



UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA MATEMÁTICA

TRANSPORTE ÓPTIMO Y PUENTES DE SCHRÖDINGER COMO GENERALIZACIÓN DE LOS MODELOS DE DIFUSIÓN

Tesis para optar al grado de Magister en Ciencia de Datos
Memoria para optar al título de Ingeniero Civil Matemático

Fernando Esteban Fêtis Riquelme

Profesor guía Felipe Tobar Henríquez

Profesor co-guía Joaquín Fontbona Torres

Comisión Jorge F. Silva

Santiago de Chile

2024

Resumen

En los últimos años, los modelos de difusión han emergido como una poderosa clase de modelos generativos, alcanzando el estado del arte en diversos dominios, particularmente en la generación de imágenes. Estos modelos operan corrompiendo gradualmente una muestra proveniente de una distribución desconocida p_{data} hasta llegar a otra distribución p_{prior} de la cual es fácil generar muestras. Para revertir este proceso y generar nuevas muestras desde p_{data} , se entrena una red neuronal que aprenda a invertir la dinámica del ruido. Así, una muestra generada desde p_{prior} puede transformarse en una muestra válida de p_{data} siguiendo el proceso inverso aprendido.

A pesar de su impresionante rendimiento y su capacidad para generar datos de alta calidad, los modelos de difusión presentan limitaciones importantes. Entre estas, destacan la incapacidad de transformar directamente la distribución inicial p_{data} en otra distribución arbitraria y las restricciones asociadas al entrenamiento en escenarios donde los datos no están emparejados. Estas limitaciones han impulsado la búsqueda de alternativas metodológicas, entre las cuales el problema del puente de Schrödinger ha surgido como una solución prometedora.

El problema del puente de Schrödinger, definido en términos del transporte óptimo regularizado, consiste en encontrar un proceso estocástico $P = (P_t)_{t \in [0,1]}$ que, manteniendo cierta proximidad a un proceso de referencia $R = (R_t)_{t \in [0,1]}$ en el sentido de la entropía relativa, cumpla con tener distribuciones marginales predefinidas μ y ν en los tiempos $t = 0$ y $t = 1$, respectivamente. Este enfoque permite superar las limitaciones de los modelos de difusión al admitir transformaciones entre distribuciones arbitrarias y al poder trabajar con datos no emparejados.

Aunque el puente de Schrödinger aborda estas limitaciones de forma efectiva, la literatura existente tiende a presentar el problema desde una perspectiva matemática abstracta, utilizando formalismos técnicos que dificultan su comprensión y adopción en la comunidad de aprendizaje automático. Por ello, el objetivo de esta tesis es cerrar esta brecha mediante un tratamiento exhaustivo y accesible que conecte de manera natural los modelos de difusión con el problema del puente de Schrödinger, utilizando la teoría del transporte óptimo como un puente conceptual y metodológico.

Esta tesis está organizada en tres capítulos principales. En el Capítulo 1, se realiza una exploración detallada de los modelos de difusión, enfatizando sus fundamentos teóricos, implementaciones prácticas y limitaciones intrínsecas. Este capítulo establece las bases necesarias para comprender los desafíos asociados con estos modelos y justifica la necesidad de enfoques alternativos.

El Capítulo 2 introduce el problema del transporte óptimo, que busca transformar una distribución de probabilidad en otra minimizando un funcional de costo. Este capítulo presenta conceptos clave como la distancia de Wasserstein, sus propiedades geométricas y su relevancia en la interpolación entre distribuciones. Estas herramientas son esenciales para la comprensión y generalización del problema del puente de Schrödinger.

Finalmente, en el Capítulo 3, se estudia una versión regularizada del problema de transporte óptimo, la cual resulta ser equivalente al problema del puente de Schrödinger en su formulación estática. Esta equivalencia permite heredar la teoría y las técnicas del transporte óptimo al análisis y resolución del puente de Schrödinger, facilitando el diseño de métodos computacionalmente eficientes para problemas generativos complejos.

Además, esta tesis incluye un apéndice con definiciones y resultados técnicos utilizados a lo largo del documento. Aunque el marco de trabajo adoptado es \mathbb{R}^d o conjuntos finitos, las definiciones y resultados presentados son fácilmente extensibles a marcos más generales como espacios topológicos polacos. Las demostraciones incluidas tienen un enfoque instructivo, priorizando la claridad conceptual sobre el rigor formal, y muchas de ellas son desarrolladas inicialmente en un contexto discreto para luego generalizarse sin demostración al caso continuo.

Por último, todos los materiales asociados a esta tesis —manuscrito, simulaciones, póster y presentación— están disponibles en el repositorio público de GitHub [fernando-fetis/mds-thesis](https://github.com/fernando-fetis/mds-thesis), con el objetivo de facilitar la reproducibilidad y la extensión de los resultados aquí presentados.

A mi mamá, a mi familia, a Anto y a nuestros gatos.

Índice general

Introducción	1
1. Modelos de difusión	3
1.1. Modelos generativos neuronales	3
1.2. Autoencoders variacionales	8
1.3. Modelos de difusión a tiempo discreto	16
1.4. Generalización a tiempo continuo	37
2. Transporte óptimo entre distribuciones	57
2.1. Problema de Monge	59
2.2. Relajación de Kantorovich	62
2.3. Formulación dinámica	81
3. Regularización y problema de Schrödinger	101
3.1. Transporte óptimo entrópico	101
3.2. Problema de Schrödinger estático	114
3.3. Formulación dinámica	124
Conclusión	135
Apéndice	139
A.1. Medidas de probabilidad	139
A.2. Procesos de Íto	144
Bibliografía	156

Índice de figuras

1.1. Dinámica de entrenamiento de una GAN. La curva negra segmentada representa a p_{data} , la curva verde representa p_g y la línea azul segmentada representa la distribución aprendida por el discriminador. En (a) se tiene el estado actual de la GAN luego de algunas iteraciones donde la convergencia del discriminador aún no es alcanzada. En (b) se observa el estado de la GAN luego de entrenar el clasificador. En (c) se ve el el estado de la GAN luego de optimizar el generador. En (d) se observa la convergencia, donde p_g es indistinguible de p_{data} y el discriminador sigue una distribución de Bernoulli (máxima entropía). Imagen obtenida desde [Goo+14].	5
1.2. Imágenes generadas por una GAN entrenada durante 50 épocas con una red fully-connected sobre el dataset MNIST (izquierda) y FashionMNIST (derecha). La implementación de este modelo se encuentra en el archivo <code>gan_images.ipynb</code> . En el archivo <code>gan.ipynb</code> se puede encontrar una implementación para datos bidimensionales.	6
1.3. (Izquierda) muestras generadas por un VAE para una distribución p_{data} bidimensional. (Derecha) interpolación en el espacio latente del VAE, lo cual es posible debido a que no hay colapso de la posterior. La implementación de este modelo se encuentra en el archivo <code>vae_1d.ipynb</code>	11
1.4. Representación gráfica de un autoencoder variacional para MNIST. Imagen obtenida desde [Haf18].	13
1.5. Muestras generadas por un autoencoder variacional condicional sobre el dataset MNIST. La implementación de este modelo se encuentra en el archivo <code>conditional_vae.ipynb</code>	14
1.6. (Izquierda) reconstrucción de imágenes realizada por un VAE sobre el dataset CIFAR-10 utilizando una red convolucional, donde las primeras 4 filas corresponden a las imágenes originales y las últimas 4 filas corresponden a la reconstrucción. (Derecha) generación incondicional de imágenes a partir del VAE. La implementación de este modelo se encuentra en el archivo <code>vae.ipynb</code>	16
1.7. Modelo gráfico para un VAE estándar (izquierda) y para un HVAE con dos niveles de jerarquía (derecha). Imágenes obtenidas desde [Tur21].	17
1.8. Modelo gráfico del proceso de denoising de un DDPM. El proceso comienza con una muestra $x_T \sim p_{\text{prior}}(x_T)$ y continúa con las transiciones de Markov dadas en (1.3.2) utilizando el modelo p_θ ya entrenado. La flecha reversa indica el proceso de inyección de ruido realizado durante el entrenamiento. Imagen obtenida desde [HJA20].	18
1.9. (Arriba) muestras del proceso forward dado por la Proposición 1.10 para $T = 5$. (Abajo) muestras del proceso reverso condicional dado por la Proposición 1.11. La implementación de estas distribuciones se encuentra en el archivo <code>ddpm.ipynb</code>	20
1.10. Proceso de generación de muestras para un modelo entrenado en CIFAR-10. Cada fila es una muestra distinta y cada columna muestra la predicción actual en un tiempo t . Imagen obtenida desde [HJA20].	27

1.11. (Izquierda) batch de entrenamiento para un modelo de difusión. (Derecha) muestras generadas por el modelo de difusión. La implementación de esta técnica se encuentra en el archivo <code>ddpm.ipynb</code>	27
1.12. Arquitectura U-Net original para segmentación de imágenes. Imagen obtenida desde [RFB15].	28
1.13. Arquitectura U-Net usada para modelos de difusión. El bloque gris representa el input (imagen RGB) mientras que los bloques rosados son bloques convolucionales. Los bloques celestes y verdes son bloques de <i>downsampling</i> y <i>upsampling</i> respectivamente. Los bloques naranjas son bloques de atención. Por último, las flechas grises son las conexiones residuales. Imagen obtenida desde [Erd23].	29
1.14. (arriba) Proceso de difusión asociado a un scheduler lineal. (abajo) Proceso de difusión asociado a un scheduler coseno. Imagen obtenida desde [ND21].	31
1.15. Desviación de la distribución aprendida por el modelo incondicional $p_\theta(x)$ según la escala de <i>guidance</i> (aquí denotado por s). El modelo discriminativo $p_\theta(y x)$ le da un alto valor a puntos cercanos a la cruz marcada en negro. La implementación de esta técnica se encuentra en el archivo <code>ddpm.ipynb</code>	34
1.16. Trade-off entre calidad y variedad que se observa al cambiar el factor de escala del gradiente. Por un lado, al aumentar el factor aumenta el FID, IS y la precisión (indicadores asociados a la calidad), mientras que disminuye el recall (indicador asociado a la diversidad). Imagen obtenida desde [Nic+22].	35
1.17. Visualización del trade-off entre calidad y variedad al usar guidance. (Izquierda) modelo GLIDE incondicional. (Derecha) modelo GLIDE con classifier-free guidance, escala $\gamma = 3,0$. Imagen obtenida desde [Nic+22].	36
1.18. Arquitectura del modelo <i>Stable Diffusion</i> . El bloque rosado corresponde al VAE mientras que el bloque verde corresponde al modelo de difusión en el espacio latente. Para condicionar la generación, un embedding aprendido τ_θ es injectado a la U-Net mediante un mecanismo de <i>cross-attention</i> . Imagen obtenida desde [Rom+22].	37
1.19. Mecanismo de condicionamiento en el modelo <i>Stable Diffusion</i> . La matriz τ_θ es obtenida a partir de y . Un mecanismo de atención cruzada entre las distintas capas de la red y τ_θ permite introducir la información de y en la red. Imagen obtenida desde [Sam22].	37
1.20. (izquierda) generación condicionada a un mapa semántico. (derecha) eliminación de objetos mediante <i>inpainting</i> . Imagen obtenida desde [Rom+22].	38
1.21. Campo vectorial de la función de score real (izquierda) y de la función de score aprendida (derecha) para una mixtura gaussiana. El mapa de calor está asociado a la densidad de la mixtura y los rectángulos muestran las zonas donde la predicción del score es cercana a la real. Se observa que la predicción solo es precisa alrededor de las modas de la mixtura.	42
1.22. Muestras generadas a partir de una mixtura gaussiana utilizando la dinámica de Langevin dada en el Algoritmo 5. El mapa de calor indica la densidad de la mixtura en cada punto del plano y las curvas muestran la trayectoria seguida por el proceso de generación $(x_t)_{t=1}^T$. Se observa que la dinámica de Langevin no es capaz de respetar los priors de cada componente de la mixtura, donde se esperaría que la componente con mayor relevancia (con prior de clase 0,6) tenga una mayor cantidad de muestras generadas. Esta simulación se encuentra en el archivo <code>langevin.ipynb</code>	43
1.23. (izquierda) muestras de entrenamiento generadas a partir de la mixtura de la Figura 1.21. (centro) muestras generadas mediante DSM con la dinámica de Langevin usual. (derecha) muestras generadas con la dinámica de Langevin propuesta por [SE20].	44

1.24. SDEs de los procesos de difusión y <i>denoising</i> para el modelo de difusión basado en una SDE, donde el proceso reverso depende únicamente del score a lo largo del proceso forward. Imagen obtenida desde [Son+21b].	49
1.25. Ilustración de los procesos forward y backward en un modelo de difusión. El proceso forward comienza con una distribución bimodal $p_0 = p_{\text{data}}$ y termina en una distribución gaussiana $p_T = p_{\text{prior}}$. El proceso backward comienza desde la distribución prior y termina en la distribución de los datos. Las curvas erráticas muestran trayectorias del proceso estocástico para cuatro condiciones iniciales obtenidas desde $x_0 \sim p_{\text{data}}(x_0)$. En el proceso forward, las curvas blancas muestran la evolución de la <i>probability flow ODE</i> al comenzar desde las muestras x_0 obtenidas. En el proceso backward las curvas blancas muestran como el proceso determinista regresa a las muestras x_0 originales al comenzar desde x_T . Imagen obtenida desde [Son+21b].	51
1.26. Transformación de una distribución de imágenes en otra utilizando una CycleGan. Imagen obtenida desde [Zhu+20].	53
1.27. Efecto de simular el proceso de difusión por un tiempo demasiado corto. Al no aproximar bien la distribución p_{prior} en x_T , el proceso reverso no podrá generar muestras coherentes con la distribución p_{data} . Imagen obtenida desde [Bor+23b].	54
1.28. Efectos de inyectar un mismo ruido sobre imágenes de distinta resolución. Se observa que para imágenes más grandes, es necesario aumentar el nivel de ruido. Imagen obtenida desde [Che23].	54
2.1. Problema del transporte de tierra, donde el montículo rojo (con densidad de masa ρ_1) debe ser trasladado al montículo azul (con densidad de masa ρ_2). Para un mapa de transporte T , un volumen $A \subset \mathcal{X}$ es transportado hacia el volumen $T(A) = \{T(x) : x \in A\} \subset \mathcal{Y}$. Inversamente, un volumen $B \subset \mathcal{Y}$ es formado por la tierra proveniente de $T^{-1}(B) = \{x \in A : T(x) \in B\} \subset \mathcal{X}$. Notar que la masa de un volumen $A \subset \mathcal{X}$ se puede calcular mediante $m_1(A) = \int_A \rho_1(x) dx$ (análogo para \mathcal{Y}). Esta figura se puede encontrar en el archivo <code>earth_mover.ipynb</code>	61
2.2. (Arriba) histogramas discretos (no normalizados) asociados a las medidas μ y ν respectivamente, donde $ \mathcal{X} = 5$ y $ \mathcal{Y} = 7$. (Abajo) solución (normalizada) para el problema de Kantorovich entre las dos distribuciones discretas. El código de esta simulación se encuentra en el archivo <code>kantorovich.ipynb</code>	64
2.3. Polítopo $\Pi_d(\mu, \nu)$ representado en el plano por un conjunto $U(a, b)$ de vectores de transporte, donde la matriz de costo C es representado por un vector M_{XY} . El costo de transporte $\langle C, P \rangle_F = M_{XY}^\top P$ aumenta al moverse en la dirección del vector de costo M_{XY} , por lo que para minimizar el costo de transporte se busca el elemento $P \in U(a, b)$ que más pueda avanzar en el sentido contrario al vector de costo M_{XY} sin salirse del polítopo $U(a, b)$, lo cual se consigue en el vértice P^* . Notar que en este caso el minimizador del problema de transporte es único. En cambio, si el vector de costo M_{XY} tuviese la inclinación precisa para ser ortogonal a alguna de las facetas que define el polítopo, entonces podrían existir infinitas soluciones. Imagen adaptada desde [Cut+17].	65
2.4. Plan de transporte determinista inducido por un mapa de Monge $T^* : \mathcal{X} \rightarrow \mathcal{Y}$ definido como $T^*(x_5) = T^*(x_6) = y_1, T^*(x_2) = y_2, T^*(x_1) = T^*(x_4) = y_3, T^*(x_3) = y_4$. Notar que la medida producto $\pi^{T^*} \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})$ asociada a este plan de transporte puede verse como la medida push-forward inducida por μ a través del mapa $x \in \mathcal{X} \mapsto (x, T(x)) \in \mathcal{X} \times \mathcal{Y}$. Esta observación será importante en la formulación continua al enunciar el Teorema 2.2. La imagen se encuentra en el archivo <code>deterministic_plan.ipynb</code>	66
2.5. Ejemplos de couplings entre las medidas de probabilidad α y β cuando se considera un marco discreto (izquierda), semidiscreto (centro) o continuo (derecha). Imagen obtenida desde [PC20].	68

2.6. (Arriba) gráficos de funciones de densidad asociadas a las medidas μ y ν respectivamente. (Abajo) solución para el problema de Kantorovich entre ambas medidas. Se observa como toda la masa de π^* se concentra en una curva, la cual corresponde al gráfico del mapa de Monge para este mismo problema (ver Teorema 2.2). El código de esta simulación se encuentra en el archivo <code>kantorovich.ipynb</code>	69
2.7. Plan de Kantorovich para distintos pares de medidas α y β . (Arriba) las flechas indican cómo se debe repartir la masa de α en β para el transporte óptimo en el sentido de Kantorovich. En el sentido de Monge, solo el 1º caso tiene solución ya que los otros casos requieren división de masa. (Abajo) coupling óptimo del problema. Imagen obtenida desde [PC20].	70
2.8. Solución para el problema de Kantorovich entre dos medidas α y β . (Izquierda) coupling óptimo entre dos medidas continuas, donde la intensidad del color negro en la curva indica la densidad de masa del <i>coupling</i> óptimo π en cada punto del plano $\mathcal{X} \times \mathcal{Y}$. Notar que, en este caso, el plan de Kantorovich está soportado precisamente en el gráfico del mapa de Monge. (Derecha) plan de Kantorovich π entre dos medidas discretas, donde el radio de cada círculo es proporcional a la masa asignada a cada átomo de $\mathcal{X} \times \mathcal{Y}$. Imagen obtenida desde [PC20].	70
2.9. La función f no es diferenciable en $x = 1$, por lo que su conjunto de subgradientes no será un único punto. Las rectas g_i , $i \in \{1, 2, 3, 4\}$ son tangentes a f en $(1, 1)$, por lo que sus pendientes corresponden a subgradientes de f . Esta imagen se encuentra en el archivo <code>subdifferential.ipynb</code>	75
2.10. Interpolación entre dos curvas gaussianas utilizando la solución cerrada del Teorema 2.5. El código de esta simulación se encuentra en el archivo <code>distr_interpolation.ipynb</code>	79
2.11. Interpolación en el espacio de Wasserstein (izquierda) y en el espacio de las muestras (derecha). Se observa interpolando en el espacio de Wasserstein se obtienen geodésicas más naturales que al hacerlo en el espacio observable. Imagen obtenida desde [Kol+17].	84
2.12. Interpolación entre dos mezclas gaussianas. En este caso, el mapa de transporte no tiene solución en forma cerrada y debe ser encontrado de forma numérica. El código se encuentra en el archivo <code>distr_interpolation.ipynb</code>	85
2.13. Interpolación baricéntrica entre 4 figuras en el plano (una figura en cada esquina). Cada imagen representa un vector de peso $\lambda \in \Sigma_4$ diferente, el cual indica el nivel de importancia de cada figura en la interpolación. Imagen obtenida desde [PC20].	86
2.14. Imágenes generadas por una Wasserstein GAN en el dataset MNIST. La calidad de las imágenes es inferior a las generadas en el Capítulo 1 debido a que se entrenó el modelo neuronal menos tiempo. En el archivo <code>wgan.ipynb</code> hay una implementación minimal de este modelo generativo.	91
2.15. campo vectorial \vec{F} en \mathbb{R}^3 atravesando una superficie cerrada $\partial\Omega \subset \mathbb{R}^3$. El flujo total que pasa a través de $\partial\Omega$ (definido por la integral de superficie) es precisamente la cantidad de fluido se está <i>creando o perdiendo</i> en cada punto del espacio dentro de Ω , lo cual viene dado por el operador divergencia. De esta forma, sumando todas estas divergencias infinitesimales se obtiene el flujo neto a través de $\partial\Omega$. Esta imagen se encuentra en el archivo <code>surface_integral.ipynb</code>	94
2.16. Simulación del problema de Benamou-Brenier para dos distribuciones. Se observa la interpolación entre ambas distribuciones para distintos tiempos. El código se encuentra en el archivo <code>benamou_brenier.ipynb</code>	95
2.17. Trayectorias seguidas por las muestras de dos distribuciones iniciales $x \sim \mu$ (una en cada gráfico) cuando la distribución de llegada $E_\mu(x)$ es una variable aleatoria gaussiana estándar. Imagen obtenida desde [Khr+22].	98
3.1. Inestabilidad del mapa de Kantorovich bajo pequeñas perturbaciones en los parámetros del problema. En este caso, todos los elementos de \mathcal{X} e \mathcal{Y} fueron perturbados con un ruido gaussiano de baja varianza. La imagen se encuentra en el archivo <code>ot_instability.ipynb</code>	102

3.2. Evolución de la solución óptima $P_\epsilon \in \Pi_d(\mu, \nu)$ del problema regularizado (3.1.6) para distintos valores de ϵ . Dado que el problema de Kantorovich discreto no regularizado ($\epsilon = 0$) es un problema de programación lineal, el argumento minimizante se encontrará siempre en un vértice del polítopo $\Pi_d(\mu, \nu)$. Al aumentar el valor de ϵ , la solución óptima para el problema entrópico tiende a alejarse de los vértices y converger hacia un centro de alta entropía. Más aún, se puede demostrar que la matriz de transporte óptima para este problema regularizado tiene todas sus coordenadas positivas (dado que no tiene ninguna restricción de desigualdad activa, lo cual ocurre en los vértices de $\Pi_d(\mu, \nu)$), por lo que todos los puntos de \mathcal{X} transportan masa a todos los elementos de \mathcal{Y} . Imagen obtenida desde [PC20].	105
3.3. (Izquierda) plan de Kantorovich entre dos medidas μ y ν con vectores de probabilidad α y β respectivamente. (Derecha) evolución de la solución entrópica $P_\epsilon \in \Pi_d(\mu, \nu)$ para distintos valores de ϵ . Solo se muestran los arcos para valores $(P_\epsilon)_{ij} \in [0, 1]$ mayores a un cierto umbral. Imagen obtenida desde [PC20].	108
3.4. Evolución de la solución $\pi_\epsilon \in \Pi(\mu, \nu)$ para distintos valores de $\epsilon > 0$ en el problema entrópico con costo cuadrático entre dos medidas de probabilidad continuas $\mu, \nu \in \mathcal{M}_+^1(\mathbb{R})$. Imagen obtenida desde [PC20].	111
3.5. Soluciones para el problema entrópico utilizando el Algoritmo 8. El grusor de los arcos es proporcional a la cantidad de masa transferida según el plan de transporte. Se observa como el plan de transporte entrópico se va volviendo determinista a medida que disminuye ϵ . Esta simulación se encuentra en el archivo <code>sinkhorn.ipynb</code>	121
3.6. Distancia de Hilbert entre dos rayos $u, u' \in \mathbb{R}_{++}/\sim$, donde la distancia depende únicamente del ángulo entre ambos rayos. Imagen obtenida desde [PC20].	122
3.7. Contracción del primer cuadrante en \mathbb{R}^2 provocada por la multiplicación consecutiva de una matriz positiva K . Se observa que a medida que se va multiplicando por K , los rayos están cada vez más cerca, convergiendo a un único rayo en el cuadrante positivo. Imagen obtenida desde [PC20].	122
3.8. Iteraciones del algoritmo de Sinkhorn entre dos distribuciones discretas. Se observa que a medida se disminuye ϵ , el error disminuye más lentamente, mientras que el plan de transporte se vuelve más determinista. La simulación se encuentra en el archivo <code>sinkhorn.ipynb</code>	125
3.9. Iteraciones del algoritmo de Sinkhorn entre dos distribuciones continuas. Se observa que la proyección baricéntrica del plan de transporte óptimo converge al mapa de Monge del problema cuando $\epsilon \rightarrow 0$. Además, se observa que para $\epsilon = 1$ la solución (similar a la medida producto) se encuentra en muy pocas iteraciones. La simulación se encuentra en el archivo <code>sinkhorn.ipynb</code>	126
3.10. Puente browniano entre los puntos $x_0 = 2$ y $x_1 = 5$. Esta simulación se encuentra en el archivo <code>sdes.ipynb</code>	128
3.11. Puentes de Schrödinger para distintos valores de ϵ . Se observa como disminuye la aleatoriedad del plan de transporte estocástico cuando disminuye ϵ . El código que genera estas imágenes se encuentra en el archivo <code>sbp.ipynb</code>	129
3.12. Imágenes generadas utilizando el enfoque del Teorema 3.5 sobre los datasets MNIST, CelebA y CIFAR-10. Imagen obtenida desde [CLT23].	133
A.1. Simulación mediante el algoritmo de Euler-Maruyama (ver Algoritmo 7) de 5 trayectorias para el proceso de Ornstein–Uhlenbeck $dx_t = 2(1 - x_t)dt + \frac{1}{2}dW_t$ en el intervalo temporal $[0, 5]$ y con condición inicial $x_0 = 10$. Esta simulación se encuentra en el archivo <code>sdes.ipynb</code>	150

Introducción

En la última década, los *modelos generativos* han cobrado gran relevancia en el campo de la inteligencia artificial, destacándose por su capacidad para modelar datos complejos y sintetizar ejemplos realistas en diversas aplicaciones. Estas capacidades han permitido avances significativos en áreas como el entretenimiento, la biomedicina, la simulación científica y el procesamiento de lenguaje natural. Dentro de este panorama, los *modelos de difusión* han surgido como una familia particularmente prometedora debido a su robustez y a la alta calidad de los datos generados.

Los modelos de difusión operan distorsionando progresivamente datos provenientes de una distribución inicial, hasta alcanzar una distribución de ruido gaussiano, y luego aprenden a revertir este proceso para generar nuevas muestras. Esta dinámica está inspirada en procesos físicos, como la difusión de partículas, y ha demostrado ser altamente efectiva en tareas como la generación de imágenes de alta fidelidad. Sin embargo, a pesar de su impresionante rendimiento, estos modelos presentan ciertas limitaciones significativas. Entre ellas, destacan el alto costo computacional asociado a las simulaciones iterativas y la dificultad para ajustar adecuadamente las transiciones reversas, especialmente en distribuciones de alta dimensión.

Estas limitaciones han motivado la exploración de enfoques alternativos, entre los cuales el *problema del puente de Schrödinger* ha ganado creciente atención. Este enfoque, basado en la teoría del transporte óptimo regularizado, proporciona una solución robusta para la interpolación estocástica entre distribuciones, superando barreras inherentes de los modelos de difusión. Al formularse como un problema de optimización con regularización entrópica, el puente de Schrödinger no solo permite transformar distribuciones de manera eficiente, sino que también extiende las capacidades generativas a contextos más amplios, como datos no emparejados y trayectorias estocásticas personalizadas.

El presente trabajo se sitúa en la intersección de estas tres áreas: los modelos de difusión, el transporte óptimo y el problema del puente de Schrödinger, con el objetivo de proporcionar un marco unificado y autocontenido para entender y conectar estas metodologías. A través de un análisis detallado y simulaciones prácticas, esta tesis busca cerrar la brecha conceptual y técnica que actualmente existe entre estas áreas, facilitando su aplicación en problemas generativos modernos.

Contribuciones

Las principales contribuciones de esta tesis son las siguientes:

1. Desarrollo exhaustivo de los aspectos teóricos y prácticos de los modelos de difusión, el transporte óptimo y el problema del puente de Schrödinger. Se presenta una integración coherente de estos conceptos, conectando los temas de forma natural y facilitando una comprensión unificada.
2. Formulación detallada del transporte óptimo para justificar de manera sólida el uso de los puentes de Schrödinger como una extensión natural de los modelos de difusión. Esta formulación permite una

transición fluida desde los modelos de difusión hacia los métodos de interpolación estocástica entre distribuciones arbitrarias.

3. Revisión crítica de la literatura reciente, incorporando resultados y técnicas modernas tanto en el desarrollo teórico como en las implementaciones numéricas de los modelos generativos.
4. Desarrollo de un conjunto extenso de simulaciones numéricas y arquitecturas neuronales, incluyendo:
 - **Modelos generativos tradicionales:** se implementó un modelo de autoencoder variacional y una red generativa adversarial para introducir los temas de generación neuronal y modelos de variable latente.
 - **Modelos de difusión:** en el estudio de este tipo de modelos, se implementaron los modelos denoising diffusion probabilistic models (DDPM), score matching (SM y DSM), dinámica de Langevin y técnicas de guidance. Además, se implementaron las arquitecturas neuronales U-Net y DiT, las cuales constituyen el estado del arte en redes neuronales para modelos de difusión.
 - **Transporte óptimo y puentes de Schrödinger:** para el estudio de este problema, se simuló una solución del problema de Kantorovich (tanto en el caso discreto como en el continuo) y se realizó una simulación de la formulación de Benamou-Brenier. Además, se implementó una interpolación de McCann, el algoritmo de Sinkhorn y una red WGAN.

Es importante destacar que uno de los objetivos principales de esta tesis es proporcionar un estudio autocontenido del problema del puente de Schrödinger, el cual presenta múltiples formulaciones equivalentes, muchas de ellas utilizando resultados técnicos avanzados. Para facilitar la comprensión y el desarrollo de estas formulaciones, se ha incluido un apéndice que cubre de manera detallada los resultados necesarios de cálculo estocástico y teoría de la medida, proporcionando un soporte técnico robusto para los desarrollos teóricos presentados.

Con este marco integral, esta tesis no solo aporta al entendimiento teórico de los modelos generativos modernos, sino que también habilita su aplicación práctica en un amplio espectro de problemas, desde la síntesis de datos realistas hasta la interpolación de distribuciones complejas.

Capítulo 1

Modelos de difusión

En este capítulo se definirán y estudiaron los modelos generativos de difusión, los cuales constituyen el estado del arte en la generación de imágenes. Estos modelos distorsionan progresivamente los datos de entrenamiento provenientes desde una distribución de probabilidad p_{data} hasta llegar a una distribución final p_{prior} . Resolviendo un problema de optimización para aprender las transiciones del proceso reverso, es posible generar nuevas muestras desde p_{data} a partir de una muestra generada desde p_{prior} mediante la simulación del proceso backward aprendido.

Dado que los modelos de difusión han sido investigados principalmente desde un punto de vista práctico, en este primer capítulo se asumirá, como es usual en el aprendizaje automático, que la medida de probabilidad que genera los datos de entrenamiento posee una función de densidad p_{data}^1 . Esta suposición se levantará en los capítulos posteriores, donde se trabajará directamente sobre medidas de probabilidad.

1.1. Modelos generativos neuronales

A modo de introducción, se explorarán dos enfoques fundamentales en el campo de los modelos generativos neuronales: las redes generativas adversarias (GANs) y los modelos basados en energía (EBMs). Estos modelos representan diferentes paradigmas para abordar el desafío de generar muestras que se asemejen a una distribución empírica que representa observaciones disponibles. Por un lado, las GANs han revolucionado el campo de la generación de imágenes mediante un enfoque de competencia entre dos redes neuronales, mientras que los EBMs ofrecen una perspectiva alternativa, modelando explícitamente la densidad de probabilidad de los datos a través de una función de energía. Como se verá a lo largo de la sección, ambos enfoques tienen sus propias fortalezas y limitaciones, y su estudio permitirá apreciar cómo los modelos de difusión abordan algunas de las limitaciones de estos enfoques más clásicos.

Por otra parte, en la Subsección 1.4.1 se verá una estrecha conexión entre los modelos de difusión y los EBMs.

1.1.1. Redes generativas adversarias

Previo al uso de modelos de difusión, las redes generativas adversarias² (GANs) constituyeron durante varios años el estado del arte de los modelos generativos para imágenes. Este tipo de modelos fueron propuestos

¹Es decir, se asumirá que la medida que genera los datos es absolutamente continua con respecto a la medida de Lebesgue en \mathbb{R}^d (ver Definición A.3), cuya consecuencia más importante es, de acuerdo al teorema de Radon-Nikodym (ver Teorema A.1), que existe una función de densidad y esta es única.

²También se les conoce como redes generativas adversativas o antagónicas.

por investigadores de la Universidad de Montreal en [Goo+14], donde los autores propusieron entrenar dos modelos independientes representados por redes neuronales. Por un lado, un modelo generador G busca aprender a generar muestras artificiales a partir de una cierta distribución p_{data} , engañando a otro modelo discriminador D que busca discernir si una muestra de entrada corresponde a una muestra proveniente de p_{data} o a una muestra artificial generada por G .

Formulación y convergencia

La descripción anterior afirma que el modelo generativo que propone una GAN consta de dos partes:

- G_θ es un modelo generativo de variable latente que transforma una distribución latente $z \sim p_z(z)$ en otra distribución $g \sim p_g(g)$ mediante $z \mapsto G_\theta(z)$. El objetivo de este modelo es que $p_g \approx p_{\text{data}}$, por lo que puede ser comparado con el decoder de un autoencoder variacional (ver Sección 1.2).
- D_ϕ es un modelo discriminativo probabilístico (clasificador) tal que, para una cierta entrada x , $D_\phi(x)$ indica la probabilidad de que x provenga de la distribución real de los datos. Es decir, si una muestra x proviene de p_{data} entonces $D_\phi(x) = 1$ mientras que si proviene de p_g , entonces $D_\phi(x) = 0$.

Con esto, se tiene una única función objetivo para ambos modelos, la cual está asociada al rendimiento del modelo discriminador. Por un lado, D busca maximizar su capacidad de predicción, mientras que G busca minimizarla:

$$\min_{\theta} \max_{\phi} C(G_\theta, D_\phi) := \mathbb{E}_{x \sim p_{\text{data}}(x)} [\log D_\phi(x)] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D_\phi(G_\theta(z)))] . \quad (1.1.1)$$

Los modelos G_θ y D_ϕ se entrena de forma alternada, comenzando por el discriminador. Además, por motivos teóricos que se verán a continuación, el discriminador se suele entrenar $K \geq 1$ veces por cada iteración del generador (es decir, por cada iteración del descenso del gradiente sobre el generador, el discriminador realiza K iteraciones de descenso del gradiente). El bucle de entrenamiento para una GAN genérica puede ser encontrado en el Algoritmo 1, donde las esperanzas son estimadas por aproximaciones de por Monte Carlo. Una ilustración de su dinámica se puede encontrar en la Figura 1.1.

Algoritmo 1 Entrenamiento de una GAN

Require: número de pasos K para entrenar el discriminador, redes neuronales G_θ y D_ϕ .

- 1: **while** no hay convergencia **do**
- 2: **for** $k = 1$ to K **do**
- 3: Obtener batch de entrenamiento $\{x_1, \dots, x_m\}$ con $x_i \sim p_{\text{data}}(x_i)$, $\forall i \in \{1, \dots, m\}$.
- 4: Generar batch de latentes $\{z_1, \dots, z_m\}$ con $z_i \sim p_z(z_i)$, $\forall i \in \{1, \dots, m\}$.
- 5: Realizar un paso del algoritmo de gradiente ascendente para el discriminador mediante el gradiente

$$\nabla_{\phi} \frac{1}{m} \sum_{i=1}^m (\log D_\phi(x_i) + \log(1 - D_\phi(G_\theta(z_i))))$$

- 6: **end for**
- 7: Generar batch de latentes $\{z_1, \dots, z_m\}$ con $z_i \sim p_z(z_i)$, $\forall i \in \{1, \dots, m\}$.
- 8: Realizar un paso del algoritmo de gradiente descendente para el generador mediante el gradiente

$$\nabla_{\theta} \frac{1}{m} \sum_{i=1}^m (\log(1 - D_\phi(G_\theta(z_i))))$$

- 9: **end while**
-

En la Figura 1.2 se pueden observar imágenes generadas a partir de una GAN sobre los dataset MNIST y FashionMNIST utilizando una red fully-connected y el algoritmo de entrenamiento indicado en el Algoritmo 1.

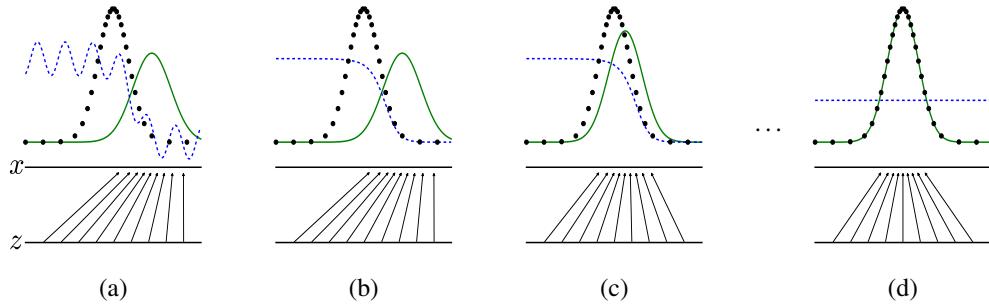


Figura 1.1: Dinámica de entrenamiento de una GAN. La curva negra segmentada representa a p_{data} , la curva verde representa p_g y la línea azul segmentada representa la distribución aprendida por el discriminador. En (a) se tiene el estado actual de la GAN luego de algunas iteraciones donde la convergencia del discriminador aún no es alcanzada. En (b) se observa el estado de la GAN luego de entrenar el clasificador. En (c) se ve el el estado de la GAN luego de optimizar el generador. En (d) se observa la convergencia, donde p_g es indistinguible de p_{data} y el discriminador sigue una distribución de Bernoulli (máxima entropía). Imagen obtenida desde [Goo+14].

Por otra parte, los autores de [Goo+14] muestran algunos resultados elementales acerca de la convergencia de las GANs. El primero de ellos afirma que, dado un generador fijo G , el discriminador óptimo tiene el comportamiento esperado:

Proposición 1.1 (discriminador óptimo). Para un generador G fijo, el discriminador óptimo D_G^* viene dado por:

$$D_G^*(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_g(x)}.$$

Por otra parte, asumiendo que siempre es posible obtener un discriminador óptimo D_G^* para un generador G , el siguiente resultado afirma que el generador óptimo para (1.1.1) se alcanza cuando $p_g = p_{\text{data}}$:

Proposición 1.2 (generador óptimo). Para una función objetivo $C(G) := C(G, D_G^*) = \max_\phi C(G, D_\phi)$, su mínimo global $G^* = \arg \min_\theta C(G_\theta)$ es alcanzado si y solo si $p_g = p_{\text{data}}$. En este caso, $C(G^*) = -\log 4$.

