# Modelos de difusión, transporte óptimo y puentes de Schrödinger

Fernando Fêtis Riquelme

DIM/MDS - Universidad de Chile

## Formulación de los modelos de difusión

Los modelos de difusión son una familia de **modelos generativos** que buscan aprender a generar muestras a partir de una distribución de datos $p_{\text{true}}$. Para esto, realizan un proceso de distorsión (1) sobre muestras de $p_{\text{true}}$, transformándolas en muestras de otra distribución $p_{\text{prior}}$. En paralelo, una red neuronal $p_\theta$ busca aprender el proceso de reconstrucción mediante (2).

### Procesos de difusión y reconstrucción

Dada una secuencia de ruidos, $(\beta_t)_{t=1}^T$, el proceso de difusión es una cadena de Markov que progresivamente va inyectando ruido gaussiano a muestras $x_0 \sim q(x_0) = p_{\text{true}}(x_0)$ hasta llegar a una muestra final $x_T \sim q(x_T) \approx p_{\text{prior}}(x_T)$ mediante el proceso forward con factorización causal

$$q(x_{1:T}|x_0) = \prod_{t=1} q(x_t|x_{t-1}), \quad \text{donde} \quad q(x_t|x_{t-1}) \sim \mathcal{N}\left(\sqrt{1-\beta}\, x_{t-1},\, \beta_t \,\mathrm{I}_d\right) \tag{1}$$

El proceso de reconstrucción es otra cadena de Markov (hacia atrás en el tiempo) que busca aprender las transiciones reversas $q(x_{t-1}|x_t)$ mediante un modelo paramétrico $p_\theta(x_{t-1}|x_t)$. Este proceso es el que permite generar nuevas muestras comenzando desde $x_T$:

$$p_\theta(x_{0:T}) = p_\theta(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t), \quad \text{donde} \quad p_\theta(x_T) = p(x_T) = p_{\text{prior}}(x_T) \sim \mathcal{N}(0,\, \mathrm{I}_d) \tag{2}$$

Eligiendo transiciones gaussianas, $p_\theta(x_{t-1}|x_t) \sim \mathcal{N}(\mu_\theta(x_t,t),\, \Sigma_\theta(x_t,t))$, basta entrenar una red neuronal que aprenda un vector de medias $\mu_\theta$ y una matriz de covarianzas $\Sigma_\theta$. Para esto, **se maximiza una cota inferior de la evidencia**: $\text{ELBO} = \log p_\theta(x_0) - \mathrm{D}_{\mathrm{KL}}\left(q(x_{1:T}|x_0) \,\|\, p_\theta(x_{1:T}|x_0)\right)$.

## Aspectos de los modelos de difusión

Es usual realizar el **proceso de difusión en el espacio latente** (mediante el uso de un VAE) y usar **arquitecturas especializadas** para modelos de difusión (U-Net y DiT), permitiendo realizar generación condicional, la cual puede ser mejorada mediante técnicas de **guidance**. Además, estos modelos pueden ser extendidos a una **formulación continua** mediante el uso de ecuaciones diferenciales estocásticas (ver Figura 1), estableciendo una conexión con los **modelos basados en score**.

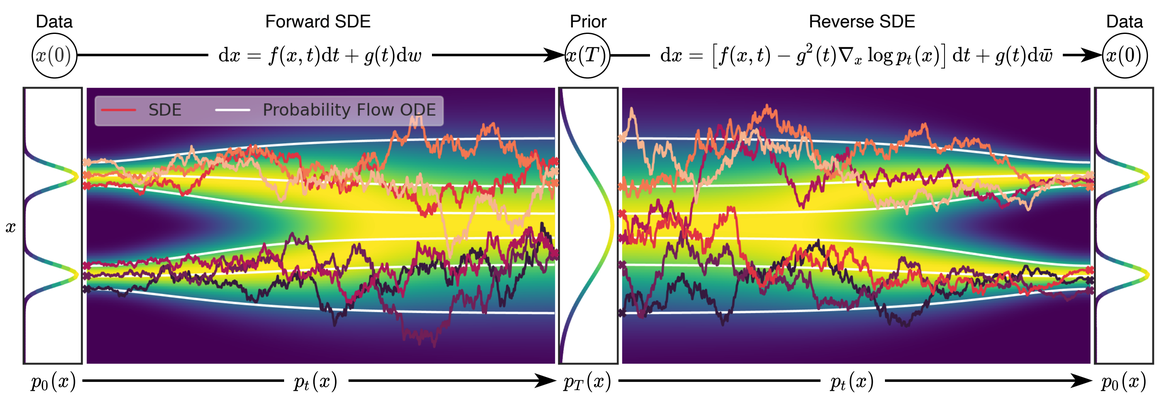


**Figura 1.** Modelo de difusión utilizando una SDE. La SDE del proceso reverso viene dado por el teorema de Anderson.

## Limitaciones

Algunas limitaciones de esta familia de modelos son las siguientes:

