

Transporte óptimo y puentes de Schrödinger como generalización de los modelos de difusión

Fernando Fêtis Riquelme

Primavera, 2024

fcm - Universidad de Chile

Tabla de contenidos

1. Modelos de difusión
2. Problema de Schrödinger
3. Transporte óptimo

Modelos de difusión

Modelos de difusión — Proceso forward

El proceso de inyección de ruido es una cadena de Markov en \mathbb{R}^d factorizada de forma causal, cuyos hiperparámetros corresponden a una secuencia finita y decreciente $(\alpha_t)_{t=1}^T \subset (0, 1)$:

Proceso forward

$$q(x_{0:T}) = q(x_0) \prod_{t=1}^T q(x_t | x_{t-1}),$$

con $q(x_0) = p_{\text{true}}(x_0)$ y transiciones gaussianas isotrópicas:

$$q(x_t | x_{t-1}) \sim \mathcal{N}(\sqrt{\alpha_t} x_{t-1}, (1 - \alpha_t) I_d).$$

La secuencia $(\alpha_t)_{t=1}^T$ debe ser tal que

$$q(x_T) = \int_{(\mathbb{R}^d)^T} q(x_{0:T}) dx_{0:(T-1)} \approx p_{\text{prior}}(x_T) \sim \mathcal{N}(0, I_d).$$

Modelos de difusión — Proceso forward

El proceso de inyección de ruido es una cadena de Markov en \mathbb{R}^d factorizada de forma causal, cuyos hiperparámetros corresponden a una secuencia finita y decreciente $(\alpha_t)_{t=1}^T \subset (0, 1)$:

Proceso forward

$$q(x_{0:T}) = q(x_0) \prod_{t=1}^T q(x_t | x_{t-1}),$$

con $q(x_0) = p_{\text{true}}(x_0)$ y transiciones gaussianas isotrópicas:

$$q(x_t | x_{t-1}) \sim \mathcal{N}(\sqrt{\alpha_t} x_{t-1}, (1 - \alpha_t) I_d).$$

La secuencia $(\alpha_t)_{t=1}^T$ debe ser tal que

$$q(x_T) = \int_{(\mathbb{R}^d)^T} q(x_{0:T}) dx_{0:(T-1)} \approx p_{\text{prior}}(x_T) \sim \mathcal{N}(0, I_d).$$

Modelos de difusión — Proceso backward

El proceso de reconstrucción es otra cadena de Markov factorizada de forma anticausal. Se demuestra que $q(x_{t-1}|x_t, x_0) \sim \mathcal{N}(\mu_q(x_0, x_t, t), \sigma_q^2(t) I_d)$, por lo que se propone aprender transiciones gaussianas:

Proceso backward

$$p_\theta(x_{0:T}) = p_\theta(x_T) \prod_{t=1}^T p_\theta(x_{t-1}|x_t),$$

con $p_\theta(x_T) = p_{\text{prior}}(x_T)$ y transiciones gaussianas:

$$p_\theta(x_{t-1}|x_t) \sim \mathcal{N}(\mu_\theta(x_t, t), \Sigma_\theta(x_t, t)).$$

- La función objetivo buscará que $p_\theta(x_{t-1}|x_t) \approx q(x_{t-1}|x_t, x_0)$. Fijando $\Sigma_\theta(x_t, t) = \sigma_q^2(t) I_d$, la función objetivo se reduce a una diferencia de cuadrados.
- Con esta elección, solo se necesita aprender el vector de medias de $p_\theta(x_{t-1}|x_t)$ mediante una red neuronal $\mu_\theta : \mathbb{R}^d \times \{0, \dots, T\} \rightarrow \mathbb{R}^d$.

Modelos de difusión — Proceso backward

El proceso de reconstrucción es otra cadena de Markov factorizada de forma anticausal. Se demuestra que $q(x_{t-1}|x_t, x_0) \sim \mathcal{N}(\mu_q(x_0, x_t, t), \sigma_q^2(t) I_d)$, por lo que se propone aprender transiciones gaussianas:

Proceso backward

$$p_\theta(x_{0:T}) = p_\theta(x_T) \prod_{t=1}^T p_\theta(x_{t-1}|x_t),$$

con $p_\theta(x_T) = p_{\text{prior}}(x_T)$ y transiciones gaussianas:

$$p_\theta(x_{t-1}|x_t) \sim \mathcal{N}(\mu_\theta(x_t, t), \Sigma_\theta(x_t, t)).$$

- La función objetivo buscará que $p_\theta(x_{t-1}|x_t) \approx q(x_{t-1}|x_t, x_0)$. Fijando $\Sigma_\theta(x_t, t) = \sigma_q^2(t) I_d$, la función objetivo se reduce a una diferencia de cuadrados.
- Con esta elección, solo se necesita aprender el vector de medias de $p_\theta(x_{t-1}|x_t)$ mediante una red neuronal $\mu_\theta : \mathbb{R}^d \times \{0, \dots, T\} \rightarrow \mathbb{R}^d$.

Modelos de difusión — Proceso backward

El proceso de reconstrucción es otra cadena de Markov factorizada de forma anticausal. Se demuestra que $q(x_{t-1}|x_t, x_0) \sim \mathcal{N}(\mu_q(x_0, x_t, t), \sigma_q^2(t) I_d)$, por lo que se propone aprender transiciones gaussianas:

Proceso backward

$$p_\theta(x_{0:T}) = p_\theta(x_T) \prod_{t=1}^T p_\theta(x_{t-1}|x_t),$$

con $p_\theta(x_T) = p_{\text{prior}}(x_T)$ y transiciones gaussianas:

$$p_\theta(x_{t-1}|x_t) \sim \mathcal{N}(\mu_\theta(x_t, t), \Sigma_\theta(x_t, t)).$$

- La función objetivo buscará que $p_\theta(x_{t-1}|x_t) \approx q(x_{t-1}|x_t, x_0)$. Fijando $\Sigma_\theta(x_t, t) = \sigma_q^2(t) I_d$, la función objetivo se reduce a una diferencia de cuadrados.
- Con esta elección, solo se necesita aprender el vector de medias de $p_\theta(x_{t-1}|x_t)$ mediante una red neuronal $\mu_\theta : \mathbb{R}^d \times \{0, \dots, T\} \rightarrow \mathbb{R}^d$.

Modelos de difusión — Entrenamiento e inferencia

La verosimilitud $p_\theta(x_0) = \int_{(\mathbb{R}^D)^T} p_\theta(x_{0:T}) dx_{1:T}$ no es computable de forma eficiente. Para entrenar μ_θ se maximiza $\mathbb{E}_{x_0 \sim p_{\text{true}}(x_0)} [\text{ELBO}(x_0)]$, donde

$$\text{ELBO}(x_0) := \log p_\theta(x_0) - D_{\text{KL}}(q(x_{1:T}|x_0) \| p_\theta(x_{1:T}|x_0)).$$

La ELBO se puede evaluar eficientemente:

ELBO para DDPM

Dada una muestra $x_0 \sim p_{\text{true}}(x_0)$, entonces:

$$\text{ELBO}(x_0) = - \sum_{t=1}^T \frac{1}{2\sigma_q^2(t)} \|\mu_q(x_0, x_t, t) - \mu_\theta(x_t, t)\|^2 + \text{constante}.$$

Para la generación de nuevas muestras desde $p_\theta(x_0)$, se simula el proceso backward comenzando con una muestra $x_T \sim p_{\text{prior}}(x_T)$.