Estos resultados permiten concluir que al seguir el algoritmo dado en el Algoritmo 1, la distribución p_g asociada al generador converge a p_{data} :

Proposición 1.3 (convergencia del Algoritmo 1). Si los modelos G_θ y D_ϕ tienen suficiente capacidad y K es lo suficientemente grande como para permitir que D_ϕ sea óptimo en cada iteración, entonces p_g converge a p_{data} .

En la proposición anterior, es importante entrenar el discriminador hasta su convergencia, lo cual justifica la asimetría dada en el Algoritmo 1.

Generación condicional

El modelo descrito anteriormente es un modelo generativo que busca aprender la distribución no condicional p_{data} , por lo que no es posible controlar la generación. En [MO14] los autores proponen realizar un pequeño



Figura 1.2: Imágenes generadas por una GAN entrenada durante 50 épocas con una red fully-connected sobre el dataset MNIST (izquierdo) y FashionMNIST (derecha). La implementación de este modelo se encuentra en el archivo `gan_images.ipynb`. En el archivo `gan.ipynb` se puede encontrar una implementación para datos bidimensionales.

cambio a la arquitectura para permitir generar muestras a partir de una distribución condicional $p_{\text{data}}(x|y)$, donde y es una etiqueta o condición sobre los datos. Esto permite, por ejemplo, generar dígitos específicos con la GAN usada para la Figura 1.2.

Para conseguir esto, solo hace falta modificar la estructura de los modelos G_θ y D_ϕ para permitir una entrada adicional y . Con esto, la nueva función objetivo es

$$\min_{\theta} \max_{\phi} C(G_\theta, D_\phi) := \mathbb{E}_{(x,y) \sim p_{\text{data}}(x,y)} [\log D_\phi(x|y)] + \mathbb{E}_{z \sim p_z(z), y \sim p_{\text{data}}(y)} [\log(1 - D_\phi(G(z|y)))] .$$

Limitaciones de las GANs

Uno de los problemas más conocidos en las GANs es que poseen un entrenamiento altamente inestable, dependiendo fuertemente de la arquitectura, hiperparámetros y datos utilizados durante el entrenamiento. Con el fin de aminorar la inestabilidad provocada por la arquitectura, los autores de [RMC16] proponen algunas reglas generales para la construcción de una arquitectura para GANs, denominada DCGAN (deep convolutional GAN). En particular, concluyen lo siguiente:

- Las capas de global average pooling estabilizan el entrenamiento pero la convergencia se vuelve más lenta. Debido a esto, proponen sustituir las capas de pooling por capas convolucionales con stride en el encoder y por capas convolucionales transpuestas con stride fraccional en el generador.
- Usar batch normalization [IS15] tanto en el generador como en el discriminador. Si bien hoy en día esto es una práctica frecuente, no era obvio su uso en la fecha de publicación de este trabajo.
- Eliminar capas fully-connected en arquitecturas profundas.
- Usar ReLU en el generador (excepto en la última capa donde se suele usar tanh) y Leaky-ReLU en el discriminador.

Por otra parte, un problema intrínseco de este tipo de modelos es que tienden a concentrar su aprendizaje en las modas de las distribuciones, perdiendo la capacidad de generar muestras más diversas. Si bien este problema no es notorio en datasets de juguete como MNIST, sí se vuelve importante cuando se busca aprender distribuciones más complejas, donde es usual tener que entrenar una GAN para cada grupo de datos de la distribución. Como se verá en la Sección 1.3, los modelos de difusión no poseen este problema, pudiendo aprender distribuciones complejas y generar muestras de alta calidad.

1.1.2. Modelos basados en energía

Los modelos basados en energía (EBMs) constituyen otra familia importante de modelos generativos neuronales. A diferencia de las GANs, que aprenden implícitamente la distribución de los datos a través de una dinámica adversativa, los EBMs aprenden explícitamente una función de energía que caracteriza la distribución de probabilidad de los datos.

Un EBM busca aproximar una distribución de probabilidad p_{data} mediante otra distribución $p_{\theta}(x)$ parametrizada por θ de la siguiente forma:

$$p_{\theta}(x) = \frac{1}{Z(\theta)} \exp(-E_{\theta}(x)), \quad (1.1.2)$$

donde $E_{\theta} : \mathbb{R}^d \rightarrow \mathbb{R}$ se denomina *función de energía* y $Z(\theta) = \int_{\mathbb{R}^d} \exp(-E_{\theta}(x)) dx$ es la constante de normalización, también conocida como *función de partición*. El nombre de este tipo de modelos viene de su interpretación física, donde (1.1.2) corresponde a una parametrización de la distribución de Boltzmann, la cual será definida posteriormente en (1.2.5).

La función de energía E_{θ} es típicamente modelada por una red neuronal, lo que permite capturar relaciones complejas entre los datos y heredar sus propiedades de diferenciabilidad. Para el entrenamiento de este tipo de modelos, se busca ajustar los parámetros de $E_{\theta}(x)$ para que la distribución aprendida, $p_{\theta}(x)$, se aproxime a la verdadera distribución de los datos, p_{data} . Para esto, es usual minimizar la divergencia de Kullback-Leibler entre p_{data} y p_{θ} :

Definición 1.1 (divergencia de Kullback-Leibler, caso absolutamente continuo). Dadas dos densidades de probabilidad p y q en \mathbb{R}^d , se define la divergencia de Kullback-Leibler entre p y q como³:

$$D_{\text{KL}}(p \parallel q) = \mathbb{E}_{p(x)} \left[\log \frac{p(x)}{q(x)} \right] = \int_{\mathbb{R}^d} \log \left(\frac{p(x)}{q(x)} \right) p(x) dx,$$

donde se define, por continuidad, $0 \cdot \log(0) = 0$ en los casos que sea necesario.

Es importante recordar que a lo largo de este capítulo se asumirá siempre la existencia de densidades de probabilidad, por lo que la divergencia de Kullback-Leibler se puede definir mediante la expresión dada en la Definición 1.1. Sin embargo, en el Capítulo 2 y en el Capítulo 3 será necesario trabajar directamente sobre medidas de probabilidad, donde la existencia de tales densidades no está garantizada. Esto exigirá redefinir la divergencia de Kullback-Leibler en función de la derivada de Radon-Nikodym (ver Teorema A.1).

Con esta definición, el parámetro óptimo para un EBM es el que minimiza la siguiente divergencia:

$$\theta^* = \arg \min_{\theta} D_{\text{KL}}(p_{\text{true}} \parallel p_{\theta}).$$

³Notar que la integral de abajo solo está bien definida si $p(x) = 0$ cuando $q(x) = 0$, es decir, si $p \ll q$ (en el sentido de las medidas que inducen estas densidades). A diferencia de la continuidad absoluta asumida para la existencia de las densidades, esta restricción es propia de la divergencia de Kullback-Leibler.

Notar que esta es una cantidad conveniente de minimizar para un modelo de energía ya que al desarrollar esta expresión, se llega a que minimizar la divergencia de Kullback-Leibler es equivalente a maximizar la log-verosimilitud de los datos, lo cual se puede escribir directamente en función de la función de energía:

$$D_{KL}(p_{\text{true}} \| p_{\theta}) = \mathbb{E}_{x \sim p_{\text{true}}} \left[\log \frac{p_{\text{data}}(x)}{p_{\theta}(x)} \right] = \mathbb{E}_{x \sim p_{\text{true}}} [\log p_{\text{data}}(x)] + \mathbb{E}_{x \sim p_{\text{true}}} [-E_{\theta}(x)] + \log Z(\theta).$$

Sin embargo, el cálculo exacto de $\log Z(\theta)$ es intratable en la mayoría de los casos prácticos, ya que implica una integral sobre todo el espacio de datos. Si bien se han propuesto métodos para un entrenamiento aproximado de este modelo, en la Subsección 1.4.1 se verá una variante de esta familia de modelos donde se aprenderá directamente $\nabla_x \log p_{\theta}(x)$ en vez de la función de energía.

Por último, para la generación de muestras a partir de un modelo de energía entrenado, E_{θ} , es usual utilizar un algoritmo de sampling denominado *Langevin sampling*, el cual requiere computar $\nabla_x E_{\theta}$. Este algoritmo está detallado en el Algoritmo 5, donde es usado para generar muestras a partir de un modelo basado en *score*.

Debido a que el algoritmo de Langevin sampling consiste en simular una cadena de Markov, el tiempo de simulación es mucho más lento que en una GAN. Además, este tipo de modelos también sufre de inestabilidades durante el entrenamiento. En [SK21] estudian en detalle este tipo de modelos junto a diversas técnicas de entrenamiento y muestran su relación con los modelos de *score matching* estudiados en la Subsección 1.4.1.

Las limitaciones mencionadas han motivado la investigación de enfoques generativos alternativos como los autoencoders variacionales (VAEs), que abordan algunas de estas limitaciones ofreciendo un entrenamiento más estable, una generación de muestras más rápida y una forma explícita de aproximar la verosimilitud.

1.2. Autoencoders variacionales

Con el fin de introducir el estudio de los modelos de difusión, se comenzará estudiando una familia de modelos generativos denominada autoencoders variacionales, los cuales permitirán obtener una construcción natural de los modelos de difusión en la Sección 1.3. Además, hoy en día es usual trabajar con ambos modelos de manera conjunta, donde un autoencoder variacional actúa como un reductor de dimensionalidad para así poder trabajar con los modelos de difusión sobre un espacio latente (ver Subsección 1.3.3).

1.2.1. Modelos de variable latente

Un enfoque frecuente en formulación de modelos generativos es el uso de variables latentes. Este tipo de modelos propone que los datos observados a partir de $x \sim p_{\text{data}}$ provienen realmente⁴ de una distribución conjunta con densidad $p_{\text{data}}(x, z)$, donde la componente z es una variable aleatoria latente (i.e., no es observada) que influye directamente en la generación de las muestras observadas (componente x). Con esto, los modelos de variable latente proponen un modelo gráfico cuyas densidad se factoriza como

$$p_{\theta}(x, z) = p_{\theta}(z)p_{\theta}(x|z). \tag{1.2.1}$$

Además, es usual suponer que la dimensión de la variable latente z es estrictamente menor que la dimensión de los datos observados x . A lo largo de este capítulo, se considerará siempre que los datos observados viven en el espacio ambiente \mathbb{R}^d , mientras que las características ocultas estarán inmersas en el espacio \mathbb{R}^l , con $l \ll d$. Esta elección permite muchas veces transformar datos de alta dimensión a representaciones de menor

⁴Si bien esto no tiene por qué ser cierto, se puede asumir sin pérdida de generalidad, ya una posible independencia entre x y z vuelve prescindible a la variable latente z .

dimensión, pudiendo ganar eficiencia computacional sin perder una cantidad significativa de información. Esto último se suele justificar mediante el siguiente resultado, el cual afirma que es posible proyectar un conjunto de n puntos en \mathbb{R}^d a cualquier espacio de dimensión mayor o igual a $\mathcal{O}\left(\frac{\log n}{\epsilon^2}\right)$ de tal forma que la distancia euclíadiana original de los puntos no se distorsione más allá de un factor $1 \pm \epsilon$. Además, el mapa de proyección puede ser encontrado aleatoriamente en tiempo polinomial (en el sentido de Turing), aunque para efectos de los modelos generativos, este último resultado es omitido ya que se suele buscar un proyector mediante el entrenamiento de una red neuronal:

Teorema 1.1 (Johnson-Lindenstrauss). Sea $\epsilon \in (0, 1)$ y $\{x^k\}_{k=1}^n \subset \mathbb{R}^d$ un conjunto de puntos. Entonces, para todo $l \in \mathbb{N}$ con $l \geq \frac{24}{3\epsilon^2 - 2\epsilon^3} \log n$, existe un mapa $f : \mathbb{R}^d \rightarrow \mathbb{R}^l$ tal que

$$(1 - \epsilon) \|x_i - x_j\|^2 \leq \|f(x_i) - f(x_j)\|^2 \leq (1 + \epsilon) \|x_i - x_j\|^2, \quad \forall i, j \in \{1, \dots, n\}.$$

Además, existe un conjunto $\{x^k\}_{k=1}^n \subset \mathbb{R}^d$ tal que la cota para $l \in \mathbb{N}$ es alcanzada. En cualquier caso, el problema de encontrar f está en la clase de complejidad BPP⁵.

Una demostración de este resultado se puede encontrar en [DG03].

Criterio de máxima verosimilitud

Un criterio natural para elegir un mejor modelo dentro de una familia de modelos generativos es el criterio de máxima verosimilitud. En [ND21] afirman que hoy en día se cree que optimizar un modelo generativo maximizando la log-verosimilitud, fuerza al modelo a capturar todas las modas de la distribución de los datos de entrenamiento, lo cual no ocurre al entrenar una GAN.

Definición 1.2 (función de log-verosimilitud). Sea $\mathcal{V} = \{p_\theta\}_{\theta \in \Theta}$ una familia paramétrica de funciones de densidad de probabilidad y $\mathcal{D} = \{x^k\}_{k=1}^n \subset \mathbb{R}^d$ un conjunto de observaciones⁶ i.i.d. provenientes de alguna distribución desconocida p_{data} . La función de log-verosimilitud $l : \Theta \rightarrow \mathbb{R}$ asociada a \mathcal{V} y a \mathcal{D} se define como:

$$l_{\mathcal{V}}(\theta | \mathcal{D}) := \sum_{k=1}^n \log p_\theta(x^k).$$

Con esta definición, el criterio de máxima verosimilitud elige, como su nombre lo indica, al modelo generativo que maximiza la (log-)verosimilitud dentro de la familia de modelos:

$$\theta_{\text{MLE}} = \arg \max_{\theta \in \Theta} l_{\mathcal{V}}(\theta | \mathcal{D}).$$

Sin embargo, para utilizar el criterio de máxima verosimilitud es necesario poder evaluar la log-densidad marginal $\log p_\theta(x)$ en el conjunto de datos de entrenamiento (de aquí en adelante, a esta cantidad se le denominará *evidencia*). En un modelo de variable latente factorizado como $p_\theta(x, z) = p_\theta(z)p_\theta(x|z)$ en general no es posible dicha evaluación, en efecto:

$$p_\theta(x) = \int_{\mathbb{R}^l} p_\theta(x, z) dz = \frac{p_\theta(z)p_\theta(x|z)}{p_\theta(z|x)}. \quad (1.2.2)$$

⁵Recordar que esta es la clase de problemas decidibles en tiempo polinomial por máquinas de Turing probabilísticas con una probabilidad de error menor a $\frac{1}{3}$ (ver [AB09]).

⁶En este trabajo se utilizará la notación de superíndice para indexar un conjunto de muestras, reservando la notación de sub-índice para una familia de variables aleatorias.

La primera igualdad corresponde a la marginalización de x , mientras que la segunda igualdad se obtiene al aplicar la definición de probabilidad condicional, $p_\theta(x, z) = p_\theta(x)p_\theta(z|x)$. Notar que la primera igualdad no es tratable para modelos p_θ complejos ya que requiere evaluar varias veces el modelo para aproximar la integral numéricamente. La segunda igualdad no es computable dado que no se conoce la posterior $p_\theta(z|x)$ debido a que z es una variable oculta.

Cota inferior de la evidencia

Debido a que no es posible calcular explícitamente la evidencia de un modelo de variable latente, se utilizará un enfoque aproximado para su evaluación, el cual está basado en técnicas de inferencia variacional (ver, por ejemplo, [BKM17]). En este caso, la densidad marginal $p_\theta(x)$ será calculada utilizando la segunda igualdad en (1.2.2), donde la posterior $p_\theta(z|x)$ se aproximarán mediante un nuevo modelo paramétrico $q_\phi(z|x)$, el cual se puede interpretar como un modelo generativo condicional para la variable latente.

Si bien se podría sustituir directamente el nuevo modelo $q_\phi(z|x)$ en (1.2.2) y obtener $p_\theta(x) \approx \frac{p_\theta(z)p_\theta(x|z)}{q_\phi(z|x)}$, no es posible maximizar dicha cantidad de forma conjunta ya que, al no tener una restricción que asegure que $q_\phi(z|x)$ aproxime a $p_\theta(z|x)$, bastaría tomar $p_\theta(x, z) \rightarrow \infty$ y $q_\phi(z|x) \rightarrow 0^+$ para la maximización. Debido a esto, se utilizará, como es usual en el aprendizaje automático, la divergencia de Kullback-Leibler (ver Definición 1.1) para comparar la densidad $q_\phi(z|x)$ con la densidad que realmente busca estimar, $p_\theta(z|x)$:

$$\text{D}_{\text{KL}}(q_\phi(z|x) \| p_\theta(z|x)) = \mathbb{E}_{q_\phi(z|x)} \left[\log \frac{q_\phi(z|x)}{p_\theta(z|x)} \right] = \mathbb{E}_{q_\phi(z|x)} \left[\log \frac{q_\phi(z|x)}{p_\theta(x, z)} + \log p_\theta(x) \right],$$

de donde se obtiene una relación explícita entre el error de aproximación de $q_\phi(z|x)$ y la evidencia:

Proposición 1.4 (descomposición de la evidencia). Dado un modelo paramétrico auxiliar $q_\phi(z|x)$ para la distribución posterior $p_\theta(z|x)$, se tiene la siguiente descomposición para la evidencia:

$$\log p_\theta(x) = \text{D}_{\text{KL}}(q_\phi(z|x) \| p_\theta(z|x)) + \underbrace{\mathbb{E}_{q_\phi(z|x)} \left[\log \frac{p_\theta(x, z)}{q_\phi(z|x)} \right]}_{\text{ELBO}}. \quad (1.2.3)$$

Si bien el primer término de la igualdad no es tratable ya que requiere conocer la posterior verdadera, $p_\theta(z|x)$, la desigualdad de Gibbs⁷ permite afirmar que el término definido como ELBO⁸ (*evidence lower bound*) es una cota inferior de la evidencia $\log p_\theta(x)$. Además, esta descomposición sugiere que maximizar la ELBO es una función objetivo natural para abordar el problema de la máxima verosimilitud ya que

$$\text{ELBO} = \log p_\theta(x) - \text{D}_{\text{KL}}(q_\phi(z|x) \| p_\theta(z|x)),$$

y, por lo tanto, al maximizar la ELBO según su definición en (1.2.3) se estará maximizando la evidencia al mismo tiempo que se minimiza la divergencia $\text{D}_{\text{KL}}(q_\phi(z|x) \| p_\theta(z|x))$. Además, el hecho de que la ELBO sea una cota inferior de la evidencia, la maximización de la ELBO se puede ver como el problema de optimizar el peor caso de la log-verosimilitud. De este modo, si los modelos p_θ y q_ϕ tienen suficiente capacidad, se estará optimizando la evidencia al mismo tiempo que se ajusta el modelo auxiliar $q_\phi(z|x)$ a la posterior real $p_\theta(z|x)$.

Por otra parte, si bien la definición de la ELBO dada en (1.2.3) es aproximable mediante estimaciones de Monte Carlo, su expresión es poco interpretable. Sin embargo, es posible trabajar más esta expresión:

⁷ Esta desigualdad afirma que la divergencia de Kullback-Leibler es siempre no negativa.

⁸ Notar que la ELBO depende de x , por lo que debería escribirse como $\text{ELBO}(x)$. A lo largo de este trabajo se escribirá, por simplicidad, simplemente ELBO, sin olvidar la dependencia en la variable observada x .

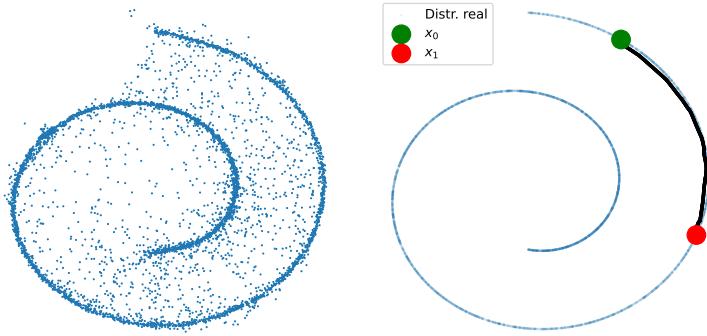


Figura 1.3: (Izquierdo) muestras generadas por un VAE para una distribución p_{data} bidimensional. (Derecha) interpolación en el espacio latente del VAE, lo cual es posible debido a que no hay colapso de la posterior. La implementación de este modelo se encuentra en el archivo `vae_1d.ipynb`.

$$\text{ELBO} = \mathbb{E}_{q_\phi(z|x)} \left[\log \frac{p_\theta(z)p_\theta(x|z)}{q_\phi(z|x)} \right] = \mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x|z)] + \mathbb{E}_{q_\phi(z|x)} \left[\log \frac{p_\theta(z)}{q_\phi(z|x)} \right].$$

En consecuencia, viendo que el segundo término de la segunda igualdad corresponde (al negativo de) una divergencia de Kullback-Leibler, se llega a la caracterización usada en general:

Proposición 1.5 (descomposición de la ELBO). Se tiene la siguiente descomposición para la ELBO:

$$\text{ELBO} = \underbrace{\mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x|z)]}_{\text{término de reconstrucción}} - \underbrace{\text{D}_{\text{KL}}(q_\phi(z|x) \| p_\theta(z))}_{\text{prior matching}}. \quad (1.2.4)$$

El término de reconstrucción⁹, el cual se estima mediante aproximaciones de Monte Carlo, mide qué tan probable es, en esperanza, que la variable latente z asociada a x (según $q_\phi(z|x)$) genere realmente a x (según $p_\theta(x|z)$). Por otra parte, el segundo término busca que la posterior aproximada $q_\phi(z|x)$ no se desvíe demasiado del prior original $p_\theta(z)$, actuando como un regularizador sobre el modelo. En la Subsección 1.2.2 se verá que, bajo ciertas elecciones de modelos, el segundo término puede tener forma cerrada.

Es importante notar que cuando el modelo $p_\theta(x|z)$ tiene demasiada capacidad, es posible que este pueda reconstruir los datos $x \sim p_{\text{data}}$ sin hacer uso de la variable latente (i.e., x se vuelve independiente de z). Este fenómeno se denomina colapso de la posterior y, si bien se sigue logrando el objetivo de tener un modelo generativo para x , ya no es posible guiar la generación de acuerdo a elecciones específicas de z , lo cual es deseable, por ejemplo, para interpolar entre muestras (ver Figura 1.3). Una posible solución es agregar la restricción de que el término *prior matching* sea mayor que un cierto valor δ , tal como se propone en [Raz+19].

Conexión entre la ELBO y la física estadística De acuerdo a la física estadística, la probabilidad de que un sistema físico esté en un estado k con energía ε_k viene dada por la distribución de Boltzmann

$$p(k) \propto \exp \left(-\frac{\varepsilon_k}{k_B T} \right), \quad (1.2.5)$$

donde k_B es la constante de Boltzmann y T es la temperatura del sistema, por lo que la fracción en la función exponencial es adimensional.

⁹Si bien este término tiene forma de entropía cruzada, no lo es ya que $p_\theta(x|\cdot)$ no es distribución de probabilidad.

Esto motiva a que el término $-\log p(k)$ sea generalmente llamado energía. Con esta analogía, es posible conectar la ELBO con la energía libre de Helmholtz. En efecto, al desarrollar la ELBO a partir de su definición original en (1.2.3):

$$\begin{aligned} \text{ELBO} &= \mathbb{E}_{q_\phi(z|x)} \left[\log \frac{p_\theta(x, z)}{q_\phi(z|x)} \right] \\ &= \mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x, z)] - \mathbb{E}_{q_\phi(z|x)} [\log q_\phi(z|x)] \\ &= \underbrace{-\mathbb{E}_{q_\phi(z|x)} [-\log p_\theta(x, z)]}_{\text{energía esperada}} + \underbrace{\mathcal{H}(q_\phi(\cdot|x))}_{\text{entropía}}, \end{aligned}$$

donde $\mathcal{H}(\cdot)$ es el operador de entropía (ver Definición 3.1 para el caso discreto). Por otra parte, si U es la energía interna de un sistema físico, T es la temperatura ambiente y S es la entropía del sistema, la energía libre de Helmholtz queda definida como:

$$F = U - TS,$$

la cual es, de acuerdo a la física estadística, minimizada en el equilibrio termodinámico. En consecuencia, a la ELBO (la cual se busca maximizar) también se le conoce como energía libre variacional negativa.

En la Sección 3.2 se volverá a utilizar la distribución de Boltzmann para conectar el problema de regularización entrópica estudiado en la Sección 3.1 con el problema del puente de Schrödinger en su versión estática.

1.2.2. Formulación de un VAE

La descomposición de la ELBO dada en (1.2.4) motiva a elegir ciertos modelos $q_\phi(z|x)$ y $p_\theta(z)$ de tal forma que la divergencia $D_{KL}(q_\phi(z|x) \| p_\theta(z))$ tenga forma cerrada. Un autoencoder variacional (VAE) es un modelo de variable latente propuesto en [KW22] que propone la siguiente estructura para los modelos paramétricos:

$$q_\phi(z|x) \sim \mathcal{N}(\mu_\phi(x), \Sigma_\phi(x)) \quad (1.2.6)$$

$$p_\theta(z) = p(z) \sim \mathcal{N}(0, I_l), \quad (1.2.7)$$

donde $\mu_\phi : \mathbb{R}^d \rightarrow \mathbb{R}^l$ y $\Sigma_\phi : \mathbb{R}^d \rightarrow \mathcal{S}_l^{++10}$ son funciones generalmente aprendidas por redes neuronales. Por otro lado, el modelo $p_\theta(x|z)$ (cuyos parámetros también serán aprendidos por una red neuronal) dependerá de la naturaleza de $x \in \mathbb{R}^d$, por lo que será definida más adelante.

Con respecto al modelo de inferencia aproximada, $p_\phi(z|x)$, la matriz de covarianzas se asume diagonal, indicando que la variable latente z tiene componentes independientes. Esta restricción, además de simplificar la expresión de la divergencia $D_{KL}(q_\phi(z|x) \| p_\theta(z))$, reduce la cantidad de parámetros y estabiliza el entrenamiento. En la Figura 1.4 se puede observar una ilustración de este modelo.

Para el cálculo de la divergencia en (1.2.4) se utilizará el siguiente resultado clásico:

Teorema 1.2 (divergencia de Kullback-Leibler entre distribuciones gaussianas). Para dos variables aleatorias gaussianas $x \sim \mathcal{N}(\mu_1, \Sigma_1)$ e $y \sim \mathcal{N}(\mu_2, \Sigma_2)$ en \mathbb{R}^l se tiene que:

$$D_{KL}(x \| y) = \frac{1}{2} \left[(\mu_2 - \mu_1)^\top \Sigma_2^{-1} (\mu_2 - \mu_1) + \text{Tr}(\Sigma_2^{-1} \Sigma_1) - \log \left(\frac{|\Sigma_1|}{|\Sigma_2|} \right) - l \right],$$

¹⁰En este caso, \mathcal{S}_l^{++} denota el conjunto de matrices de tamaño $l \times l$ que son simétricas y definidas positivas.

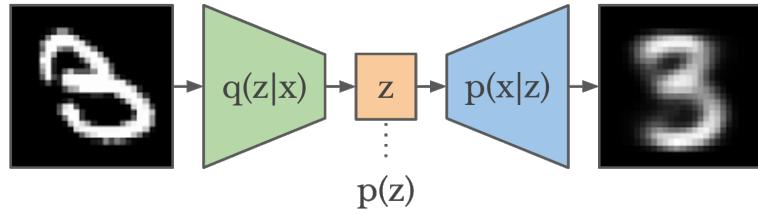


Figura 1.4: Representación gráfica de un autoencoder variacional para MNIST. Imagen obtenida desde [Haf18].

donde se ha denotado $D_{KL}(x \| y)$ para indicar la divergencia entre las medidas asociadas a las variables aleatorias x e y respectivamente.

Por lo tanto, considerando una matriz de covarianzas diagonal, $\Sigma_\phi(x) = \text{diag}(\sigma_\phi^2(x))$ con $\sigma_\phi^2(x) \in \mathbb{R}_{++}^l$ parámetros aprendibles¹¹, la divergencia en (1.2.4) en el caso de un VAE queda de la forma:

$$D_{KL}(q_\phi(z|x) \| p(z)) = \frac{1}{2} \left[\|\mu_\phi(x)\|^2 + \|\sigma_\phi(x)\|^2 - \log \left(\prod_{k=1}^l \sigma_\phi^2(x)_k \right) - l \right].$$

Luego, la función objetivo usada en un VAE viene dada en la siguiente proposición:

Proposición 1.6 (descomposición de la ELBO para un VAE). Bajo los modelos propuestos en (1.2.6) y (1.2.7), la ELBO dada en (1.2.4) se reformula como:

$$\text{ELBO} = \mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x|z)] + \frac{1}{2} \left[\sum_{k=1}^l \mu_\phi^2(x)_k + \sigma_\phi^2(x)_k - \log (\sigma_\phi^2(x)_k) - 1 \right], \quad (1.2.8)$$

donde el término de reconstrucción se aproxima utilizando una estimación de Monte Carlo con $K \leq 1$ muestras (en [KW22] utilizan $K = 1$):

$$\mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x|z)] \approx \frac{1}{K} \sum_{k=1}^K \log p_\theta(x | \mu_\phi(x) + \sigma_\phi(x) \odot \epsilon_k), \quad \epsilon_k \sim \mathcal{N}(0, I_l), \forall k \in \{1 \dots K\}, \quad (1.2.9)$$

con \odot representando el producto de Hadamard (producto coordenada a coordenada).

Es importante destacar que en (1.2.9) es necesario condicionar sobre $\mu_\phi(x) + \sigma_\phi(x) \odot \epsilon_k$ (con $\epsilon_k \sim \mathcal{N}(0, I_l)$) y no directamente sobre z_k con $z_k \sim \mathcal{N}(\mu_\phi(x), \sigma_\phi^2(x))$, ya que de esta última forma se pierde la diferenciabilidad necesaria para el algoritmo de descenso del gradiente. Esta técnica se conoce como *truco de la reparametrización*.

Por otro lado, al igual que para las GANs, es posible extender este enfoque generativo a uno condicional agregando una entrada adicional a las redes neuronales. En la

Cálculo del término de reconstrucción

En esta subsección se verá la forma que toma el término de reconstrucción $\mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x|z)]$ en (1.2.8) cuando se trabaja con datos continuos y con imágenes.

¹¹En la práctica las redes neuronales suelen aprender parámetros asociados a $\log(\sigma_\phi^2(x)_k) \in \mathbb{R}$ para eliminar la restricción de positividad en la salida de red neuronal.



Figura 1.5: Muestras generadas por un autoencoder variacional condicional sobre el dataset MNIST. La implementación de este modelo se encuentra en el archivo `conditional_vae.ipynb`

Término de reconstrucción en datos continuos Cuando la variable observada x tiene soporte continuo, se suele proponer un modelo

$$p_\theta(x|z) \sim \mathcal{N}(f_\theta(z), \sigma_0^2 I_d). \quad (1.2.10)$$

Donde σ_0^2 es una varianza fija, usualmente pequeña. En este caso, $\log p_\theta(x|z) = \log(2\pi\sigma_0^2)^{-\frac{d}{2}} - \frac{1}{2\sigma_0^2} \|x - f_\theta(z)\|^2$. En consecuencia, se tiene el siguiente resultado:

Proposición 1.7 (término de reconstrucción en datos continuos). Bajo el modelo (1.2.10), el término de reconstrucción en (1.2.8) toma la siguiente forma:

$$\mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x|z)] = -\frac{1}{2\sigma_0^2} \|x - f_\theta(z)\|^2 + \text{constante},$$

es decir, se recupera el error cuadrático generalmente usado en un autoencoder convencional (i.e., no variacional) por lo que, en este caso, un VAE utiliza la función de un autoencoder común pero le agregan un término adicional de regulación (prior matching).

Término de reconstrucción en imágenes Por simplicidad, se considerará que se está trabajando con imágenes monocromáticas, las cuales usualmente son representadas como matrices en $\mathcal{M}_{p,q}([0, 1]) \cong [0, 1]^{pq}$. En el caso de trabajar con más canales, el siguiente análisis se extrae de forma natural a tensores de orden 3 o más.

Cada pixel $x_{ij} \in \{0, 1\}$ de la imagen puede ser interpretado como una distribución Bernoulli (con soporte en $\{0, 1\}$), donde los estados indican si el pixel está desactivado (0) o encendido (1). Suponiendo que los pixeles son independientes entre sí dada la variable latente que genera la imagen, es natural proponer el modelo

$$p_\theta(x|z) = \prod_{i=1}^p \prod_{j=1}^q p_\theta(x_{ij}|z), \quad p_\theta(x_{ij}|z) \sim \text{Bernoulli}(r_\theta(z)_{ij}), \quad (1.2.11)$$

donde $r_\theta(z) \in \mathcal{M}_{p,q}([0, 1])$ son parámetros entrenables. En el caso de usar una red neuronal para aprender los modelos, es usual usar una capa sigmoidal en la salida para restringir el codominio. Bajo este modelo paramétrico:

$$\log p_\theta(x|z) = \log \left(\prod_{i=1}^p \prod_{j=1}^q p_\theta(x_{ij}|z) \right) = \sum_{i=1}^p \sum_{j=1}^q \log p_\theta(x_{ij}|z).$$

Dado que $p_\theta(x_{ij}|z) = r_\theta(z)_{ij}^{x_{ij}} (1 - r_\theta(z)_{ij})^{1-x_{ij}}$, se tiene el siguiente resultado:

Proposición 1.8 (término de reconstrucción en imágenes). Bajo el modelo (1.2.11), el término de reconstrucción en (1.2.8) toma la siguiente forma:

$$\mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x|z)] = \mathbb{E}_{q_\phi(z|x)} \left[\sum_{i=1}^p \sum_{j=1}^q x_{ij} \log (r_\theta(z)_{ij}) + (1 - x_{ij}) \log (1 - r_\theta(z)_{ij}) \right].$$

La esperanza con respecto a $q_\phi(z|x)$ se sigue aproximando con estimaciones de Monte Carlo (y utilizando el truco de la reparametrización), mientras que lo que está dentro de la suma se suele programar como una entropía cruzada binaria negativa $-H(x_{ij}, r_\theta(z)_{ij})$ con x_{ij} y $r_\theta(z)_{ij}$ como distribuciones Bernoulli.

Un detalle importante, que muchas veces es pasado por alto, es que los pixeles tienen intensidad de color, por lo que realmente toman valores en $[0, 1]$ y no en $\{0, 1\}$, por lo que modelar los pixeles como distribuciones Bernoulli es incorrecto. En [LC19] muestran que este error no es únicamente un problema técnico y proponen una nueva distribución (*continuous Bernoulli*) con soporte en $[0, 1]$ que busca solucionar este problema, mejorando notablemente la calidad en la generación.

En la Figura 1.6 (izquierda) se puede observar la reconstrucción realizada por un VAE a partir de las variables latentes de un grupo de imágenes.

Autoencoders variacionales jerárquicos markovianos

Una generalización natural de los VAEs consiste en considerar una familia ordenada de variables latentes z_1, \dots, z_T , donde la t -ésima variable latente depende de $z_{(t+1):T}$ ¹². Este tipo de modelos se conocen como autoencoders variacionales jerárquicos (HVAE) y corresponden a un paso previo para introducir los modelos de difusión a tiempo discreto en la Sección 1.3. La distribución conjunta del decoder de este tipo de modelos se factoriza de la siguiente forma:

$$p_\theta(x, z_{1:T}) = p_\theta(x|z_{1:T}) \underbrace{\left(\prod_{t=2}^T p_\theta(z_{t-1}|z_{t:T}) \right)}_{p_\theta(z_{1:T})} p_\theta(z_T). \quad (1.2.12)$$

Notando que el único cambio en la formulación de un HVAE con respecto al modelo de variable latente simple (1.2.1) fue pasar de z a $z_{1:T}$, se puede obtener directamente una expresión para la ELBO en este nuevo tipo de modelos:

Proposición 1.9 (ELBO para un HVAE). Para el modelo dado en (1.2.12) y considerando una posterior aproximada $q_\phi(z_{1:T}|x)$, la ELBO toma la siguiente forma:

¹²La notación $z_{t_1:t_2}$ es una forma compacta de escribir $(z_t)_{t=t_1}^{t_2}$, y será utilizada generalmente para representar modelos jerárquicos.

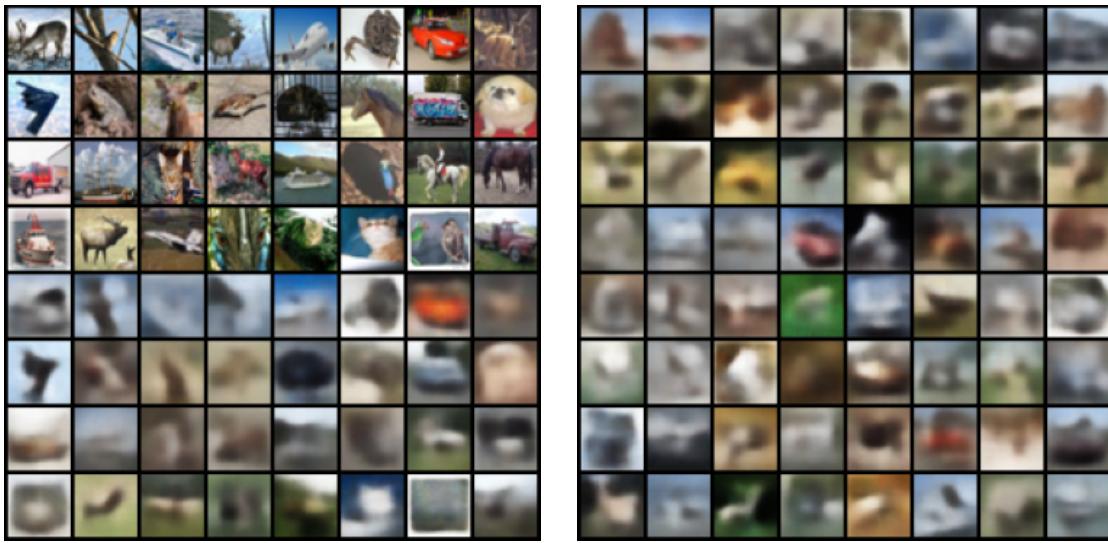


Figura 1.6: (Izquierda) reconstrucción de imágenes realizada por un VAE sobre el dataset CIFAR-10 utilizando una red convolucional, donde las primeras 4 filas corresponden a las imágenes originales y las últimas 4 filas corresponden a la reconstrucción. (Derecha) generación incondicional de imágenes a partir del VAE. La implementación de este modelo se encuentra en el archivo `vae.ipynb`.

$$\text{ELBO} = \mathbb{E}_{q_\phi(z_{1:T}|x)} \left[\log \frac{p_\theta(x, z_{1:T})}{q_\phi(z_{1:T}|x)} \right]. \quad (1.2.13)$$

Al imponer una condición de independencia markoviana, $p_\theta(z_{t-1}|z_{t:T}) = p_\theta(z_{t-1}|z_t)$, se obtiene un modelo gráfico $z_T \rightarrow z_{T-1} \rightarrow \dots \rightarrow z_1 \rightarrow x$ denominado autoencoder variacional jerárquico markoviano (MHVAE). En este caso, la distribución conjunta del decoder se reduce a

$$p_\theta(x, z_{1:T}) = p_\theta(x|z_1) \left(\prod_{t=2}^T p_\theta(z_{t-1}|z_t) \right) p_\theta(z_T),$$

mientras que el modelo de inferencia aproximada para la posterior $p_\theta(z_{1:T}|x)$ toma la forma

$$q_\phi(z_{1:T}|x) = q_\phi(z_1|x) \left(\prod_{t=2}^T q_\phi(z_t|z_{t-1}) \right).$$

En la Figura 1.7 se puede observar el modelo gráfico asociado a un VAE estándar y un HVAE con un nivel adicional de jerarquía. Si bien esta familia de modelos son interesantes por sí solos, aquí son introducidos únicamente para definir lo que es un modelo de difusión en la siguiente subsección.

