	—	2457.61 (16.25)