- **Distribución final:** los modelos de difusión están limitados a terminar en $p_{\text{prior}}(x_T) \sim \mathcal{N}(0,\, \mathrm{I}_d)$.
- **Convergencia asintótica:** la igualdad $q(x_T) = p_{\text{prior}}(x_T)$ solo se alcanza cuando $T \to \infty$, provocando discrepancias al generar muestras usando el proceso reverso (2).
- **Tiempo de simulación muy largo:** la limitación anterior obliga a que el proceso de difusión sea simulado hasta tiempos finales $T$ muy grandes.
- **Sensibilidad a los procesos de difusión:** la elección de los niveles de ruido $(\beta_t)_{t=1}^T$ (o de la SDE de difusión en el caso continuo) tiene un impacto significativo sobre el rendimiento del modelo.
- **Dificultad en la interpretabilidad:** dado el exceso de heurísticas usadas en los modelos de difusión y la falta de garantías teóricas, estos modelos son difíciles de interpretar.

## Enfoque alternativo mediante la teoría del transporte óptimo

La *probability flow ODE* (ver Figura 1) induce un mapa determinista entre muestras de $p_{\text{true}}$ y muestras de $p_{\text{prior}}$. El **problema de Monge** permite extender este escenario a medidas de probabilidad arbitrarias, $\mu \in \mathcal{M}_+^1(\mathcal{X})$ y $\nu \in \mathcal{M}_+^1(\mathcal{Y})$, que no necesariamente posean funciones de densidad. En este nuevo problema se busca un *mapa de transporte* que vincule los soporte de ambas medidas de forma eficiente de acuerdo a un funcional de costo $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$:

$$\inf_{\substack{T:\mathcal{X}\to\mathcal{Y} \\ T_\#\mu=\nu}} \int_{\mathcal{X}} c(x, T(x))\, \mathrm{d}\mu(x), \quad \text{donde} \quad T_\#\mu = \nu \quad \text{indica que} \quad \mu\left(T^{-1}(B)\right) = \nu(B), \quad \forall B \in \mathcal{B}(\mathcal{Y})$$

Dado que la restricción $T_\#\mu = \nu$ es altamente no lineal, el problema de Monge es difícil de estudiar. Por este motivo, se suele trabajar con la **relajación de Kantorovich**, donde se busca un *plan de transporte* $\pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})$ que distribuya eficientemente la masa de $\mu$ (interpretada como oferta) de acuerdo a la masa de $\nu$ (interpretada como demanda) pudiendo, eventualmente, realizar división de masa:

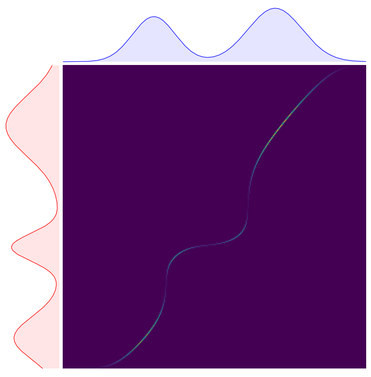

**Figura 2.** El soporte del plan de transporte óptimo $\pi^*$ está ubicado sobre el grafo del mapa de transporte óptimo $T^*$.

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\, \mathrm{d}\pi(x,y) \tag{3}$$

donde el conjunto factible es el conjunto de *couplings* entre las medidas de probabilidad $\mu$ y $\nu$:

$$\Pi(\mu,\nu) := \left\{\pi \in \mathcal{M}_+^1(\mathcal{X}\times\mathcal{Y}) \,:\, \pi_1 = \mu,\, \pi_2 = \nu\right\}$$

Bajo hipótesis razonables, el teorema de Brenier indica que **la relajación de Kantorovich y el problema de Monge son equivalentes**, en el sentido de que es posible obtener un mapa de transporte óptimo $T^*$ a partir de un plan de transporte óptimo $\pi^*$. Esto ocurre gracias a que las soluciones de la relajación de Kantorovich son *deterministas* tal como lo muestra la Figura 2.