1.3. Modelos de difusión a tiempo discreto

Los modelos de difusión fueron introducidos originalmente el año 2015 en [Soh+15]. En esa época, los modelos generativos predominantes eran las GANs (ver Subsección 1.1.1), introducidas un año antes en [Goo+14]. Debido al gran progreso que se estaba obteniendo en los modelos tipo GAN, el modelo de difusión propuesto en [Soh+15] no obtuvo mayor interés hasta el año 2020, cuando investigadores de la universidad de California, en Berkeley, propusieron el modelo *denoising diffusion probabilistic models* (DDPM) en [HJA20]. Los modelos

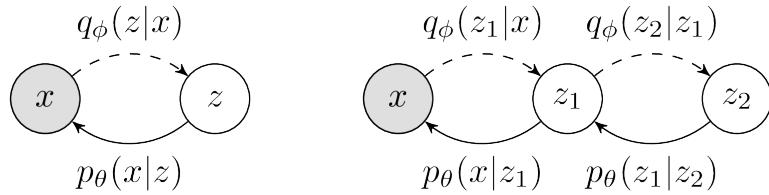


Figura 1.7: Modelo gráfico para un VAE estándar (izquierda) y para un HVAE con dos niveles de jerarquía (derecha). Imágenes obtenidas desde [Tur21].

de este tipo actualmente constituyen el estado del arte de los modelos generativos para imágenes, así como en otras modalidades como la generación de audio y video.

De aquí en adelante, a la familia de modelos relacionados con el modelo DDPM se les denominará modelos de difusión.

1.3.1. Formulación

El modelo de difusión propuesto por [HJA20], puede ser construido como un MHVAE con dos grandes diferencias o imposiciones.

Primero, la dimensión de las variables latentes $z_{1:T}$ es igual a la dimensión de los datos (es decir, $l = d$). Esto contrasta con la noción de cuello de botella vista en el VAE clásico, donde la variable latente z , motivada por la *manifold assumption*¹³, tenía dimensión estrictamente menor a la dimensión observable. Esta nueva restricción permite denotar, con un leve abuso de notación, $(x, z_{1:T})$ como $x_{0:T}$, donde $x_0 = x$ y $x_t = z_t$.

La segunda gran diferencia es que ahora el modelo $q_\phi(x_{1:T}|x_0)$ (que originalmente buscaba aproximar la posterior intratable $p_\theta(x_{1:T}|x_0)$) ahora está fijo, por lo que ya no posee parámetros entrenables ϕ (por lo que se escribirá únicamente q en vez de q_ϕ). Más aún, en un modelo de difusión se impone que las transiciones del encoder $q(x_t|x_{t-1})$ sean gaussianas con parámetros (fijos) cuya función será inyectar ruido de tal forma que $q(x_T) \approx p_{\text{prior}}(x_T)$, donde p_{prior} es una distribución de la cual es fácil generar muestras, por lo que se suele fijar $p_{\text{prior}}(x_T) \sim \mathcal{N}(0, \mathbf{I}_d)$.

Para cumplir estas restricciones, se suele considerar una secuencia finita y decreciente $(\alpha_t)_{t=1}^T \subset [0, 1]$ y se definen las siguientes probabilidades de transición de forma autorregresiva para un $x_0 \sim p_{\text{data}}(x_0)$ dado:

$$q(x_{1:T}|x_0) = \prod_{t=1}^T q(x_t|x_{t-1}), \quad \text{donde} \quad q(x_t|x_{t-1}) \sim \mathcal{N}(\sqrt{\alpha_t}x_{t-1}, (1 - \alpha_t)\mathbf{I}_d), \quad t \in \{1, \dots, T\}. \quad (1.3.1)$$

Para un tiempo de difusión T lo suficientemente grande y elecciones de $(\alpha_t)_{t=1}^T \subset [0, 1]$ adecuadas (en particular, $\alpha_T \approx 0$), la densidad marginal de x_T , $q(x_T) = \int q(x_{0:T}) dx_{0:(T-1)}$, será aproximadamente $p_{\text{prior}}(x_T) \sim \mathcal{N}(0, \mathbf{I}_d)$. Notar que la igualdad solo se alcanza cuando $T \rightarrow \infty$, donde la distribución gaussiana corresponde a la distribución estacionaria del proceso de difusión.

Por otra parte, el decoder $p_\theta(x_{0:T})$ sigue teniendo la misma estructura que antes:

¹³Esta hipótesis afirma que los datos observados realmente provienen de un espacio de menor dimensión (más precisamente, una variedad), el cual está inmerso en el espacio ambiente \mathbb{R}^d . Esta suposición permite explicar por qué, aparentemente, las redes neuronales no sufren de la maldición de la dimensionalidad.

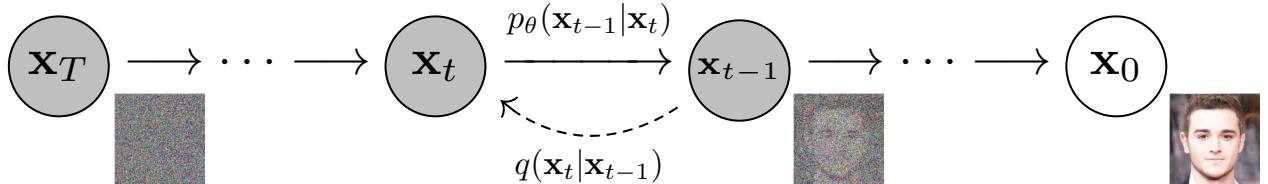


Figura 1.8: Modelo gráfico del proceso de denoising de un DDPM. El proceso comienza con una muestra $x_T \sim p_{\text{prior}}(x_T)$ y continúa con las transiciones de Markov dadas en (1.3.2) utilizando el modelo p_θ ya entrenado. La flecha reversa indica el proceso de inyección de ruido realizado durante el entrenamiento. Imagen obtenida desde [HJA20].

$$p_\theta(x_{0:T}) = p_\theta(x_T) \prod_{t=1}^T p_\theta(x_{t-1}|x_t), \quad \text{donde } p_\theta(x_T) = p(x_T) = p_{\text{prior}}(x_T) \sim \mathcal{N}(0, \mathbf{I}_d). \quad (1.3.2)$$

Este modelo se interpreta diciendo que el proceso forward $q(x_{1:T}|x_0)$ agrega progresivamente ruido a una muestra x_0 generada desde p_{data} , mientras que el proceso backward $p_\theta(x_{0:T})$ aprende a deshacer dicha inyección de ruido (*denoising*) comenzando desde $p(x_T) = p_{\text{prior}}(x_T) \approx q(x_T)$. Una ilustración del modelo gráfico asociado al modelo DDPM se puede encontrar en la Figura 1.8.

Por lo tanto, para generar una nueva muestra desde p_{data} teniendo el modelo p_θ ya entrenado, bastará con generar una muestra $x_T \sim p_{\text{prior}}(x_T)$ y aplicar iterativamente las transiciones de denoising $p_\theta(x_{t-1}|x_t)$ hasta llegar a una muestra $x_0 \sim p(x_0) \approx p_{\text{data}}(x_0)$.

Antes de pasar a la distribución basal que se utilizará para modelar las transiciones $p_\theta(x_{t-1}|x_t)$ se verán dos propiedades útiles que motivarán su estructura.

Propiedades de los procesos forward y backward

El proceso de inyección de ruido definido en (1.3.1) es una cadena de Markov $(x_t)_{t=0}^T$ con distribución inicial $x_0 \sim q(x_0) = p_{\text{data}}(x_0)$. Por lo tanto, para obtener la muestra ruidosa en tiempo t es necesario conocer la muestra ruidosa en tiempo $t-1$. Si bien esto se puede conseguir iterando mediante $q(x_t|x_{t-1})$, este procedimiento no es eficiente cuando se necesita la distribución en un único tiempo $t \gg 1$. Dado que la cadena tiene transiciones gaussianas, es posible obtener una forma cerrada para $q(x_t|x_0)$:

Proposición 1.10 (marginal condicional para el proceso forward). Bajo las transiciones definidas en (1.3.1), el proceso forward posee la siguiente propiedad para $t \in \{1, \dots, T\}$:

$$q(x_t|x_0) \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)\mathbf{I}_d), \quad \text{donde } \bar{\alpha}_t := \prod_{k=1}^t \alpha_k. \quad (1.3.3)$$

Demostración. De acuerdo al proceso forward (1.3.1), $q(x_t|x_{t-1}) \sim \mathcal{N}(\sqrt{\alpha_t}x_{t-1}, (1 - \alpha_t)\mathbf{I}_d)$, por lo que una muestra a partir de $(x_t|x_{t-1})$ se puede obtener mediante

$$x_t = \sqrt{\alpha_t}x_{t-1} + \sqrt{1 - \alpha_t}\epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \mathbf{I}_d).$$

Si x_{t-1} se obtiene del mismo modo a partir de x_{t-2} , se llega a que

$$\begin{aligned}
 x_t &= \sqrt{\alpha_t} \left(\sqrt{\alpha_{t-1}} x_{t-2} + \sqrt{1 - \alpha_{t-1}} \epsilon_{t-1} \right) + \sqrt{1 - \alpha_t} \epsilon_t, \quad \epsilon_{t-1}, \epsilon_t \sim \mathcal{N}(0, \mathbf{I}_d) \\
 &= \sqrt{\alpha_t \alpha_{t-1}} x_{t-2} + \sqrt{\alpha_t(1 - \alpha_{t-1})} \epsilon_{t-1} + \sqrt{1 - \alpha_t} \epsilon_t \\
 &= \sqrt{\alpha_t \alpha_{t-1}} x_{t-2} + \sqrt{[\alpha_t - \alpha_t \alpha_{t-1}] + [1 - \alpha_t]} \tilde{\epsilon}_{t-1}, \quad \tilde{\epsilon}_{t-1} \sim \mathcal{N}(0, \mathbf{I}_d) \\
 &= \sqrt{\alpha_t \alpha_{t-1}} x_{t-2} + \sqrt{1 - \alpha_t \alpha_{t-1}} \tilde{\epsilon}_{t-1},
 \end{aligned}$$

donde en la penúltima igualdad se usó la fórmula de suma de gaussianas. Inductivamente, si todos los x_t se generaron de este modo, se llega a que

$$x_t = \sqrt{\prod_{i=1}^t \alpha_i} x_0 + \sqrt{1 - \prod_{i=1}^t \alpha_i} \tilde{\epsilon}_0, \quad \tilde{\epsilon}_0 \sim \mathcal{N}(0, \mathbf{I}_d).$$

Es decir,

$$x_t \sim \mathcal{N} \left(\sqrt{\prod_{i=1}^t \alpha_i} x_0, \left(1 - \prod_{i=1}^t \alpha_i \right) \mathbf{I}_d \right).$$

□

Esta propiedad es considerada, generalmente, como una de las más importantes dentro de los modelos de difusión, al punto que se le suele citar, sin hacer referencia explícita, como la *nice property*¹⁴. Por otra parte, algunas variantes del modelo original de DDPM, como DDIM (ver Subsección 1.3.3) se enfocan en realizar modificaciones al modelo DDPM original buscando preservar precisamente esta propiedad.

Por otra parte, si bien el proceso backward dado en (1.3.2) no es conocido (ya que no se conoce $q(x_0)$), es posible obtener las transiciones inversas cuando se condiciona a x_0 . Esta propiedad es importante ya que, como se verá más adelante, motivará la forma del proceso reverso $p_\theta(x_{t-1}|x_t)$ en la función de costo de un modelo de difusión aparece una divergencia con esta distribución.

Proposición 1.11 (proceso inverso condicional a x_0). Dado el proceso forward definido en (1.3.1), entonces el proceso inverso dado x_0 también es gaussiano y tiene distribución

$$q(x_{t-1}|x_t, x_0) \sim \mathcal{N}(\mu_q(x_0, x_t, t), \sigma_q^2(t) \mathbf{I}_d), \quad (1.3.4)$$

donde la media μ_q es una combinación lineal entre x_0 y x_t , mientras que la varianza σ_q^2 solo depende de t . Para $t \geq 2$:

$$\mu_q(x_0, x_t, t) := \frac{\sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)}{1 - \bar{\alpha}_t} x_0 + \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} x_t \quad y \quad \sigma_q^2(t) := \frac{(1 - \alpha_t)(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t}, \quad (1.3.5)$$

con $\bar{\alpha}_t := \prod_{k=1}^t \alpha_k$. Para $t = 1$, se define $\mu_q(x_0, x_t, t) = x_0$ y $\sigma_q^2(t) = 0$ ¹⁵.

Demostración. Para $t = 1$ es directo ya que $q(x_0|x_1, x_0) \sim \delta_{x_0}$, por lo que se puede considerar como una variable aleatoria gaussiana de media x_0 y varianza nula. Para $t > 1$, por regla de Bayes y propiedad de Markov:

¹⁴Dado que para el entrenamiento se utiliza esta propiedad, algunos trabajos entregan los niveles de ruidos como secuencias asociadas a la varianza de $q(x_t|x_0)$ y no de $q(x_t|x_{t-1})$. Ambos métodos son identificables entre sí.

¹⁵En la práctica, se define como $\sigma_q^2(t) = \epsilon > 0$ para que la función de densidad exista y se pueda trabajar con ella posteriormente.

$$\begin{aligned} q(x_{t-1}|x_t, x_0) &= \frac{q(x_t|x_{t-1})q(x_{t-1}|x_0)}{q(x_t|x_0)} \propto q(x_t|x_{t-1})q(x_{t-1}|x_0) \\ &\propto \exp \left[-\frac{1}{2} \left(\frac{\|x_t - \sqrt{\alpha_t}x_{t-1}\|^2}{1 - \alpha_t} + \frac{\|x_{t-1} - \sqrt{\bar{\alpha}_{t-1}}x_0\|^2}{1 - \bar{\alpha}_{t-1}} \right) \right]. \end{aligned}$$

Usando las propiedades $\|a \pm b\|^2 = \|a\|^2 \pm 2\langle a, b \rangle + \|b\|^2$ y $\langle \lambda a, b \rangle = \langle a, \lambda b \rangle$ para el término dentro del paréntesis y dejando fuera las constantes que no dependen de x_{t-1} :

$$\begin{aligned} &\frac{\|x_t - \sqrt{\alpha_t}x_{t-1}\|^2}{1 - \alpha_t} + \frac{\|x_{t-1} - \sqrt{\bar{\alpha}_{t-1}}x_0\|^2}{1 - \bar{\alpha}_{t-1}} \\ &= \frac{\alpha_t \|x_{t-1}\|^2 - 2 \langle \sqrt{\alpha_t}x_{t-1}, x_t \rangle}{1 - \alpha_t} + \frac{\|x_{t-1}\|^2 - 2 \langle x_{t-1}, \sqrt{\bar{\alpha}_{t-1}}x_0 \rangle}{1 - \bar{\alpha}_{t-1}} + \text{constante} \\ &= \|x_{t-1}\|^2 \left(\frac{\alpha_t}{1 - \alpha_t} + \frac{1}{1 - \bar{\alpha}_{t-1}} \right) - 2 \left\langle x_{t-1}, \left(\frac{\sqrt{\alpha_t}}{1 - \alpha_t} x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1 - \bar{\alpha}_{t-1}} x_0 \right) \right\rangle + \text{constante} \\ &= \frac{1 - \bar{\alpha}_t}{(1 - \alpha_t)(1 - \bar{\alpha}_{t-1})} \left(\|x_{t-1}\|^2 - 2 \left\langle x_{t-1}, \frac{\sqrt{\alpha_t}(1 - \sqrt{\alpha_{t-1}})}{1 - \bar{\alpha}_t} x_t + \frac{\sqrt{\alpha_{t-1}}(1 - \alpha_t)}{1 - \bar{\alpha}_t} x_0 \right\rangle \right) + \text{constante}. \end{aligned}$$

Por lo tanto:

$$q(x_{t-1}|x_t, x_0) \propto \exp \left(-\frac{1}{2\sigma_q^2(t)} \|x_{t-1} - \mu_q(x_0, x_t, t)\|^2 \right).$$

La demostración concluye notando que esto define completamente $q(x_{t-1}|x_t, x_0)$ salvo constantes multiplicativas de normalización. \square

Una ilustración de las distribuciones encontradas se puede ver en la Figura 1.9.

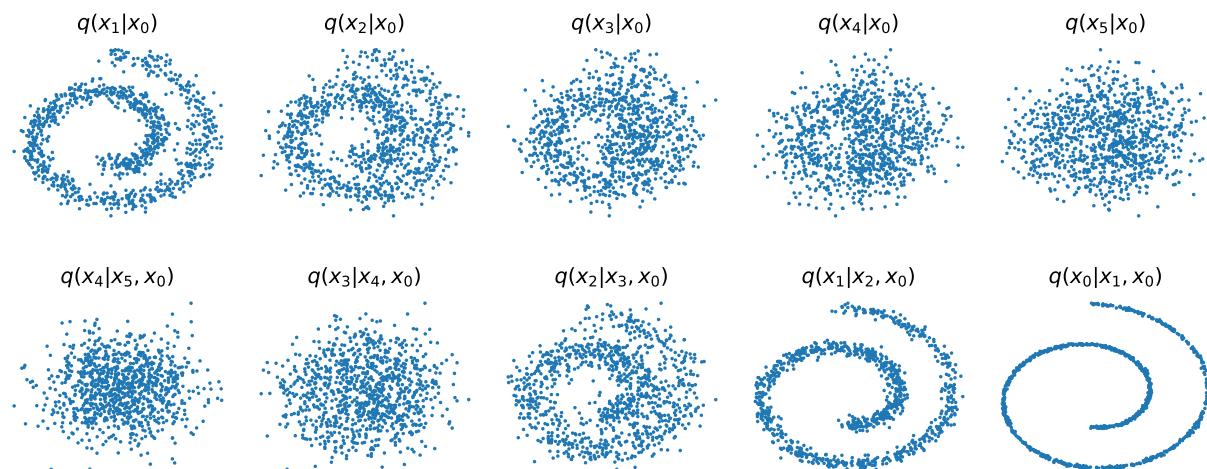


Figura 1.9: (Arriba) muestras del proceso forward dado por la Proposición 1.10 para $T = 5$. (Abajo) muestras del proceso reverso condicional dado por la Proposición 1.11. La implementación de estas distribuciones se encuentra en el archivo `ddpm.ipynb`.

Formulación para el proceso inverso

Motivado por la Proposición 1.11, se propone elegir un modelo paramétrico gaussiano para el proceso reverso (incondicional) del modelo de difusión:

$$p_\theta(x_{t-1}|x_t) \sim \mathcal{N}(\mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), \quad t \in \{1, \dots, T\}, \quad (1.3.6)$$

donde μ_θ y Σ_θ son modelos paramétricos entrenados mediante redes neuronales que reciben una muestra ruidosa x_t en un tiempo t y entregan la media y covarianza para la muestra x_{t-1} en el tiempo anterior. Notando que la matriz de covarianza de $q(x_{t-1}|x_t, x_0)$ en (1.3.4) es independiente de x_0 (solo depende de t), los autores de [HJA20] decidieron fijar la varianza de $p_\theta(x_{t-1}|x_t)$ como la varianza de $p_\theta(x_{t-1}|x_t, x_0)$:

$$\Sigma_\theta(x_t, t) := \Sigma(t) := \sigma_q^2(t) \mathbf{I}_d, \quad t \in \{1, \dots, T\}. \quad (1.3.7)$$

Como se verá en el Teorema 1.3, esta elección permitirá simplificar la función objetivo de los modelos de difusión. Los autores también probaron usar $\Sigma(t) := (1 - \alpha_t) \mathbf{I}_d$ basándose en la covarianza de $q(x_t|x_{t-1})$ según (1.3.1), que también es independiente de x_0 . El primer caso corresponde a una cota inferior de la entropía para el proceso reverso, donde x_0 concentra toda su masa en un único punto, mientras que el segundo caso corresponde a la cota superior. Ambas elecciones de $\Sigma_\theta(x_t, t)$ entregan resultados similares, necesitando únicamente entrenar una red neuronal para predecir la media $\mu_\theta(x_t, t) \in \mathbb{R}^d$. En el archivo `ddpm.ipynb` se entrena un modelo de difusión utilizando la varianza dada en (1.3.7).

Es importante notar que nada garantiza, en ninguno de los dos casos, que la covarianza fijada $\Sigma_\theta(x_t, t)$ sea igual a la covarianza realmente buscada, $\text{Var}(q(x_{t-1}|x_t))$, la cual es desconocida¹⁶. En [ND21] muestran que aprender también la varianza $\Sigma_\theta(x_t, t)$ permite generar muestras en menos pasos sin afectar considerablemente la calidad de las muestras generadas (ver Subsección 1.3.3).

Por otra parte, al igual que los autoencoders variacionales, los modelos de difusión serán entrenados mediante la ELBO. Como se verá a continuación, esta función objetivo puede tomar diferentes formas equivalentes, por lo que, si bien ahora se está entrenando un modelo para predecir la media de x_{t-1} conociendo x_t , también se podrá formular el problema para buscar el ruido ϵ_t insertado a x_0 en el tiempo t , e incluso se podrá buscar directamente x_0 .

ELBO y función objetivo

Recordando que los modelos de difusión son una MHVAE con condiciones específicas, se puede utilizar la expresión dada en (1.2.13) para obtener una ELBO conveniente para los modelos de difusión.

Teorema 1.3 (ELBO para DDPM). Para un modelo de difusión genérico¹⁷ factorizado como en (1.3.1) y (1.3.2), su ELBO viene dada por:

$$\underbrace{\mathbb{E}_{q(x_1|x_0)} [\log p_\theta(x_0|x_1)]}_{\text{término de reconstrucción}} - \underbrace{\text{D}_{\text{KL}}(q(x_T|x_0) \| p(x_T))}_{\text{prior matching}} - \sum_{t=2}^T \underbrace{\mathbb{E}_{q(x_t|x_0)} [\text{D}_{\text{KL}}(q(x_{t-1}|x_t, x_0) \| p_\theta(x_{t-1}|x_t))]}_{\text{denoising matching}}. \quad (1.3.8)$$

Demostración. Sustituyendo la factorización forward (1.3.1) y la factorización backward (1.3.2) en el ELBO (1.2.13):

¹⁶Ya que no se conoce la distribución $q_0 = p_{\text{data}}$. Si se conociera, se podría obtener $q(x_{t-1}|x_t)$ usando $q(x_{t-1}|x_t, x_0)$

¹⁷No necesariamente haciendo uso de las transiciones dadas en el lado derecho de (1.3.1).

$$\begin{aligned}
 \text{ELBO} &= \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log \left(\frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} \right) \right] \\
 &= \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log \left(\frac{p(x_T) \prod_{t=1}^T p_\theta(x_{t-1}|x_t)}{\prod_{t=1}^T q(x_t|x_{t-1})} \right) \right] \\
 &= \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log \left(p(x_T) \prod_{t=1}^T \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} \right) \right]. \tag{1.3.9}
 \end{aligned}$$

Dado que $q(x_t|x_0)$ y $q(x_{t-1}|x_t, x_0)$ se pueden conocer en forma cerrada, se realiza la siguiente sustitución:

$$q(x_t|x_{t-1}) = q(x_t|x_{t-1}, x_0) = \frac{q(x_{t-1}|x_t, x_0)q(x_t|x_0)}{q(x_{t-1}|x_0)}. \tag{1.3.10}$$

Sustituyendo (1.3.10) en (1.3.9) y considerando que $q(x_0|x_0) = 1$ y $q(x_0|x_t, x_0) = 1$:

$$\begin{aligned}
 \text{ELBO} &= \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log \left(p(x_T) \prod_{t=1}^T \frac{p_\theta(x_{t-1}|x_t)q(x_{t-1}|x_0)}{q(x_{t-1}|x_t, x_0)q(x_t|x_0)} \right) \right] \\
 &= \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log p(x_T) + \sum_{t=1}^T \log \left(\frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} \right) + \underbrace{\sum_{t=1}^T (\log q(x_{t-1}|x_0) - \log q(x_t|x_0))}_{\log q(x_0|x_0) - \log q(x_T|x_0)} \right] \\
 &= \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log \left(\frac{p(x_T)}{q(x_T|x_0)} \right) + \log \left(\frac{p_\theta(x_0|x_1)}{1} \right) + \sum_{t=2}^T \log \left(\frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} \right) \right] \\
 &= \mathbb{E}_{q(x_T|x_0)} \left[\log \frac{p(x_T)}{q(x_T|x_0)} \right] + \mathbb{E}_{q(x_1|x_0)} [\log p_\theta(x_0|x_1)] + \sum_{t=2}^T \mathbb{E}_{q(x_t, x_{t-1}|x_0)} \left[\log \frac{p_\theta(x_{t-1}|x_t)}{q(x_{t-1}|x_t, x_0)} \right] \\
 &= -D_{KL}(q(x_T|x_0) \| p(x_T)) + \mathbb{E}_{q(x_1|x_0)} [\log p_\theta(x_0|x_1)] - \sum_{t=2}^T \mathbb{E}_{q(x_t|x_0)} [D_{KL}(q(x_{t-1}|x_t, x_0) \| p_\theta(x_{t-1}|x_t))],
 \end{aligned}$$

donde en la última igualdad se factorizó $q(x_t, x_{t-1}|x_0) = q(x_t|x_0)q(x_{t-1}|x_t, x_0)$ para descomponer la esperanza $\mathbb{E}_{q(x_t, x_{t-1}|x_0)} [\cdot]$ como $\mathbb{E}_{q(x_t|x_0)} [\mathbb{E}_{q(x_{t-1}|x_t, x_0)} [\cdot]]$. \square

Notar que el término de reconstrucción es el mismo que aparece en un VAE convencional, por lo que puede ser aproximado mediante estimaciones de Monte Carlo. Sin embargo, como se verá a continuación, no se considerará en la optimización debido a que se trabajará con una función objetivo simplificada. Por otra parte, el término de prior matching puede omitirse en el entrenamiento ya que es constante con respecto a los parámetros del modelo p_θ .

Por último, los términos de denoising matching confirman la elección de las transiciones backward $p_\theta(x_{t-1}|x_t)$ gaussianas, ya que permitirán tener una forma cerrada para este término. Más aún, al elegir $\Sigma_\theta(x_t, t) = \sigma_q^2(t) I_d$ de acuerdo a (1.3.7), la varianza de $q(x_{t-1}|x_t, x_0)$ y $p_\theta(x_{t-1}|x_t)$ coinciden, simplificando aún más la función objetivo. Utilizando el Teorema 1.2 para este caso particular, se obtiene que:

$$\begin{aligned} \text{D}_{\text{KL}}(q(x_{t-1}|x_t, x_0) \| p_\theta(x_{t-1}|x_t)) &= \frac{1}{2} \left(\frac{\|\mu_q(x_0, x_t, t) - \mu_\theta(x_t, t)\|^2}{\sigma_q^2(t)} + \text{Tr}(\mathbf{I}_d) - \log(1) - d \right) \\ &= \frac{1}{2\sigma_q^2(t)} \|\mu_q(x_0, x_t, t) - \mu_\theta(x_t, t)\|^2. \end{aligned}$$

Por otro lado, para el término de reconstrucción $\mathbb{E}_{q(x_1|x_0)} [\log p_\theta(x_0|x_1)]$, es útil recordar de la Proposición 1.11 que, $q(x_{t-1}|x_t, x_0) \sim \mathcal{N}(\mu_q(x_0, x_t, t), \sigma_q^2(t) \mathbf{I}_d)$. Por lo tanto:

$$\log p_\theta(x_0|x_1) = \frac{-1}{2\sigma_q^2(1)} \|\mu_q(x_0, x_1, 1) - \mu_\theta(x_1, 1)\|^2 + \text{constante}.$$

En consecuencia, los términos de reconstrucción y denoising matching se pueden agrupar en una única suma (comenzando desde 1), mientras que el término de prior matching se puede omitir dado que es constante con respecto a los parámetros:

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

Recordando que se busca maximizar la ELBO en esperanza sobre $x_0 \sim p_{\text{data}}(x_0)$, el problema de optimización asociado a DDPM se reduce a resolver una diferencia de cuadrados:

Proposición 1.12 (problema de optimización DDPM, μ -prediction). Para los modelos propuestos en (1.3.1) y (1.3.6), la función de media del proceso reverso, $\mu_\theta(x_t, t)$, que maximiza la ELBO se encuentra resolviendo el siguiente problema de optimización:

$$\mu_\theta^* = \arg \min_{\theta} \sum_{t=1}^T \frac{1}{2\sigma_q^2(t)} \mathbb{E}_{q(x_0, x_t)} [\|\mu_q(x_0, x_t, t) - \mu_\theta(x_t, t)\|^2],$$

donde se recuerda que $q(x_0, x_t) = q(x_0)q(x_t|x_0)$ con $q(x_0) \sim p_{\text{data}}(x_0)$ y $q(x_t|x_0)$ dado en (1.3.3).

En la función objetivo anterior, la esperanza $\mathbb{E}_{q(x_0, x_t)} [\cdot] = \mathbb{E}_{x_0 \sim p_{\text{data}}} [\mathbb{E}_{q(x_t|x_0)} [\cdot]]$ se puede aproximar mediante un conjunto de muestras $(x_0^k)_{k=1}^n \subset \mathbb{R}^d$ para $\mathbb{E}_{p_{\text{data}}} [\cdot]$ y con un conjunto de muestras ruidosas $(x_t^k)_{k=1}^m$ generadas a partir de cada $x_0^k \sim p_{\text{data}}(x_0)$ de acuerdo a $q(x_t|x_0)$.

Si bien la expresión a optimizar es bastante simple de implementar, es posible trabajar un poco más la expresión para disminuirle la carga al modelo paramétrico $\mu_\theta(x_t, t)$. Para esto, se verán dos reformulaciones equivalentes de este problema de optimización.

Enfoque x_0 -prediction Notando que $\mu_q(x_0, x_t, t)$ en (1.3.5) es de la forma $c_1(t)x_0 + c_2(t)x_t$, entonces el modelo $\mu_\theta(x_t, t)$ se puede reparametrizar como $\mu_\theta(x_t, t) = c_1(t)x_\theta(x_t, t) + c_2(t)x_t$, donde ahora x_θ es otro modelo paramétrico que busca predecir x_0 a partir de x_t :

Proposición 1.13 (problema de optimización DDPM, x_0 -prediction). Para los modelos propuestos en (1.3.1) y (1.3.6), la función de media del proceso reverso, $\mu_\theta(x_t, t)$, que maximiza la ELBO es una combinación lineal entre x_θ y x_t :

$$\mu_\theta(x_t, t) = \frac{\sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)}{1 - \bar{\alpha}_t} x_\theta(x_t, t) + \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} x_t,$$

donde el modelo $x_\theta(x_t, t)$ resuelve el siguiente problema de optimización:

$$x_\theta^* = \arg \min_{x_\theta} \sum_{t=1}^T \frac{\bar{\alpha}_{t-1}(1-\alpha_t)^2}{2\sigma_q^2(t)(1-\bar{\alpha}_t)^2} \mathbb{E}_{q(x_0, x_t)} [\|x_0 - x_\theta(x_t, t)\|^2],$$

Demostración. Considerando que $\mu_q(x_0, x_t, t) = c_1(t)x_0 + c_2(t)x_t$ y $\mu_\theta(x_t, t) = c_1(t)x_\theta(x_t, t) + c_2(t)x_t$ con c_1, c_2 dadas en (1.3.5):

$$\|\mu_q(x_0, x_t, t) - \mu_\theta(x_t, t)\|^2 = \|c_2(t)(x_0 - x_\theta(x_t, t))\|^2 = c_2(t)^2 \|(x_0 - x_\theta(x_t, t))\|^2,$$

Luego, basta sustituir esta expresión en la Proposición 1.12. \square

Enfoque ϵ -prediction Al igual que para el enfoque x_0 -prediction, es posible reparametrizar el modelo μ_θ usando otro modelo que aprenda el ruido $\epsilon(x_t)$ que se le inyectó a x_0 para obtener x_t . Para esto es útil notar que, de acuerdo a la Proposición 1.10, $q(x_t|x_0) \sim \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$ con $\epsilon \sim \mathcal{N}(0, I_d)$:

Proposición 1.14 (problema de optimización DDPM, ϵ -prediction). Para los modelos propuestos en (1.3.1) y (1.3.6), la función de media del proceso reverso, $\mu_\theta(x_t, t)$, que maximiza la ELBO es una combinación lineal entre x_t y ϵ_θ :

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\bar{\alpha}_t}}x_t - \frac{1-\alpha_t}{\sqrt{\alpha_t(1-\bar{\alpha}_t)}}\epsilon_\theta(x_t, t), \quad (1.3.11)$$

donde el modelo $\epsilon_\theta(x_t, t)$ resuelve el siguiente problema de optimización:

$$\epsilon_\theta^* = \arg \min_{\epsilon_\theta} \sum_{t=1}^T \frac{(1-\alpha_t)^2}{2\sigma_q^2(t)(1-\bar{\alpha}_t)\alpha_t} \mathbb{E}_{q(x_0, x_t)} [\|\epsilon(x_t) - \epsilon_\theta(x_t, t)\|^2].$$

Donde se ha definido $\epsilon(x_t) = \frac{x_t - \sqrt{\bar{\alpha}_t}x_0}{\sqrt{1-\bar{\alpha}_t}}$.

Demostración. Basta probar que $\|\mu_\theta(x_t, t) - \mu_q(x_0, x_t, t)\|^2 = \frac{(1-\alpha_t)^2}{(1-\bar{\alpha}_t)\alpha_t} \|\epsilon - \epsilon_\theta(x_t, t)\|^2$ para μ_θ descrito en el enunciado. Considerando que $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon(x_t)$, es posible despejar x_0 y obtener $x_0 = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon(x_t)}{\sqrt{\bar{\alpha}_t}}$. Sustituyendo esta expresión en $\mu_q(x_0, x_t, t)$:

$$\begin{aligned} \mu_q(x_0, x_t, t) &= \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t} \cdot \frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon(x_t)}{\sqrt{\bar{\alpha}_t}} + \frac{\sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t \\ &= \left(\frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{(1-\bar{\alpha}_t)\sqrt{\bar{\alpha}_t}} + \frac{\sqrt{\bar{\alpha}_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} \right) x_t - \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\sqrt{1-\bar{\alpha}_t}}{(1-\bar{\alpha}_t)\sqrt{\bar{\alpha}_t}}\epsilon(x_t) \\ &= \left(\frac{1-\alpha_t}{(1-\bar{\alpha}_t)\sqrt{\bar{\alpha}_t}} + \frac{\alpha_t - \alpha_t\bar{\alpha}_{t-1}}{(1-\bar{\alpha}_t)\sqrt{\bar{\alpha}_t}} \right) x_t - \frac{(1-\alpha_t)\sqrt{1-\bar{\alpha}_t}}{(1-\bar{\alpha}_t)\sqrt{\bar{\alpha}_t}}\epsilon(x_t) \\ &= \frac{1-\bar{\alpha}_t}{(1-\bar{\alpha}_t)\sqrt{\bar{\alpha}_t}}x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\bar{\alpha}_t}}\epsilon(x_t) \\ &= \frac{x_t}{\sqrt{\bar{\alpha}_t}} - \frac{1-\alpha_t}{\sqrt{(1-\bar{\alpha}_t)\alpha_t}}\epsilon(x_t), \end{aligned}$$

donde en la tercera y cuarta igualdad se usó que $\alpha_t\bar{\alpha}_{t-1} = \bar{\alpha}_t$. Dado que μ_θ es función de x_t , se puede reparametrizar como:

$$\mu_\theta(x_t, t) = \frac{x_t}{\sqrt{\bar{\alpha}_t}} - \frac{1-\alpha_t}{\sqrt{(1-\bar{\alpha}_t)\alpha_t}}\epsilon_\theta(x_t, t).$$

Luego:

$$\begin{aligned}\|\mu_\theta(x_t, t) - \mu_q(x_0, x_t, t)\|^2 &= \left\| \left(\frac{x_t}{\sqrt{\alpha_t}} - \frac{1 - \alpha_t}{\sqrt{(1 - \bar{\alpha}_t)\alpha_t}} \epsilon_\theta(x_t, t) \right) - \left(\frac{x_t}{\sqrt{\alpha_t}} - \frac{1 - \alpha_t}{\sqrt{(1 - \bar{\alpha}_t)\alpha_t}} \epsilon(x_t) \right) \right\|^2 \\ &= \frac{(1 - \alpha_t)^2}{(1 - \bar{\alpha}_t)\alpha_t} \|\epsilon(x_t) - \epsilon_\theta(x_t, t)\|^2,\end{aligned}$$

que es lo que se quería probar. \square

Si bien este enfoque parece ser el más complicado, en [HJA20] afirman que es el que entregó mejores resultados. Además, en este trabajo los autores consideran una versión simplificada de la función objetivo:

- Los ponderadores de las esperanzas son omitidos para que todos los sumandos tengan el mismo peso. En [SME22] justifican esta decisión notando que si los modelos $\epsilon_\theta(x_t, t)$ son independientes en la segunda variable (i.e., se tiene un modelo independiente para cada tiempo), entonces el óptimo global es alcanzado optimizando localmente cada sumando, por lo que la función objetivo se vuelve independiente de los ponderadores.
- Con el fin de reducir el costo de entrenamiento, la suma sobre $t \in \{2, \dots, T\}$ es sustituida por una evaluación aleatoria del tiempo. Para esto consideran una variable discreta uniforme $t \sim \text{Uniform}(\{1, \dots, T\})$ y calculan un único sumando de forma aleatoria en cada iteración del entrenamiento.