Modelos de difusión — Entrenamiento e inferencia

La verosimilitud $p_\theta(x_0) = \int_{(\mathbb{R}^D)^T} p_\theta(x_{0:T}) dx_{1:T}$ no es computable de forma eficiente. Para entrenar μ_θ se maximiza $\mathbb{E}_{x_0 \sim p_{\text{true}}(x_0)} [\text{ELBO}(x_0)]$, donde

$$\text{ELBO}(x_0) := \log p_\theta(x_0) - D_{\text{KL}}(q(x_{1:T}|x_0) \| p_\theta(x_{1:T}|x_0)).$$

La ELBO se puede evaluar eficientemente:

ELBO para DDPM

Dada una muestra $x_0 \sim p_{\text{true}}(x_0)$, entonces:

$$\text{ELBO}(x_0) = - \sum_{t=1}^T \frac{1}{2\sigma_q^2(t)} \|\mu_q(x_0, x_t, t) - \mu_\theta(x_t, t)\|^2 + \text{constante}.$$

Para la generación de nuevas muestras desde $p_\theta(x_0)$, se simula el proceso backward comenzando con una muestra $x_T \sim p_{\text{prior}}(x_T)$.

Modelos de difusión — Entrenamiento e inferencia

La verosimilitud $p_\theta(x_0) = \int_{(\mathbb{R}^D)^T} p_\theta(x_{0:T}) dx_{1:T}$ no es computable de forma eficiente. Para entrenar μ_θ se maximiza $\mathbb{E}_{x_0 \sim p_{\text{true}}(x_0)} [\text{ELBO}(x_0)]$, donde

$$\text{ELBO}(x_0) := \log p_\theta(x_0) - D_{\text{KL}}(q(x_{1:T}|x_0) \| p_\theta(x_{1:T}|x_0)).$$

La ELBO se puede evaluar eficientemente:

ELBO para DDPM

Dada una muestra $x_0 \sim p_{\text{true}}(x_0)$, entonces:

$$\text{ELBO}(x_0) = - \sum_{t=1}^T \frac{1}{2\sigma_q^2(t)} \|\mu_q(x_0, x_t, t) - \mu_\theta(x_t, t)\|^2 + \text{constante}.$$

Para la generación de nuevas muestras desde $p_\theta(x_0)$, se simula el proceso backward comenzando con una muestra $x_T \sim p_{\text{prior}}(x_T)$.

Modelos de difusión — Formulación basada en score

La red neuronal $\mu_\theta(x_t, t)$ busca aprender $\mu_q(x_0, x_t, t)$.

Se puede demostrar que

$$\mu_q(x_0, x_t, t) = \frac{1}{\sqrt{\alpha_t}}x_t - \frac{1 - \alpha_t}{\sqrt{\alpha_t}}\nabla_{x_t} \log q(x_t),$$

por lo que $\mu_\theta(x_t, t)$ puede ser reparametrizada por una red neuronal $s_\theta(x_t, t)$ que aprenda directamente la función de score $\nabla_{x_t} \log q(x_t)$.

- Esto conecta los modelos de difusión con SM y con EBM.
- Entrega un método de generación condicional (guidance):

$$\underbrace{\nabla_{x_t} \log p_\theta(x_t|y)}_{\text{score condicional}} = \underbrace{\nabla_{x_t} \log p_\theta(y|x_t)}_{\text{modelo discriminativo}} - \underbrace{\nabla_{x_t} \log p_\theta(x_t)}_{\text{score incondicional}},$$

con $p_\theta(y|x_t)$ un clasificador o un modelo tipo CLIP.

Modelos de difusión — Formulación basada en score

La red neuronal $\mu_\theta(x_t, t)$ busca aprender $\mu_q(x_0, x_t, t)$.

Se puede demostrar que

$$\mu_q(x_0, x_t, t) = \frac{1}{\sqrt{\alpha_t}}x_t - \frac{1 - \alpha_t}{\sqrt{\alpha_t}}\nabla_{x_t} \log q(x_t),$$

por lo que $\mu_\theta(x_t, t)$ puede ser reparametrizada por una red neuronal $s_\theta(x_t, t)$ que aprenda directamente la función de score $\nabla_{x_t} \log q(x_t)$.

- Esto conecta los modelos de difusión con SM y con EBM.
- Entrega un método de generación condicional (guidance):

$$\underbrace{\nabla_{x_t} \log p_\theta(x_t|y)}_{\text{score condicional}} = \underbrace{\nabla_{x_t} \log p_\theta(y|x_t)}_{\text{modelo discriminativo}} - \underbrace{\nabla_{x_t} \log p_\theta(x_t)}_{\text{score incondicional}},$$

con $p_\theta(y|x_t)$ un clasificador o un modelo tipo CLIP.

Modelos de difusión — Formulación basada en score

La red neuronal $\mu_\theta(x_t, t)$ busca aprender $\mu_q(x_0, x_t, t)$.

Se puede demostrar que

$$\mu_q(x_0, x_t, t) = \frac{1}{\sqrt{\alpha_t}}x_t - \frac{1 - \alpha_t}{\sqrt{\alpha_t}}\nabla_{x_t} \log q(x_t),$$

por lo que $\mu_\theta(x_t, t)$ puede ser reparametrizada por una red neuronal $s_\theta(x_t, t)$ que aprenda directamente la función de score $\nabla_{x_t} \log q(x_t)$.

- Esto conecta los modelos de difusión con SM y con EBM.
- Entrega un método de generación condicional (guidance):

$$\underbrace{\nabla_{x_t} \log p_\theta(x_t|y)}_{\text{score condicional}} = \underbrace{\nabla_{x_t} \log p_\theta(y|x_t)}_{\text{modelo discriminativo}} - \underbrace{\nabla_{x_t} \log p_\theta(x_t)}_{\text{score incondicional}},$$

con $p_\theta(y|x_t)$ un clasificador o un modelo tipo CLIP.

Modelos de difusión — Formulación basada en score

La red neuronal $\mu_\theta(x_t, t)$ busca aprender $\mu_q(x_0, x_t, t)$.

Se puede demostrar que

$$\mu_q(x_0, x_t, t) = \frac{1}{\sqrt{\alpha_t}}x_t - \frac{1 - \alpha_t}{\sqrt{\alpha_t}}\nabla_{x_t} \log q(x_t),$$

por lo que $\mu_\theta(x_t, t)$ puede ser reparametrizada por una red neuronal $s_\theta(x_t, t)$ que aprenda directamente la función de score $\nabla_{x_t} \log q(x_t)$.

- Esto conecta los modelos de difusión con SM y con EBM.
- Entrega un método de generación condicional (guidance):

$$\underbrace{\nabla_{x_t} \log p_\theta(x_t|y)}_{\text{score condicional}} = \underbrace{\nabla_{x_t} \log p_\theta(y|x_t)}_{\text{modelo discriminativo}} - \underbrace{\nabla_{x_t} \log p_\theta(x_t)}_{\text{score incondicional}},$$

con $p_\theta(y|x_t)$ un clasificador o un modelo tipo CLIP.

Modelos de difusión — Formulación continua

La formulación basada en score permite extender los modelos de difusión a tiempo continuo usando SDEs:

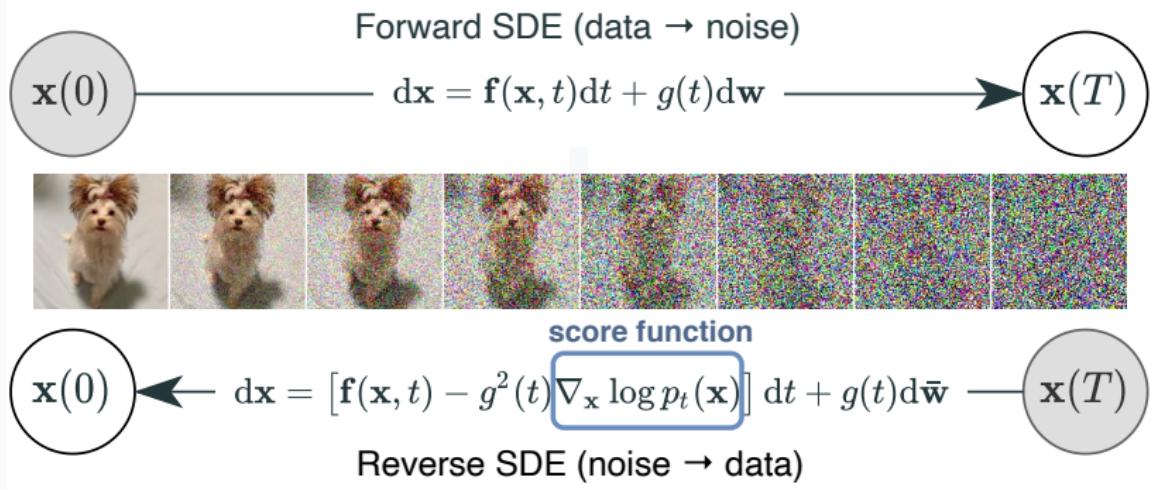


Figura 1: Imagen obtenida desde Song et al., 2021.