## Propiedades del problema de Kantorovich y regularización entrópica

### Ventajas del problema de Kantorovich

- **Convexidad:** (3) es un problema de optimización lineal y convexo.
- **Dualidad:** su formulación dual reduce cuadráticamente la cantidad de incógnitas y es interpretable.
- **Metrizabilidad:** cuando $\mathcal{X} = \mathcal{Y}$ es compacto y $c$ es una distancia en $\mathcal{X}$, el valor óptimo de (3) induce una distancia entre $\mu$ y $\nu$. Más aún, esta distancia metriza la convergencia débil de medidas en $\mathcal{M}_+^1(\mathcal{X})$, por lo que es una distancia más débil que la distancia en variación total.
- **Espacio geodésico:** la distancia que induce este problema (conocida como distancia de Wasserstein) vuelve a $\mathcal{M}_+^1(\mathcal{X})$ un espacio métrico geodésico, permitiendo hacer interpolación entre medidas de probabilidad, lo cual se puede observar en la Figura 3.

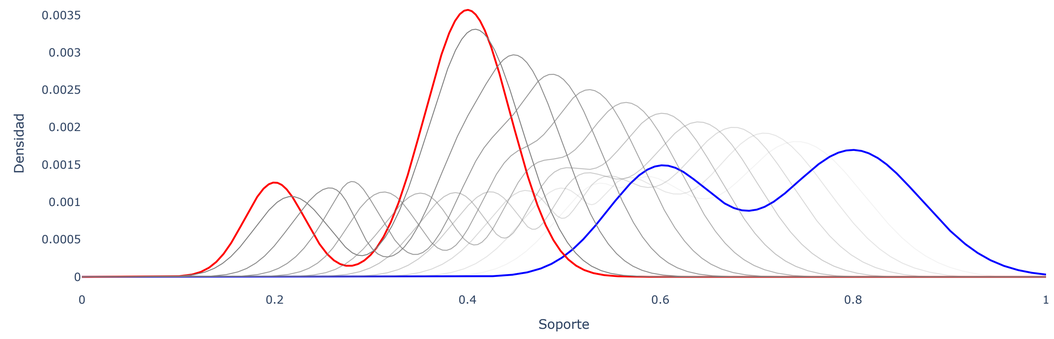


**Figura 3.** Interpolaciones $\mu_t = (T_t)_\#\mu$, donde $T_t = (1-t)\,\mathrm{Id} + tT^*$ y $T^*$ es el mapa de Monge entre dos mixturas gaussianas.

### Regularización entrópica

Si bien el problema de Kantorovich tiene buenas propiedades, es costoso de resolver y sufre de la **maldición de la dimensionalidad**. Agregando un término de regularización basado en entropía, se obtiene un problema estrictamente convexo, por lo que **su solución es única**. Además el problema dual de este nuevo problema permite obtener la solución del problema primal y da paso al **algoritmo de Sinkhorn**, con el cual se puede resolver eficientemente el problema de Kantorovich regularizado:

$$\inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c\, \mathrm{d}\pi + \epsilon \cdot \mathrm{D}_{\mathrm{KL}}\left(\pi \,\|\, \mu\otimes\nu\right) \quad \text{donde} \quad \mathrm{D}_{\mathrm{KL}}\left(\pi \,\|\, \mu\otimes\nu\right) = \int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}(\mu\otimes\nu)}\right) \mathrm{d}\pi$$

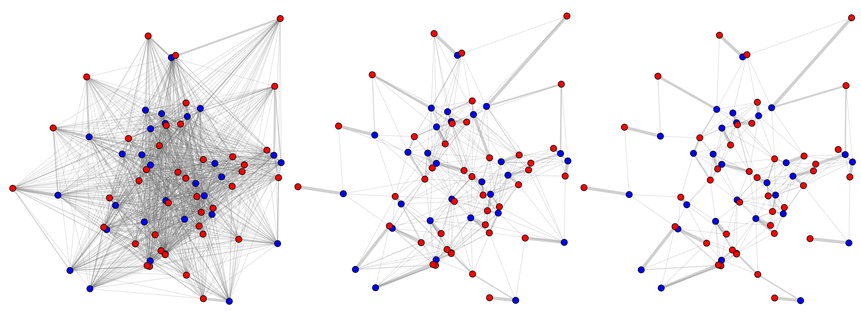

**Figura 4.** Solución del problema entrópico discreto para $\epsilon \in \{1,0; 0,01; 0,005\}$ (descendente hacia la derecha).