Con estas modificaciones, la función objetivo que proponen minimizar en [HJA20] con el enfoque ϵ -prediction es la siguiente:

$$L_{\text{simple}}(\theta) = \mathbb{E}_{x_0 \sim p_{\text{data}}, t \sim \text{Uniform}(\{1, \dots, T\}), \epsilon \sim \mathcal{N}(0, I_d)} \left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, t)\|^2 \right], \quad (1.3.12)$$

donde se consideró que $q(x_t | x_0) \sim \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$ con $\epsilon \sim \mathcal{N}(0, I_d)$ para sustituir x_t en $\epsilon_\theta(x_t, t)$ y cambiar la esperanza $\mathbb{E}_{q(x_t | x_0)} [\cdot]$ por $\mathbb{E}_{\epsilon \sim \mathcal{N}(0, I_d)} [\cdot]$.

Algoritmos de entrenamiento y sampling

Utilizando el enfoque ϵ -prediction con la función de pérdida simplificada (1.3.12) el algoritmo de entrenamiento consiste en aproximar todas las esperanzas de L_{simple} con una estimación de Monte Carlo simple, tal como lo muestra el Algoritmo 2.

Algoritmo 2 Entrenamiento del modelo DDPM

Require: muestras de entrenamiento provenientes de p_{data} .

Require: Modelo ϵ_θ (no entrenado) y secuencia $(\alpha_t)_{t=1}^T$.

- 1: **while** no haya convergencia **do**
 - 2: Obtener muestra de entrenamiento $x_0 \sim p_{\text{data}}(x_0)$.
 - 3: Generar $t \sim \text{Uniform}(\{1, \dots, T\})$.
 - 4: Generar $\epsilon \sim \mathcal{N}(0, I_d)$.
 - 5: Actualizar parámetros de ϵ_θ según $\nabla_\theta \|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, t)\|$.
 - 6: **end while**
-

Por otra parte, como se comentó anteriormente, para la generación de muestras es suficiente generar una muestra desde la distribución prior $p(x_T) = p_{\text{prior}}$ y luego aplicar el proceso backward aprendido. Para esto es necesario recordar que, de acuerdo a la Proposición 1.14, la media μ_θ de $p_\theta(x_{t-1} | x_t)$ viene dada por:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}x_t - \frac{1 - \alpha_t}{\sqrt{\alpha_t(1 - \bar{\alpha}_t)}}\epsilon_\theta(x_t, t),$$

mientras que la matriz de covarianzas para $p_\theta(x_{t-1}|x_t)$ se fijó en (1.3.7) como $\Sigma_\theta(x_t, t) = \sigma_q^2(t) I_d$. A este procedimiento de generación se le suele llamar *ancestral sampling* (nombre atribuido en [Son+21b]) y se encuentra detallado en el Algoritmo 3. Es importante mencionar que al comenzar el proceso reverso desde la distribución prior se está induciendo un error en la generación, ya que el proceso de difusión en el Algoritmo 2 solo se ejecuta hasta un horizonte de tiempo finito, mientras que la distribución invariante $\mathcal{N}(0, I_d)$ es alcanzada únicamente cuando $T \rightarrow \infty$. Este es un problema intrínseco de los modelos de difusión y puede generar problemas en la generación de muestras tal como se estudia en [Bor+23a]. En el Capítulo 3 se entregará una formulación más robusta mediante el problema del puente de Schrödinger, donde se estudiará el problema de transformar una distribución en otra en un horizonte de tiempo finito, permitiendo, además, considerar una distribución terminal $q(x_T)$ diferente a $\mathcal{N}(0, I_d)$.

Algoritmo 3 Ancestral sampling

Require: modelo ϵ_θ (entrenado) y secuencia $(\alpha_t)_{t=1}^T$ que se usó durante el entrenamiento.

```

1: Generar  $x_T \sim p_{\text{prior}}(x_T) = \mathcal{N}(0, I_d)$ .
2: for  $t$  desde  $T$  hasta 1 do
3:   if  $t > 1$  then
4:     generar  $z \sim \mathcal{N}(0, I_d)$ 
5:   else
6:      $z \leftarrow 0$ 
7:   end if
8:    $x_{t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}} \left( x_t - \frac{1-\alpha_t}{\sqrt{1-\alpha_t}} \epsilon_\theta(x_t, t) \right) + \sigma_q(t) z$ 
9: end for
10: return  $x_0$ 
```

En la Figura 1.10 se puede observar la evolución del proceso de denoising en cada etapa del Algoritmo 3, mientras que en la Figura 1.11 se observan muestras generadas por un modelo de difusión sobre un dataset bidimensional.

1.3.2. Arquitectura U-Net para modelos de difusión

Como se ha visto hasta el momento, para aprender $p_\theta(x_{t-1}|x_t) \sim \mathcal{N}(\mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$ se suele entrenar un modelo paramétrico que aprenda directamente μ_q , la muestra original x_0 o el ruido ϵ . En cualquiera de estos casos, el modelo es una función de la forma $\mathbb{R}^d \times [0, T] \rightarrow \mathbb{R}^d$. Si bien, a priori, es posible utilizar cualquier arquitectura neuronal con estas características, al trabajar con imágenes se suele utilizar una modificación de la arquitectura U-Net [RFB15], cuya arquitectura original se puede observar en la Figura 1.12.

En esta subsección se verán las modificaciones usuales que se aplican sobre la arquitectura U-Net para adaptarla a modelos de difusión. En el archivo `diffusion_unet.ipynb` se encuentra una implementación detallada de esta arquitectura neuronal junto a las componentes de atención. Es importante destacar que se han propuesto distintas variantes para esta arquitectura, donde el factor común entre ellas son los bloques que las componen. Las principales características son las siguientes:

- Se utiliza una red completamente convolucional, donde la parte descendente de la U (ver Figura 1.12) reduce la resolución a medida que aumenta la cantidad de canales, mientras que la parte ascendente de la U realiza el proceso reverso utilizando convoluciones transpuestas.
- En la parte ascendente se realizan conexiones residuales (*skip-connections*) usando los bloques homólogos de la parte descendente.

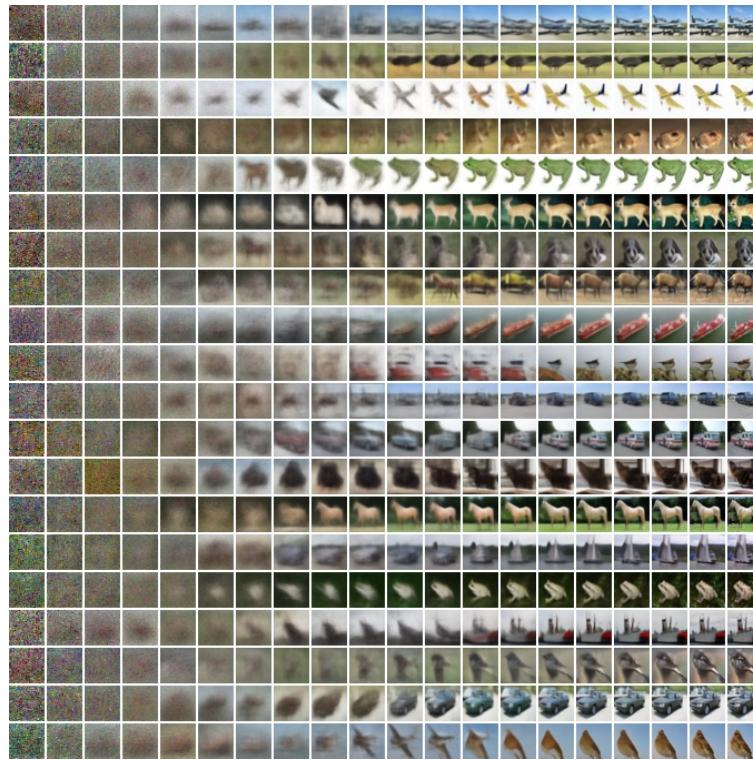


Figura 1.10: Proceso de generación de muestras para un modelo entrenado en CIFAR-10. Cada fila es una muestra distinta y cada columna muestra la predicción actual en un tiempo t . Imagen obtenida desde [HJA20].

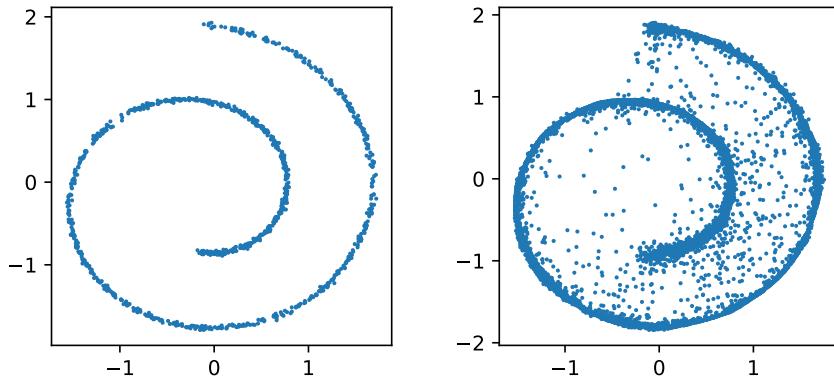


Figura 1.11: (Izquierda) batch de entrenamiento para un modelo de difusión. (Derecha) muestras generadas por el modelo de difusión. La implementación de esta técnica se encuentra en el archivo `ddpm.ipynb`.

- El tiempo de denoising t es añadido al final de cada bloque convolucional. Para esto se utiliza un embedding temporal sinusoidal similar al usado en la arquitectura *Transformer* [Vas+23].
- Posterior a cada bloque convolucional, luego de sumar el embedding temporal y antes de la conexión residual, se aplica una capa de *multi-head self-attention* similar a la usada en la arquitectura de *Vision Transformer* (ViT) [Dos+21].

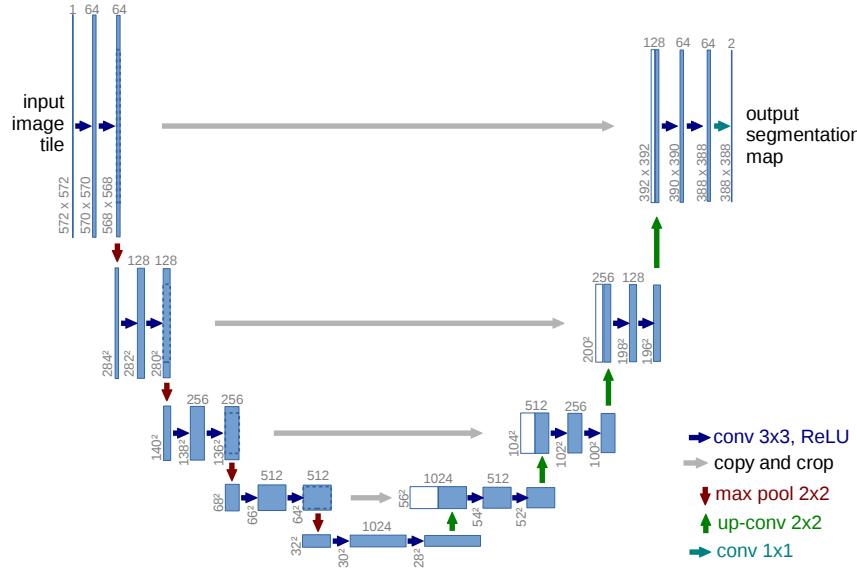


Figura 1.12: Arquitectura U-Net original para segmentación de imágenes. Imagen obtenida desde [RFB15].

- Como función de activación se suele usar GELU [HG23] y SiLU [EUD17], mientras que se suele aplicar normalización por grupos [Haf18] luego de cada convolución.

Una visualización de esta arquitectura modificada puede verse en Figura 1.13. En [ND21] y [DN21] se proponen algunas mejoras adicionales sobre esta arquitectura. En particular, los autores de [DN21] proponen los siguientes cambios con respecto a la arquitectura usada en el modelo original de DDPM:

- Aumentar la profundidad a cambio de ancho, manteniendo una cantidad de parámetros similar.
- Aumentar el número de cabezales de atención.
- Usar capas de atención en múltiples resoluciones de la U-Net.
- Usar los bloques residuales usados en la BigGAN [BDS19].
- Normalizar las conexiones residuales por $\frac{1}{\sqrt{2}}$.

Al hacer estos cambios notaron que, si bien el intercambio de ancho por profundidad mejoraba el rendimiento del modelo, también aumentaba el tiempo de entrenamiento necesario para alcanzar el mismo rendimiento en un modelo más ancho. Además, notaron que al aumentar el número de cabezales mejoraba considerablemente el FID¹⁸. De todos modos, es importante destacar que prácticamente todos los modelos tienen diferentes implementaciones de las arquitecturas U-Net y no existe una arquitectura por defecto (aunque es usual entrenar este modelo usando Adam y *exponential moving average* (EMA) sobre los valores de los parámetros).

Por otra parte, en el último tiempo se ha cambiado el uso de la U-Net por una arquitectura más moderna basada en transformers (*diffusion Transformer* o DiT) [PX23], la cual está basada en la arquitectura de *Vision Transformer* (ViT) [Dos+21]. En el archivo `dit.ipynb` hay una implementación minimal de esta arquitectura.

¹⁸El FID (*Fréchet inception distance*) es una métrica que mide la similitud entre la distribución real de los datos y la distribución aprendida por un modelo generativo.

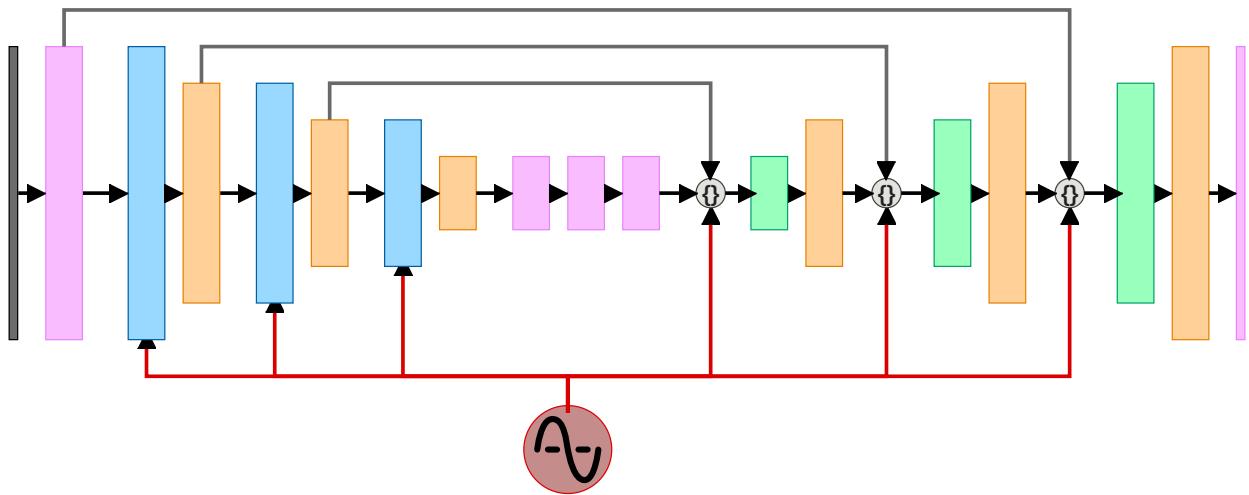


Figura 1.13: Arquitectura U-Net usada para modelos de difusión. El bloque gris representa el input (imagen RGB) mientras que los bloques rosados son bloques convolucionales. Los bloques celestes y verdes son bloques de *downsampling* y *upsampling* respectivamente. Los bloques naranjos son bloques de atención. Por último, las flechas grises son las conexiones residuales. Imagen obtenida desde [Erd23].

1.3.3. Mejoras al modelo DDPM

En esta subsección se entregarán algunas variantes importantes del modelo propuesto por [HJA20], las cuales han permitido mejorar aún más la capacidad de los modelos de difusión.

IDDPM

En [ND21], investigadores de OpenAI realizaron un estudio conocido como *Improved DDPM* donde muestran que aprender la varianza $\Sigma_\theta(x_t, t)$ del proceso backward $p_\theta(x_{t-1}|x_t)$ en (1.3.6) y utilizar un *variance scheduler* coseno mejora el proceso de inferencia y aumenta la verosimilitud. Además, dan un primer indicio de una *neural scaling law* para DDPM.

Varianza $\Sigma_\theta(x_t, t)$ aprendible Como se vio anteriormente, el modelo original de DDPM [HJA20] fija la varianza $\Sigma_\theta(x_t, t)$ del modelo reverso $p_\theta(x_{t-1}|x_t)$ a ser isotrópica: $\Sigma_\theta(x_t, t) = \sigma_\theta^2(x_t, t) I_d$, donde para $\sigma_\theta^2(x_t, t)$ se prueban dos opciones, ambas independientes de x_t :

- $\sigma_\theta^2(x_t, t) = \sigma_q^2(t)$ de acuerdo a la varianza del proceso reverso condicional a x_0 , $q(x_{t-1}|x_t, x_0)$ (ver (1.3.7)). Fijar esta covarianza permite simplificar la divergencia $D_{KL}(q(x_{t-1}|x_t, x_0) \| p_\theta(x_{t-1}|x_t))$ en la función de costo (1.3.8), llegando a un problema de mínimos cuadrados entre las medias de ambas distribuciones.
- $\sigma_\theta^2(x_t, t) = 1 - \alpha_t$ de acuerdo a la varianza del proceso forward $q(x_t|x_{t-1})$ (ver (1.3.1)).

Dado que ninguna de las dos varianzas es la verdadera, en [ND21] deciden modelar $\sigma_\theta^2(x_t, t)$ como una interpolación de estos dos valores en el espacio logarítmico:

$$\sigma_\theta^2(x_t, t) := \exp(v_\theta(x_t, t) \log(1 - \alpha_t) + (1 - v_\theta(x_t, t)) \log \sigma_q^2(t)),$$

donde $v_\theta(x_t, t)$ es un modelo neuronal. A priori, la varianza aprendida $\sigma_\theta^2(x_t, t)$ podría estar fuera del intervalo formado por $\sigma_q^2(t)$ y $1 - \alpha_t$, pero eso no ocurre en la práctica, mostrando que interpolar entre estos valores

es una buena parametrización.

Sampling con menos iteraciones Dado que el proceso forward por lo general tiene muchas iteraciones ($T = 1000$ en DDPM), el proceso reverso tendrá la misma cantidad de iteraciones, haciendo que la generación sea un proceso lento. En [ND21] descubrieron de forma inesperada que al permitir entrenar la varianza $\Sigma_\theta(x_t, t)$ es posible obtener muestras a partir de un modelo ya entrenado usando menos pasos de los que se usaron durante el entrenamiento y sin la necesidad de hacer *fine-tuning*. Las muestras que se generan de esta forma siguen siendo de alta calidad y permite generar imágenes en segundos en vez de minutos.

Para disminuir la cantidad de iteraciones durante la generación, consideraron una sub-secuencia decreciente de tiempos $S = (S_t)_{t=1}^{|S|} \subset \{0, \dots, T\}$ con $\{0, T\} \in S$. Dado que $v_\theta(x_{S_t}, S_t)$ es el parámetro que define $\sigma_\theta(x_{S_t}, S_t)$ mediante interpolación de las varianzas $1 - \alpha_{S_t}$ y σ_q^2 , este parámetro será invariante al reescalar los rangos en procesos de difusión más cortos, por lo que se puede hacer el proceso reverso utilizando únicamente los tiempos de S y generando desde $p_\theta(x_{S_{t-1}} | x_{S_t})$ mediante $\mathcal{N}(\mu_\theta(x_{S_t}, S_t), \Sigma_\theta(x_{S_t}, S_t))$.

Con esta técnica, se puede mantener la calidad de generación incluso tomando subconjuntos S de largo 100. Al intentar hacer esto mismo en el modelo original de DDPM hay una pérdida considerable de calidad al disminuir el número de iteraciones.

Función objetivo híbrida La función objetivo L_{simple} (ver (1.3.12)) usada en DDPM solo es válida cuando se fija $\Sigma_\theta(x_t, t) = \sigma_q^2(t) I_d$. Para un modelo que aprende la varianza, se deben volver a computar las expresiones para las divergencias en la ELBO (ver (1.3.8)). Sin embargo, en [ND21] obtienen dificultades para optimizar esta nueva función objetivo y observan que la media del proceso reverso, μ_θ , es más importante que la varianza Σ_θ . Por lo tanto, proponen mantener la función simplificada L_{simple} (que estima bien la media) y combinarla cónicamente con (el negativo de) la ELBO real:

$$L_{\text{híbrido}} := L_{\text{simple}} - \lambda \cdot \text{ELBO},$$

donde $\lambda > 0$ es un ponderador y la ELBO de un DM está dada en (1.3.8).

Por otra parte, si bien es esperable alcanzar mayores verosimilitudes optimizando directamente la ELBO, los autores observan que se obtienen mejores resultados optimizando esta función objetivo híbrida.

Cosine scheduler Otra observación importante es que al usar un *variance scheduler* lineal como en [HJA20] la difusión ocurre muy rápido, destruyendo una gran cantidad de información al comienzo del proceso. En la Figura 1.14 (arriba) se puede observar este fenómeno. Para corregir este problema, proponen un scheduler para $(\bar{\alpha}_t)_{t=1}^T$ ¹⁹ que inyecte el ruido más lento:

$$\bar{\alpha}_t = \frac{f(t)}{f(0)}, \quad f(t) = \cos^2 \left(\frac{\frac{t}{T} + s}{1+s} \cdot \frac{\pi}{2} \right),$$

donde $s > 0$ (de *smoothing*) es un parámetro de suavizado que permite que la inyección de ruido ocurra de forma más suave al comienzo. Este funcional de ruido tiene un comportamiento lineal en el centro, mientras que en los extremos del intervalo $[0, T]$ el decaimiento es más suave. Es importante notar que el nivel de corrupción visual obtenido en una muestra de $q(x_t | x_0)$ (para un cierto nivel de ruido $1 - \bar{\alpha}_t$) depende de la resolución de la imagen, por lo que este scheduler puede dejar de ser útil en otras resoluciones. En [Che23] estudian más en detalle el efecto del *scheduler* en el proceso de difusión.

¹⁹Esto ya que al entrenar usando (1.10), α_t es un parámetro más natural para definir que α_t . Sin embargo, ambas formas son equivalentes ya que $\alpha_t = \frac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}}$.



Figura 1.14: (arriba) Proceso de difusión asociado a un scheduler lineal. (abajo) Proceso de difusión asociado a un scheduler coseno. Imagen obtenida desde [ND21].

Neural scaling law para los modelos de difusión En los últimos años se ha visto que algunos modelos basados en redes neuronales siguen un patrón claro al comparar el rendimiento del modelo con el número de parámetros y la cantidad de datos. Por ejemplo, en [Kap+20] estudian cómo cambia la entropía cruzada en un *large language model* (LLM) cuando se utiliza una arquitectura tipo *transformer* y encuentran relaciones en forma de potencia entre esta y el número de parámetros, el tamaño del dataset y la capacidad de cómputo del entrenamiento. En [FKW22] generalizan este resultado al estudiar el transformer como modelo autorregresivo en otras modalidades como imagen o video.

Para modelos de difusión, en [ND21] notaron que, usando la función objetivo $L_{\text{híbrido}}$, el modelo DDPM mejora de forma predecible al aumentar el cómputo de entrenamiento. Más específicamente, observan que el FID escala siguiendo una ley de potencia, igual que para la entropía cruzada en un LLM. Para la log-verosimilitud, no se observó un patrón claro, sugiriendo que el aumento de cómputo durante el entrenamiento no favorece tan notoriamente al aumento de la verosimilitud. Es importante mencionar que tener una ley que prediga el rendimiento de una familia de modelos generativos es una propiedad sumamente valiosa, lo que puede verse como otra motivación para usar las buenas técnicas empleadas en los modelos de difusión en enfoques más generales como el problema puente de Schrödinger estudiado en el Capítulo 3.

Procesos de difusión no markovianos

Como se mencionó anteriormente, una de las principales desventajas de los modelos de difusión es su velocidad para generar muestras. Dado que el proceso de denoising está compuesto por varias etapas, la generación de muestras es mucho más lenta que la generación con GANs.

En un trabajo paralelo a iDDPM, los autores de [SME22] notaron que durante el cálculo de la función de pérdida del modelo DDPM, el proceso (markoviano) forward solo es usado a través de las marginales condicionales $q(x_t|x_0)$, $t \in \{1, \dots, T\}$ y no a través de la distribución conjunta $q(x_{1:T}|x_0)$ (ver Algoritmo 2). Con esta observación, en [SME22] definieron una nueva familia de procesos forward, $\mathcal{Q} = \{q_\sigma(x_{1:T}|x_0)\}_{\sigma \in \mathbb{R}_+^T}$, tales que sus marginales condicionales $q_\sigma(x_t|x_0)$ coincidieran con las del modelo DDPM, es decir, tales que

$$q_\sigma(x_t|x_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)I_d).$$

Los procesos forward de la familia \mathcal{Q} son dados de forma implícita mediante la imposición de la forma que deben tener sus procesos reversos:

$$q_\sigma(x_{t-1}|x_t, x_0) \sim \mathcal{N}(f_{\sigma_t}(x_0, x_t), \sigma_t^2 I_d),$$

donde $f_{\sigma_t} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ es una función con una forma precisa (conocida) para que se logre la igualdad de las marginales condicionales $q_\sigma(x_t|x_0)$ que se mencionó más arriba. Notar que no es necesario conocer estos procesos forward ya el enfoque DDIM se utiliza cuando el modelo ya está entrenado.

En este trabajo argumentan que, dado que el nuevo proceso forward es indistinguible del proceso forward de DDPM durante el cálculo de la función de pérdida (ya que $q_\sigma(x_t|x_0)$ y $q(x_t|x_0)$ coinciden), entonces el modelo entrenado sirve tanto para el denoising de DDPM como para el denoising de este nuevo proceso de difusión (más específicamente, ellos muestran una equivalencia en el problema de optimización bajo ciertas hipótesis razonables). Por lo tanto, se puede usar la red neuronal ya entrenada en DDPM como si hubiese sido entrenada para este nuevo proceso.

De esta forma, considerando que la red neuronal $\epsilon_\theta(x_t, t)$ busca predecir el ruido ϵ inyectado a x_0 en la etapa t , el proceso de generación de muestras que ellos proponen se puede ver en el Algoritmo 4.

Algoritmo 4 Generación usando DDIM

Require: modelo ϵ_θ (entrenado), secuencia $(\alpha_t)_{t=1}^T$, subsecuencia temporal $\tau = (\tau_1, \dots, \tau_S)$ donde $\tau_S = T$.

- 1: Generar $x_T \sim q_{\text{prior}}(x_T)$
- 2: **for** t desde S hasta 1 **do**
- 3: Estimar x_0 : $\hat{x}_0 \leftarrow \frac{x_{\tau_t} - \sqrt{1 - \bar{\alpha}_{\tau_t}} \epsilon_\theta(x_{\tau_t}, \tau_t)}{\sqrt{\bar{\alpha}_{\tau_t}}}$
- 4: **if** $t > 1$ **then**
- 5: $x_{\tau_{t-1}} \leftarrow \sqrt{\bar{\alpha}_{\tau_{t-1}}} \hat{x}_0 + \sqrt{1 - \bar{\alpha}_{\tau_{t-1}} - \sigma_t^2} \cdot \frac{x_{\tau_t} - \sqrt{\bar{\alpha}_{\tau_t}} \hat{x}_0}{\sqrt{1 - \bar{\alpha}_{\tau_t}}}$
- 6: **end if**
- 7: **end for**
- 8: **return** $x_0 \leftarrow \hat{x}_0$

La formulación de $q_\sigma(x_{t-1}|x_t, x_0)$ permite definir diferentes procesos de denoising (y respectivos procesos forward). Más aún, eligiendo un cierto parámetro de proceso $\sigma \in \mathbb{R}_+^d$ es posible hacer el proceso markoviano y recuperar la formulación original de DDPM. Sin embargo, el modelo de interés se consigue en el caso límite cuando $\sigma_t \rightarrow 0$, $\forall t \in \{1, \dots, T\}$ ya que de este modo, el proceso reverso se vuelve determinista. Este es el modelo que en el paper denominan DDIM. Este determinismo permite, al menos en teoría, obtener directamente x_0 a partir de x_T . Si bien esto no funciona bien en la práctica debido a la alta complejidad del proceso de difusión, es posible realizar el proceso reverso en menos pasos bajo la justificación de que x_T define totalmente x_0 desde un comienzo.

Por otra parte, este determinismo está íntimamente relacionado con la *probability flow ODE* de los modelos de difusión en tiempo continuo (ver Subsección 1.4.2), la cual a su vez es usada para mostrar que los modelos de difusión aprenden (o al menos aproximan) un mapa de transporte óptimo entre la distribución de los datos y la distribución prior, lo cual es estudiado en la Subsección 2.3.4. En consecuencia, esta es una primera conexión directa con el problema estático del puente de Schrödinger estudiado en la Sección 3.2.

Generación condicional

Es posible modificar el modelo de difusión estudiado hasta ahora para que funcione como un modelo generativo condicional. Una primera opción para hacer esto es entrenar un modelo de difusión que aprenda una distribución condicional $p_{\text{data}}(x|y)$, donde y es una entrada adicional al modelo que indica información acerca de la muestra x que se está generando. Por ejemplo, y podría ser una clase o una descripción de texto de x , la cual se puede inyectar al modelo utilizando un embedding vectorial de y concatenado al embedding temporal de t . En el caso que y sea un prompt de texto, el embedding se puede obtener mediante un modelo tipo BERT (ver [Dev+19]) o tipo CLIP (ver [Rad+21]).

Si bien esta es una técnica usual que se puede emplear en otros modelos como GANs (Subsección 1.1.1) y VAEs (Sección 1.2), en esta sección se revisará una técnica de condicionamiento llamada *guidance* que es específica para los modelos de difusión (o más generalmente, para los modelos basados en score).

Classifier-guidance Para introducir la técnica de *guidance*, se usará un resultado posterior que afirma que los modelos de difusión (que buscan aprender la función de media $\mu_q(x_t, x_0, t)$) se pueden reparametrizar por otro modelo $s_\theta(x_t, t)$ que aprenda la función de score $\nabla_{x_t} \log q(x_t)$. En efecto, de acuerdo a la Proposición 1.18 se puede reparametrizar el modelo como

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} x_t - \frac{1 - \alpha_t}{\sqrt{\alpha_t}} s_\theta(x_t, t).$$

Esta propiedad permite trabajar directamente con la función de score de un modelo de difusión condicional, el cual busca aprender, para cada tiempo $t \in \{1, \dots, T\}$, una función de score $\nabla_{x_t} \log p_\theta(x_t|y)$ para luego poder generar muestras desde $p_{\text{data}}(x|y)$. Para conseguir esto es importante notar lo siguiente:

$$\begin{aligned} p_\theta(x_t|y) &= \frac{p_\theta(y|x_t)p_\theta(x_t)}{p_\theta(y)} \implies \log p_\theta(x_t|y) = \log p_\theta(y|x_t) + \log p_\theta(x_t) - \log p_\theta(y) \\ &\implies \nabla_{x_t} \log p_\theta(x_t|y) = \nabla_{x_t} \log p_\theta(y|x_t) + \nabla_{x_t} \log p_\theta(x_t). \end{aligned} \quad (1.3.13)$$

En consecuencia, si $\nabla_{x_t} \log p_\theta(x_t)$ es un modelo de difusión incondicional ya entrenado (el cual se obtiene mediante (1.3.3)), bastaría con tener un modelo discriminativo $p_\theta(y|x_t)$ para poder generar muestras desde $p_{\text{data}}(x|y)$ a partir del modelo de difusión incondicional. Notar que la distribución $p_\theta(y|x_t)$ corresponde a la de un clasificador que es entrenado sobre pares (x_t, y) para cada tiempo $t \in \{1, \dots, T\}$, lo que motiva el nombre *classifier-guidance* para esta técnica. Notar que la condición y no necesariamente debe tener naturaleza categórica, si no que puede ser un texto, donde $p(y|x)$ suele ser calculado por un modelo tipo CLIP [Rad+21].

Por otra parte, notar que la cantidad $\nabla_{x_t} \log p_\theta(y|x_t)$ no es una función de score ya que el gradiente es tomado con respecto a x_t y no y . Sin embargo, $p(y|x_t)$ sí es una densidad de probabilidad, por lo que se puede aplicar la idea de temperatura usada en los modelos de lenguaje²⁰, donde $p_\theta(y|x_t)$ es cambiada por una densidad proporcional a $p(y|x_t)^\gamma$ con $\gamma > 0$. Con este cambio, se puede obtener el score para una nueva distribución $\tilde{p}_\theta(x_t|y)$ (dependiente de γ), diferente a la distribución $p_\theta(x_t|y)$ en (1.3.13):

$$\tilde{p}_\theta(x_t|y) \propto p_\theta(y|x_t)^\gamma p_\theta(x_t) \implies \nabla_{x_t} \log \tilde{p}_\theta(x_t|y) = \gamma \nabla_{x_t} \log p_\theta(y|x_t) + \nabla_{x_t} \log p_\theta(x_t). \quad (1.3.14)$$

Es decir, utilizando este nuevo score condicional es posible generar muestras de la distribución $p_{\text{data}}(x|y)$, pero asignándole un peso relativo a la señal y mediante el ponderador γ . En la Figura 1.15 se puede observar el efecto de aumentar el *guidance scale* γ .

Notar que este factor, al igual como ocurre en el modelamiento del lenguaje, implica un trade-off entre calidad y diversidad ya que al aumentar el parámetro γ , se están concentrando aún más las modas del modelo discriminativo $p_\theta(y|x_t)$. Esto se puede observar en la Figura 1.16 y en la Figura 1.17.

Classifier-free guidance Una limitación del enfoque anterior es que se vuelve necesario entrenar un clasificador $p_\theta(x_t|y)$ para cada tiempo $t \in \{0, \dots, T\}$ del proceso de difusión ya que, al ser este un modelo discriminativo, nada garantiza poder clasificar bien para diferentes tiempos usando un único clasificador entrenado para $t = 0$. Hacer este entrenamiento puede ser extremadamente costoso, por lo que en [HS22] proponen una alternativa que no requiere entrenar un clasificador separado, si no que lo entrena a medida que se entrena el modelo de difusión. Para esto, vuelven a aplicar la regla de Bayes, pero ahora sobre el clasificador $p_\theta(y|x_t)$:

²⁰Esta técnica consiste en aumentar ($\gamma > 1$) o aminorar ($0 < \gamma < 1$) la concentración de la masa en las modas de una distribución de probabilidad elevando la función densidad a una potencia γ . En rigor, el parámetro de temperatura es $\frac{1}{\gamma}$, por lo que γ se conoce como temperatura inversa.

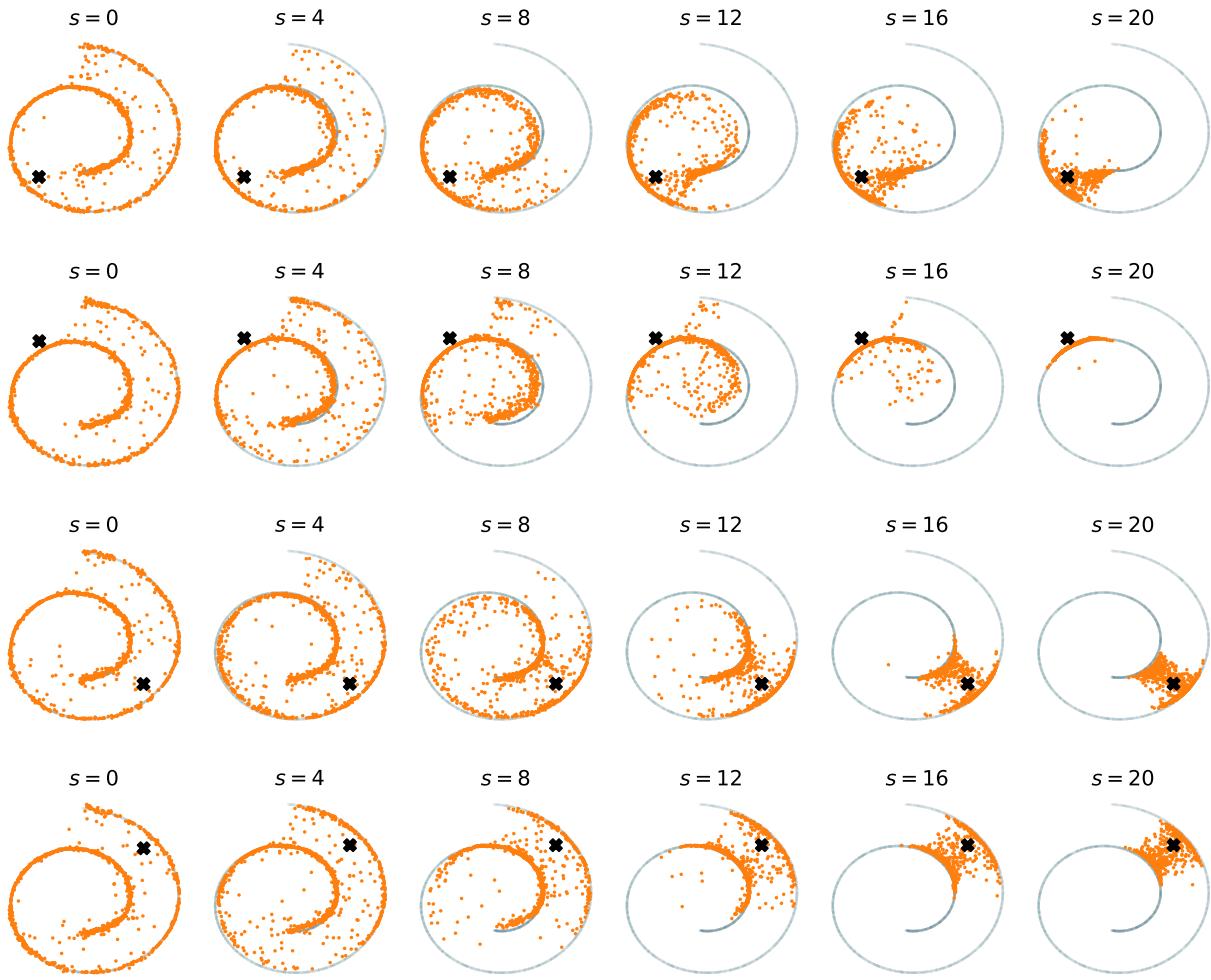


Figura 1.15: Desviación de la distribución aprendida por el modelo incondicional $p_\theta(x)$ según la escala de *guidance* (aquí denotado por s). El modelo discriminativo $p_\theta(y|x)$ le da un alto valor a puntos cercanos a la cruz marcada en negro. La implementación de esta técnica se encuentra en el archivo `ddpm.ipynb`.

$$p_\theta(y|x_t) = \frac{p_\theta(x_t|y)p_\theta(y)}{p_\theta(x_t)} \implies \nabla_{x_t} \log p_\theta(y|x_t) = \nabla_{x_t} \log p_\theta(x_t|y) - \nabla_{x_t} \log p_\theta(x_t). \quad (1.3.15)$$

Logrando obtener el término discriminativo necesario en (1.3.13). Otra observación que realizan en [HS22] es que todos los términos al lado derecho de (1.3.15) se pueden obtener si se entrena un modelo de difusión condicional $p_\theta(x|y)$ que permita entregar condiciones vacías para realizar generaciones incondicionales $p_\theta(x)$. Para esto:

- Se entrena un modelo de difusión condicional $p_\theta(x_t|y)$, donde y es una entrada adicional a la red neuronal.
- Durante el entrenamiento, de forma aleatoria se eligen iteraciones donde se entrena el modelo en modo incondicional. Esto es, se entrena $p_\theta(x|y)$ considerando como condición a un token especial $y = \emptyset$ que representa la ausencia de condición, simulando el aprendizaje de un modelo incondicional $p_\theta(x)$.