La función de costo se puede extender de forma *natural*. También es posible encontrar una expresión análoga a la ELBO en tiempo discreto.

Modelos de difusión — Formulación continua

- DDPM y DSM son discretizaciones de SDEs específicas.
- Se pueden usar diferentes solvers para el proceso backward durante la generación.
- Conexión con CNF: existe un proceso determinista con las mismas distribuciones marginales que los procesos de difusión y denoising.
- En particular, se puede calcular la log-verosimilitud de forma exacta.

Modelos de difusión — Formulación continua

- DDPM y DSM son discretizaciones de SDEs específicas.
- Se pueden usar diferentes solvers para el proceso backward durante la generación.
- Conexión con CNF: existe un proceso determinista con las mismas distribuciones marginales que los procesos de difusión y denoising.
- En particular, se puede calcular la log-verosimilitud de forma exacta.

Modelos de difusión — Formulación continua

- DDPM y DSM son discretizaciones de SDEs específicas.
- Se pueden usar diferentes solvers para el proceso backward durante la generación.
- Conexión con CNF: existe un proceso determinista con las mismas distribuciones marginales que los procesos de difusión y denoising.

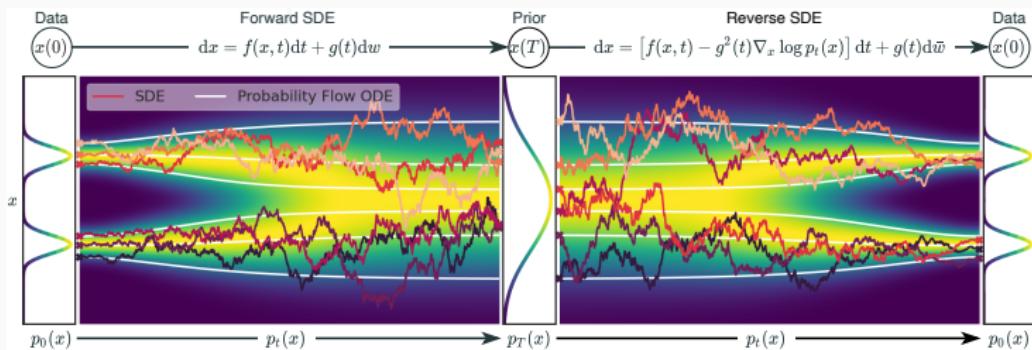


Figura 2: Imagen obtenida desde Song et al., 2021.

- En particular, se puede calcular la log-verosimilitud de forma exacta.

Modelos de difusión — Formulación continua

- DDPM y DSM son discretizaciones de SDEs específicas.
- Se pueden usar diferentes solvers para el proceso backward durante la generación.
- Conexión con CNF: existe un proceso determinista con las mismas distribuciones marginales que los procesos de difusión y denoising.
- En particular, se puede calcular la log-verosimilitud de forma exacta.

Modelos de difusión — Limitaciones

- Sensibilidad a la elección del proceso de difusión.
- No permite transformación entre distribuciones (p_{prior} fija).
- Convergencia asintótica a p_{prior} y generación lenta.



(a) 64×64



(b) 128×128



(c) 256×256



(d) 512×512



(e) 1024×1024

Figura 2: Imágenes obtenidas desde Nichol y Dhariwal, 2021 y Chen, 2023.

Modelos de difusión — Limitaciones

- Sensibilidad a la elección del proceso de difusión.
- No permite transformación entre distribuciones (p_{prior} fija).
- Convergencia asintótica a p_{prior} y generación lenta.



Figura 2: Imagen obtenida desde Zhu et al., 2020.

Modelos de difusión — Limitaciones

- Sensibilidad a la elección del proceso de difusión.
- No permite transformación entre distribuciones (p_{prior} fija).
- Convergencia asintótica a p_{prior} y generación lenta.

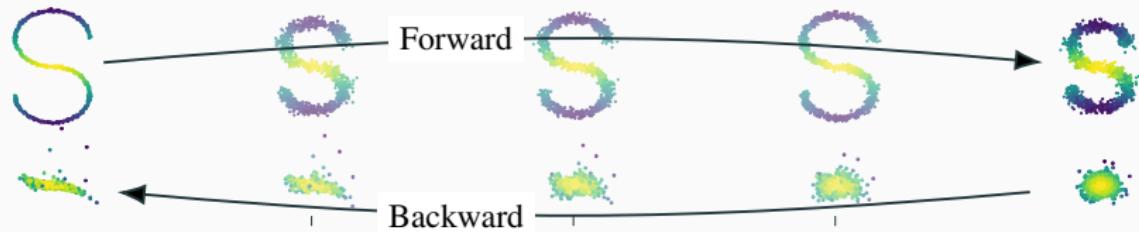


Figura 2: Imagen obtenida desde Bortoli et al., 2023.

Problema de Schrödinger

Problema de Schrödinger — Notación

- Por simplicidad, se asumirá que $\mathcal{X} = \mathcal{Y} \subset \mathbb{R}^d$ es compacto.
- Si una medida de probabilidad $\mu \in \mathcal{M}_+^1(\mathcal{X})$ tiene función de densidad $p : \mathcal{X} \rightarrow \mathbb{R}_+$, entonces

$$\int_{\mathcal{X}} f(x) \, d\mu(x) = \int_{\mathcal{X}} f(x) p(x) \, dx.$$

Problema de Schrödinger — Notación

- Por simplicidad, se asumirá que $\mathcal{X} = \mathcal{Y} \subset \mathbb{R}^d$ es compacto.
- Si una medida de probabilidad $\mu \in \mathcal{M}_+^1(\mathcal{X})$ tiene función de densidad $p : \mathcal{X} \rightarrow \mathbb{R}_+$, entonces

$$\int_{\mathcal{X}} f(x) \, d\mu(x) = \int_{\mathcal{X}} f(x) p(x) \, dx.$$

Problema de Schrödinger — Formulación dinámica

El siguiente problema consiste en encontrar un proceso estocástico *natural* que transforme una distribución de probabilidad en otra en un horizonte de tiempo finito:

SBP (formulación dinámica)

El puente de Schrödinger entre dos medidas $\mu, \nu \in \mathcal{M}_+^1(\mathcal{X})$ es el (único) proceso estocástico P^* que resuelve el problema

$$\min_{P \in \Gamma(\mu, \nu)} D_{KL}(P \| W^\epsilon),$$

donde $\Gamma(\mu, \nu) := \{P \in \mathcal{C}([0, 1], \mathcal{X}) : (P_0 \sim \mu) \wedge (P_1 \sim \nu)\}$, mientras que W^ϵ es un movimiento browniano con difusividad ϵ .

- Notar que se debería definir bien la cantidad $D_{KL}(P \| W^\epsilon)$. No se hará por simplicidad.
- La formulación se puede extender a otras medidas de referencias.

Problema de Schrödinger — Formulación dinámica

El siguiente problema consiste en encontrar un proceso estocástico *natural* que transforme una distribución de probabilidad en otra en un horizonte de tiempo finito:

SBP (formulación dinámica)

El puente de Schrödinger entre dos medidas $\mu, \nu \in \mathcal{M}_+^1(\mathcal{X})$ es el (único) proceso estocástico P^* que resuelve el problema

$$\min_{P \in \Gamma(\mu, \nu)} D_{KL}(P \| W^\epsilon),$$

donde $\Gamma(\mu, \nu) := \{P \in \mathcal{C}([0, 1], \mathcal{X}) : (P_0 \sim \mu) \wedge (P_1 \sim \nu)\}$, mientras que W^ϵ es un movimiento browniano con difusividad ϵ .