Esta regularización soluciona los problemas mencionados anteriormente y se puede probar que cuando $\epsilon \to 0$, su solución converge a la solución de (3) de máxima entropía. Además, esta regularización **vuelve al problema de Kantorovich diferenciable**, pudiendo ser resuelto usando redes neuronales.

## Formulación dinámica y problema del puente de Schrödinger

El problema de Kantorovich regularizado resulta ser equivalente al **problema de Schrödinger estático**, el cual busca un coupling $\pi \in \Pi(\mu,\nu)$ que esté lo más cercano posible (en el sentido de la entropía relativa) a una medida de referencia $\mathcal{W} \in \mathcal{M}_+^1(\mathcal{X}\times\mathcal{X})$:

$$\operatorname*{arg\,min}_{\pi\in\Pi(\mu,\nu)} \mathrm{D}_{\mathrm{KL}}\left(\pi \,\|\, \mathcal{W}\right)$$

Este problema puede ser extendido a una versión dinámica, donde en vez de buscar un coupling $\pi^*$, se busca un proceso estocástico $\mathbb{P}$ (en un horizonte temporal $t \in [0,1]$) que sea lo más similar a un proceso estocástico de referencia $\mathbb{W}$ (usualmente un movimiento browniano). Denotando por $\mathcal{F}(\mu,\nu)$ al conjunto de procesos estocásticos cuyas distribuciones marginales en $t=0$ y $t=1$ son $\mu$ y $\nu$ respectivamente, se puede demostrar que la solución $\mathbb{P}^*$ de este problema dinámico, conocida como **puente de Schrödinger**, corresponde a interpolar un puente browniano de acuerdo a $\pi^*$:

$$\mathbb{P}^* = \operatorname*{arg\,min}_{\mathbb{P}\in\mathcal{F}(\mu,\nu)} \mathrm{D}_{\mathrm{KL}}\left(\mathbb{P} \,\|\, \mathbb{W}\right) \implies \mathbb{P}^*(\cdot) = \int_{\mathcal{X}\times\mathcal{X}} \mathbb{W}_{xy}(\cdot)\, \mathrm{d}\pi^*(x,y)$$

donde $\mathbb{W}_{xy}$ es un movimiento browniano que empieza en $x$ (para $t=0$) y termina en $y$ (para $t=1$). En la Figura 5 se muestra un ejemplo de este problema dinámico.

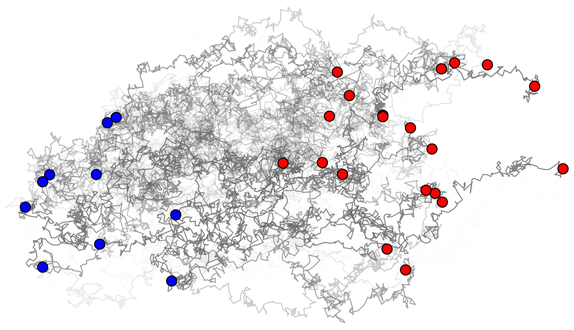

**Figura 5.** Puente de Schrödinger entre dos distribuciones discretas cuando se considera un proceso de referencia browniano con difusividad $\epsilon = 0{,}1$.

El problema de Schrödinger también se puede estudiar desde la perspectiva del **control óptimo**, donde se obtiene una problema de control que generaliza la **formulación de Benamou-Brenier** del transporte óptimo no regularizado. Además, es posible trabajar directamente con las SDEs que modelan los procesos estocásticos $\mathbb{P}$ y $\mathbb{R}$, donde el término de drift de $\mathbb{P}$ puede ser encontrado resolviendo un sistema acoplado de PDEs. Además, la SDE asociada a $\mathbb{P}$ tiene la misma forma que el proceso reverso de un modelo de difusión basado en SDEs (ver Figura 1), mostrando que los modelos de difusión son un caso particular del problema del puente de Schrödinger.