De esta forma, durante un único entrenamiento de un modelo de difusión condicional, es posible entrenar

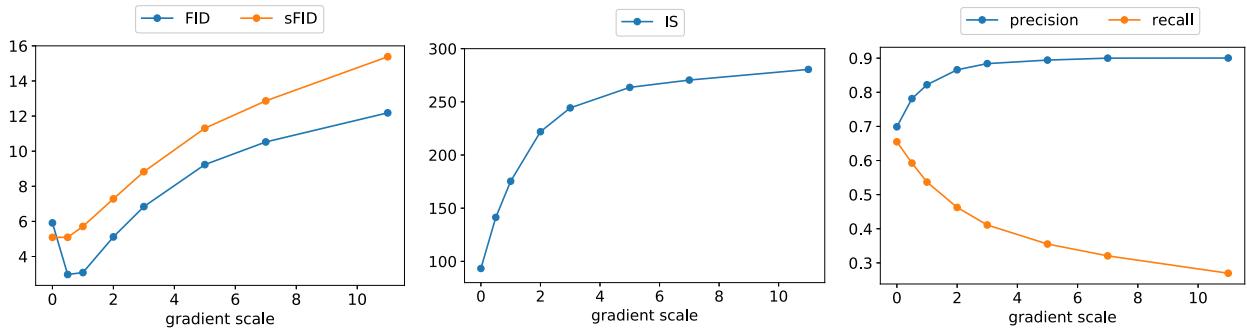


Figura 1.16: Trade-off entre calidad y variedad que se observa al cambiar el factor de escala del gradiente. Por un lado, al aumentar el factor aumenta el FID, IS y la precisión (indicadores asociados a la calidad), mientras que disminuye el recall (indicador asociado a la diversidad). Imagen obtenida desde [Nic+22].

un modelo incondicional $p_\theta(x)$, lo que permite obtener la expresión $\nabla_{x_t} \log p_\theta(y|x_t)$ necesaria en classifier-guidance sin haber entrenado un clasificador independiente. Por este motivo, esta técnica recibe el nombre de *classifier-free guidance*.

Por otra parte, dado que es necesario entrenar un generador condicional para aplicar esta técnica, podría parecer que no se obtienen beneficios al aplicar classifier-free guidance (técnica que busca precisamente obtener un generador condicional). Sin embargo, al descomponer el score condicional, es posible aplicar el concepto de temperatura usado en classifier-guidance. En efecto, sustituyendo el lado derecho de (1.3.15) en (1.3.14) se obtiene

$$\begin{aligned}\nabla_{x_t} \log \tilde{p}_\theta(x_t|y) &= \gamma \nabla_{x_t} \log p_\theta(y|x_t) + \nabla_{x_t} \log p_\theta(x_t) \\ &= \gamma (\nabla_{x_t} \log p_\theta(x_t|y) - \nabla_{x_t} \log p_\theta(x_t)) + \nabla_{x_t} \log p_\theta(x_t) \\ &= (1 - \gamma) \nabla_{x_t} \log p_\theta(x_t) + \gamma \nabla_{x_t} \log p_\theta(x_t|y).\end{aligned}$$

Es decir, el score modificado que se debe usar para la generación condicional es una combinación baricéntrica (pero no necesariamente convexa) entre el score incondicional y el score condicional. En consecuencia, variando el parámetro γ se obtienen 3 régimenes usuales:

- **Temperatura infinita ($\gamma = 0$):** se recupera el modelo incondicional $p_\theta(x)$ ya que la distribución condicional se vuelve uniforme (máxima entropía).
- **Temperatura unitaria ($\gamma = 1$):** se obtiene el modelo condicional estándar.
- **Temperatura baja ($\gamma > 1$):** la condición y toma más relevancia, pudiendo controlar la influencia del prompt. Esto se observa en la Figura 1.17.

Por otra parte, al entrenar el modelo discriminativo $p(y|x_t)$ utilizando el mismo modelo de difusión, se suele obtener un clasificador mucho más robusto, por lo que la técnica de classifier-free guidance entrega mejores resultados que classifier-guidance. Más aún, en el trabajo de [HS22] muestran que con esta técnica, los modelos de difusión son estrictamente superiores que las GANs en cuanto a calidad y versatilidad de las muestras.

Difusión en el espacio latente

Una subfamilia importante dentro del grupo de los modelos de difusión son los *latent diffusion models* (LDM). El trabajo más conocido en esta familia es *Stable Diffusion*, cuya versión inicial se encuentra en [Rom+22].



Figura 1.17: Visualización del trade-off entre calidad y variedad al usar guidance. (Izquierda) modelo GLIDE incondicional. (Derecha) modelo GLIDE con classifier-free guidance, escala $\gamma = 3,0$. Imagen obtenida desde [Nic+22].

Este trabajo es un proyecto colaborativo de código abierto cuyos pesos (modelos entrenados) están disponibles para la comunidad.

La principal característica de este modelo es que trabaja en el espacio latente de las imágenes. Para esto, en [Rom+22] primero entranan un VAE para luego entrenar un modelo de difusión en el espacio latente de dicho VAE. Con esto, se consigue un entrenamiento más eficiente y se libera la opción de poder escalar el modelo a mayores resoluciones. El diagrama de la arquitectura del modelo *Stable Diffusion* se puede ver en la Figura 1.18.

Este modelo busca un espacio latente de menor dimensión²¹ que sea perceptualmente equivalente al espacio de los datos mediante el uso de un VAE. A diferencia de propuestas anteriores, aquí se entrena este modelo primero, sin considerar un modelo de difusión en particular. Esto permite reutilizar el VAE para entrenar distintos modelos de difusión, permitiendo explorar una mayor cantidad de variantes. Además, con el fin de evitar varianzas extremadamente largas en el espacio latente, los autores propusieron dos regularizaciones al VAE, que hoy se conocen como *KL-reg* y *VQ-reg*.

Por otra parte, la arquitectura del modelo *Stable Diffusion* permite condicionar información externa y de forma nativa (i.e., entrenando un modelo condicional) para poder realizar classifier-free guidance. Para esto, el modelo introduce un encoder especializado τ_θ , para así representar el condicionamiento y mediante $\tau_\theta(y)$. Esta nueva representación es transmitida a cada capa de la U-Net mediante un mecanismo de *cross-attention* (ver [Vas+23]), por lo que es necesario usar representaciones secuenciales:

- $\mathcal{U}_i(z_t) \in \mathcal{M}_{N,d_u^i}(\mathbb{R})$ es una representación aplanada de la capa i de la U-Net que realizará atención sobre la condición $\tau_\theta(y)$. Por lo tanto, en esa capa de atención será la secuencia que actúa como *query*.
- $\tau_\theta(y) \in \mathcal{M}_{M,d_\tau}(\mathbb{R})$ actuará como el par *key-value*. En el caso de que y sea un texto, τ_θ es el encoder

²¹Al asumir que este espacio es de dimensión menor que la dimensión de los datos, implícitamente se está aceptando la *manifold assumption*. En [Rom+22] muestran empíricamente que la mayoría de los bits de una imagen digital corresponden a detalles imperceptibles.

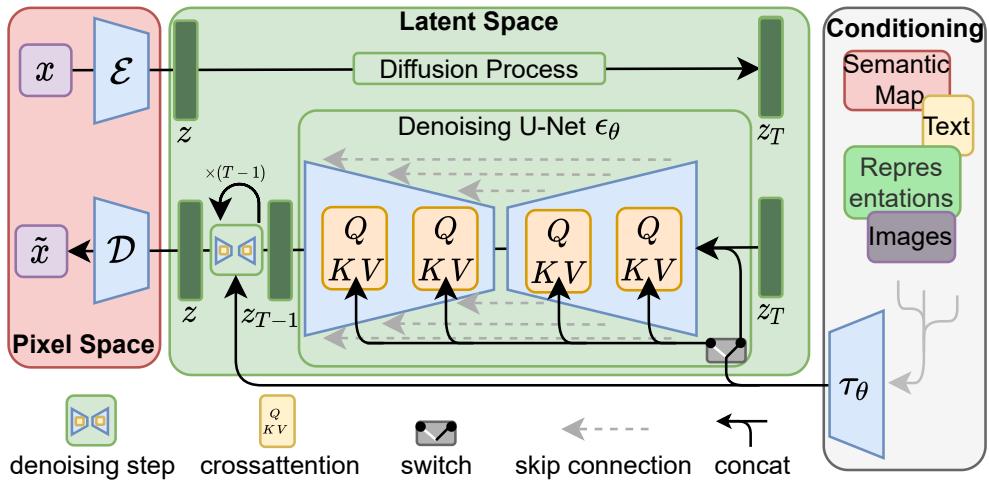


Figura 1.18: Arquitectura del modelo *Stable Diffusion*. El bloque rosado corresponde al VAE mientras que el bloque verde corresponde al modelo de difusión en el espacio latente. Para condicionar la generación, un embedding aprendido τ_θ es inyectado a la U-Net mediante un mecanismo de *cross-attention*. Imagen obtenida desde [Rom+22].

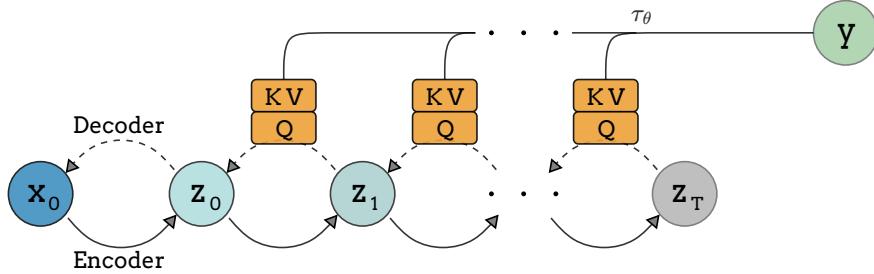


Figura 1.19: Mecanismo de condicionamiento en el modelo *Stable Diffusion*. La matriz τ_θ es obtenida a partir de y . Un mecanismo de atención cruzada entre las distintas capas de la red y τ_θ permite introducir la información de y en la red. Imagen obtenida desde [Sam22].

de un transformer, M es el largo de la secuencia y d_τ la dimensión de cada embedding. Es importante mencionar que el modelo τ_θ se entrena al mismo tiempo que el modelo de difusión (*Stable Diffusion* usa el enfoque ϵ -prediction).

En la Figura 1.19 se ve una ilustración de este mecanismo de atención. Condicionando sobre imágenes, los autores de [Rom+22] logran realizar síntesis semántica (creando imágenes realistas asociadas a un mapa semántico), super-resolución e *inpainting*²². En la Figura 1.20 se pueden ver ejemplos de estas técnicas.

1.4. Generalización a tiempo continuo

En esta sección se revisará una generalización del modelo de difusión propuesto por [HJA20], donde la cadena de Markov utilizada para los procesos de inyección de ruido y denoising será sustituida por un proceso estocástico a tiempo continuo. En consecuencia, en esta sección se utilizarán varios resultados acerca de ecua-

²²En términos generales, *inpainting* consiste en llenar regiones de una imagen siguiendo una máscara. Esto sirve para reparar imágenes corruptas o sustituir partes con contenido no deseado en la imagen.

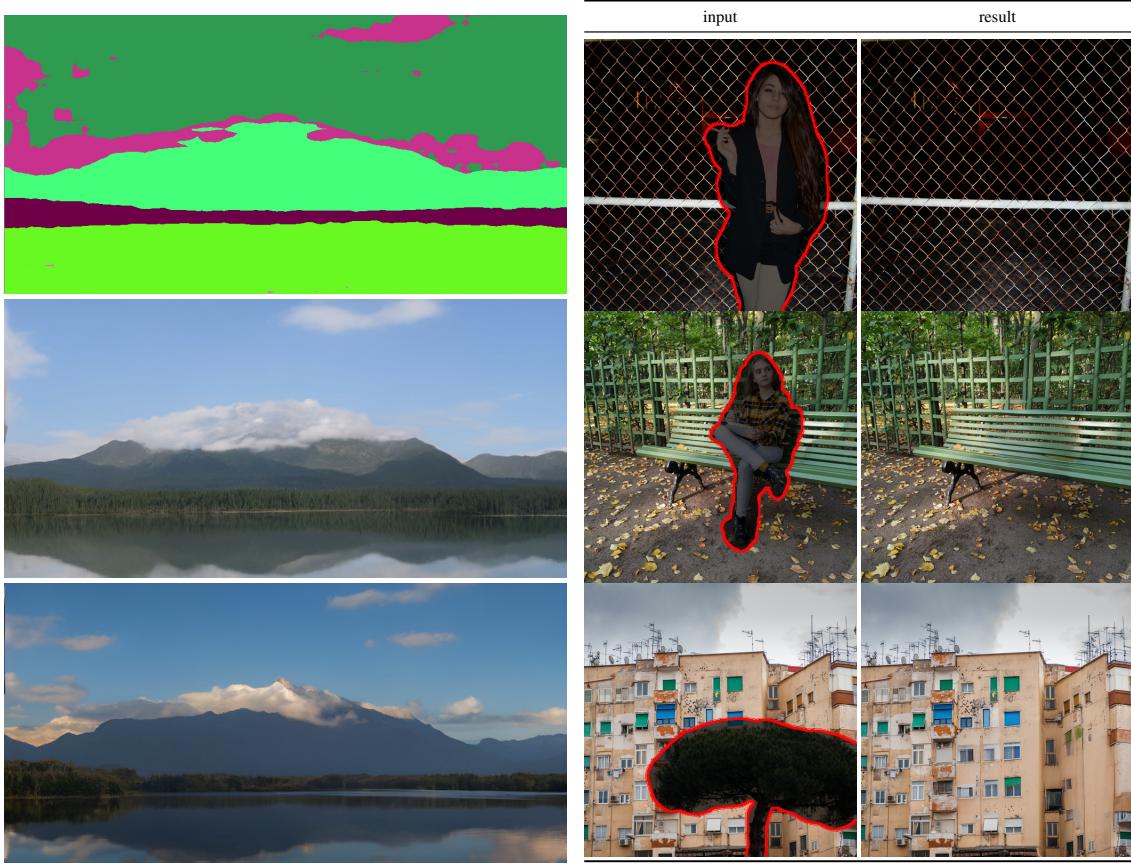


Figura 1.20: (izquierda) generación condicionada a un mapa semántico. (derecha) eliminación de objetos mediante *inpainting*. Imagen obtenida desde [Rom+22].

ciones diferenciales estocásticas (SDEs), los cuales se pueden encontrar en la Sección A.2. Esta formulación continua de los modelos de difusión será la que permitirá conectarlos con el problema del puente de Schrödinger en el Capítulo 3. Además, este nuevo enfoque permite trabajar con diferentes procesos de inyección de ruido sobre un mismo marco teórico. Por otro lado, en [Son+21b] notaron que, al ver el proceso de difusión en el continuo, es posible reducir la cantidad de iteraciones disminuyendo la precisión del *solver* que se utilice para simular la SDE o la *probability flow ODE* del proceso de generación.

Antes de introducir esta nueva formulación se revisará una formulación equivalente de los modelos de difusión a tiempo discreto donde se verá que aprender la media del proceso reverso es equivalente a aprender la derivada logarítmica de la distribución marginal $p_\theta(x_t)$ en un tiempo $t \in \{0, \dots, T\}$, lo que permitirá, además, conectar los modelos de difusión con otra familia de modelos que al comienzo parecían independientes. Esta propiedad permitirá, posteriormente, aprender un proceso reverso a tiempo continuo gracias a un teorema importante del cálculo estocástico conocido como *teorema de inversión de Anderson* (ver Teorema A.6).

1.4.1. Modelos generativos basados en score

En esta subsección se revisará otra familia de modelos generativos denominada *score-based models*. Si bien en un inicio este enfoque se propuso de manera independiente a los modelos de difusión, un trabajo posterior mostró que ambos paradigmas pueden ser vistos como un *modelo de difusión* más general cuya dinámica de

difusión y denoising viene dada por una SDE. El concepto principal de este tipo de modelos es el siguiente:

Definición 1.3 (función de score). Dada una distribución de probabilidad con densidad p , se define la *función de score*²³ de dicha distribución, $\text{score}_p : \mathbb{R}^d \rightarrow \mathbb{R}^d$, como la derivada logarítmica de su densidad:

$$\text{score}_p(x) := \nabla_x \log p(x).$$

Los modelos que buscan aprender la función de score en vez de aprender directamente la densidad p (que es lo que hacen los modelos basados en verosimilitud) se denominan *modelos basados en score* y pueden ser vistos como una variante de los modelos basados en energía [SK21]. Es importante notar que una función de score define totalmente una única función verosimilitud $\log p(x)$. En efecto, si $\nabla_x \log p_1(x) = \nabla_x \log p_2(x)$ entonces, por el teorema fundamental del cálculo, $\log p_1(x) = \log p_2(x) + c$, donde necesariamente $c = 0$ cuando p_1 y p_2 son densidades de probabilidad en \mathbb{R}^d , en efecto:

$$\log p_1(x) = \log p_2(x) + c \implies \int_{\mathbb{R}^d} e^{\log p_1(x)} dx = \int_{\mathbb{R}^d} e^{\log p_2(x)+c} dx \implies 1 = e^c \implies c = 0.$$

De esta forma, una función de score caracteriza totalmente a su densidad de probabilidad asociada.

Modelo de score matching

Para comenzar el estudio de los modelos basados en score, se revisará el modelo generativo de *score matching*, el cual consiste en aprender la función de score de una distribución p_{data} aprovechando el hecho de que los modelos basados en redes neuronales se pueden derivar fácilmente mediante diferenciación automática.

Dada una densidad desconocida p_{data} y un modelo paramétrico s_θ que busque aprender el score de p_{data} , la función objetivo natural en este caso es la divergencia de Fisher entre la densidad real p_{data} y la densidad aprendida a partir de s_θ , p_θ :

$$\begin{aligned} D_F(p_{\text{data}} \| p_\theta) &:= \frac{1}{2} \mathbb{E}_{x \sim p_{\text{data}}(x)} \left[\|\nabla_x \log p_{\text{data}}(x) - \nabla_x \log p_\theta(x)\|^2 \right] \\ &= \frac{1}{2} \mathbb{E}_{x \sim p_{\text{data}}(x)} \left[\|\nabla_x \log p_{\text{data}}(x) - s_\theta(x)\|^2 \right], \end{aligned}$$

donde p_θ es la densidad asociada al score s_θ . Al igual que antes, la esperanza es aproximada según la distribución empírica asociada a un conjunto de muestras de entrenamiento $(x^k)_{k=1}^n$ proveniente de p_{data} . Si bien el score real $\nabla_x \log p_{\text{data}}(x)$ no es conocido, en [Hyv05] prueban que la divergencia de Fisher es equivalente al siguiente objetivo:

Proposición 1.15 (divergencia de Fisher para score matching). Dada una densidad desconocida p_{data} y otra densidad p_θ con función de score s_θ conocida, se tiene que:

$$D_F(p_{\text{data}} \| p_\theta) = \mathbb{E}_{x \sim p_{\text{data}}(x)} \left[\frac{1}{2} \|s_\theta(x)\|^2 + \text{Tr}(\nabla_x s_\theta(x)) \right] + \text{constante},$$

donde $\text{Tr}(\cdot)$ es el operador traza y $\nabla_x s_\theta(x) \in \mathcal{M}_{d,d}(\mathbb{R})$ es la matriz jacobiana de s_θ (evaluada en x).

²³A esta función también se le denomina *score de Hyvärinen* o *score de Stein*, para diferenciarla de la función score usada en la definición de la *información de Fisher*. En esta última, el gradiente es tomado con respecto a los parámetros del modelo, mientras que aquí se está considerando el gradiente con respecto a la entrada.

De esta forma se obtiene una función de pérdida tratable²⁴, lo que permite aprender la función de score de p_{data} dado un conjunto de muestras para el entrenamiento. Este método se conoce en la literatura como *score-matching* (SM).

Una vez aprendido un modelo de score, lo único que queda pendiente es ver cómo generar muestras a partir de p_{θ} teniendo únicamente la función de score s_{θ} . Para esto, es usual usar un algoritmo denominado *Langevin sampling* o *Langevin Monte Carlo* (LMC), donde este último nombre se debe a que este algoritmo puede interpretarse como un caso particular de una variante del método de MCMC, denominada *Hamiltonian Monte Carlo* (HMC). Para estas variantes ver, por ejemplo, [Mur23].

La idea del algoritmo de Langevin sampling es simular un proceso dinámico aleatorio específico, $x = (x_t)_{t \geq 0}$, del cual se sabe que para tiempos de simulación $t \gg 1$ muy largos, la distribución marginal de x_t , $p(x_t)$, es precisamente la distribución que se quiere simular. Un proceso que tiene esta propiedad es conocido como dinámica de Langevin, cuya SDE (ver Sección A.2) es

$$dx_t = \frac{1}{2} \nabla_{x_t} \log \pi(x_t) dt + dw, \quad (1.4.1)$$

donde $\pi(x)$ es alguna función de densidad de la que se quiere generar muestras. La siguiente proposición indica que este proceso estocástico tiene la propiedad buscada:

Proposición 1.16 (distribución estacionaria de la dinámica de Langevin). Si $p(x_t)$ denota la densidad marginal en tiempo t del proceso estocástico $(x_t)_{t \geq 1}$ que resuelve la SDE (1.4.1), entonces π es la distribución estacionaria del proceso estocástico x , es decir, $p(x_t) \rightarrow \pi(x_t)$ cuando $t \rightarrow \infty$.

Demostración. Por simplicidad, se demostrará el resultado en el caso unidimensional ($d = 1$). En este caso, la SDE asociada al proceso estocástico $x = (x_t)_{t \geq 1}$ toma la forma

$$dx_t = \frac{1}{2} (\log \pi)'(x_t) dt + dw. \quad (1.4.2)$$

Por otra parte, si $p(x, t) = p(x_t)$ es la densidad marginal en tiempo t , entonces $p(x, t)$ cumple ecuación de Fokker-Planck (ver Teorema A.4) asociada a la SDE (1.4.2). Denotando las derivadas como subíndices:

$$p_x = -\frac{\partial}{\partial x} \left(\frac{1}{2} (\log \pi)' p \right) + \frac{1}{2} \frac{\partial^2}{\partial x^2} (1^2 p) = -\frac{1}{2} ((\log \pi)'' p + (\log \pi)' p_x) + \frac{1}{2} p_{xx}.$$

Tomando $p_x = 0$ (donde la densidad marginal ya no cambia) se obtiene la SDE de la distribución estacionaria:

$$\begin{aligned} p_{xx} - (\log \pi)' p' - (\log \pi)'' p &= 0 \\ p_{xx} - \frac{\pi'}{\pi} p' - \frac{\pi'' \pi - (\pi')^2}{\pi^2} p &= 0. \end{aligned}$$

Se verá que la densidad $p(x, t) = \pi(x)$ satisface esta ecuación:

$$\pi_{xx} - \frac{\pi'}{\pi} \pi' - \frac{\pi'' \pi - (\pi')^2}{\pi^2} \pi = \frac{\pi'' \pi - (\pi')^2}{\pi^2} \pi - \frac{\pi'' \pi - (\pi')^2}{\pi^2} \pi = 0.$$

Por lo tanto, π es la distribución estacionaria de la dinámica de Langevin. \square

²⁴Dado que la función de score s_{θ} es modelada mediante una red neuronal, sus derivadas pueden ser encontradas mediante diferenciación automática. En el archivo `sgm.ipynb` se implementa un modelo con esta función objetivo.

El resultado anterior nos garantiza que simular el proceso estocástico asociado a la SDE (1.4.1) permite obtener una muestra desde la distribución π si la simulación se realiza hasta un tiempo $T \gg 1$ lo suficientemente grande²⁵. Aplicando este método para generar muestras desde una densidad p_θ de la cual se conoce su función de score $s_\theta = \log p_\theta$, se obtiene el procedimiento de generación de muestras para un método basado en score, el cual se conoce como Langevin sampling y se detalla en el Algoritmo 5, donde la SDE (1.4.1) es simulada utilizando el algoritmo de Euler-Maruyama (ver Algoritmo 7).

Algoritmo 5 Langevin sampling

Require: Distribución inicial p_0 , cantidad de iteraciones T y largo de paso ϵ .

Require: Función de score aprendida $s(x) = \nabla_x \log p(x)$

- 1: Generar iteración inicial $x_0 \sim p_0(x_0)$.
 - 2: **for** t desde 1 a hasta T **do**
 - 3: Generar $z_t \sim \mathcal{N}(0, I_d)$.
 - 4: $x_t \leftarrow x_{t-1} + \frac{\epsilon}{2} s(x_{t-1}) + \sqrt{\epsilon} z_t$
 - 5: **end for**
 - 6: **return** x_T
-

El Algoritmo 5 puede verse como un método de gradiente ascendente perturbado que dirige x_t hacia los máximos de p_θ . Al agregar aleatoriedad mediante el ruido z_t se está garantizando que $x = (x_t)_{t=1}^T$ no colapse hacia alguna de las modas (máximos locales) de p_θ . En el archivo `sgm.ipynb` se encuentra implementado este modelo para un dataset de juguete.

Score matching mediante la inyección de ruido

Una de las principales limitaciones del enfoque anterior es el tiempo cuadrático que toma computar el término $\text{Tr}(\mathbf{J}_x s_\theta(x))$ durante el entrenamiento en (1.15) ya que se vuelve intratable en altas dimensiones. Para evitar computar este término directamente, se han propuesto diferentes variantes de score-matching como *sliced score matching* (SSM²⁶) en [Son+19] y *Denoising score matching* (DSM) en [Vin11]. Este último enfoque realiza score matching sobre los datos perturbados por la inyección de un ruido aditivo $\tilde{x} = x + \epsilon$, estimando el score de $q(\tilde{x}) = \int q(\tilde{x}|x)p_{\text{data}}(x) dx$ en vez de estimar directamente el score de p_{data} . Con esto, en [Vin11] muestran que la divergencia de Fisher toma la siguiente forma:

Proposición 1.17 (divergencia de Fisher para DSM). Dado un modelo paramétrico p_θ con función de score s_θ conocida, si p_θ busca aproximar una densidad desconocida $q = \int q(\cdot|x)p_{\text{data}}(x) dx$, se tiene que:

$$\begin{aligned} D_F(q \| p_\theta) &= \frac{1}{2} \mathbb{E}_{\tilde{x} \sim q(\tilde{x})} \left[\|\nabla_{\tilde{x}} \log q(\tilde{x}) - s_\theta(\tilde{x})\|^2 \right] \\ &= \frac{1}{2} \mathbb{E}_{x \sim p_{\text{data}}(x), \tilde{x} \sim q(\tilde{x}|x)} \left[\|\nabla_{\tilde{x}} \log q(\tilde{x}|x) - s_\theta(\tilde{x})\|^2 \right] + \text{constante}. \end{aligned}$$

Dado que se considera un kernel de perturbación gaussiano $q(\tilde{x}|x) \sim \mathcal{N}(x, \sigma^2 I_d)$, el score $\nabla_{\tilde{x}} \log q(\tilde{x}|x)$ se puede obtener en forma cerrada y la función objetivo se simplifica a una cantidad tratable:

$$D_F(q \| p_\theta) = \frac{1}{2} \mathbb{E}_{x \sim p_{\text{data}}(x), \tilde{x} \sim \mathcal{N}(x, \sigma^2 I_d)} \left[\left\| \frac{x - \tilde{x}}{\sigma^2} - s_\theta(\tilde{x}) \right\|^2 \right]. \quad (1.4.3)$$

²⁵En rigor, la distribución estacionaria se alcanza en un horizonte de tiempo infinito, por lo que siempre se obtienen muestras de una distribución $p(x_t)$ cercana pero no igual a la distribución estacionaria. El problema del puente de Schrödinger estudiado en el Capítulo 3 soluciona este problema.

²⁶No confundir con los *state space models* [GGR22], los cuales no tienen relación directa con los modelos basados en score.

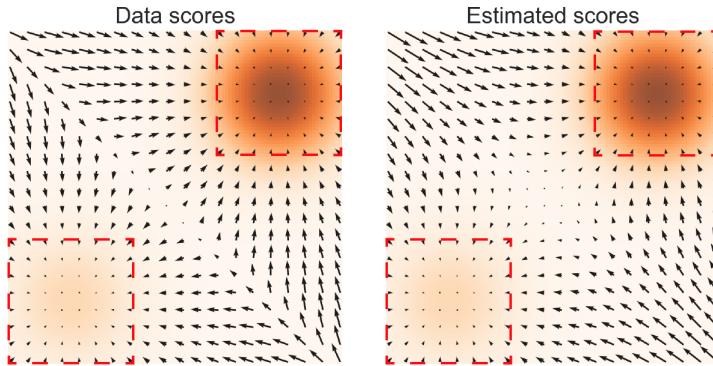


Figura 1.21: Campo vectorial de la función de score real (izquierda) y de la función de score aprendida (derecha) para una mixtura gaussiana. El mapa de calor está asociado a la densidad de la mixtura y los rectángulos muestran las zonas donde la predicción del score es cercana a la real. Se observa que la predicción solo es precisa alrededor de las modas de la mixtura.

De esta forma, estimando el score sobre los datos perturbados se evita la dificultad de computar $\text{Tr}(\mathbf{J}_x s_\theta(x))$ en (1.15) y, si bien el modelo está aprendiendo el score de $q(\tilde{x})$, se puede considerar que $q \approx p_{\text{data}}$ cuando σ^2 es suficientemente pequeño.

Esta técnica, conocida como *denoising score matching* (DSM), también soluciona otros problemas observados en el modelo SM original:

- Considerando la *manifold assumption*, el score $\nabla_x \log p_{\text{data}}(x)$ está indefinido en gran parte del dominio de p_{data} ya que el soporte de p_{data} está confinado en una variedad de dimensión estrictamente menor. Más aún, esta hipótesis implica que la medida de probabilidad real asociada a los datos de entrenamiento, μ_{true} , no debería poseer una función de densidad p_{data} (ver Subsección A.1.2). Al usar una perturbación gaussiana, el soporte vuelve a ser todo \mathbb{R}^d .
- Este kernel de perturbación asegura que la densidad condicional $q(\tilde{x}|x)$ sea diferenciable, lo cual no es necesariamente cierto al tratar directamente con p_{data} .

En el archivo `sgm.ipynb` se encuentra una implementación minimal de este tipo de modelos donde se puede comparar (1.4.3) con la función objetivo (1.15).

Por otra parte, si bien la técnica presentada en [Vin11] estabiliza el entrenamiento, un trabajo posterior muestra algunas de sus limitaciones y propone hacer DSM utilizando distintos niveles de ruido. En particular, los problemas adicionales que reconocen en [SE20] son los siguientes:

- Considerando que las muestras de entrenamiento fueron generadas directamente desde p_{data} , es altamente probable no tener muestras de entrenamiento en regiones de baja densidad. En consecuencia, dado que la esperanza de la divergencia de Fisher en (1.4.3) es aproximada mediante la distribución empírica, la función de score aprendida s_θ no será correcta en zonas de baja densidad. Esto puede observarse en la Figura 1.21.
- Dada una mixtura $p_{\text{mixture}} = \pi p_1 + (1 - \pi)p_2$, donde p_1 y p_2 son densidades con soporte disjunto, su score vendrá dado por:

$$\nabla_x \log p_{\text{mixture}}(x) = \begin{cases} \nabla_x(\log \pi + \log p_1(x)) = \nabla_x \log p_1(x), & \text{si } x \in \text{Supp}(p_1) \\ \nabla_x(\log(1 - \pi) + \log p_2(x)) = \nabla_x \log p_2(x), & \text{si } x \in \text{Supp}(p_2). \end{cases}$$

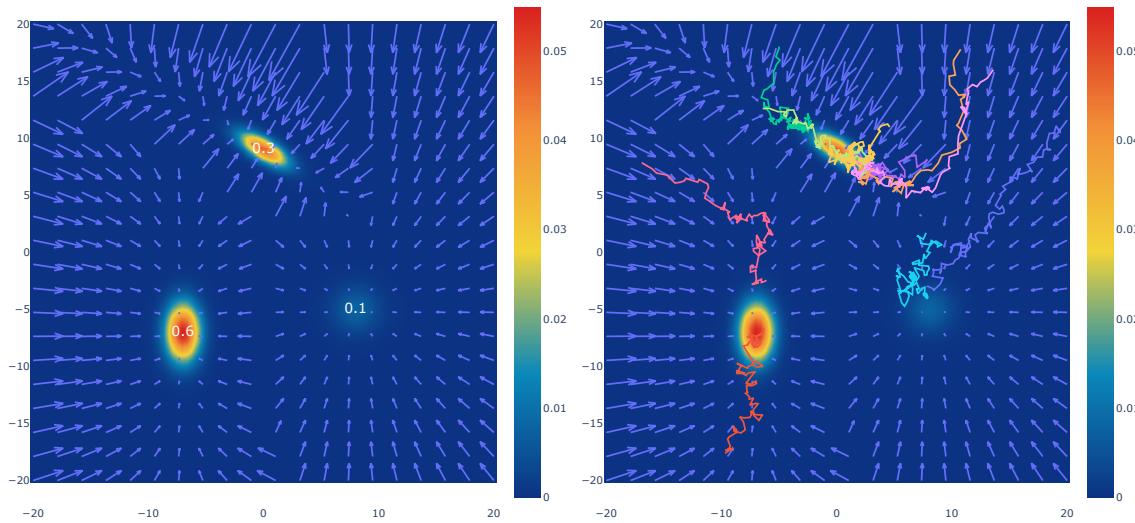


Figura 1.22: Muestras generadas a partir de una mixtura gaussiana utilizando la dinámica de Langevin dada en el Algoritmo 5. El mapa de calor indica la densidad de la mixtura en cada punto del plano y las curvas muestran la trayectoria seguida por el proceso de generación $(x_t)_{t=1}^T$. Se observa que la dinámica de Langevin no es capaz de respetar los priors de cada componente de la mixtura, donde se esperaría que la componente con mayor relevancia (con prior de clase 0,6) tenga una mayor cantidad de muestras generadas. Esta simulación se encuentra en el archivo `langevin.ipynb`.

Por lo tanto, el score de la mixtura no depende de π y la dinámica de Langevin no podrá generar muestras respetando los ponderadores de la mixtura, dándole igual probabilidad a muestras generadas desde p_1 o p_2 . Si bien al agregar ruido gaussiano el soporte pasa a ser todo \mathbb{R}^d , el fenómeno se sigue observando cuando se intentan conectar dos regiones separadas por zonas de baja densidad, donde la dinámica de Langevin requerirá más iteraciones para converger. Este problema se observa en la Figura 1.22 y en la Figura 1.23 (centro).

Para atacar estos problemas, los autores de [SE20] proponen dos mejoras. En primer lugar, continúan usando el enfoque de DSM ,pero utilizando distintos niveles de ruido. Por un lado, al usar un ruido de alta varianza se soluciona el problema de tener regiones de muy baja densidad y la aproximación del score puede ser más exacta, pero la distribución aprendida se aleja de la distribución buscada p_{data} . Con esta observación, los autores proponen utilizar una secuencia creciente²⁷ de niveles de ruidos $(\sigma_k)_{k=1}^L \subset \mathbb{R}_{++}$, cada uno asociado a un kernel de perturbación q_{σ_k} definido, al igual que antes, como un ruido aditivo $x \mapsto x + z$, $z \sim \mathcal{N}(0, \sigma_k^2 I_d)$. Los autores proponen usar una secuencia geométrica de niveles de ruido tal que σ_1 sea lo suficientemente pequeño para que $q_{\sigma_1} \approx p_{\text{data}}$, y σ_L sea lo suficientemente grande para que $q_{\sigma_L} \approx \mathcal{N}(0, \sigma_L I_d)$ y que además se logre mitigar el problema de las zonas de baja densidad.

De esta forma, para cada $\sigma \in (\sigma_k)_{k=1}^L$ se entrena un modelo de score $s_\theta(x, \sigma)$ que aproxime a $\nabla_x \log q_\sigma(x)$. Con esto, se tendrá una función de score-matching (ver (1.4.3)) para cada nivel de ruido. La función objetivo final será un promedio ponderado de las pérdidas individuales²⁸:

²⁷Originalmente, en [SE20], se define la secuencia en orden decreciente. Aquí se considerará en orden creciente para seguir el mismo orden que el modelo DDPM.

²⁸Se puede justificar el mejor desempeño de este modelo sobre el original considerando el promedio como un *ensemble* (ver *bagging* en [Bis06]).