- Notar que se debería definir bien la cantidad $D_{KL}(P \| W^\epsilon)$. No se hará por simplicidad.
- La formulación se puede extender a otras medidas de referencias.

Problema de Schrödinger — Formulación dinámica

El siguiente problema consiste en encontrar un proceso estocástico *natural* que transforme una distribución de probabilidad en otra en un horizonte de tiempo finito:

SBP (formulación dinámica)

El puente de Schrödinger entre dos medidas $\mu, \nu \in \mathcal{M}_+^1(\mathcal{X})$ es el (único) proceso estocástico P^* que resuelve el problema

$$\min_{P \in \Gamma(\mu, \nu)} D_{KL}(P \| W^\epsilon),$$

donde $\Gamma(\mu, \nu) := \{P \in \mathcal{C}([0, 1], \mathcal{X}) : (P_0 \sim \mu) \wedge (P_1 \sim \nu)\}$, mientras que W^ϵ es un movimiento browniano con difusividad ϵ .

- Notar que se debería definir bien la cantidad $D_{KL}(P \| W^\epsilon)$. No se hará por simplicidad.
- La formulación se puede extender a otras medidas de referencias.

Problema de Schrödinger — Formulación dinámica

Puentes de Schrödinger para $\epsilon = 0.05$

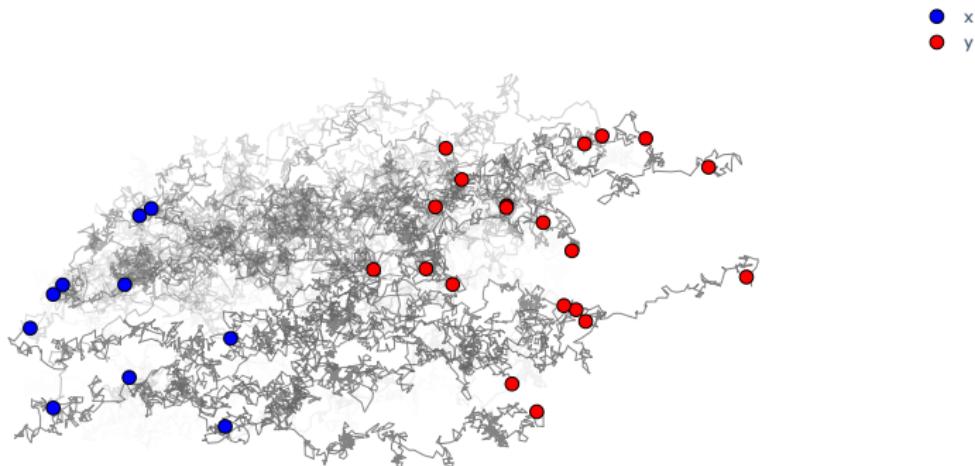


Figura 3: Puente de Schrödinger (dinámico) entre dos distribuciones discretas.

Problema de Schrödinger — Formulación dinámica

Este problema se puede reformular como un problema de control óptimo:

$$\min_{u \in \mathcal{U}} \mathbb{E}_x \left[\int_0^1 \frac{1}{2} \|u_t(x)\|^2 dt \right] \quad \text{sujeto a} \quad \begin{cases} dx_t = u_t(x) dt + dW_t^\epsilon \\ (x_0 \sim \mu) \wedge (x_1 \sim \nu) \end{cases},$$

donde \mathcal{U} es el conjunto de controles admisibles.

- Se puede reinterpretar el problema como uno de fluidodinámica cambiando la SDE por su ecuación de Fokker-Planck.
- La optimalidad también se puede caracterizar por un sistema acoplado de PDEs (sistema de Schrödinger). Esto permite entrenar un modelo neuronal para el SBP mediante máxima verosimilitud.
- En este caso, la función objetivo generaliza a la de los modelos de difusión a tiempo continuo, y los modelos de difusión pueden verse como un caso particular del SBP.

Problema de Schrödinger — Formulación dinámica

Este problema se puede reformular como un problema de control óptimo:

$$\min_{u \in \mathcal{U}} \mathbb{E}_x \left[\int_0^1 \frac{1}{2} \|u_t(x)\|^2 dt \right] \quad \text{sujeto a} \quad \begin{cases} dx_t = u_t(x) dt + dW_t^\epsilon \\ (x_0 \sim \mu) \wedge (x_1 \sim \nu) \end{cases},$$

donde \mathcal{U} es el conjunto de controles admisibles.

- Se puede reinterpretar el problema como uno de fluidodinámica cambiando la SDE por su ecuación de Fokker-Planck.
- La optimalidad también se puede caracterizar por un sistema acoplado de PDEs (sistema de Schrödinger). Esto permite entrenar un modelo neuronal para el SBP mediante máxima verosimilitud.
- En este caso, la función objetivo generaliza a la de los modelos de difusión a tiempo continuo, y los modelos de difusión pueden verse como un caso particular del SBP.

Problema de Schrödinger — Formulación dinámica

Este problema se puede reformular como un problema de control óptimo:

$$\min_{u \in \mathcal{U}} \mathbb{E}_x \left[\int_0^1 \frac{1}{2} \|u_t(x)\|^2 dt \right] \quad \text{sujeto a} \quad \begin{cases} dx_t = u_t(x) dt + dW_t^\epsilon \\ (x_0 \sim \mu) \wedge (x_1 \sim \nu) \end{cases},$$

donde \mathcal{U} es el conjunto de controles admisibles.

- Se puede reinterpretar el problema como uno de fluidodinámica cambiando la SDE por su ecuación de Fokker-Planck.
- La optimalidad también se puede caracterizar por un sistema acoplado de PDEs (sistema de Schrödinger). Esto permite entrenar un modelo neuronal para el SBP mediante máxima verosimilitud.
- En este caso, la función objetivo generaliza a la de los modelos de difusión a tiempo continuo, y los modelos de difusión pueden verse como un caso particular del SBP.

Problema de Schrödinger — Formulación dinámica

Este problema se puede reformular como un problema de control óptimo:

$$\min_{u \in \mathcal{U}} \mathbb{E}_x \left[\int_0^1 \frac{1}{2} \|u_t(x)\|^2 dt \right] \quad \text{sujeto a} \quad \begin{cases} dx_t = u_t(x) dt + dW_t^\epsilon \\ (x_0 \sim \mu) \wedge (x_1 \sim \nu) \end{cases},$$

donde \mathcal{U} es el conjunto de controles admisibles.

- Se puede reinterpretar el problema como uno de fluidodinámica cambiando la SDE por su ecuación de Fokker-Planck.
- La optimalidad también se puede caracterizar por un sistema acoplado de PDEs (sistema de Schrödinger). Esto permite entrenar un modelo neuronal para el SBP mediante máxima verosimilitud.
- En este caso, la función objetivo generaliza a la de los modelos de difusión a tiempo continuo, y los modelos de difusión pueden verse como un caso particular del SBP.

Problema de Schrödinger — Formulación estática

Se puede demostrar la siguiente descomposición:

$$D_{KL}(P \parallel W^\epsilon) = D_{KL}(P_{01} \parallel W_{01}^\epsilon) + \mathbb{E}_{(x,y) \sim P_{01}} \left[D_{KL}\left(P_{|xy} \parallel W_{|xy}^\epsilon\right) \right].$$

Luego, el SBP dinámico se puede reducir a un problema estático enfocado únicamente en los extremos del proceso:

$$\underbrace{\min_{P \in \Gamma(\mu, \nu)} D_{KL}(P \parallel W^\epsilon)}_{\text{problema dinámico}} = \underbrace{\min_{P_{01} \in \Pi(\mu, \nu)} D_{KL}(P_{01} \parallel W_{01}^\epsilon)}_{\text{problema estático}},$$

donde $\Pi(\mu, \nu) := \{\pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y}) : (\pi_0 \sim \mu) \wedge (\pi_1 \sim \nu)\}$.

Por lo tanto, si P_{01}^* es la (única) solución del SBP estático, la (única) solución del SBP dinámico es

$$P^*(\cdot) = \int_{\mathcal{X} \times \mathcal{Y}} W_{|xy}^\epsilon(\cdot) dP_{01}^*(x, y).$$

Problema de Schrödinger — Formulación estática

Se puede demostrar la siguiente descomposición:

$$D_{KL}(P \parallel W^\epsilon) = D_{KL}(P_{01} \parallel W_{01}^\epsilon) + \mathbb{E}_{(x,y) \sim P_{01}} \left[D_{KL}(P_{|xy} \parallel W_{|xy}^\epsilon) \right].$$

Luego, el SBP dinámico se puede reducir a un problema estático enfocado únicamente en los extremos del proceso:

$$\underbrace{\min_{P \in \Gamma(\mu, \nu)} D_{KL}(P \parallel W^\epsilon)}_{\text{problema dinámico}} = \underbrace{\min_{P_{01} \in \Pi(\mu, \nu)} D_{KL}(P_{01} \parallel W_{01}^\epsilon)}_{\text{problema estático}},$$

donde $\Pi(\mu, \nu) := \{\pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y}) : (\pi_0 \sim \mu) \wedge (\pi_1 \sim \nu)\}$.

Por lo tanto, si P_{01}^* es la (única) solución del SBP estático, la (única) solución del SBP dinámico es

$$P^*(\cdot) = \int_{\mathcal{X} \times \mathcal{Y}} W_{|xy}^\epsilon(\cdot) dP_{01}^*(x, y).$$

Problema de Schrödinger — Formulación estática

Se puede demostrar la siguiente descomposición:

$$D_{KL}(P \parallel W^\epsilon) = D_{KL}(P_{01} \parallel W_{01}^\epsilon) + \mathbb{E}_{(x,y) \sim P_{01}} \left[D_{KL}(P_{|xy} \parallel W_{|xy}^\epsilon) \right].$$

Luego, el SBP dinámico se puede reducir a un problema estático enfocado únicamente en los extremos del proceso:

$$\underbrace{\min_{P \in \Gamma(\mu, \nu)} D_{KL}(P \parallel W^\epsilon)}_{\text{problema dinámico}} = \underbrace{\min_{P_{01} \in \Pi(\mu, \nu)} D_{KL}(P_{01} \parallel W_{01}^\epsilon)}_{\text{problema estático}},$$

donde $\Pi(\mu, \nu) := \{\pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y}) : (\pi_0 \sim \mu) \wedge (\pi_1 \sim \nu)\}$.

Por lo tanto, si P_{01}^* es la (única) solución del SBP estático, la (única) solución del SBP dinámico es

$$P^*(\cdot) = \int_{\mathcal{X} \times \mathcal{Y}} W_{|xy}^\epsilon(\cdot) dP_{01}^*(x, y).$$

Problema de Schrödinger — Formulación dinámica

Plan de transporte óptimo para $\epsilon = 0.1$

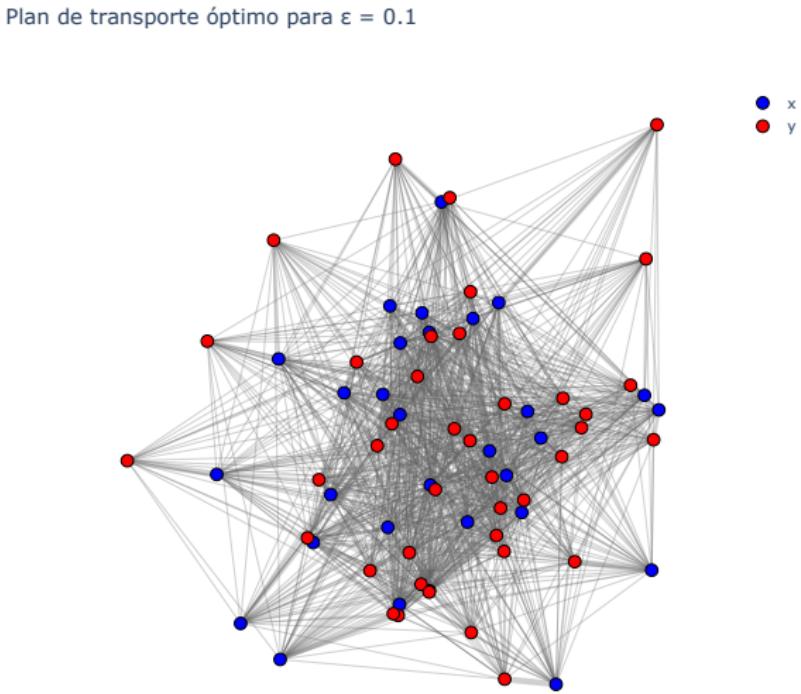


Figura 4: Puente de Schrödinger (estático) entre dos distribuciones discretas.

Problema de Schrödinger — Formulación estática

El SBP estático es equivalente al problema de transporte óptimo con regularización entrópica:

$$\begin{aligned} & D_{\text{KL}}(P_{01} \parallel W_{01}^\epsilon) \\ &= \frac{1}{\epsilon} \left[\underbrace{\int_{\mathcal{X} \times \mathcal{Y}} \frac{1}{2} \|x - y\|^2 dP_{01}(x, y)}_{\text{costo de transporte}} + \underbrace{-\epsilon \cdot \mathcal{H}(P_{01})}_{\text{regularizador}} \right] + \text{constante}, \end{aligned}$$

donde $\mathcal{H}(P_{01})$ es la entropía (diferencial) de P_{01} .

Más aún, todo SBP (con una cierta medida de referencia) puede ser transformado a un problema de EOT (con una cierta función de costo), y viceversa.

Problema de Schrödinger — Formulación estática

El SBP estático es equivalente al problema de transporte óptimo con regularización entrópica:

$$\begin{aligned} & D_{KL}(P_{01} \parallel W_{01}^\epsilon) \\ &= \frac{1}{\epsilon} \left[\underbrace{\int_{\mathcal{X} \times \mathcal{Y}} \frac{1}{2} \|x - y\|^2 dP_{01}(x, y)}_{\text{costo de transporte}} + \underbrace{-\epsilon \cdot \mathcal{H}(P_{01})}_{\text{regularizador}} \right] + \text{constante}, \end{aligned}$$

donde $\mathcal{H}(P_{01})$ es la entropía (diferencial) de P_{01} .

Más aún, todo SBP (con una cierta medida de referencia) puede ser transformado a un problema de EOT (con una cierta función de costo), y viceversa.

Problema de Schrödinger — Formulación estática

El SBP estático es equivalente al problema de transporte óptimo con regularización entrópica:

$$\begin{aligned} & D_{\text{KL}}(P_{01} \parallel W_{01}^\epsilon) \\ &= \frac{1}{\epsilon} \left[\underbrace{\int_{\mathcal{X} \times \mathcal{Y}} \frac{1}{2} \|x - y\|^2 dP_{01}(x, y)}_{\text{costo de transporte}} + \underbrace{-\epsilon \cdot \mathcal{H}(P_{01})}_{\text{regularizador}} \right] + \text{constante}, \end{aligned}$$

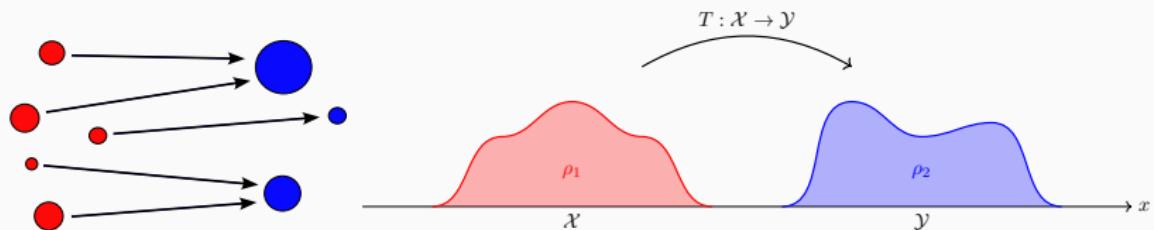
donde $\mathcal{H}(P_{01})$ es la entropía (diferencial) de P_{01} .