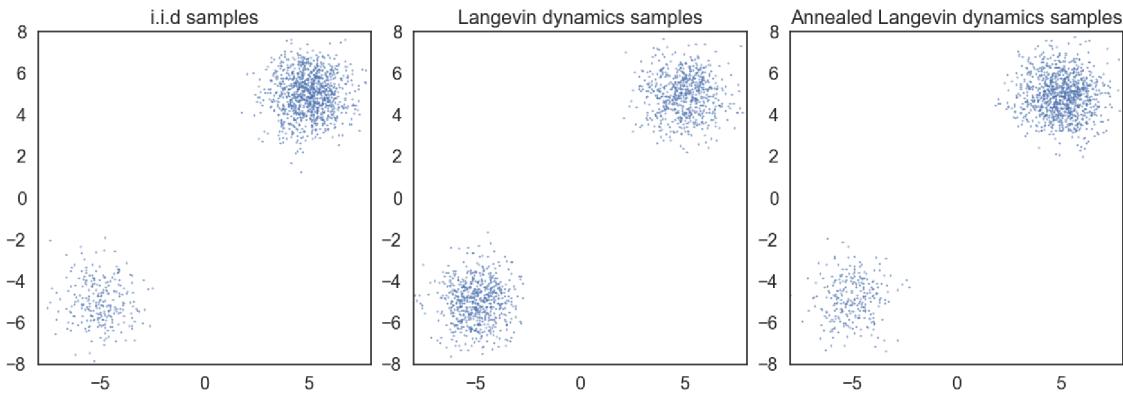


Figura 1.23: (izquierda) muestras de entrenamiento generadas a partir de la mixtura de la Figura 1.21. (centro) muestras generadas mediante DSM con la dinámica de Langevin usual. (derecha) muestras generadas con la dinámica de Langevin propuesta por [SE20].

$$l_\sigma(\theta) := \frac{1}{2} \mathbb{E}_{x \sim p_{\text{data}}, \tilde{x} \sim \mathcal{N}(x, \sigma^2 \mathbf{I}_d)} \left[\left\| \frac{x - \tilde{x}}{\sigma^2} - s_\theta(\tilde{x}, \sigma) \right\|^2 \right], \quad \forall \sigma \in (\sigma_k)_{k=1}^L$$

$$L(\theta) = \frac{1}{L} \sum_{k=1}^L \lambda(\sigma_k) l_{\sigma_k}(\theta). \quad (1.4.4)$$

Si bien se podrían utilizar modelos de score s_θ independientes para cada nivel de ruido σ , es usual considerar una única red neuronal $s_\theta(x, \sigma)$ entrenada de forma conjunta para todos los niveles de ruido. En [SE20] le llaman a esta red *noise conditional score network* (NCSN).

Para la elección del ponderador $\lambda(\sigma)$, los autores notaron empíricamente que, en el óptimo, $\|s_\theta(\cdot, \sigma)\| \propto \sigma^{-1}$, por lo que tomando $\lambda(\sigma) = \sigma^2$, el término $\lambda(\sigma_k) l_{\sigma_k}(\theta)$ se vuelve independiente de σ y así, ningún sumando de L gana o pierde relevancia únicamente por el nivel de ruido σ .

En segundo lugar, para la generación de muestras, los autores de [SE20] proponen una modificación al algoritmo de Langevin descrito en el Algoritmo 5, donde ahora computan una cadena de Langevin $(x_t)_{t=1}^T$ para cada uno de los niveles de ruido (de forma descendente en la cantidad de ruido), y el estado inicial de la cadena para el ruido σ_k viene dado por el estado final de la cadena para el ruido σ_{k+1} . El Algoritmo 6 detalla el procedimiento.

Si bien esta variante de DSM recibe distintos nombres en la literatura, aquí se le seguirá denominando simplemente DSM ya que se ha vuelto la variante predominante en esta familia de modelos y serán los que se conectarán con el modelo DDPM. En el archivo `sgm.ipynb` hay una implementación de SM y DSM.

DDPM como modelo basado en score

Hasta el momento, los modelos basados en score han sido trabajados como modelos totalmente independientes de los modelos de difusión descritos en la Subsección 1.3.1. En este apartado se mostrará que los modelos de difusión pueden ser formulados como un modelo DSM con ciertos niveles de ruido. Posteriormente, en la Subsección 1.4.2 se verá que ambos modelos corresponden a una familia de modelos de difusión más general, donde el proceso forward vendrá dado por un proceso estocástico a tiempo continuo. Por otra parte, es importante recordar que la técnica de *guidance* para la generación condicional es posible gracias el hecho de poder escribir el modelo DDPM como un modelo basado en score.

Algoritmo 6 Annealed Langevin sampling

Require: Niveles de ruido $(\sigma_k)_{k=1}^L$ usandos durante el entrenamiento.

Require: Distribución inicial π , cantidad de iteraciones T por nivel de ruido y largo de paso ϵ .

Require: Modelo $s_\theta(\cdot, \cdot)$ ya entrenado.

- 1: Generar $x_0 \sim \pi(x_0)$.
- 2: **for** k desde L hasta 1 **do**
- 3: $\epsilon_k \leftarrow \epsilon \frac{\sigma_k^2}{\sigma_L^2}$ ▷ Largo de paso adaptado para el nivel de ruido.
- 4: **for** t desde 1 hasta T **do**
- 5: Generar $z_t \sim \mathcal{N}(0, I_d)$.
- 6: $x_t \leftarrow x_{t-1} + \frac{\epsilon_k}{2} s_\theta(x_{t-1}, \sigma_k) + \sqrt{\epsilon_k} z_t$
- 7: **end for**
- 8: $x_0 \leftarrow x_T$
- 9: **end for**
- 10: **return** x_T

Para un modelo de difusión, en la Subsección 1.3.1 se mostró que lo único que se necesita para invertir el proceso de inyección de ruido mediante transiciones $p_\theta(x_{t-1}|x_t) \sim \mathcal{N}(\mu_\theta(x_t, t), \sigma_q^2(t) I_d)$ es aprender una función de media $\mu_\theta(x_t, t)$ que aproxime a la media de $q(x_{t-1}|x_t, x_0)$, $\mu_q(x_0, x_t, t)$ (ver (1.3.5)). Para esto, se estudiaron distintas formulaciones equivalentes:

- Aprender $\mu_q(x_0, x_t, t)$ directamente mediante un modelo $\mu_\theta(x_t, t)$.
- Aprender la muestra x_0 que generó x_t en la etapa t mediante un modelo $x_\theta(x_t, t)$.
- Aprender el ruido ϵ inyectado a x_0 en la etapa t mediante un modelo $\epsilon_\theta(x_t, t)$.

Para ver que un modelo de difusión se puede escribir como un modelo basado en score, se deducirá una cuarta forma de formular el problema de optimización, donde ahora un modelo paramétrico $s_\theta(x_t, t)$ tendrá que aprender el score $\nabla_x \log p(x_t)$ para obtener un modelo $\mu_\theta(x_t, t)$ para aprender la función de media $\mu_q(x_t, t, t)$. Para introducir la función de score en el modelo se utilizará el siguiente resultado:

Teorema 1.4 (fórmula de Tweedie, caso gaussiano). Sea $p(z|\mu) \sim \mathcal{N}(\mu, \Sigma)$, con $\mu \sim p(\mu)$ una distribución desconocida y Σ un parámetro fijo y conocido. Dada una muestra $z \sim p(z) = \int p(z|\mu)p(\mu) d\mu$, entonces:

$$\mathbb{E}_{p(\mu|z)} [\mu] = z + \underbrace{\nabla_z \log p(z)}_{\text{corrección de Bayes}}.$$

Demostración. Usando que $\int z p(\mu|z) = z \int p(\mu|z) = z$ y por teorema de Bayes:

$$\mathbb{E}_{p(\mu|z)} [\mu] = \int \mu p(\mu|z) d\mu = z - \int (z - \mu) p(\mu|z) d\mu = z - \int (z - \mu) \frac{p(z|\mu)p(\mu)}{p(z)} d\mu.$$

Notando que $\nabla_z \log p(z|\mu) = -\Sigma^{-1}(z - \mu)$, el integrando se escribe como:

$$(z - \mu) \frac{p(z|\mu)p(\mu)}{p(z)} = -\Sigma \nabla_z \log p(z|\mu) \frac{p(z|\mu)p(\mu)}{p(z)} = -\Sigma \frac{\nabla_z p(z|\mu)}{p(z|\mu)} \frac{p(z|\mu)p(\mu)}{p(z)} = -\Sigma \frac{\nabla_z p(z, \mu)}{p(z)}.$$

Por lo tanto, por regla de integración de Leibniz:

$$\mathbb{E}_{p(\mu|z)} [\mu|z] = z + \Sigma \frac{\int \nabla_z p(z, \mu) d\mu}{p(z)} = z + \Sigma \frac{\nabla_z \int p(z, \mu) d\mu}{p(z)} = z + \Sigma \frac{\nabla_z p(z)}{p(z)}.$$

Donde el cociente es precisamente el score de $p(z)$. □

Este resultado busca corregir la estimación empírica de la media real μ^{29} mediante un desplazamiento según el score de z . Esta corrección es importante en casos donde la muestra z puede ser un outlier, en cuyo caso el score $\nabla_z \log p(z)$ buscará desplazar la estimación hacia el centro.

En la Proposición 1.14 se usó que $q(x_t|x_0) \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}_d)$ para escribir $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t$ y realizar la sustitución

$$x_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}x_t - \frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}}\epsilon_t \quad (1.4.5)$$

en $\mu_q(x_0, x_t, t)$. Con esto, se pudo reparametrizar el modelo $\mu_\theta(x_t, t)$ por otro modelo $\epsilon_\theta(x_t, t)$. Ahora se sustituirá x_0 por otra cantidad equivalente para reparametrizar $\mu_\theta(x_t, t)$ por un modelo que aprenda la función de score. Gracias a la fórmula de Tweedie:

$$\sqrt{\bar{\alpha}_t} \underbrace{\mathbb{E}_{q(x_0|x_t)}[x_0]}_{x_0} = x_t + (1-\bar{\alpha}_t)\nabla_{x_t} \log q(x_t) \implies x_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}x_t + \frac{1-\bar{\alpha}_t}{\sqrt{\bar{\alpha}_t}}\nabla_{x_t} \log q(x_t). \quad (1.4.6)$$

Sustituyendo esta nueva estimación de x_0 en $\mu_q(x_0, x_t, t)$ de forma análoga a lo hecho en la Proposición 1.14, se obtiene el siguiente resultado:

Proposición 1.18 (problema de optimización DDPM (score-prediction)). Para los modelos propuestos en (1.3.1) y (1.3.6), la función de media del proceso reverso, $\mu_\theta(x_t, t)$, que maximiza la ELBO es combinación lineal entre x_t y s_θ (ver similitud con (1.3.11)):

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\bar{\alpha}_t}}x_t - \frac{1-\bar{\alpha}_t}{\sqrt{\bar{\alpha}_t}}s_\theta(x_t, t),$$

donde el modelo $s_\theta(x_t, t)$ resuelve el siguiente problema de optimización:

$$s_\theta^* = \arg \min_{\epsilon_\theta} \sum_{t=1}^T \frac{(1-\bar{\alpha}_t)^2}{2\sigma_q^2(t)\alpha_t} \mathbb{E}_{x_0 \sim p_{\text{data}}, x_t \sim q(x_t|x_0)} \left[\|\nabla_{x_t} \log q(x_t) - s_\theta(x_t, t)\|^2 \right].$$

Esta formulación permite afirmar que el modelo DDPM también puede ser visto como un modelo basado en score y más precisamente, un modelo tipo DSM, heredando de forma automática las propiedades que se puedan obtener sobre este tipo de modelos y modelos basados en energía en general.

Por último, igualando las ecuaciones (1.4.5) y (1.4.6) se puede observar que el score $\nabla_{x_t} \log q(x_t)$ y el ruido inyectado $\epsilon(x_t)$ están en proporción directa:

$$\frac{\epsilon(x_t)}{\nabla_{x_t} \log q(x_t)} = -\sqrt{1-\bar{\alpha}_t}.$$

En consecuencia, es posible obtener un modelo neuronal entrenado para DSM a partir un modelo entrenado para DDPM y viceversa.

1.4.2. Modelos de difusión mediante el uso de SDEs

Algunos meses después de los trabajos de DDPM y DSM, los autores de [Son+21b] notaron que ambos procesos de inyección de ruido correspondían a la discretización una SDE distinta. Con esta observación,

²⁹Recordar que el estimador de máxima verosimilitud de la media es la media empírica, que en este caso es z .

lograron generalizar el proceso de inyección de ruido mediante el uso de otras SDEs, aprovechando el hecho de que para obtener la SDE asociada al proceso reverso solo hace falta entrenar un modelo de score.

SDEs asociadas a DDPM y DSM

Para el modelo DDPM, de acuerdo a (1.3.1), la cadena de Markov asociada al proceso de difusión es

$$x_t = \sqrt{1 - \beta_t} x_{t-1} + \sqrt{\beta_t} \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, I_d), \quad (1.4.7)$$

donde se ha usado $\beta_t := 1 - \alpha_t$ para trabajar directamente con las varianzas de las transiciones. La cadena anterior puede verse como la discretización de una SDE, la cual se entrega en el siguiente resultado:

Proposición 1.19 (SDE asociada a DDPM). Considerando que la cadena de Markov (1.4.7) es a tiempo continuo, su SDE asociada es la siguiente:

$$dx_t = -\frac{1}{2} \beta_t x_t dt + \sqrt{\beta_t} dw_t, \quad x_0 \sim p_{\text{data}}(x_0). \quad (1.4.8)$$

Además, la varianza de la distribución marginal $p(x_t)$ está siempre acotada por la varianza de $p(x_0) = p_{\text{data}}(x_0)$. Más aún, se demuestra que si $\text{Var}(x_0) = I_d$ entonces $\text{Var}(x_t) = I_d$. Dado que x_t tiene varianza acotada, a esta SDE se le conocerá como *variance preserving SDE* o *VP-SDE*.

La demostración de esta propiedad se puede encontrar en [Son+21b]. Sin embargo, una manera informal de obtener la SDE (1.4.8) es considerando la aproximación de Taylor $\sqrt{1 - \beta_t} \approx 1 - \frac{\beta_t}{2}$. Para β_t pequeño y considerando $\epsilon \sim \mathcal{N}(0, I_d)$ se obtiene que

$$\begin{aligned} x_t &= \sqrt{1 - \beta_t} x_{t-1} + \sqrt{\beta_t} \epsilon \\ &\implies x_t - x_{t-1} = (\sqrt{1 - \beta_t} - 1) x_{t-1} + \sqrt{\beta_t} \epsilon \\ &\implies x_t - x_{t-1} \approx -\frac{1}{2} \beta_t x_{t-1} + \sqrt{\beta_t} \epsilon. \end{aligned}$$

Tomando el límite cuando el tamaño del paso tiende a cero, se llega a que

$$dx_t = -\frac{1}{2} \beta_t x_t dt + \sqrt{\beta_t} dw_t, \quad x_0 \sim p_{\text{data}}(x_0).$$

Por lo tanto, (1.4.2) es la ecuación del proceso de difusión (a tiempo continuo) correspondiente al modelo discreto DDPM. Por otra parte, si bien el modelo DSM no entrega una cadena de Markov explícita, también puede ser asociado a una SDE. Para esto, se usa el hecho de que este modelo utiliza kernels de perturbación gaussianos para construir una cadena de Markov. En [Son+21b] prueban el siguiente resultado:

Proposición 1.20 (SDE asociada a DSM). El modelo DSM con kernels de perturbación gaussianos isotrópicos $q_\sigma(\tilde{x}|x) = \mathcal{N}(x, \sigma^2 I_d)$ indexados por una familia de ruidos $(\sigma_k)_{k=1}^n$ está asociado a la siguiente cadena de Markov:

$$x_t = x_{t-1} + \sqrt{\sigma_t^2 - \sigma_{t-1}^2} \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, I_d).$$

Por otra parte, esta cadena de Markov está asociada a la siguiente SDE:

$$dx_t = \sqrt{\frac{d\sigma_t^2}{dt}} dw_t, \quad x_0 \sim p_{\text{data}}(x_0), \quad (1.4.9)$$

donde $\frac{d\sigma_t^2}{dt}$ es una derivada (recordar que, en el continuo, σ_t^2 es función).

En consecuencia, el modelo DSM puede verse como un modelo de difusión, el cual también es generalizable a tiempo continuo.

Generalización a otras SDEs

Los resultados anteriores motivan a definir una nueva familia de modelos generativos, donde ahora el proceso de difusión sobre los datos corresponderá a la solución de una SDE. La idea será, al igual que en el caso discreto, transformar la distribución de los datos p_{data} en una distribución p_{prior} para luego, generar muestras utilizando el proceso inverso, el cual se podrá obtener entrenando un modelo de score.

Sea $(x_t)_{t \geq 1}$ el proceso de Itô que resuelve la SDE

$$dx_t = f(x_t, t) dt + g(t) dw, \quad x_0 \sim p(x_0) = p_{\text{data}}(x_0), \quad (1.4.10)$$

donde la distribución marginal $p_{\text{prior}}(x_T) = p(x_T)$ es una distribución de la que es fácil generar muestras. Por simplicidad en la notación, en esta sección se considerará que g es una función escalar que solo depende de t . Para simular nuevas muestras desde p_{data} desde p_{prior} , se necesita conocer la SDE asociada al proceso de difusión inverso, el cual comienza desde p_{prior} . Un resultado fundamental para esto es el teorema de inversión de Anderson (ver Teorema A.6), el cual afirma que el proceso inverso de (1.4.10) viene dado por la solución de la siguiente SDE:

$$dx_t = [f(x_t, t) - g(t)^2 \nabla_{x_t} \log p(x_t)] dt + g(t) d\bar{w}_t. \quad (1.4.11)$$

Donde $p(x_t)$ es la densidad marginal en tiempo t ³⁰, \bar{w} es un movimiento browniano fluyendo hacia atrás en el tiempo y dt es un paso temporal infinitesimal negativo. Por lo tanto, para conocer completamente la SDE del proceso inverso de un proceso de difusión genérico, y así poder generar nuevas muestras mediante la simulación de este proceso, basta conocer la función score de la densidad marginal $p(x_t)$ en cada tiempo del proceso, por lo que este nuevo tipo de modelos también es de tipo SM. En la Figura 1.24 hay un diagrama de este modelo, donde la cadena de difusión del modelo DDPM es sustituida por un proceso de difusión a tiempo continuo. Una vez teniendo un modelo de score $s_\theta(x, t)$ entrenado, este puede ser usado en la SDE inversa para luego generar muestras a partir de él.

Para entrenar el modelo de score $s_\theta(x_t, t)$ que aproxime al score $\nabla_{x_t} \log q(x_t | x_0)$, se puede considerar como función de costo un funcional de score matching continuo:

$$\frac{1}{2} \int_0^T \mathbb{E}_{x_t \sim p(x_t)} [\lambda(t) \|\nabla_{x_t} \log p(x_t) - s_\theta(x_t, t)\|^2]$$

Dado que el score $\nabla_{x_t} \log p(x_t)$ es intratable, en [Son+21b] aplican una técnica de denoising score matching similar a la vista en la Proposición 1.17 para obtener la siguiente función de costo equivalente (salvo constante aditiva) para los modelos de difusión a tiempo continuo:

$$L_{\text{SDE}}(\theta) := \frac{1}{2} \mathbb{E}_{t \sim \text{Unif}([0, 1])} [\lambda(t) \mathbb{E}_{x_0 \sim p_0, x_t \sim p(x_t | x_0)} [\|\nabla_{x_t} \log p(x_t | x_0) - s_\theta(x_t, t)\|^2]], \quad (1.4.12)$$

donde $p(x_t | x_0)$ es el kernel de transición desde x_0 hacia x_t .

Cuando $f(\cdot, t)$ es afín, el kernel de transición es gaussiano y su media y covarianza se pueden obtener de forma cerrada, por lo que la función de score en $L_{\text{SDE}}(\theta)$ se conoce de forma cerrada. En particular, las

³⁰Las densidades marginales $p(x_t)$ son iguales para los procesos forward y backward.

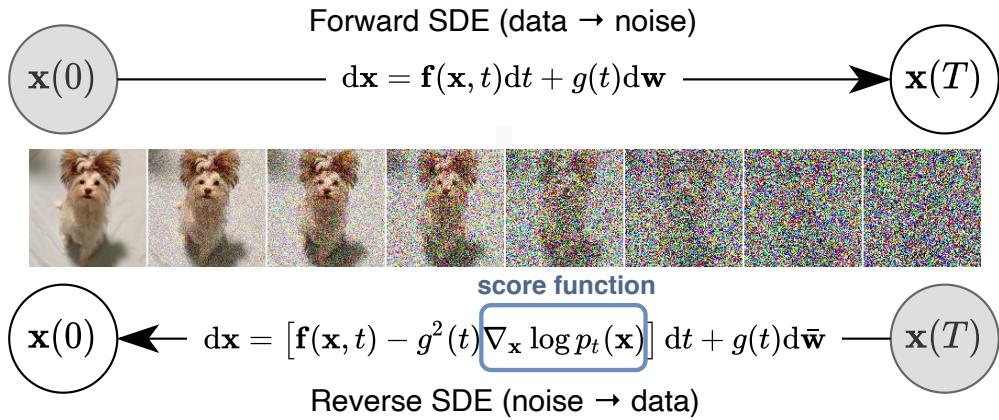


Figura 1.24: SDEs de los procesos de difusión y *denoising* para el modelo de difusión basado en una SDE, donde el proceso reverso depende únicamente del score a lo largo del proceso forward. Imagen obtenida desde [Son+21b].

SDEs (1.4.8) y (1.4.9) tienen kernels gaussianos. En SDEs más complejas, se puede resolver la ecuación de Fokker-Planck para obtener $p_{0t}(x_t|x_0)$ (ver Teorema A.4).

Por otra parte, para elegir la función de peso $\lambda(t)$ se intentará obtener un patrón común entre la forma continua de DDPM y DSM. De acuerdo a la Proposición 1.18, la función objetivo del modelo DDPM es de la forma:

$$L_{\text{DDPM}}(\theta) := \sum_{t=1}^T (1 - \bar{\alpha}_t) \mathbb{E}_{x_0 \sim p_{\text{data}}, x_t \sim q(x_t|x_0)} \left[\|\nabla_{x_t} \log q(x_t|x_0) - s_\theta(x_t, t)\|^2 \right].$$

Mientras que para DSM, la función a minimizar (ver (1.4.4)) es de la forma:

$$L_{\text{DSM}}(\theta) := \sum_{t=1}^T \sigma_t^2 \mathbb{E}_{x \sim p_{\text{data}}(x), \tilde{x} \sim q_{\sigma_t}(\tilde{x}|x)} \left[\|\nabla_{\tilde{x}} \log q_{\sigma_t}(\tilde{x}|x) - s_\theta(\tilde{x}, \sigma_t)\|^2 \right].$$

Notar que los poderadores de las sumas en $L_{\text{DDPM}}(\theta)$ y $L_{\text{DSM}}(\theta)$ tienen la misma forma estructural:

- En DDPM: $(1 - \bar{\alpha}_t) \propto 1/\mathbb{E} [\|\nabla_{x_0} \log q(x_t|x_0)\|^2]$.
- En DSM: $\sigma_t^2 \propto 1/\mathbb{E}_{x \sim p_{\text{data}}(x), \tilde{x} \sim q_{\sigma_t}(\tilde{x}|x)} [\|\nabla_{\tilde{x}} \log q_{\sigma_t}(\tilde{x}|x)\|^2]$.

Por lo tanto, los autores de [Son+21b] proponen normalizar usando la siguiente función de peso:

$$\lambda(t)^{-1} \propto \mathbb{E} [\|\nabla_{x_t} \log p(x_t|x_0)\|^2].$$

En la Subsección 1.4.3 se verá que la elección $\lambda(t) = g(t)^2$ funciona como la generalización de la ELBO para modelos de difusión a tiempo continuo.

Por otra parte, basándose en las SDEs de los modelos DDPM y DSM, los autores de [Son+21b] también proponen una tercera SDE que funciona bien para el cálculo de verosimilitud:

$$dx_t = -\frac{1}{2} \beta_t x dt + \sqrt{\beta_t \left(1 - \exp \left(-2 \int_0^t \beta(s) ds \right) \right)} dw_t. \quad (1.4.13)$$

El término de drift de esta SDE también es lineal, por lo que este proceso de difusión también posee kernels de transición gaussianos, cuyos parámetros tienen forma cerrada. Al igual que para la VP-SDE, la varianza se estabiliza en I_d cuando $t \rightarrow \infty$. Sin embargo, al comparar ambas SDEs usando la misma función de ruido β_t , se observa que la varianza de este proceso es menor a la varianza de la VP-SDE. Por este motivo, a esta SDE se le suele denominar *sub-VP-SDE*.

Para la generación de muestras, el modelo DDPM utiliza el algoritmo de *ancestral sampling* (Algoritmo 3), mientras que DSM utiliza *Langevin sampling* (Algoritmo 6). En el caso de un modelo con una SDE genérica, es necesario poder simular una trayectoria del proceso inverso para poder llegar a una muestra de $p(x_0) = p_{\text{data}}(x_0)$. En esta familia de modelos, el proceso de generación de muestras es más flexible ya que se puede utilizar cualquier técnica para generar trayectorias asociadas a la SDE (1.4.11) una vez entrenado el modelo de score $s_\theta(x_t, t)$.

Para una SDE genérica, es posible generar una muestra del proceso reverso utilizando métodos numéricos, lo que permite explícitamente manejar el trade-off entre precisión y eficiencia. Algunos métodos que se utilizan en [Son+21b] son el algoritmo de Euler-Maruyama (ver Algoritmo 7) y los métodos estocásticos de Runge-Kutta. En particular, los métodos *ancestral sampling* y *Langevin sampling* corresponden a una discretización particular de la SDE asociada a DDPM y a SGM respectivamente. Por otra parte, en los últimos años se han propuesto diversos métodos numéricos especializados en simular la SDE asociada al proceso backward de los modelos de difusión. Entre los métodos más conocidos, están DPM-Solver [Lu+22] y [Lu+23], los cuales mejoran considerablemente el tiempo de inferencia.

En este trabajo se mencionarán dos métodos, los cuales tienen relevancia en los próximos capítulos.

Método de Euler-Maruyama El método más simple para simular una SDE, y el método que se utilizará para todas las simulaciones de este trabajo, es el algoritmo de Euler-Maruyama, el cual corresponde a la extensión natural del método de Euler para ecuaciones diferenciales ordinarias (ODEs) al contexto estocástico.

Dada una SDE de la forma $dx = \mu(x, t) dt + \sigma(x, t) dW_t$ con condición inicial $x_0 \sim p(x_0)$ (determinista o aleatoria), el algoritmo de Euler-Maruyama construye la siguiente cadena de Markov como aproximación del proceso estocástico x :

Algoritmo 7 Euler-Maruyama

- 1: **Entrada:** funciones μ y σ , distribución inicial $p(x_0)$, tiempo final T y número de pasos N .
 - 2: Generar y definir $x_0 \sim p(x_0)$.
 - 3: Calcular $\Delta t = \frac{T}{N}$.
 - 4: **for** $i = 0$ to $N - 1$ **do**
 - 5: Generar una muestra $z_i \sim \mathcal{N}(0, \Delta t)$.
 - 6: Calcular $x_{i+1} = x_i + \mu(x_i, t_i)\Delta t + \sigma(x_i, t_i)z_i$
 - 7: **end for**
 - 8: **return** x_T
-

Es importante notar que, aunque el método de Euler-Maruyama es conceptualmente sencillo, puede tener problemas de estabilidad y precisión, especialmente para SDEs con términos de drift o dispersión poco regulares. Por lo tanto, para SDEs más complejas, pueden ser necesarios métodos numéricos más avanzados.

Probability flow ODE Dado el proceso de difusión forward (1.4.10), este tiene asociada su propia ecuación de Fokker-Planck (ver Teorema A.4), la cual es una ecuación que indica cómo evoluciona la función de densidad marginal $p(x_t)$ del proceso de difusión. En [Son+21b] reordenan los términos de la ecuación de Fokker-Planck

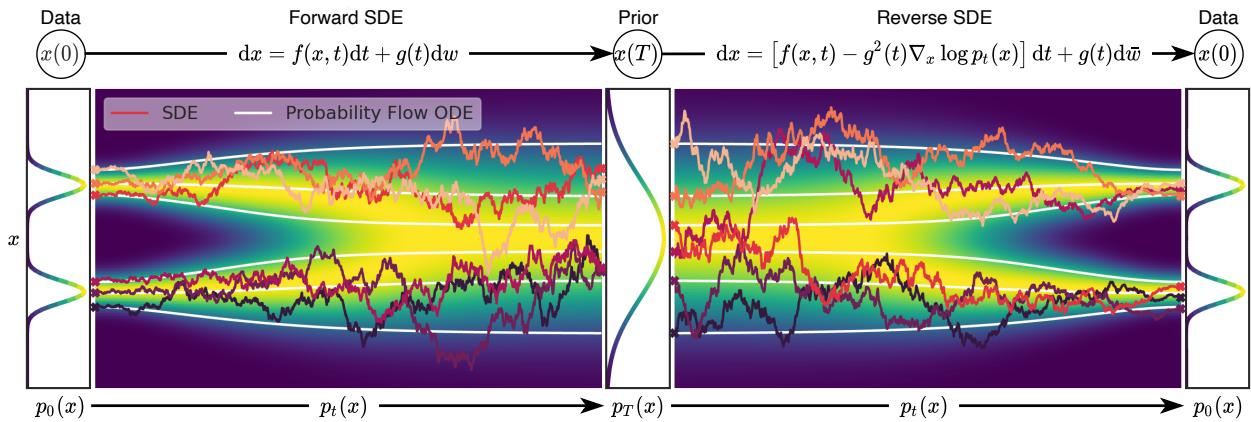


Figura 1.25: Ilustración de los procesos forward y backward en un modelo de difusión. El proceso forward comienza con una distribución bimodal $p_0 = p_{\text{data}}$ y termina en una distribución gaussiana $p_T = p_{\text{prior}}$. El proceso backward comienza desde la distribución prior y termina en la distribución de los datos. Las curvas erráticas muestran trayectorias del proceso estocástico para cuatro condiciones iniciales obtenidas desde $x_0 \sim p_{\text{data}}(x_0)$. En el proceso forward, las curvas blancas muestran la evolución de la *probability flow ODE* al comenzar desde las muestras x_0 obtenidas. En el proceso backward las curvas blancas muestran como el proceso determinista regresa a las muestras x_0 originales al comenzar desde x_T . Imagen obtenida desde [Son+21b].

de este proceso estocástico para formar otra ecuación de Fokker-Planck, con la misma solución, asociada a un proceso determinista (ver Teorema A.5). De esta forma, se logra obtener un nuevo proceso de inyección de ruido pero totalmente determinista (es decir, guiado por una ODE y no una SDE) que evolucione con las mismas distribuciones marginales $p(x_t)$ que el proceso estocástico (1.4.10)³¹. Este proceso viene guiado por la siguiente ODE:

$$dx_t = \left(f(x_t, t) - \frac{1}{2}g(t)^2 \nabla_{x_t} \log p_t(x_t) \right) dt. \quad (1.4.14)$$

Notar que este proceso determinista induce, naturalmente, un mapa entre muestras $x_0 \sim p_{\text{data}}(x_0)$ y muestras $x_T \sim p_{\text{prior}}(x_T)$. Por otra parte, es importante mencionar que esta ODE corresponde a la formulación continua del proceso determinista usado en DDIM (ver Subsección 1.3.3), lo que nuevamente permite conectar formulaciones discretas con formulaciones continuas en los modelos de difusión.

Por lo tanto, es posible generar muestras a partir de la misma distribución que la SDE reversa comenzando con una muestra $x_T \sim p_T(x_T)$ y resolviendo la EDO hacia atrás en el tiempo, utilizando el modelo de score aprendido. En la Figura 1.25 se puede ver una comparación del proceso de difusión y denoising seguido de forma estocástica y usando esta versión determinista.

Esta ODE es conocida como *probability flow ODE* y puede ser resuelta con cualquier algoritmo clásico, evitando tener que usar simuladores estocásticos. La demostración de esta propiedad en una versión más general se encuentra en [Son+21b].

Notar que, al igual que para la SDE reversa, solo es necesario conocer el score de las densidades marginales $(p_t)_{t \in [0,T]}$, por lo que se puede utilizar un modelo entrenado para el proceso de difusión basado en una SDE para generar muestras a partir de la *probability flow ODE*. Por otra parte, esta formulación determinista

³¹Toda la aleatoriedad del proceso ocurre en la generación de la muestra inicial $x_0 \sim p(x_0) = p_{\text{data}}(x_0)$ con la que comienza el proceso de difusión.

de los modelos de difusión será utilizada en la Subsección 2.3.4, donde la asignación de una muestra inicial $x_0 \in \mathbb{R}^d$ con su posición final $x_T \in \mathbb{R}^d$ será interpretada como un mapa de transporte óptimo en el sentido de Monge, mientras que la versión estocástica original podrá ser comparada con la regularización entrópica en el Capítulo 3, generando así una estrecha relación entre los modelos de difusión y el problema del puente de Schrödinger.

1.4.3. Entrenamiento basado en verosimilitud

En esta subsección se extenderá el concepto de ELBO (introducido en los autoencoders variacionales) a los modelos de difusión a tiempo continuo, permitiendo hacer un entrenamiento de máxima verosimilitud aproximada. Es importante destacar que, a través de la probability flow ODE, los modelos de difusión a tiempo continuo pueden verse como instancias de flujos normalizantes a tiempo continuo (ver [Che+19]), por lo que la verosimilitud se puede computar de forma cerrada. Sin embargo, esto es costoso computacionalmente ya que requiere múltiples evaluaciones de un solver de ecuaciones diferenciales.

Para entregar el resultado principal de esta subsección, es necesario indicar que el modelo generativo induce 2 distribuciones de probabilidades sobre las cuales se puede querer calcular verosimilitud:

- p_θ^{SDE} : distribución marginal $p(x_0)$ para el proceso reverso (1.4.11) comenzando desde $p(x_T) \sim p_{\text{prior}}(x_T)$ y usando el modelo aprendido $s_\theta(x_t, t)$.
- p_θ^{ODE} : distribución marginal $p(x_0)$ para el proceso reverso determinista (1.4.14) comenzando desde $p(x_T) \sim p_{\text{prior}}(x_T)$ y usando el modelo aprendido $s_\theta(x_t, t)$.

Es decir, una verosimilitud cuando se considera el proceso reverso estocástico y otra cuando se considera el proceso reverso seguido por la probability flow ODE. Notar que la verosimilitud que se puede obtener de forma exacta viendo al modelo como un flujo normalizante continuo es p_θ^{ODE} , mientras que para p_θ^{SDE} no es posible aplicar este enfoque. Sin embargo, en [Son+21a] muestran que una elección particular de ponderadores $\lambda(t) > 0$ (ver (1.4.12)) permite interpretar la suma ponderada en (1.4.12) como una cota de la verosimilitud para p_θ^{SDE} . Específicamente, se tiene el siguiente resultado:

Teorema 1.5. Para el proceso de difusión a tiempo continuo (1.4.10) se tiene la siguiente cota para la verosimilitud esperada:

$$\mathbb{E}_{x_0 \sim p_{\text{data}}(x_0)} [\log p_\theta^{\text{SDE}}(x_0)] \geq - \int_0^T \mathbb{E}_{x_0 \sim p(x_0), x_t \sim p(x_t|x_0)} \left[g^2(t) \|\nabla_{x_t} \log p(x_t|x_0) - s_\theta(x_t, t)\|^2 \right] + \text{constante}, \quad (1.4.15)$$

donde $p(x_t|x_0)$ es el kernel de transición desde x_0 hacia x_t . Por otra parte, para una estimación puntual de la verosimilitud se tiene que:

$$\log p_\theta^{\text{SDE}}(x_0) \geq - \int_0^T \mathbb{E}_{x_t \sim p(x_t|x_0)} \left[\frac{1}{2} \|g(t)s_\theta(x_t, t)\|^2 + \nabla_{x_t} \cdot (g^2(t)s_\theta(x_t, t) - f(x_t, t)) \right] dt$$

Notar que todos los términos al lado derecho de (1.4.15) son computables, por lo que se puede usar esta función objetivo como función proxy para maximizar la verosimilitud $\log p_\theta^{\text{SDE}}$. Notar que el único cambio que sugiere este resultado con respecto a la función objetivo (1.4.12) es considerar $\lambda(t) = g^2(t)$.



Figura 1.26: Transformación de una distribución de imágenes en otra utilizando una CycleGAN. Imagen obtenida desde [Zhu+20].

1.4.4. Limitaciones de los modelos de difusión

Para concluir este capítulo, se nombrarán algunas limitaciones tanto teóricas como prácticas que poseen los modelos de difusión, las cuales serán abordadas desde distintos puntos de vista en los siguientes capítulos.

Limitación en las transformaciones Por construcción, los modelos de difusión están diseñados para transformar una distribución inicial p_{data} en otra distribución final p_{prior} (generalmente gaussiana), y viceversa. Sin embargo, muchas veces se desea trabajar en un marco más flexible, donde la distribución final $p(x_T)$ pueda ser cualquier otra distribución, no necesariamente gaussiana (ver Figura 1.26). Esto motiva a estudiar el problema de transformar una distribución de probabilidad en otra y de forma óptima (en algún sentido). Para esto, en el Capítulo 2 se comenzará el estudio del transporte óptimo, el cual busca resolver precisamente este problema desde una perspectiva estática, es decir, sin considerar un proceso que realice el transporte a lo largo del tiempo. Posteriormente, en la Sección 2.3 se extenderá este problema estático a una perspectiva dinámica y se buscará un sistema dinámico (i.e., un sistema que evoluciona con el tiempo) que realice el transporte entre las distribuciones y se verá que, sorpresivamente, el modelo DDPM resuelve este problema.

Convergencia asintótica a la distribución a priori Durante el proceso de difusión en el entrenamiento del modelo DDPM (descrito en la Subsección 1.3.1) la densidad marginal en el tiempo final, $q(x_T)$, solo es una aproximación de la distribución asintótica $p_{\text{prior}}(x_T) = \mathcal{N}(0, I_d)$, por lo que es necesario simular el proceso de difusión hasta un tiempo $T \gg 1$ lo suficientemente grande como para poder tener $p(x_T) \approx p_{\text{prior}}(x_T)$ con suficiente precisión. Esto vuelve a los modelos de difusión extremadamente lentos comparados con otros modelos como las GANs (ver Subsección 1.1.1). Más aún, el error de aproximación puede generar malas aproximaciones en la densidad aprendida $p_\theta(x_0) \approx p_{\text{data}}(x_0)$ tal como se discute en [Bor+23a].

Por otra parte, dada la naturaleza de la formulación de los modelos de difusión, la distribución p_{prior} solo es alcanzada en un horizonte de tiempo infinito (ver Figura 1.27), lo cual es intratable en la práctica. Para abordar este problema, es posible plantear el problema considerando un horizonte de tiempo finito desde el comienzo, lo cual se consigue con la formulación dinámica estudiada Sección 2.3. Sin embargo, si bien la solución a este problema tiene buenas propiedades matemáticas, es muy costosa de obtener en alta dimensión, donde además, el problema de transporte óptimo sufre de la maldición de la dimensionalidad en cuanto a la cantidad de muestras necesarias durante el entrenamiento.

Para solucionar esto, en el Capítulo 3 se estudia una versión entrópica del problema, la cual se puede resolver eficientemente. Este nuevo problema regularizado resulta ser equivalente al problema de Schrödinger estático, el cual puede ser extendido a su formulación dinámica muy fácilmente.