Más aún, todo SBP (con una cierta medida de referencia) puede ser transformado a un problema de EOT (con una cierta función de costo), y viceversa.

Transporte óptimo

Transporte óptimo — Problema de Monge

El problema de transformar una distribución en otra puede modelarse como un problema de transporte óptimo:



Problema de Monge

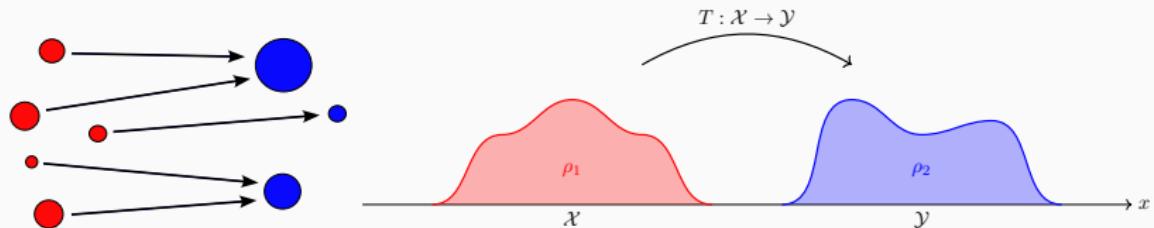
Si $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ es una función continua que mide la *discrepancia* entre dos puntos, el problema de Monge entre $\mu \in \mathcal{M}_+^1(\mathcal{X})$ y $\nu \in \mathcal{M}_+^1(\mathcal{Y})$ es:

$$\inf_{\substack{T: \mathcal{X} \rightarrow \mathcal{Y} \\ T_\# \mu = \nu}} \int_{\mathcal{X}} c(x, T(x)) d\mu(x),$$

donde la igualdad $T_\# \mu = \nu$ indica que $T(x) \sim \nu$ cuando $x \sim \mu$.

Transporte óptimo — Problema de Monge

El problema de transformar una distribución en otra puede modelarse como un problema de transporte óptimo:



Problema de Monge

Si $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ es una función continua que mide la *discrepancia* entre dos puntos, el problema de Monge entre $\mu \in \mathcal{M}_+^1(\mathcal{X})$ y $\nu \in \mathcal{M}_+^1(\mathcal{Y})$ es:

$$\inf_{\substack{T: \mathcal{X} \rightarrow \mathcal{Y} \\ T_\# \mu = \nu}} \int_{\mathcal{X}} c(x, T(x)) d\mu(x),$$

donde la igualdad $T_\# \mu = \nu$ indica que $T(x) \sim \nu$ cuando $x \sim \mu$.

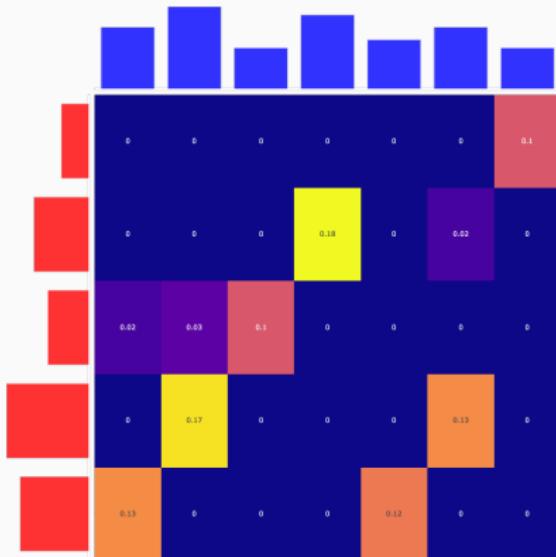
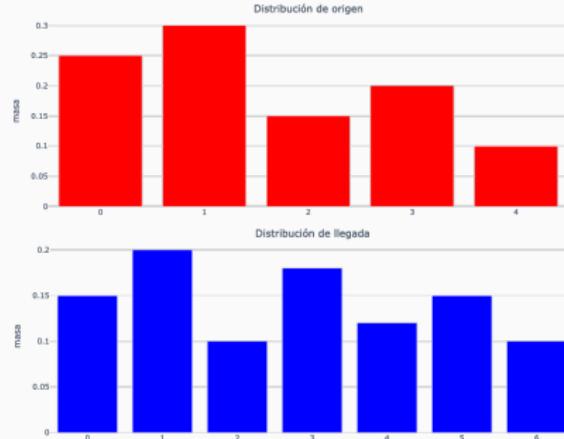
Transporte óptimo — Relajación de Kantorovich

El problema de Monge es altamente no lineal, no es convexo ni posee necesariamente solución.

Transporte óptimo — Relajación de Kantorovich

El problema de Monge es altamente no lineal, no es convexo ni posee necesariamente solución.

Estas limitaciones se pueden suprimir si se permite división de masa.



Transporte óptimo — Relajación de Kantorovich

Relajación de Kantorovich

Si $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ es una función continua que mide la *discrepancia* entre dos puntos, el problema de Kantorovich entre $\mu \in \mathcal{M}_+^1(\mathcal{X})$ y $\nu \in \mathcal{M}_+^1(\mathcal{Y})$ es:

$$\inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \, d\pi(x, y),$$

donde $\Pi(\mu, \nu) = \{\pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y}) : \pi_0 = \mu, \pi_1 = \nu\}$.

- Es un problema convexo y cumple dualidad fuerte.
- Este problema, más un término regularizador, equivale al SBP estático.

Transporte óptimo — Relajación de Kantorovich

Relajación de Kantorovich

Si $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ es una función continua que mide la *discrepancia* entre dos puntos, el problema de Kantorovich entre $\mu \in \mathcal{M}_+^1(\mathcal{X})$ y $\nu \in \mathcal{M}_+^1(\mathcal{Y})$ es:

$$\inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \, d\pi(x, y),$$

donde $\Pi(\mu, \nu) = \{\pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y}) : \pi_0 = \mu, \pi_1 = \nu\}$.

- Es un problema convexo y cumple dualidad fuerte.
- Este problema, más un término regularizador, equivale al SBP estático.

Transporte óptimo — Relajación de Kantorovich

Relajación de Kantorovich

Si $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ es una función continua que mide la *discrepancia* entre dos puntos, el problema de Kantorovich entre $\mu \in \mathcal{M}_+^1(\mathcal{X})$ y $\nu \in \mathcal{M}_+^1(\mathcal{Y})$ es:

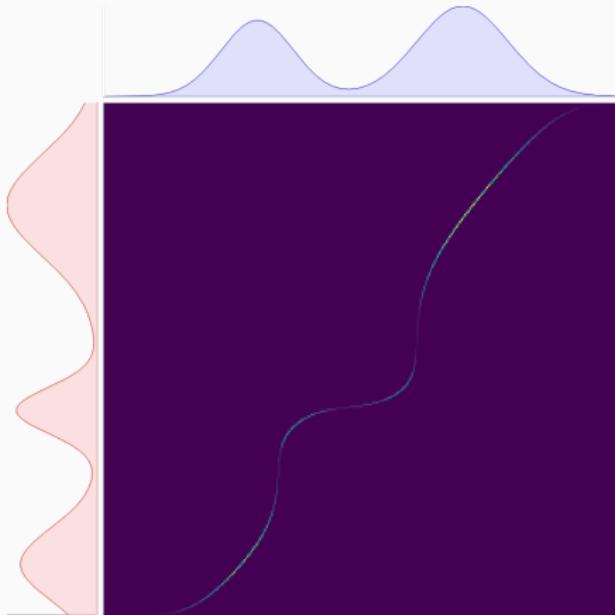
$$\inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \, d\pi(x, y),$$

donde $\Pi(\mu, \nu) = \{\pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y}) : \pi_0 = \mu, \pi_1 = \nu\}$.

- Es un problema convexo y cumple dualidad fuerte.
- Este problema, más un término regularizador, equivale al SBP estático.

Transporte óptimo — Relajación de Kantorovich

Bajo hipótesis razonables, ambos problemas son equivalentes en el caso continuo: si $\pi^* \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})$ es la solución del problema de Kantorovich, toda su masa está concentrada en una curva, que resulta ser el grafo de la solución $T^* : \mathcal{X} \rightarrow \mathcal{Y}$ del problema de Monge.



Transporte óptimo — Distancia de Wasserstein

Si $d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$ es una distancia en \mathcal{X} , entonces

$$\mathcal{W}_p(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{X}} d(x, y)^p \, d\pi \right)^{\frac{1}{p}}$$

es una distancia en $\mathcal{M}_+^1(\mathcal{X})$, llamada *distancia de Wasserstein*.

Transporte óptimo — Distancia de Wasserstein

Si $d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$ es una distancia en \mathcal{X} , entonces

$$\mathcal{W}_p(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{X}} d(x, y)^p \, d\pi \right)^{\frac{1}{p}}$$

es una distancia en $\mathcal{M}_+^1(\mathcal{X})$, llamada *distancia de Wasserstein*.

El espacio métrico $(\mathcal{M}_+^1(\mathcal{X}), \mathcal{W}_p)$ es geodésico. Esto permite obtener una formulación dinámica del transporte óptimo.

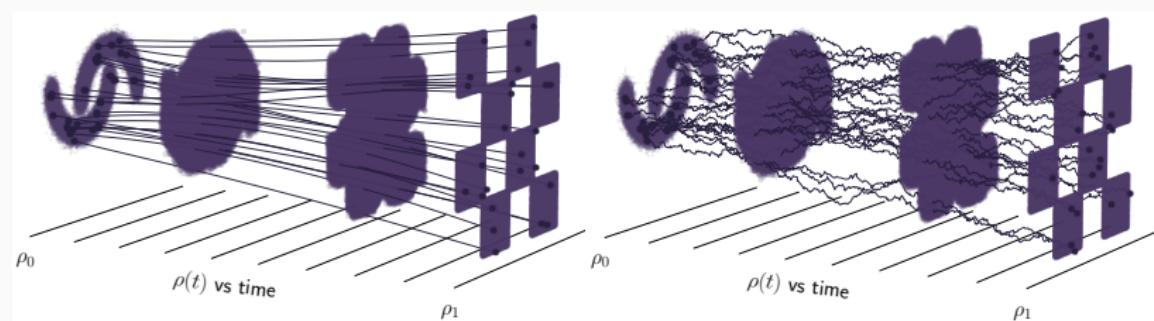
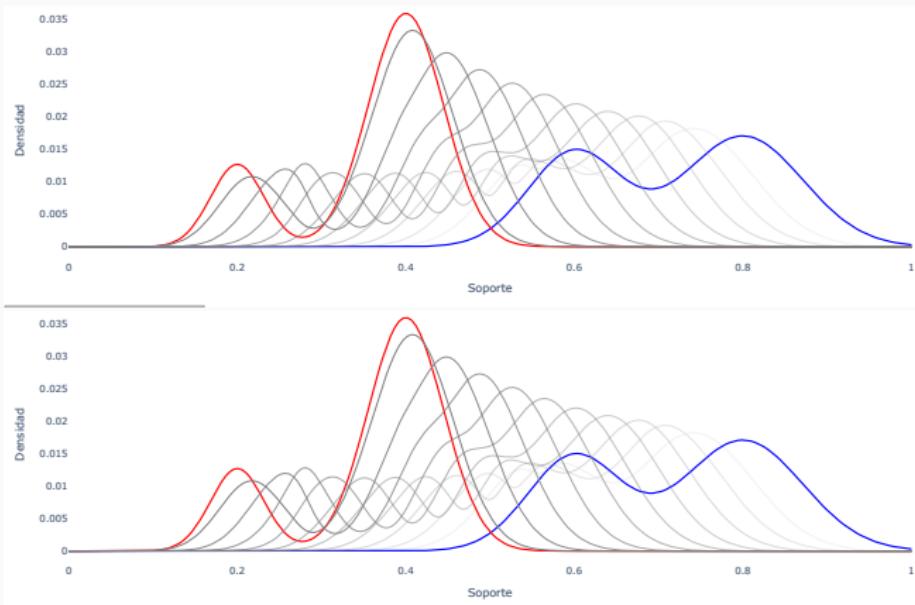


Figura 5: Imagen obtenida desde...

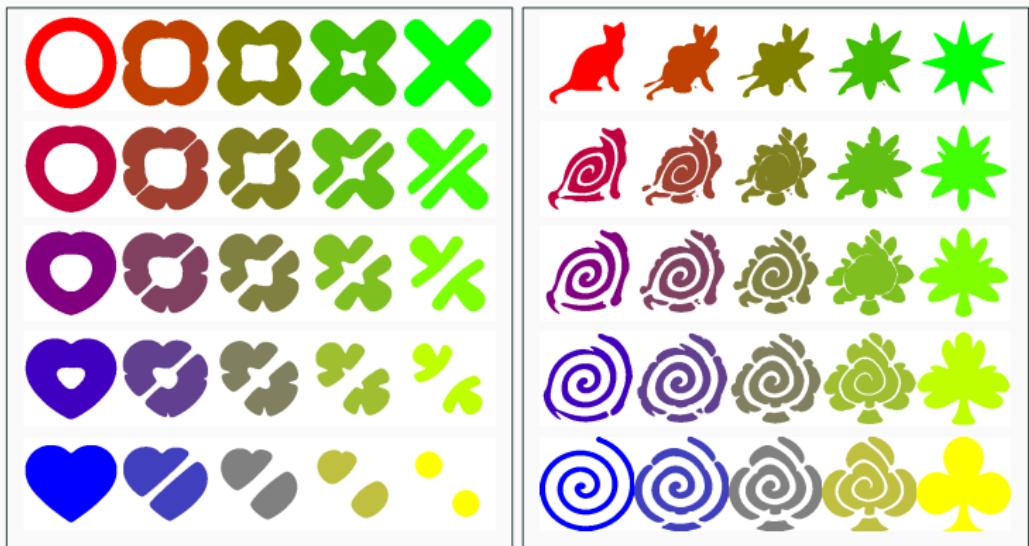
Transporte óptimo — Distancia de Wasserstein

Esta distancia permite interpolar entre distribuciones de probabilidad a través de geodésicas en $\mathcal{M}_+^1(\mathcal{X})$, generando interpolaciones más realistas que la interpolación euclíadiana.



Transporte óptimo — Distancia de Wasserstein

Más aún, se puede realizar interpolación baricéntrica de forma eficiente:



Transporte óptimo — Distancia de Wasserstein

La distancia de Wasserstein tiene buenas propiedades matemáticas:

- Es más débil que la distancia en variación total. Más aún, metriza la convergencia débil de medidas.
- Puede ser optimizada por redes neuronales: si $g_\theta(z) \sim \mu_\theta$ es un modelo generativo neuronal de variable latente $z \sim \mathcal{N}(0, I_I)$, entonces $\theta \mapsto \mathcal{W}_1(\mu_\theta, \mu_{\text{true}})$ es diferenciable (c.t.p.).
- El problema se puede reformular como uno de control óptimo:

$$\min_{u \in \mathcal{U}} \mathbb{E}_x \left[\int_0^1 \frac{1}{2} \|u_t(x)\|^2 dt \right] \quad \text{sujeto a} \quad \begin{cases} dx_t = u_t(x) dt \\ (x_0 \sim \mu) \wedge (x_1 \sim \nu), \end{cases}$$

donde \mathcal{U} es el conjunto de controles admisibles.

- Formulación de Benamou-Brenier: se puede sustituir la dinámica de x por la ecuación de continuidad, transformando el problema en uno de fluidodinámica.