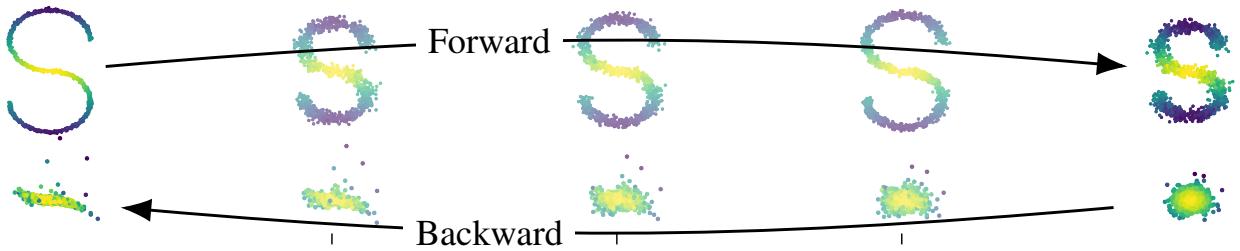


Figura 1.27: Efecto de simular el proceso de difusión por un tiempo demasiado corto. Al no aproximar bien la distribución p_{prior} en x_T , el proceso inverso no podrá generar muestras coherentes con la distribución p_{data} . Imagen obtenida desde [Bor+23b].

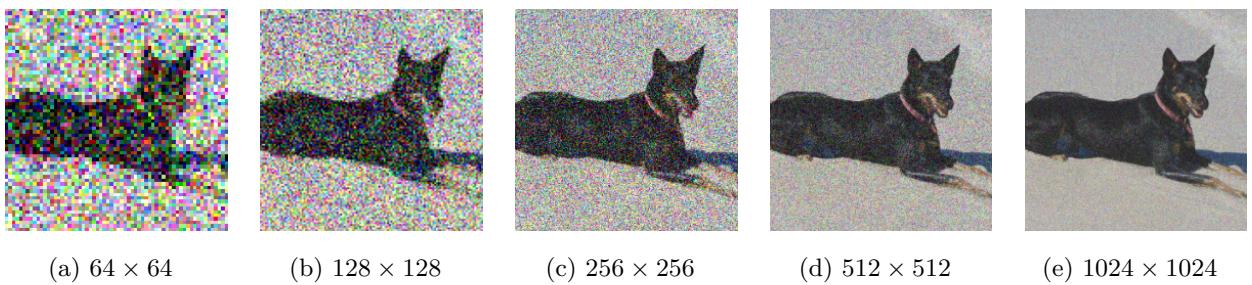


Figura 1.28: Efectos de injectar un mismo ruido sobre imágenes de distinta resolución. Se observa que para imágenes más grandes, es necesario aumentar el nivel de ruido. Imagen obtenida desde [Che23].

Sensibilidad a la elección de la SDE La elección de la SDE para el proceso de difusión puede tener un impacto significativo en el rendimiento del modelo tal como lo demostró el trabajo de [Son+21b]. Sin embargo, no existe un marco teórico completo que guíe la selección óptima de la SDE para una tarea dada (aunque hay trabajos en esa dirección como el de [Kar+22]), lo que ha llevado a una búsqueda de SDES únicamente basada en heurísticas. Más aún, si bien se han comenzado a obtener algunos resultados de convergencia para los modelos de difusión (ver [Bor23]), aún falta un entendimiento más profundo de las propiedades de este tipo de modelos en general. En la Figura 1.28 se puede ver que la elección del noise scheduler también depende de la resolución de las imágenes que se usan para el entrenamiento del modelo de difusión.

Dificultad en la interpretabilidad A diferencia de algunos métodos de transporte óptimo que ofrecen interpretaciones geométricas claras (como las geodésicas en el espacio de Wasserstein estudiadas en la Subsección 2.3.1), los modelos de difusión pueden ser menos interpretables en términos de la transformación que realizan entre las distribuciones inicial y final. Esta falta de interpretabilidad puede dificultar la comprensión de cómo el modelo está generando o transformando los datos, lo cual muchas veces es necesario, por ejemplo, para estudios de sesgos.

Complejidad computacional A pesar de que los modelos de difusión suelen generar resultados de muy alta calidad, su entrenamiento y muestreo pueden ser computacionalmente costosos, especialmente para procesos de difusión largos. Esto contrasta con algunos métodos de transporte óptimo que, una vez resueltos, permiten una transformación más directa entre distribuciones. Si bien propuestas recientes como [Son+23] y [SH22] han permitido disminuir considerablemente la cantidad de pasos necesarios para la generación, aún no se logra alcanzar el mapeo uno a uno entre los elementos de la distribución de los datos y la distribución prior como sí ocurre al usar transporte óptimo. Sin embargo, existe una conexión profunda entre los modelos

de difusión y ciertos problemas de transporte óptimo entrópico, lo que sugiere que un estudio conjunto de ambos enfoques puede llevar a avances significativos en la comprensión y aplicación de los modelos generativos, ganando las buenas propiedades empíricas de los modelos de difusión junto a las buenas garantías teóricas que entrega el marco teórico del transporte óptimo y el problema de Schrödinger.

De todas estas limitaciones, quizás la más importante desde la perspectiva del aprendizaje automático es no poder transformar una distribución de probabilidad en otra arbitraria, lo cual muchas veces es deseable en tareas de transferencia como super-resolución o *deblurring*. Si bien es posible adaptar el mecanismo de los modelos de difusión para este tipo de tareas, se siguen heredando el resto de limitaciones de esta familia de modelos. Más aún, un problema intrínseco de hacer transferencia con modelos de difusión es que estos no permiten trabajar en un marco no supervisado, donde muchas veces solo se tienen muestras de dos distribuciones $p_0(x)$ y $p_1(x)$ pero no se conoce como se emparejan muestras de p_0 con muestras de p_1 .

Por lo anterior, resulta natural plantear el problema como uno de transporte óptimo, donde se busca un mapa $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$ que empareje muestras de la distribución p_0 con muestras de la distribución p_1 . Este problema se formula de la siguiente forma:

$$T^* = \arg \min_{T \text{ transporta } p_0 \text{ a } p_1} \int_{\mathbb{R}^d} c(x, T(x)) p_0(x) \, dx,$$

donde $c : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}_+$ es una medida de (no) similitud que busca que x sea cercano a su pareja $T(x)$ (en el sentido que indique c).

El próximo capítulo está enfocado en estudiar este problema (y su relajación), la cual permitirá obtener procesos de transferencia entre distribuciones con mejores propiedades que los modelos de difusión, permitiendo, por ejemplo, interpolar de forma natural entre dos distribuciones de probabilidad.

Capítulo 2

Transporte óptimo entre distribuciones

En este capítulo se revisarán algunos conceptos y resultados fundamentales del transporte óptimo, un campo esencial dentro de la ingeniería y la matemática aplicada que últimamente ha ganado un gran interés en la comunidad del aprendizaje de máquinas y, en particular, de los modelos generativos. Desde un punto de vista práctico, el tipo de problemas de optimización que busca resolver el transporte óptimo pueden ser adaptados a una infinidad de contextos diferentes, permitiendo utilizar resultados importantes de la teoría en la toma de decisiones en escenarios complejos. Por otra parte, desde un punto de vista teórico, el transporte óptimo es un área bien estudiada desde la matemática, permitiendo obtener garantías teóricas y propiedades e interpretaciones geométricas relevantes.

Antes de conectar el transporte óptimo con los modelos de difusión y ver las ventajas que esto implica, se comenzará estudiando el transporte óptimo de forma general, sin limitar su aplicación a los modelos generativos. Posteriormente, en la Subsección 2.3.4 se conectará el transporte óptimo con los modelos de difusión y se verá que los modelos de difusión, en el mejor de los casos, resuelven un caso particular de transporte óptimo.

Con el fin de construir los problemas de transporte óptimo de una forma más natural, primero serán motivados en su versión discreta y luego generalizados a su versión continua en \mathbb{R}^d . Esto permite obtener resultados de manera sencilla al trabajar en el caso discreto, los cuales después serán extendidos al caso continuo sin demostración. Además, es importante mencionar que tener un entendimiento más acabado de las formulaciones discretas es más beneficioso al momento de realizar implementaciones numéricas, las que constituyen gran parte de las aplicaciones del transporte óptimo.

Por otra parte, de aquí en adelante ya no se asumirá que las medidas de probabilidad poseen una función de densidad (ver Subsección A.1.2) al trabajar en formulaciones continuas dado que esta suposición solo limita el alcance de los problemas de transporte óptimo y no aporta significativamente a la simplificación de la teoría. Además, cuando se hable de *solución* de un problema de optimización, se estará haciendo referencia a un elemento del conjunto factible que alcance el valor ínfimo de la función objetivo (i.e., que sea un mínimo)¹. En consecuencia, todos los problemas serán escritos utilizando \inf (resp. \sup) en vez de \min (resp. \max), esto

¹Recordar que la existencia del valor ínfimo para un problema de optimización no garantiza a priori la existencia de una solución. Considerar, por ejemplo, un problema de optimización donde el conjunto factible es un conjunto abierto y el valor ínfimo del problema de optimización se alcanza en la frontera del conjunto factible.

con el fin de mostrar el mérito de los resultados de existencia de soluciones.

Para introducir el problema de transporte óptimo, se propone considerar dos conjuntos finitos $\mathcal{X} = \{x_i\}_{i=1}^n$ e $\mathcal{Y} = \{y_j\}_{j=1}^m$, por ahora con $n = m$, donde el objetivo es encontrar una biyección $T : \mathcal{X} \rightarrow \mathcal{Y}$ que minimice algún funcional de costo, el cual suele ser una suma de costos individuales de emparejamiento². Con esta idea, dado $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$, donde $c(x_i, y_j)$ indica el costo de emparejar x_i con y_j , el problema de optimización se plantea de la siguiente forma:

$$\inf_{\substack{T : \mathcal{X} \rightarrow \mathcal{Y} \\ T \text{ es biyección}}} \sum_{i=1}^n c(x_i, T(x_i)). \quad (2.0.1)$$

Notar que el conjunto de biyecciones es finito y no vacío cuando $n = m$, por lo que este problema siempre tiene solución (aunque no necesariamente es única). Por otra parte, es importante indicar que el funcional de costo discreto $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ siempre se puede identificar con una matriz $C \in \mathcal{M}_{n,m}(\mathbb{R})$ mediante $C_{ij} = c(x_i, y_j)$.

Para ilustrar este problema en el área del aprendizaje automático, se puede considerar el problema de asignar particiones de datos de entrenamiento a diferentes GPUs para realizar un entrenamiento distribuido. En este escenario, \mathcal{X} representa el conjunto de particiones de datos, mientras que \mathcal{Y} representa el conjunto de GPUs disponibles para el entrenamiento. Para la función de costo se podría considerar buscar una asignación que minimice la suma del tiempo total de transporte de datos para cada par partición-GPU, $(x, T(x)) \in \mathcal{X} \times \mathcal{Y}$, que asigne el mapa T .

Si bien esta formulación es lo suficientemente general como para ser utilizada en diferentes problemas, también tiene múltiples desventajas. La primera de ellas es que al ser un problema de optimización discreto, debe ser tratado como un problema combinatorial, siendo difícil de resolver para valores muy grandes de $n = m$ ya que el conjunto factible tiene $n!$ elementos. Además, esta formulación no tiene solución en casos donde \mathcal{X} e \mathcal{Y} tienen distinto cardinal (ya que no es posible construir una biyección) y no permite asociar varios elementos de \mathcal{X} a un mismo elemento de \mathcal{Y} , lo cual es útil, por ejemplo, si las GPUs tienen suficiente memoria en el problema de entrenamiento distribuido.

Para resolver estas limitaciones, en la Sección 2.1 se relajará la formulación anterior permitiendo asignar varios elementos de \mathcal{X} a un mismo elemento de \mathcal{Y} , obteniendo versión más general del problema (2.0.1) conocida como *problema de Monge*. Posteriormente, en la Sección 2.2 se mostrarán algunas dificultades de este nuevo enfoque, por lo que se modificará el problema de Monge para obtener la *relajación de Kantorovich*, la cual es otro problema de optimización que tiene buenas propiedades tanto teóricas como prácticas y que, bajo ciertas condiciones, es equivalente al problema de Monge. Posteriormente, en la Sección 2.3 se incluirá la variable temporal en estos problemas para dar una interpretación dinámica del problema de transporte óptimo, la cual, en particular, será comparada con los modelos de difusión en la Subsección 2.3.4 y luego generalizada a una versión estocástica en el Capítulo 3, la cual resultará ser equivalente al problema del puente de Schrödinger.

Para poder realizar el estudio del transporte óptimo, son necesarias algunas definiciones elementales de teoría de la medida, las cuales se pueden encontrar, por completitud, en el Apéndice 3.3.1.

²Notar que este problema es equivalente al problema de buscar un matching óptimo en un grafo bipartito cuyas particiones son \mathcal{X} e \mathcal{Y} . Equivalentemente, las soluciones pueden identificarse con permutaciones del conjunto $\{1, \dots, n\}$.

2.1. Problema de Monge

Si bien el problema de la infactibilidad del problema (2.0.1) para $n \neq m$ se puede solucionar eliminando la restricción de biyectividad para T ³, el problema sigue teniendo la limitación de tratar a todos los elementos de \mathcal{X} e \mathcal{Y} por igual. En el problema de entrenamiento distribuido, esto implica que el problema (2.0.1) no es consciente de atributos relevantes como el tamaño de la partición o la capacidad de la GPU. Para resolver limitaciones de este tipo, se modificará el marco de trabajo del problema anterior y se considerará adicionalmente que cada elemento $x_i \in \mathcal{X}$ tiene asignado un valor $a_i \in \mathbb{R}$ que podrá ser usado dentro de la función de costo. Similarmente, cada elemento $y_j \in \mathcal{Y}$ tendrá asignado un valor $b_j \in \mathbb{R}$.

2.1.1. Formulación del problema

El problema de Monge será introducido en su versión discreta y luego generalizado a su versión continua en \mathbb{R}^d . De aquí en adelante, ya no se considerará la restricción $n = m$.

Formulación discreta

Con el fin de mostrar otra perspectiva del transporte óptimo, se motivará el problema de Monge usando el clásico problema de asignación de recursos mineros. En este escenario, una compañía posee un conjunto $\mathcal{X} = \{x_i\}_{i=1}^n$ de yacimientos mineros, donde cada yacimiento $x_i \in \mathcal{X}$ produce una cantidad $a_i > 0$ de cierta materia prima. Por otra parte, un conjunto de fábricas $\mathcal{Y} = \{y_j\}_{j=1}^m$ espera recibir una cantidad $b_j > 0$ en cada una de las fábricas $y_j \in \mathcal{Y}$, pudiendo abastecerse de diferentes yacimientos de \mathcal{X} ⁴. El problema de transporte óptimo aquí consiste en satisfacer la demanda de todas las fábricas minimizando el costo de transportar la materia prima desde los yacimientos mineros hacia las fábricas.

Bajo este marco, el problema de asignar la oferta de \mathcal{X} entre la demanda de \mathcal{Y} minimizando un funcional de transporte $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ se formula de la siguiente forma:

$$\inf_{T: \mathcal{X} \rightarrow \mathcal{Y}} \sum_{i=1}^n c(x_i, T(x_i)) \quad \text{s.a.} \quad \sum_{i: T(x_i)=y_j} a_i = b_j, \quad \forall j \in \{1, \dots, m\}, \quad (2.1.1)$$

donde la restricción adicional busca que se cumpla la demanda para cada $y_j \in \mathcal{Y}$. Es importante notar que en esta restricción, la suma es sobre (la oferta de) los elementos de \mathcal{X} que envían masa a y_j , es decir, el conjunto $T^{-1}(\{y_j\}) = \{x_i \in \mathcal{X} : T(x_i) = y_j\}$.

Notar que el mapa buscado T ya no necesita ser biyectivo pero sí es necesariamente sobreyectivo ya que, de lo contrario, de acuerdo a la restricción, habría un $b_j = 0$, lo cual contradice que $\{b_j\}_{j=1}^m \subset \mathbb{R}_{++}$. Por otra parte, al considerar $n = m$ y $a = b = \mathbf{1}$, se recupera la biyectividad de T y la restricción se cumple trivialmente, concluyendo que (2.0.1) es un caso particular del problema (2.1.1).

Por otra parte, la solución no tiene por qué existir ni ser única: la sobreyectividad de T no se puede alcanzar cuando $n < m$, por lo que en este caso el problema es infactible y no hay solución. Sin embargo, se verá que para el caso $n = m$ sí existe solución cuando la cantidad ofertada total es igual a la cantidad demandada total. Por otro lado, en algunos casos es posible encontrar varias soluciones cuando el problema presenta algún tipo de simetría.

Si bien la formulación del problema (2.1.1) es bastante clara, es usual escribir el problema como un problema

³En particular, pidiendo solo sobreyectividad se soluciona la infactibilidad para $n > m$ mientras que pidiendo solo inyectividad se soluciona la infactibilidad para $n < m$.

⁴La restricción $a_i, b_j > 0$ no limita la generalidad del problema ya que si algún a_i (resp. b_j) fuese nulo, se podría eliminar de la familia \mathcal{X} (resp. \mathcal{Y}) el punto x_i (resp. y_j) asociado.

de transporte entre medidas de probabilidad⁵. Bajo este enfoque, los elementos de \mathcal{X} se interpretarán como puntos en un espacio medible discreto (ver Sección A.1.1), mientras que los vectores $a = (a_1, \dots, a_n)^\top$ y $b = (b_1, \dots, b_m)^\top$ representarán medidas discretas en dichos espacios.

Para poder considerar únicamente medidas de probabilidad, se pedirá que los vectores $a \in \mathbb{R}^n$ y $b \in \mathbb{R}^m$ sean vectores de probabilidad⁶ (i.e., $a \in \Sigma_n$ y $b \in \Sigma_m$, donde Σ_n está definido en (A.1.2)). Con estas observaciones, se construyen las medidas discretas $\mu \in \mathcal{M}_+^1(\mathcal{X})$ y $\nu \in \mathcal{M}_+^1(\mathcal{Y})$ definidas como:

$$\mu = \sum_{i=1}^n a_i \delta_{x_i} \quad \nu = \sum_{j=1}^m b_j \delta_{y_j}.$$

Bajo esta interpretación, (2.1.1) se entiende como el problema de buscar un mapa $T : \mathcal{X} \rightarrow \mathcal{Y}$ que transporta de forma óptima la distribución masa μ de los elementos de \mathcal{X} hacia una distribución de masa ν en \mathcal{Y} . Por otra parte, la restricción en (2.1.1) se interpreta como una condición de preservación de medida después del transporte, donde la masa de cada elemento $y_j \in \mathcal{Y}$ debe ser igual a la masa total de todos los elementos $x_i \in \mathcal{X}$ que fueron transportados hacia y_j . En el enfoque de teoría de la medida, esta restricción indica precisamente que $T_\# \mu = \nu$ (ver Definición A.4), donde la medida push-forward $T_\# \mu \in \mathcal{M}_+^1(\mathcal{Y})$ toma la siguiente forma en el caso discreto:

$$T_\# \mu := \sum_{i=1}^n a_i \delta_{T(x_i)}. \quad (2.1.2)$$

Observar que esta medida transforma el vector de probabilidad $a \in \Sigma_n$ asociado a μ en un vector de probabilidad en Σ_m cuya j -ésima coordenada es $\sum_{i:T(x_i)=y_j} a_i$. Con esta nueva notación se puede formular lo que se conoce como el problema de Monge:

Definición 2.1 (problema de Monge, versión discreta). Dados dos conjuntos finitos, $\mathcal{X} = \{x_i\}_{i=1}^n$ e $\mathcal{Y} = \{y_j\}_{j=1}^m$, y dos medidas de probabilidad $\mu \in \mathcal{M}_+^1(\mathcal{X})$ y $\nu \in \mathcal{M}_+^1(\mathcal{Y})$ con vectores de probabilidad $a \in \Sigma_n$ y $b \in \Sigma_m$ respectivamente, el problema de Monge entre μ y ν para un funcional de costo $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ es:

$$\inf_{\substack{T : \mathcal{X} \rightarrow \mathcal{Y} \\ T_\# \mu = \nu}} \sum_{i=1}^n c(x_i, T(x_i)). \quad (2.1.3)$$

Además, a una solución óptima T^* del problema (2.1.3), en caso de existir, se le denomina *mapa de Monge*.

Como este problema es solo una reformulación del problema (2.1.1), la existencia de un mapa $T : \mathcal{X} \rightarrow \mathcal{Y}$ que satisface $T_\# \mu = \nu$ no es segura y la solución del problema, en caso de existir, no tiene por qué ser única. Sin embargo, cuando $n = m$ y las medidas μ y ν son uniformes (i.e., $a = b = \frac{1}{n} \mathbf{1}$), se puede demostrar que el problema (2.1.3), equivalente en este caso al problema de matching (2.0.1), posee solución (ver Proposición 2.2).

Formulación continua

El problema de Monge discreto (2.1.1) (o (2.1.3) en su versión medible) puede ser extendido a un problema de transporte continuo, donde se busca trasladar una cierta cantidad de masa, no necesariamente puntual, de un lugar a otro. Para motivar su formulación es usual considerar el problema de trasladar eficientemente

⁵Si bien en el caso discreto no son evidentes las ventajas de esta formulación, sí se vuelve más natural al estudiar el problema de Kantorovich en la Sección 2.2.

⁶Esto no limita el alcance de la formulación ya que cualquier vector $v \in \mathbb{R}_+^d$ puede ser normalizado al simplex Σ_d .

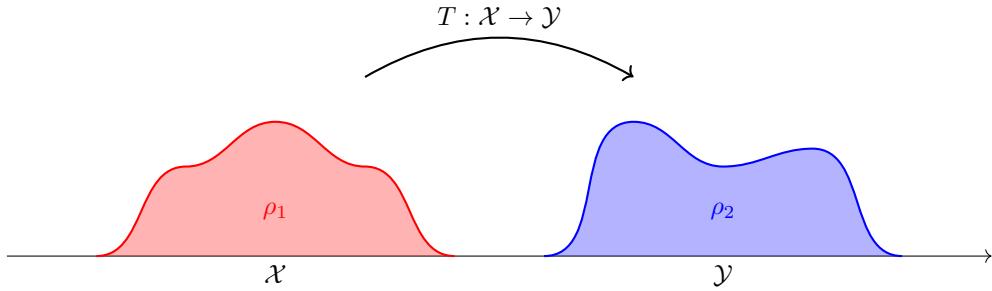


Figura 2.1: Problema del transporte de tierra, donde el montículo rojo (con densidad de masa ρ_1) debe ser trasladado al montículo azul (con densidad de masa ρ_2). Para un mapa de transporte T , un volumen $A \subset \mathcal{X}$ es transportado hacia el volumen $T(A) = \{T(x) : x \in A\} \subset \mathcal{Y}$. Inversamente, un volumen $B \subset \mathcal{Y}$ es formado por la tierra proveniente de $T^{-1}(B) = \{x \in A : T(x) \in B\} \subset \mathcal{X}$. Notar que la masa de un volumen $A \subset \mathcal{X}$ se puede calcular mediante $m_1(A) = \int_A \rho_1(x) dx$ (análogo para \mathcal{Y}). Esta figura se puede encontrar en el archivo `earth_mover.ipynb`.

un montículo de tierra con una forma y densidad específica hacia otro lugar con una forma y densidad no necesariamente igual a la original. Este problema se ilustra en la Figura 2.1.

En este caso, el problema de transporte consiste en encontrar un mapa de transporte $T : \mathcal{X} \rightarrow \mathcal{Y}$ (con $\mathcal{X} = \mathcal{Y} = \mathbb{R}^2$) que minimice un cierto costo de transporte (p.g. la distancia recorrida durante el transporte) pero que también respete una condición de preservación de masa. Para esto último, es necesario, al igual que en el caso discreto, exigir que la masa de cada volumen de tierra $B \subset \mathbb{R}^2$ en el montículo de destino sea igual a la masa total que llega desde el montículo de origen hacia B . En consecuencia, si los montículos de origen y destino tienen densidades de masa $\rho_1, \rho_2 : \mathbb{R}^2 \rightarrow \mathbb{R}_+$ respectivamente, el mapa de transporte T debe cumplir

$$\int_B \rho_2(y) dy = \int_{T^{-1}(B)} \rho_1(x) dx, \quad \forall B \in \mathcal{B}(\mathcal{Y}), \quad (2.1.4)$$

donde $T^{-1}(B) = \{x \in \mathcal{X} : T(x) \in B\} \in \mathcal{B}(\mathcal{X})$ es el volumen de masa en el montículo de origen que es transportado hacia B y $\mathcal{B}(\mathcal{X})$ (resp. $\mathcal{B}(\mathcal{Y})$) es el conjunto de conjuntos medibles en \mathcal{X} (resp. \mathcal{Y} ; ver Definición A.1). Notar que esta restricción es análoga a la restricción del problema discreto (2.1.1). Además, considerando las funciones de masa $m_1(A) := \int_A \rho_1(x) dx$ y $m_2(B) := \int_B \rho_2(y) dy$, se vuelve claro que la condición (2.1.4) es una condición de preservación de masa:

$$m_2(B) = m_1(T^{-1}(B)), \quad \forall B \in \mathcal{B}(\mathcal{Y}). \quad (2.1.5)$$

Al igual que en la formulación discreta, es posible interpretar las funciones de masa m_1 y m_2 como medidas (en este caso en \mathbb{R}^2) μ y ν respectivamente donde, además, se observa que la existencia de las densidades ρ_1 y ρ_2 no es necesaria para la formulación ya que se puede trabajar directamente con la restricción (2.1.5) en vez de (2.1.4). En ambos casos, la condición de conservación de masa en (2.1.5) o (2.1.4) se denota, al igual que en el caso discreto, como $T_\# \mu = \nu$, donde $T_\# \mu \in \mathcal{M}_+^1(\mathcal{Y})$ es la medida push-forward inducida por μ a través de T .

Esto motiva a construir la siguiente definición, la cual también es válida en el caso discreto:

Definición 2.2 (mapa de transporte). Una función $T : \mathcal{X} \rightarrow \mathcal{Y}$ es un *mapa de transporte* entre dos medidas de probabilidad $\mu \in \mathcal{M}_+^1(\mathcal{X})$ y $\nu \in \mathcal{M}_+^1(\mathcal{Y})$ si $T_\# \mu = \nu$.

Equivalentemente, si se tienen dos variables aleatorias con $x \sim \mu$ e $y \sim \nu$, entonces T es un mapa de transporte

si $y = T(x)$.

Notar que en el caso discreto es posible escribir explícitamente la medida push-forward $T_\# \mu$ de acuerdo a (2.1.2). En la Proposición A.2 se encuentran algunas propiedades del operador push-forward en el caso general. Con esta reinterpretación, el problema de Monge toma la siguiente forma en su versión continua:

Definición 2.3 (problema de Monge, versión continua). Dados los conjuntos $\mathcal{X} \subset \mathbb{R}^n$ e $\mathcal{Y} \subset \mathbb{R}^m$, el problema de Monge entre dos medidas de probabilidad, $\mu \in \mathcal{M}_+^1(\mathcal{X})$ y $\nu \in \mathcal{M}_+^1(\mathcal{Y})$, asociado a un funcional de costo integrable $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ es:

$$\inf_{\substack{T: \mathcal{X} \rightarrow \mathcal{Y} \\ T_\# \mu = \nu}} \int_{\mathcal{X}} c(x, T(x)) \, d\mu(x). \quad (2.1.6)$$

Además, a una solución óptima T^* del problema (2.1.6), en caso de existir, se le denomina *mapa de Monge*.

Notar que el conjunto factible es precisamente el conjunto de mapas de transporte. Por otra parte, cuando las medidas $\mu \in \mathcal{M}_+^1(\mathcal{X})$ y $\nu \in \mathcal{M}_+^1(\mathcal{Y})$ tienen funciones de densidad p_μ y p_ν respectivamente (ver Teorema A.1), el problema (2.1.6) se puede reescribir como:

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

donde, de acuerdo a la Proposición A.1, la restricción $T_\# \mu = \nu$ equivale a:

$$\int_{\mathcal{Y}} h(y) p_\nu(y) \, dy = \int_{\mathcal{X}} h(T(x)) p_\mu(x) \, dx, \quad \forall h \in \mathcal{C}(\mathcal{Y}),$$

donde $\mathcal{C}(\mathcal{Y})$ es el conjunto de funciones $f : \mathcal{Y} \rightarrow \mathbb{R}$ continuas.

En la próxima sección se verán limitaciones de este problema y se estudiará una formulación alternativa para el problema de transporte óptimo, la cual tendrá mejores propiedades tanto teóricas como prácticas.

2.2. Relajación de Kantorovich

Si bien el problema de Monge es más general que el problema inicial (2.0.1), no tiene propiedades deseables como ser un problema de optimización convexo. En efecto, en el caso continuo con $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, si las medidas μ y ν poseen funciones de densidad p_μ y p_ν respectivamente, se puede demostrar que la restricción $T_\# \mu = \nu$ es equivalente a que se cumpla la ecuación $p_\mu(x) = p_\nu(T(x)) |\det(\nabla_x T(x))|$, la cual es altamente no lineal.

En esta sección se relajará el problema de Monge, permitiendo repartir la masa de los elementos de \mathcal{X} en múltiples elementos de \mathcal{Y} , lo cual se puede interpretar como una asignación probabilística para el problema de transporte de Monge. Este nuevo problema, en su versión discreta, tendrá la propiedad de ser convexo y, más importante aún, permitirá definir una métrica en (un subconjunto de) el espacio de medidas de probabilidad, $\mathcal{M}_+^1(\mathcal{X})$.

2.2.1. Formulación primal

Al igual que para el problema de Monge, el problema de Kantorovich será formulado inicialmente en una versión discreta y luego extendido a su formulación continua. Además, al ser un problema de optimización convexo, será natural estudiar su problema dual, lo cual se realizará en la Subsección 2.2.2.

Problema discreto

En el problema de transporte de materia prima introducido anteriormente, es razonable considerar que un mismo yacimiento pueda enviar parte de su materia prima a distintas fábricas, lo cual no está permitido en la formulación discreta del problema de Monge, (2.1.3). Notar que esta relajación es análoga a la realizada en el problema de Monge donde, a diferencia del problema inicial (2.0.1), se permitió que elementos de \mathcal{Y} estén enlazados a más de un elemento de \mathcal{X} .

Sin embargo, permitir una asignación múltiple cambia considerablemente el problema de Monge discreto (2.1.3). En efecto, el mapa buscado $T : \mathcal{X} \rightarrow \mathcal{Y}$ deja de ser una función, perdiendo todas las propiedades que esto implicaba. Es por este motivo que para formular este nuevo problema, ahora se buscará una función $\tilde{T} : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_+$, donde $\tilde{T}(x, y)$ indica la cantidad de masa que se transportará desde $x \in \mathcal{X}$ hacia $y \in \mathcal{Y}$. Dado que se está en un marco discreto, esto es equivalente a buscar una matriz de transporte $P \in \mathcal{M}_{n,m}([0, 1])$ donde P_{ij} indica la cantidad de masa que se debe trasladar desde $x_i \in \mathcal{X}$ hacia $y_j \in \mathcal{Y}$.

Por otra parte, es necesario que la matriz de transporte P respete la conservación de masa (equilibrio ofertademanda) tanto en la medida de origen $\mu \in \mathcal{M}_+^1(\mathcal{X})$ como en la medida de destino $\nu \in \mathcal{M}_+^1(\mathcal{Y})$. En el problema de transporte desde yacimientos hacia fábricas, la i -ésima fila de la matriz P representa cómo se reparte la materia prima de $x_i \in \mathcal{X}$, por lo que se debe imponer que la suma de dicha fila sea igual al valor $a_i \in [0, 1]$. Del mismo modo, la j -ésima columna de P representa la cantidad de materia prima que llega a la fábrica $y_j \in \mathcal{Y}$, por lo que la suma de dicha columna debe ser igual a $b_j \in [0, 1]$.

Por último, para formular el problema de forma clara, el funcional de costo $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ usado en el problema de Monge (2.1.3) se identificará con una matriz $C \in \mathcal{M}_{n,m}(\mathbb{R})$ mediante $C_{ij} = c(x_i, y_j)$, por lo que $C_{ij} \in \mathbb{R}$ indica el costo de transportar una unidad de masa desde $x_i \in \mathcal{X}$ hacia $y_j \in \mathcal{Y}$. De esta forma, el costo total de transporte consistirá en el producto Frobenius entre la matriz de asignación de masa $P \in \mathcal{M}_{n,m}([0, 1])$ y la matriz de costo $C \in \mathcal{M}_{n,m}(\mathbb{R})$:

Definición 2.4 (problema de Kantorovich, caso discreto). Dados dos conjuntos finitos, $\mathcal{X} = \{x_i\}_{i=1}^n$ e $\mathcal{Y} = \{y_j\}_{j=1}^m$, y dos medidas de probabilidad $\mu \in \mathcal{M}_+^1(\mathcal{X})$ y $\nu \in \mathcal{M}_+^1(\mathcal{Y})$ con vectores de probabilidad $a \in \Sigma_n$ y $b \in \Sigma_m$ respectivamente, el problema de Kantorovich entre μ y ν para una matriz de costo $C \in \mathcal{M}_{n,m}(\mathbb{R})$ es:

$$\inf_{P \in \Pi_d(\mu, \nu)} \langle C, P \rangle_F := \sum_{i=1}^n \sum_{j=1}^m C_{ij} P_{ij}, \quad (2.2.1)$$

donde la región factible $\Pi_d(\mu, \nu)$ se define como:

$$\Pi_d(\mu, \nu) := \left\{ P \in \mathcal{M}_{n,m}([0, 1]) : \sum_{j=1}^m P_{ij} = a_i, \forall i \in \{1, \dots, n\}, \quad \sum_{i=1}^n P_{ij} = b_j, \forall j \in \{1, \dots, m\} \right\}.$$

Además, a una solución óptima P^* del problema (2.2.1), en caso de existir, se le denomina *plan de Kantorovich*.

A modo de ejemplo, en la Figura 2.2 se puede ver la solución para un problema de transporte discreto específico.

Notar que el conjunto $\Pi_d(\mu, \nu)$ es no vacío (por lo que el problema de optimización siempre es factible) ya que siempre se puede considerar la matriz $P \in \mathcal{M}_{n,m}([0, 1])$ definida como $P_{ij} = a_i b_j$, la cual pertenece a $\Pi_d(\mu, \nu)$. En efecto, dado que $a \in \Sigma_n$ y $b \in \Sigma_m$:

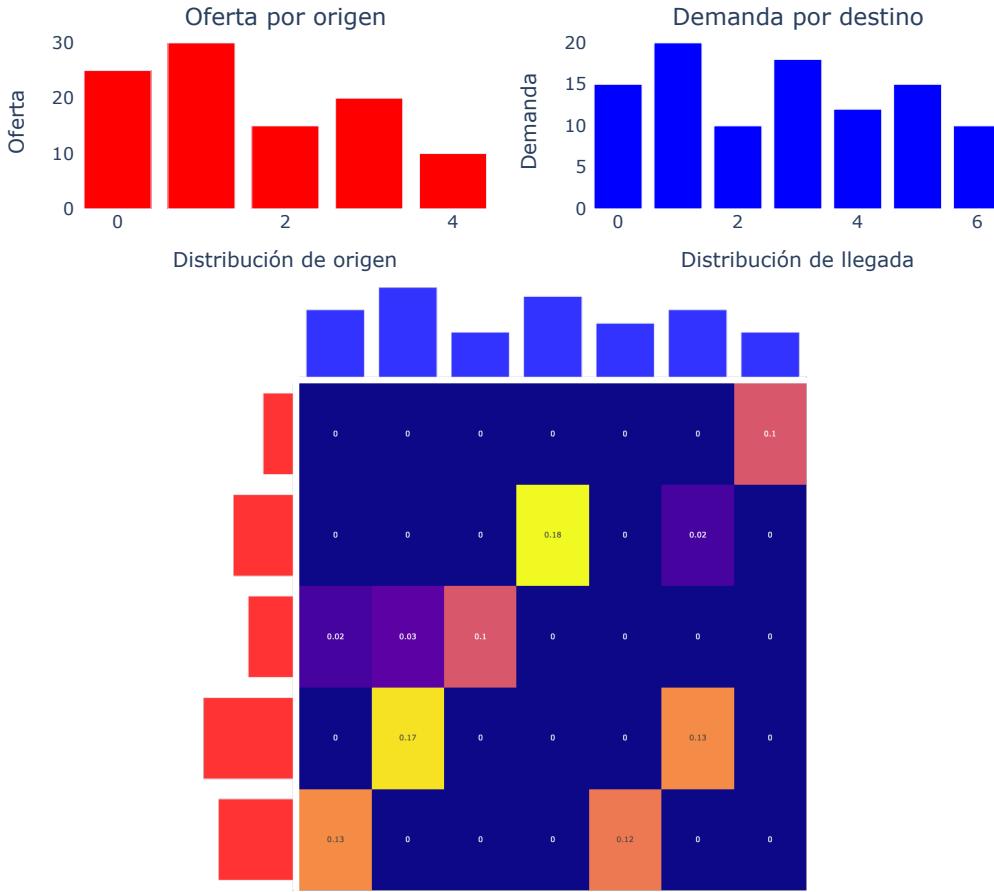


Figura 2.2: (Arriba) histogramas discretos (no normalizados) asociados a las medidas μ y ν respectivamente, donde $|X| = 5$ y $|Y| = 7$. (Abajo) solución (normalizada) para el problema de Kantorovich entre las dos distribuciones discretas. El código de esta simulación se encuentra en el archivo `kantorovich.ipynb`.

$$\sum_{j=1}^m P_{ij} = a_i \sum_{j=1}^m b_j = a_i, \forall i \in \{1, \dots, n\}, \quad \sum_{i=1}^n P_{ij} = b_j \sum_{i=1}^n a_i = b_j, \forall j \in \{1, \dots, m\}.$$

Más adelante será útil recordar que esta matriz se puede escribir como $P = ab^\top$. Por otra parte, la restricción de conservación de masa para μ se puede escribir vectorialmente como $P\mathbf{1}_m = a$, mientras que para ν , se puede escribir como $P^\top\mathbf{1}_n = b$, es decir:

$$\Pi_d(\mu, \nu) = \{P \in \mathcal{M}_{n,m}([0, 1]) : P\mathbf{1}_m = a, P^\top\mathbf{1}_n = b\}. \quad (2.2.2)$$

Por otro lado, a diferencia de lo que ocurre en el problema de Monge, el conjunto factible es simétrico en el sentido que $P \in \Pi_d(\mu, \nu)$ si y solo si $P^\top \in \Pi_d(\nu, \mu)$. Sin embargo, este problema a gran escala puede ser costoso de resolver ya que para encontrar el plan de transporte óptimo P^* es necesario encontrar $n \cdot m$ incógnitas. Al estudiar la formulación dual de este problema en la Subsección 2.2.2 se podrá reducir significativamente la cantidad de incógnitas.