Transporte óptimo — Distancia de Wasserstein

La distancia de Wasserstein tiene buenas propiedades matemáticas:

- Es más débil que la distancia en variación total. Más aún, metriza la convergencia débil de medidas.
- Puede ser optimizada por redes neuronales: si $g_\theta(z) \sim \mu_\theta$ es un modelo generativo neuronal de variable latente $z \sim \mathcal{N}(0, I_I)$, entonces $\theta \mapsto \mathcal{W}_1(\mu_\theta, \mu_{\text{true}})$ es diferenciable (c.t.p.).
- El problema se puede reformular como uno de control óptimo:

$$\min_{u \in \mathcal{U}} \mathbb{E}_x \left[\int_0^1 \frac{1}{2} \|u_t(x)\|^2 dt \right] \quad \text{sujeto a} \quad \begin{cases} dx_t = u_t(x) dt \\ (x_0 \sim \mu) \wedge (x_1 \sim \nu), \end{cases}$$

donde \mathcal{U} es el conjunto de controles admisibles.

- Formulación de Benamou-Brenier: se puede sustituir la dinámica de x por la ecuación de continuidad, transformando el problema en uno de fluidodinámica.

Transporte óptimo — Distancia de Wasserstein

La distancia de Wasserstein tiene buenas propiedades matemáticas:

- Es más débil que la distancia en variación total. Más aún, metriza la convergencia débil de medidas.
- Puede ser optimizada por redes neuronales: si $g_\theta(z) \sim \mu_\theta$ es un modelo generativo neuronal de variable latente $z \sim \mathcal{N}(0, I_I)$, entonces $\theta \mapsto \mathcal{W}_1(\mu_\theta, \mu_{\text{true}})$ es diferenciable (c.t.p.).
- El problema se puede reformular como uno de control óptimo:

$$\min_{u \in \mathcal{U}} \mathbb{E}_x \left[\int_0^1 \frac{1}{2} \|u_t(x)\|^2 dt \right] \quad \text{sujeto a} \quad \begin{cases} dx_t = u_t(x) dt \\ (x_0 \sim \mu) \wedge (x_1 \sim \nu), \end{cases}$$

donde \mathcal{U} es el conjunto de controles admisibles.

- Formulación de Benamou-Brenier: se puede sustituir la dinámica de x por la ecuación de continuidad, transformando el problema en uno de fluidodinámica.

Transporte óptimo — Distancia de Wasserstein

La distancia de Wasserstein tiene buenas propiedades matemáticas:

- Es más débil que la distancia en variación total. Más aún, metriza la convergencia débil de medidas.
- Puede ser optimizada por redes neuronales: si $g_\theta(z) \sim \mu_\theta$ es un modelo generativo neuronal de variable latente $z \sim \mathcal{N}(0, I_I)$, entonces $\theta \mapsto \mathcal{W}_1(\mu_\theta, \mu_{\text{true}})$ es diferenciable (c.t.p.).
- El problema se puede reformular como uno de control óptimo:

$$\min_{u \in \mathcal{U}} \mathbb{E}_x \left[\int_0^1 \frac{1}{2} \|u_t(x)\|^2 dt \right] \quad \text{sujeto a} \quad \begin{cases} dx_t = u_t(x) dt \\ (x_0 \sim \mu) \wedge (x_1 \sim \nu), \end{cases}$$

donde \mathcal{U} es el conjunto de controles admisibles.

- Formulación de Benamou-Brenier: se puede sustituir la dinámica de x por la ecuación de continuidad, transformando el problema en uno de fluidodinámica.

Transporte óptimo — Distancia de Wasserstein

La distancia de Wasserstein tiene buenas propiedades matemáticas:

- Es más débil que la distancia en variación total. Más aún, metriza la convergencia débil de medidas.
- Puede ser optimizada por redes neuronales: si $g_\theta(z) \sim \mu_\theta$ es un modelo generativo neuronal de variable latente $z \sim \mathcal{N}(0, I_I)$, entonces $\theta \mapsto \mathcal{W}_1(\mu_\theta, \mu_{\text{true}})$ es diferenciable (c.t.p.).
- El problema se puede reformular como uno de control óptimo:

$$\min_{u \in \mathcal{U}} \mathbb{E}_x \left[\int_0^1 \frac{1}{2} \|u_t(x)\|^2 dt \right] \quad \text{sujeto a} \quad \begin{cases} dx_t = u_t(x) dt \\ (x_0 \sim \mu) \wedge (x_1 \sim \nu), \end{cases}$$

donde \mathcal{U} es el conjunto de controles admisibles.

- Formulación de Benamou-Brenier: se puede sustituir la dinámica de x por la ecuación de continuidad, transformando el problema en uno de fluidodinámica.

Transporte óptimo — Modelos de difusión como OT

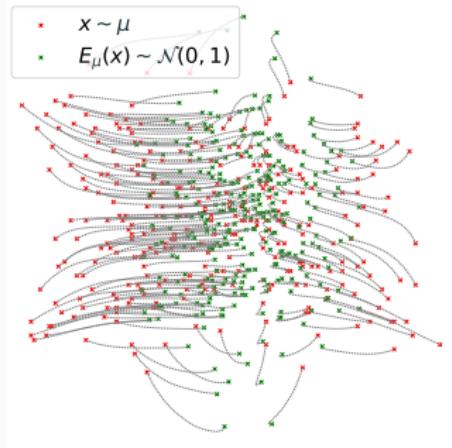


Figura 6: Imagen obtenida desde

- Para un modelo de difusión hasta tiempo $T > 0$, se denotará por $E_T(x) \in \mathcal{Y}$ el lugar al que llega un punto $x \sim p_{\text{true}}(x)$ que fluye a través de la *probability flow ODE*.
- Se ha probado empíricamente que E_T también converge al mapa de Monge entre p_{true} y p_{prior} .
- Se han dado contraejemplos teóricos donde esto no ocurre.

Transporte óptimo — Modelos de difusión como OT

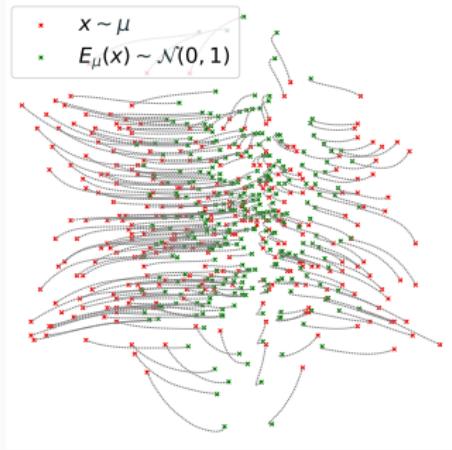


Figura 6: Imagen obtenida desde

- Para un modelo de difusión hasta tiempo $T > 0$, se denotará por $E_T(x) \in \mathcal{Y}$ el lugar al que llega un punto $x \sim p_{\text{true}}(x)$ que fluye a través de la *probability flow ODE*.
- Se ha probado empíricamente que E_T también converge al mapa de Monge entre p_{true} y p_{prior} .
- Se han dado contraejemplos teóricos donde esto no ocurre.

Transporte óptimo — Modelos de difusión como OT

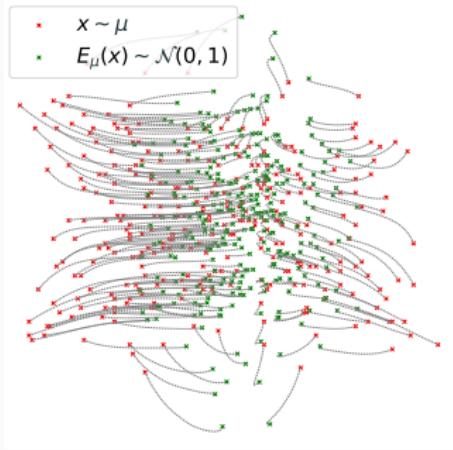


Figura 6: Imagen obtenida desde

- Para un modelo de difusión hasta tiempo $T > 0$, se denotará por $E_T(x) \in \mathcal{Y}$ el lugar al que llega un punto $x \sim p_{\text{true}}(x)$ que fluye a través de la *probability flow ODE*.
- Se ha probado empíricamente que E_T también converge al mapa de Monge entre p_{true} y p_{prior} .
- Se han dado contraejemplos teóricos donde esto no ocurre.