Para el análisis de la existencia y unicidad de la solución, es útil notar que las $n + m$ restricciones de igualdad que definen $\Pi_d(\mu, \nu)$ son lineales y, en consecuencia, $\Pi_d(\mu, \nu)$ es un polítopo⁷ (ver Figura 2.3), por

⁷En este trabajo se definirá un polítopo como un poliedro convexo y acotado (i.e., sin parte cónica en la descomposición de

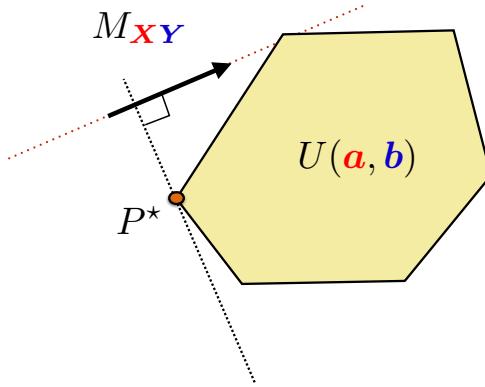


Figura 2.3: Polítopo $\Pi_d(\mu, \nu)$ representado en el plano por un conjunto $U(a, b)$ de vectores de transporte, donde la matriz de costo C es representado por un vector M_{XY} . El costo de transporte $\langle C, P \rangle_F = M_{XY}^\top P$ aumenta al moverse en la dirección del vector de costo M_{XY} , por lo que para minimizar el costo de transporte se busca el elemento $P \in U(a, b)$ que más pueda avanzar en el sentido contrario al vector de costo M_{XY} sin salirse del polítopo $U(a, b)$, lo cual se consigue en el vértice P^* . Notar que en este caso el minimizador del problema de transporte es único. En cambio, si el vector de costo M_{XY} tuviese la inclinación precisa para ser ortogonal a alguna de las facetas que define el polítopo, entonces podrían existir infinitas soluciones. Imagen adaptada desde [Cut+17].

lo que (2.2.1) es un problema de programación lineal estándar. En consecuencia, este problema siempre posee solución ya que, como se vio anteriormente, siempre es factible ($\Pi_d(\mu, \nu)$ es no vacío). Sin embargo, si bien este nuevo problema es un problema de optimización convexo, no es estrictamente convexo, por lo que no se podrá garantizar la unicidad de la solución. En el Capítulo 3 se regularizará este problema agregando un término regularizador a la función objetivo, transformando el problema en uno estrictamente convexo. Este nuevo problema regularizado resultará ser equivalente al problema del puente de Schrödinger en su versión estática.

Con respecto a la relación entre los valores óptimos del problema de Monge (2.1.3) y del problema de Kantorovich (2.2.1), si bien ambos valores no coinciden en general, sí se puede afirmar que el valor óptimo para el problema de Kantorovich es una cota inferior del valor óptimo para el problema de Monge ya que, como se verá en la demostración, siempre se puede construir un plan de transporte $P^{T^*} \in \Pi_d(\mu, \nu)$ a partir de un mapa de Monge $T^* : \mathcal{X} \rightarrow \mathcal{Y}$.

Proposición 2.1. Se tiene la siguiente desigualdad cuando el problema de Monge es factible:

$$\inf_{P \in \Pi_d(\mu, \nu)} \langle C, P \rangle_F \leq \inf_{\substack{T: \mathcal{X} \rightarrow \mathcal{Y} \\ T_\# \mu = \nu}} \sum_{i=1}^n c(x_i, T(x_i)).$$

Demostración. Dado un plan de Monge $T^* : \mathcal{X} \rightarrow \mathcal{Y}$, este permite definir un plan de transporte $P^{T^*} \in \Pi_d(\mu, \nu)$ determinista en el sentido que $x_i \in \mathcal{X}$ le entrega toda su masa a $T(x_i) \in \mathcal{Y}$, sin realizar división de masa (ver Figura 2.4):

$$P_{ij}^{T^*} = \begin{cases} a_i & \text{si } T^*(x_i) = y_j, \\ 0 & \text{si no.} \end{cases} \quad (2.2.3)$$

Minkowski-Weyl).

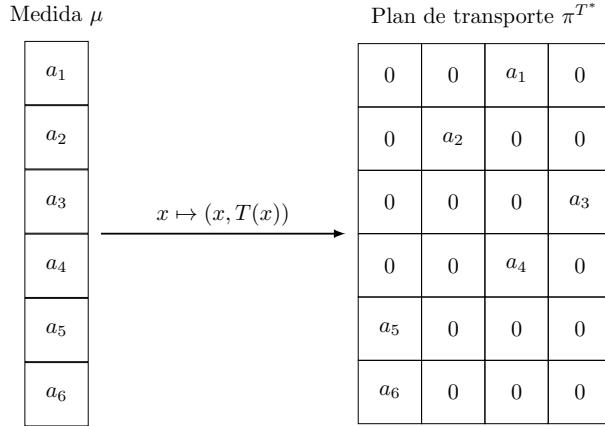


Figura 2.4: Plan de transporte determinista inducido por un mapa de Monge $T^* : \mathcal{X} \rightarrow \mathcal{Y}$ definido como $T^*(x_5) = T^*(x_6) = y_1$, $T^*(x_2) = y_2$, $T^*(x_1) = T^*(x_4) = y_3$, $T^*(x_3) = y_4$. Notar que la medida producto $\pi^{T^*} \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})$ asociada a este plan de transporte puede verse como la medida push-forward inducida por μ a través del mapa $x \in \mathcal{X} \mapsto (x, T(x)) \in \mathcal{X} \times \mathcal{Y}$. Esta observación será importante en la formulación continua al enunciar el Teorema 2.2. La imagen se encuentra en el archivo `deterministic_plan.ipynb`.

Notar que este plan de transporte efectivamente está en $\Pi_d(\mu, \nu)$, ya que respeta las distribuciones marginales. Para ver esto, se usará la notación $j(i)$ para indicar el valor $j \in \{1, \dots, m\}$ tal que $T(x_i) = y_j$. En particular, (2.2.3) indica que $P_{i,j(i)}^{T^*} = a_i$. Luego:

$$\sum_{j=1}^m P_{ij}^{T^*} = P_{i,j(i)}^{T^*} = a_i$$

$$\sum_{i=1}^n P_{ij}^{T^*} = \sum_{i: T^*(x_i)=y_j} P_{i,j(i)}^{T^*} = \sum_{i: T^*(x_i)=y_j} a_i = b_j.$$

donde en la última igualdad se usó la condición de preservación de masa, $T_\# \mu = \nu$ (ver (2.1.1)). Por otra parte, el costo de Kantorovich para este plan de transporte es igual al costo de Monge para el mapa T^* :

$$\sum_{i=1}^n \sum_{j=1}^m C_{ij} P_{ij}^{T^*} = \sum_{i=1}^n C_{i,j(i)} P_{i,j(i)}^{T^*} = \sum_{i=1}^n C_{i,j(i)} a_i = \sum_{i=1}^n c(x_i, T^*(x_i)),$$

donde en la última igualdad se usó que la matriz de costo en el problema de Kantorovich corresponde al costo marginal obtenido a partir del funcional de costo para Monge. En consecuencia, el valor óptimo para el problema de Kantorovich es una cota para el valor óptimo del problema de Monge:

$$\inf_{P \in \Pi_d(\mu, \nu)} \langle C, P \rangle_F \leq \left\langle C, P^{T^*} \right\rangle_F = \sum_{i=1}^n c(x_i, T^*(x_i)) = \inf_{T: \mathcal{X} \rightarrow \mathcal{Y} \atop T_\# \mu = \nu} \sum_{i=1}^n c(x_i, T(x_i)).$$

□

De la demostración es importante notar que, si bien la desigualdad anterior es en general estricta, para alcanzar la igualdad basta encontrar un plan de Kantorovich P^* que sea determinista (en el sentido que lo es P^{T^*} en (2.2.3)). En este caso, el mapa de transporte $T : \mathcal{X} \rightarrow \mathcal{Y}$ que induce este plan determinista es óptimo para el problema de Monge. Esta propiedad se observa al considerar medidas uniformes: en la Sección 2.1 se

indicó que el problema de Monge discreto tiene una solución $T^* : \mathcal{X} \rightarrow \mathcal{Y}$ cuando $n = m$ y $a = b = \frac{1}{n}\mathbf{1}$. En este escenario, el plan de transporte determinista que induce T^* , P^{T^*} , resulta ser óptimo para el problema de Kantorovich:

Proposición 2.2. Dados dos conjuntos discretos, \mathcal{X} y \mathcal{Y} , con el mismo cardinal ($n = m$) y dos medidas uniformes

$$\mu = \sum_{i=1}^n \frac{1}{n} \delta_{x_i} \in \mathcal{M}_+^1(\mathcal{X}), \quad \nu = \sum_{j=1}^n \frac{1}{n} \delta_{y_j} \in \mathcal{M}_+^1(\mathcal{Y}).$$

Entonces, tanto el problema de Monge discreto, (2.1.3), como el problema de Kantorovich discreto, (2.2.1) tienen solución. Más aún, una mapa de Kantorovich P^* corresponde a la matriz de permutación (escalada) asociada a un mapa de Monge T^* :

$$P_{ij}^* = \begin{cases} \frac{1}{n} & \text{si } T^*(x_i) = y_j, \\ 0 & \text{si no.} \end{cases}$$

Es decir, P envía toda la masa desde $x_i \in \mathcal{X}$ hasta $T(x_i) \in \mathcal{Y}$, para todo $i \in \{1, \dots, n\}$. Es importante indicar que, si bien el escenario que cubre esta proposición es uno de los más simples, la existencia del mapa de Monge T^* no es directa ya que la demostración de esta propiedad hace uso del teorema de Birkhoff–von Neumann⁸ (ver, por ejemplo, [Tho18]). Por otra parte, en el Teorema 2.2 se obtiene un resultado similar para el caso continuo en \mathbb{R}^d .

Por otra parte, al igual que para el problema de Monge, se le puede dar una interpretación de medida a este problema. En efecto, cada plan de transporte $P \in \Pi_d(\mu, \nu)$ en el problema discreto (2.2.1) se puede identificar con una medida discreta $\pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})$, donde P_{ij} corresponde a la masa que μ le entrega al elemento $(x_i, y_j) \in \mathcal{X} \times \mathcal{Y}$. En consecuencia, debido a las restricciones de perservación de masa, esta medida π tendrá primera marginal $\mu \in \mathcal{M}_+^1(\mathcal{X})$ y segunda marginal $\nu \in \mathcal{M}_+^1(\mathcal{Y})$. De este modo, la identificación del conjunto de planes de transporte $\Pi_d(\mu, \nu)$ en el conjunto de medidas $\mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})$ ocurre mediante la inyección

$$P \in \Pi_d(\mu, \nu) \mapsto \pi = \sum_{i=1}^n \sum_{j=1}^m P_{ij} \delta_{(x_i, y_j)} \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y}). \quad (2.2.4)$$

Esta identificación permitirá extender el problema de Kantorovich discreto al caso general.

Problema continuo

El conjunto factible discreto $\Pi_d(\mu, \nu)$ se puede extender al caso continuo $\mathcal{X} \subset \mathbb{R}^n$, $\mathcal{Y} \subset \mathbb{R}^m$ como las medidas en $\mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})$ que tienen primera marginal $\mu \in \mathcal{M}_+^1(\mathcal{X})$ y segunda marginal $\nu \in \mathcal{M}_+^1(\mathcal{Y})$. Dicho conjunto, denotado por $\Pi(\mu, \nu)$, se denomina conjunto de *couplings*⁹ entre μ y ν y, al igual que para $\Pi_d(\mu, \nu)$ en el caso discreto, es no vacío ya que, al igual que en el caso discreto, siempre contiene la medida producto $\mu \otimes \nu \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})$ definida como $(\mu \otimes \nu)(A \times B) = \mu(A)\nu(B)$ para $A \in \mathcal{B}(\mathcal{X})$, $B \in \mathcal{B}(\mathcal{Y})$. En efecto, la primera marginal para esta medida es

$$(\mu \otimes \nu)_1(A) = (\mu \otimes \nu)(A \times \mathcal{Y}) = \int_{A \times \mathcal{Y}} d(\mu \otimes \nu)(x, y) = \int_A \left(\int_{\mathcal{Y}} d\nu(y) \right) d\mu(x) = \int_A d\mu(x) = \mu(A).$$

⁸Este teorema afirma que los vértices del polítopo formado por las matrices doblemente estocásticas corresponden precisamente a matrices de permutación.

⁹Un coupling también se conoce como *plan de transporte* debido a su interpretación en transporte óptimo. Por lo tanto, se utilizarán ambos términos de forma equivalente.

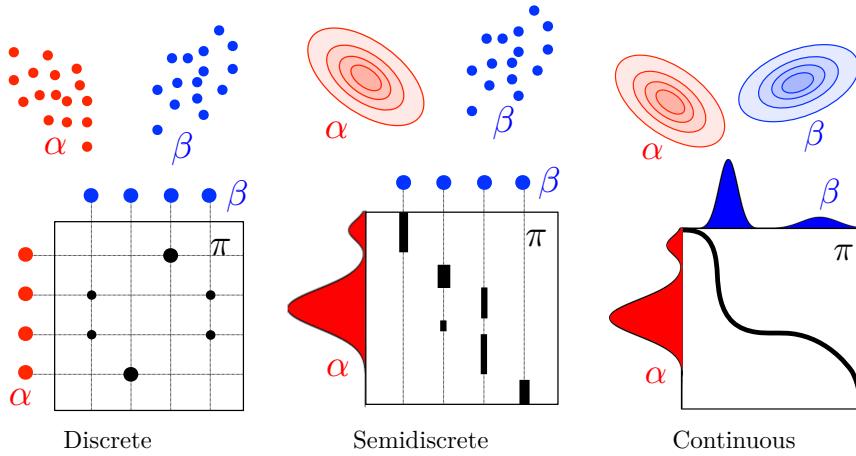


Figura 2.5: Ejemplos de couplings entre las medidas de probabilidad α y β cuando se considera un marco discreto (izquierda), semidiscreto (centro) o continuo (derecha). Imagen obtenida desde [PC20].

Donde se usó que $\nu \in \mathcal{M}_+^1(\mathcal{Y})$ es medida de probabilidad. Del mismo modo se muestra que $(\mu \otimes \nu)_2 = \nu$ por lo que, efectivamente, $\mu \otimes \nu \in \Pi(\mu, \nu)$.

Notar que en el caso discreto (i.e., cuando \mathcal{X} e \mathcal{Y} son conjuntos finitos), el conjunto factible $\Pi_d(\mu, \nu)$ se puede identificar con $\Pi(\mu, \nu)$ vía la inyección dada en (2.2.4) (que restringida al codominio $\Pi(\mu, \nu) \subset \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})$ es biyección). En este caso, la matriz de probabilidad $P = ab^\top \in \Pi_d(\mu, \nu)$ se identifica precisamente con la medida producto $\mu \otimes \nu \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})$. En la Figura 2.5 se puede observar ejemplos de couplings cuando se trabaja con medidas discretas, continuas o mixtas.

Por otro lado, el funcional de costo para la formulación continua ahora es una función con dominio $\mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^n \times \mathbb{R}^m$, por lo que ya no es posible identificarlo con una matriz. Con estas observaciones, el problema de Kantorovich discreto (2.2.1) se generaliza al caso continuo de la siguiente forma:

Definición 2.5 (problema de Kantorovich, caso continuo). Dados los conjuntos $\mathcal{X} \subset \mathbb{R}^n$ e $\mathcal{Y} \subset \mathbb{R}^m$, el problema de Kantorovich entre dos medidas de probabilidad, $\mu \in \mathcal{M}_+^1(\mathcal{X})$ y $\nu \in \mathcal{M}_+^1(\mathcal{Y})$, asociado a un funcional de costo integrable $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ es:

$$\inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \, d\pi(x, y), \quad (2.2.5)$$

donde el conjunto factible es el conjunto de couplings entre μ y ν :

$$\Pi(\mu, \nu) := \left\{ \pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y}) : \pi_1(A) = \mu(A), \forall A \in \mathcal{B}(\mathcal{X}), \quad \pi_2(B) = \nu(B), \forall B \in \mathcal{B}(\mathcal{Y}) \right\}. \quad (2.2.6)$$

Además, a una solución óptima π^* del problema (2.2.6), en caso de existir, se le denomina *plan de Kantorovich*.

La restricción $\pi_1 = \mu$ para $\pi \in \Pi(\mu, \nu)$ indica que para cualquier volumen de origen $A \in \mathcal{B}(\mathcal{X})$, la cantidad de masa transferida desde A hacia el espacio de llegada \mathcal{Y} debe ser igual a la masa original de A . Análogamente, la restricción $\pi_2 = \nu$ indica que para cualquier volumen de llegada $B \in \mathcal{B}(\mathcal{Y})$, la cantidad de masa transferida desde \mathcal{X} hacia B debe ser igual a la masa original de B .

En la Figura 2.6 se puede ver un ejemplo de transporte óptimo continuo entre dos mixturas gaussianas, mientras que en la Figura 2.7 y la Figura 2.8 se pueden ver ejemplos de solución para el problema de

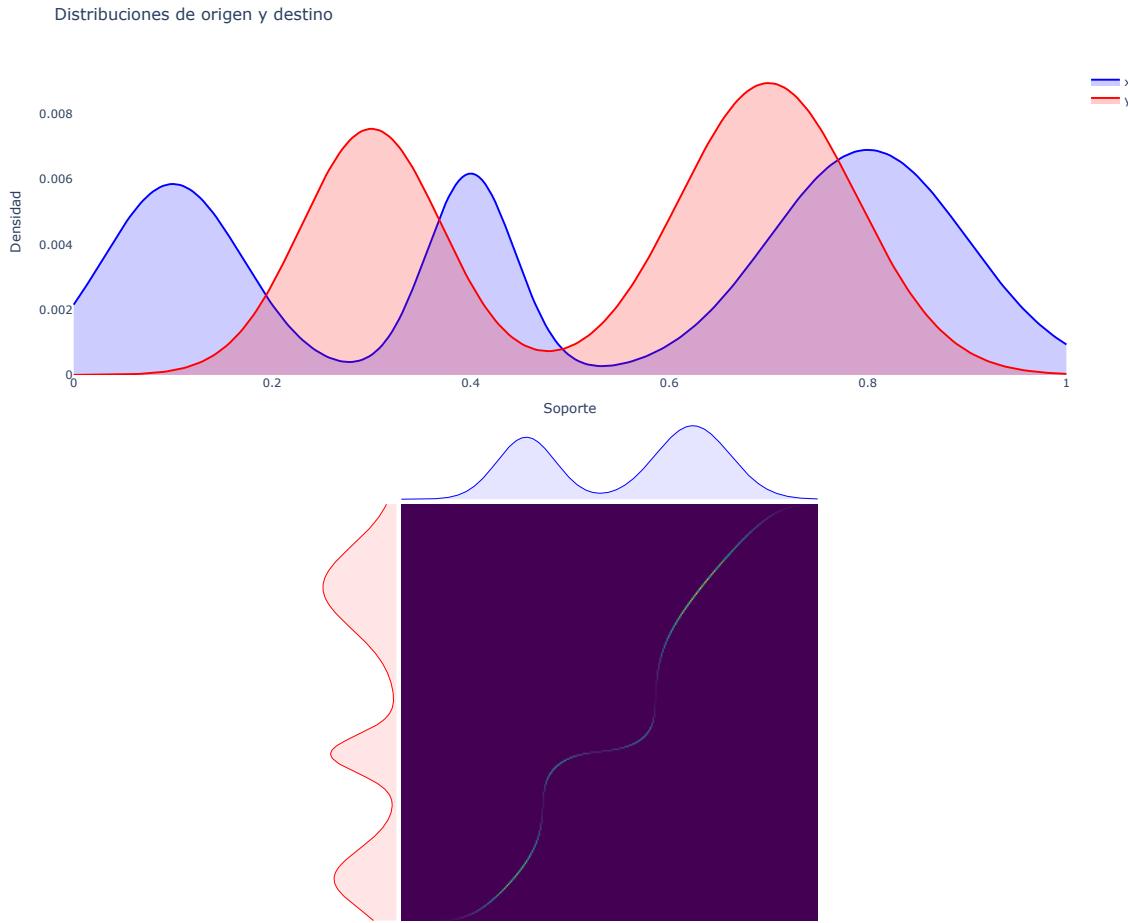


Figura 2.6: (Arriba) gráficos de funciones de densidad asociadas a las medidas μ y ν respectivamente. (Abajo) solución para el problema de Kantorovich entre ambas medidas. Se observa como toda la masa de π^* se concentra en una curva, la cual corresponde al gráfico del mapa de Monge para este mismo problema (ver Teorema 2.2). El código de esta simulación se encuentra en el archivo `kantorovich.ipynb`.

Kantorovich en su formulación discreta, continua y mixta.

Es importante notar que los planes de transporte $\pi \in \Pi(\mu, \nu)$ se interpretan, al igual que en el caso discreto, considerando que $d\pi(x, y)$ es la cantidad infinitesimal de masa transferida desde dx hacia dy . Es decir, para un volumen $A \in \mathcal{B}(\mathcal{X})$ en el espacio de origen y un volumen $B \in \mathcal{B}(\mathcal{Y})$ en el espacio de llegada, la cantidad de masa transferida desde A hacia B es precisamente $\pi(A \times B)$. En particular, al igual que para el caso discreto, la medida producto $\mu \otimes \nu \in \Pi(\mu, \nu)$ es el plan de transporte que envía cada porción $A \in \mathcal{B}(\mathcal{X})$ a todo el espacio de llegada \mathcal{Y} , repartiendo su masa de forma proporcional a la distribución que asigna ν sobre los elementos de \mathcal{Y} .

Cuando se plantea el problema de transporte óptimo sobre un espacio común $\mathcal{X} = \mathcal{Y} \subset \mathbb{R}^d$ (como en el problema de transporte de tierra de la Figura 2.1), es usual trabajar con funcionales de costo $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ de la forma $c(x, y) = d(x - y)$, donde $d : \mathbb{R}^d \rightarrow \mathbb{R}$ una función convexa por lo que, en estos casos, el problema de Kantorovich general sigue siendo un problema de optimización convexo. Un caso importante es considerar a d como una norma en \mathbb{R}^d ya que bajo este funcional de costo, el problema de Kantorovich permite definir una métrica entre las medidas μ y ν , denominada distancia de Wasserstein y estudiada en la Subsección 2.3.1.

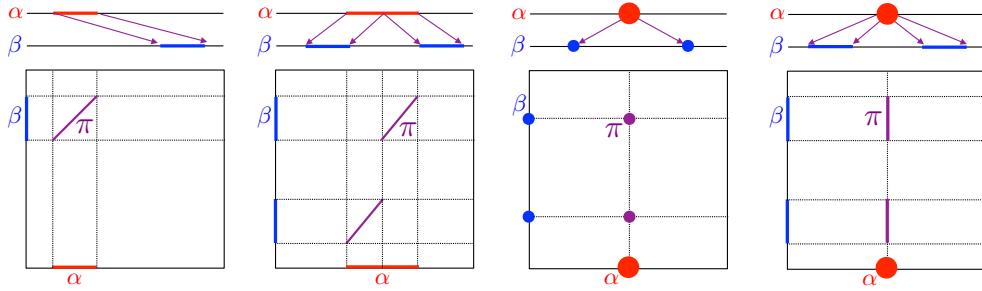


Figura 2.7: Plan de Kantorovich para distintos pares de medidas α y β . (Arriba) las flechas indican cómo se debe repartir la masa de α en β para el transporte óptimo en el sentido de Kantorovich. En el sentido de Monge, solo el 1º caso tiene solución ya que los otros casos requieren división de masa. (Abajo) coupling óptimo del problema. Imagen obtenida desde [PC20].

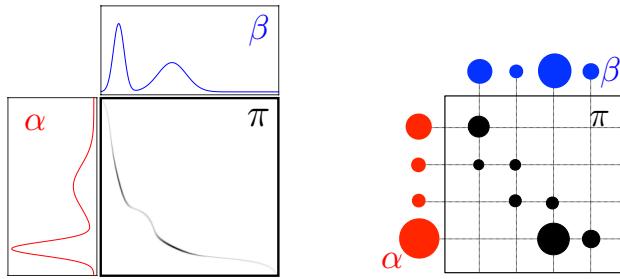


Figura 2.8: Solución para el problema de Kantorovich entre dos medidas α y β . (Izquierda) coupling óptimo entre dos medidas continuas, donde la intensidad del color negro en la curva indica la densidad de masa del coupling óptimo π en cada punto del plano $\mathcal{X} \times \mathcal{Y}$. Notar que, en este caso, el plan de Kantorovich está soportado precisamente en el gráfico del mapa de Monge. (Derecha) plan de Kantorovich π entre dos medidas discretas, donde el radio de cada círculo es proporcional a la masa asignada a cada átomo de $\mathcal{X} \times \mathcal{Y}$. Imagen obtenida desde [PC20].

En particular, a lo largo de este trabajo será usual considerar un costo cuadrático mediante $d(z) = \frac{1}{2} \|z\|^2$.

Con respecto a la existencia de una solución, si bien el conjunto factible $\Pi(\mu, \nu)$ es siempre no vacío, es necesario agregar una hipótesis de regularidad sobre la función de costo para garantizar que el ínfimo efectivamente se alcanza en algún plan de transporte óptimo:

Proposición 2.3. Dadas dos medidas de probabilidad, $\mu \in \mathcal{M}_+^1(\mathcal{X})$, $\nu \in \mathcal{M}_+^1(\mathcal{Y})$ sobre los conjuntos $\mathcal{X} = \mathbb{R}^n$ e $\mathcal{Y} = \mathbb{R}^m$, si $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ es un funcional continuo, entonces el problema de Kantorovich continuo (2.2.5) posee solución.

Para demostrar esta propiedad es usual estudiar el problema dual continuo (2.2.9), el cual suele ser más fácil de trabajar. Sin embargo, en [VS03] realizan una demostración bajo hipótesis más débiles haciendo uso de resultados de topología y teoría de la medida¹⁰.

Con respecto a la relación entre el problema de Monge y el problema de Kantorovich, se tiene la misma

¹⁰El teorema de Prokhorov permite concluir que $\Pi(\mu, \nu)$ es compacto para la convergencia débil (ver Definición 2.14), mientras que una semicontinuidad inferior para c permite concluir que el funcional $\pi \mapsto \int c d\pi$ es continuo bajo esta topología, concluyendo el resultado mediante el teorema de Weierstrass.

desigualdad que en el caso discreto (ver Proposición 2.1). Además, se precisará una condición suficiente para alcanzar la igualdad, la cual será útil en el Teorema 2.2:

Proposición 2.4. Se tiene la siguiente desigualdad para los problemas de Monge y Kantorovich:

$$\inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y) \leq \inf_{\substack{T: \mathcal{X} \rightarrow \mathcal{Y} \\ T_\# \mu = \nu}} \int_{\mathcal{X}} c(x, T(x)) d\mu(x).$$

Además, para que la igualdad se alcance basta que exista un plan de Kantorovich *determinista* π^* de la forma $\pi^* = (\text{Id} \otimes T)_\# \mu$, donde $T : \mathcal{X} \rightarrow \mathcal{Y}$ es algún un mapa de transporte. Más aún, este mapa de transporte T es óptimo para el problema de Monge.

En la proposición anterior, la igualdad $\pi^* = (\text{Id} \otimes T)_\# \mu$ indica que el plan de transporte π^* es la medida push-forward que induce μ en $\mathcal{X} \times \mathcal{Y}$ a través de la función $x \in \mathcal{X} \mapsto (x, T(x)) \in \mathcal{X} \times \mathcal{Y}$ (ver Figura 2.4). Este plan de transporte se considera determinista debido a que no realiza división de masa. En efecto, por la propia definición de π^* , su soporte está contenido en el grafo del mapa de transporte T :

$$\text{Supp}(\pi^*) \subset \text{gr}(T) := \{(x, T(x)) : x \in \mathcal{X}\} \subset \mathcal{X} \times \mathcal{Y}.$$

Por lo que es posible recuperar un mapa de transporte (el mismo mapa T) a partir de π^* , lo cual no es posible para planes de transporte generales que no tienen esta propiedad de determinismo. En estos casos, se debe recurrir a otras técnicas como las proyecciones baricéntricas, las cuales son usadas en el Capítulo 3 y estudiadas en más detalle en [PC20].

En el caso particular $\mathcal{X} = \mathcal{Y} \subset \mathbb{R}^d$, bajo algunas hipótesis adicionales, sí es posible encontrar planes de Kantorovich deterministas. Sin embargo, para obtener este resultado será necesario estudiar la formulación dual del problema de Kantorovich la cual será utilizada posteriormente, en su versión regularizada, para desarrollar el algoritmo de Sinkhorn, el cual es el algoritmo por defecto para resolver el problema del puente de Schrödinger.

Para el estudio de planes de transporte deterministas, es útil que una medida conjunta de la forma $\pi = (\text{Id} \otimes T)_\# \mu$ se puede expresar equivalentemente como $d\pi(x, y) = d\mu(x)\delta_{y=T(x)}$, lo cual evidencia el determinismo de este tipo de planes de transporte: la masa transportada desde x hacia y (indicada por $d\pi(x, y)$) es, o bien toda la masa $d\mu(x)$ si $y = T(x)$, o bien cero.

Por otra parte, en algunos casos es útil la siguiente propiedad: la pertenencia al conjunto de couplings $\Pi(\mu, \nu)$ se puede caracterizar mediante funciones test.

Proposición 2.5 (caracterización de couplings). Dadas dos medidas de probabilidad, $\mu \in \mathcal{M}_+^1(\mathcal{X})$ y $\nu \in \mathcal{M}_+^1(\mathcal{Y})$, definidas sobre $\mathcal{X} = \mathbb{R}^n$ e $\mathcal{Y} = \mathbb{R}^m$. Entonces, para $\pi \in \mathcal{M}_+^1(\mu, \nu)$ se cumple que $\pi \in \Pi(\mu, \nu)$ si y solo si

$$\int_{\mathcal{X} \times \mathcal{Y}} (\phi \oplus \psi)(x, y) d\pi(x, y) = \int_{\mathcal{X}} \phi(x) d\mu(x) + \int_{\mathcal{Y}} \psi(y) d\nu(y), \quad \forall (\phi, \psi) \in L^1(\mathcal{X}) \times L^1(\mathcal{Y}),$$

donde $\phi \oplus \psi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ es la suma tensorial: $(\phi \oplus \psi)(x, y) := \phi(x) + \psi(y)$.

Por último, dado que los planes de transporte son elementos de $\mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})$, siempre se puede asociar una variable aleatoria conjunta $(x, y) \sim \pi$ a un plan de transporte $\pi \in \Pi(\mu, \nu)$, donde la primera marginal, x tendrá distribución μ , mientras que la segunda marginal, y , tendrá distribución ν . Además, esta interpretación permite escribir el problema de Kantorovich como

$$\min_{\pi \in \Pi(\mu, \nu)} \mathbb{E}_{(x, y) \sim \pi} [c(x, y)].$$

2.2.2. Formulación dual

Dado que el problema de Kantorovich es un problema de minimización convexo (al menos su versión discreta (2.2.1)), es natural preguntarse por su problema dual, el cual es un problema de maximización convexo. Como se verá a continuación, la formulación dual reduce significativamente la cantidad de incógnitas buscadas en el problema de optimización. Además, el hecho de trabajar con un problema primal convexo permitirá concluir que hay dualidad fuerte tanto en el caso discreto como en el caso continuo.

Problema dual discreto

Dado que el problema de Kantorovich discreto (2.2.1) es un problema de programación lineal, su problema dual es otro problema de programación lineal cuya forma es conocida (ver, por ejemplo, [BV04]):

Proposición 2.6 (dualidad fuerte para Kantorovich, caso discreto). El problema dual del problema de Kantorovich discreto (2.2.1) es:

$$\sup_{\substack{(\phi, \psi) \in \mathbb{R}^n \times \mathbb{R}^m \\ \phi \oplus \psi \leq C}} \langle \phi, a \rangle + \langle \psi, b \rangle = \sum_{i=1}^n \phi_i a_i + \sum_{j=1}^m \psi_j b_j, \quad (2.2.7)$$

donde la restricción $\phi \oplus \psi \leq C$ indica que $\phi_i + \psi_j \leq C_{ij}$, $\forall i \in \{1, \dots, n\}$, $\forall j \in \{1, \dots, m\}$. Además, hay dualidad fuerte:

$$\inf_{P \in \Pi_d(\mu, \nu)} \langle C, P \rangle_F = \sup_{\substack{(\phi, \psi) \in \mathbb{R}^n \times \mathbb{R}^m \\ \phi \oplus \psi \leq C}} \langle \phi, a \rangle + \langle \psi, b \rangle. \quad (2.2.8)$$

Por otra parte, a un par de vectores duales óptimos, $(\phi^*, \psi^*) \in \mathbb{R}^n \times \mathbb{R}^m$, se les denomina *potenciales de Kantorovich*.

Notar que en esta proposición, la notación $\phi \oplus \psi \in \mathcal{M}_{n,m}(\mathbb{R})$ representa la suma tensorial entre ϕ y ψ : $(\phi \oplus \psi)_{ij} = \phi_i + \psi_j$ o, equivalentemente, $\phi \oplus \psi = \phi \mathbf{1}_m^\top + \mathbf{1}_n \psi^\top$. Por otra parte, la propiedad de dualidad permite caracterizar las soluciones primales-duales mediante las condiciones de Karush-Kuhn-Tucker (ver [BV04]). Más precisamente, se puede concluir el siguiente resultado:

Corolario 2.1. Sea $(\phi^*, \psi^*) \in \mathbb{R}^n \times \mathbb{R}^m$ un par de potenciales de Kantorovich asociados al problema dual (2.2.7). Entonces, cualquier solución $P^* \in \Pi_d(\mu, \nu)$ para el problema primal (2.2.1) tiene su soporte en las coordenadas donde se cumple la restricción dual con igualdad:

$$\{(i, j) : P_{ij}^* > 0\} \subset \{(i, j) : \phi_i^* + \psi_j^* = C_{ij}\}, \subset \{1, \dots, n\} \times \{1, \dots, m\}.$$

Demostración. Esta propiedad es consecuencia directa de la propiedad de holgura complementaria. En efecto, la restricción dual asociada a la variable primal P_{ij} es $\phi_i + \psi_j - C_{ij} \leq 0$, por lo que la propiedad de holgura complementaria indica que $P_{ij} (\phi_i + \psi_j - C_{ij}) = 0$. En particular, si $P_{ij} > 0$ entonces $\phi_i + \psi_j - C_{ij} = 0$. \square

Una ventaja importante que se obtiene al usar la formulación dual se observa al momento de intentar resolver numéricamente el problema de Kantorovich. En efecto, para resolver el problema de Kantorovich discreto (2.2.1) es necesario encontrar una matriz $P \in \Pi_d(\mu, \nu)$ ($n \cdot m$ incógnitas), mientras que al usar la formulación discreta dual (2.2.7) solo es necesario encontrar dos vectores $(\phi, \psi) \in \mathbb{R}^n \times \mathbb{R}^m$ ($n + m$ incógnitas). Sin

embargo, una solución dual (ϕ^*, ψ^*) no permite obtener directamente una solución primal como sí ocurre en la formulación regularizada del problema de Kantorovich (ver Proposición 3.5) ya que no es posible *despejar* una solución primal desde la formulación dual.

Por otra parte, una forma usual de interpretar el problema de Kantorovich discreto dual es mediante el problema de transporte desde yacimientos hacia fábricas. Para esto, se puede considerar que el dueño de los yacimientos decide contratar una empresa externa que realice el servicio de transporte de la materia prima, pagando una cantidad ϕ_i por cargar una unidad de materia prima desde el yacimiento i , y una cantidad ψ_j por descargar una unidad de materia prima en la fábrica j , sin importar la forma en cómo se distribuyen los recursos de acuerdo a un plan de transporte. Desde el punto de vista de la empresa externa, esta buscará vectores de cobro $\phi \in \mathbb{R}^n$ y $\psi \in \mathbb{R}^m$ que maximicen su ganancia y considerando que $\phi_i + \psi_j \leq C_{ij}$ ya que, de lo contrario, el dueño de los yacimientos no tendrá motivos para realizar el transporte con esta empresa. La dualidad fuerte entregada en (2.2.8) afirma que la empresa externa puede encontrar vectores de cobro (ϕ^*, ψ^*) tales que su ganancia sea igual al costo de transporte original que hubiese pagado el dueño de los yacimientos si realizara el transporte sin la empresa externa.

Problema dual continuo

En el caso continuo, se tiene un resultado análogo al presentado en la Proposición 2.6. Sin embargo, al igual que para la formulación primal continua, hay que agregar hipótesis de regularidad extras sobre la función de costo para poder garantizar la dualidad fuerte.

Teorema 2.1 (dualidad fuerte para Kantorovich, caso continuo). Si $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ es un funcional de costo continuo y acotado inferiormente, entonces, el problema dual del problema de Kantorovich continuo (2.2.5) es:

$$\sup_{\substack{(\phi, \psi) \in L^1(\mathcal{X}) \times L^1(\mathcal{Y}) \\ \phi \oplus \psi \leq c}} \int_{\mathcal{X}} \phi(x) \, d\mu(x) + \int_{\mathcal{Y}} \psi(y) \, d\nu(y), \quad (2.2.9)$$

donde la restricción $\phi \oplus \psi \leq c$ indica que $(\phi \oplus \psi)(x, y) := \phi(x) + \psi(y) \leq c(x, y)$, $\forall x \in \mathcal{X}$, $\forall y \in \mathcal{Y}$ ¹¹.

Además, hay dualidad fuerte:

$$\inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \, d\pi(x, y) = \sup_{\substack{(\phi, \psi) \in L^1(\mathcal{X}) \times L^1(\mathcal{Y}) \\ \phi \oplus \psi \leq c}} \int_{\mathcal{X}} \phi(x) \, d\mu(x) + \int_{\mathcal{Y}} \psi(y) \, d\nu(y).$$

Por otra parte, a un par de funcionales duales óptimos, $(\phi^*, \psi^*) \in L^1(\mathcal{X}) \times L^1(\mathcal{Y})$, se les denomina *potenciales de Kantorovich*.

Una demostración de esta propiedad se puede encontrar en [VS03], donde además muestran que los potenciales de Kantorovich $(\phi^*, \psi^*) \in L^1(\mathcal{X}) \times L^1(\mathcal{Y})$ se pueden elegir continuos y acotados, pudiendo optimizar el problema dual sobre $\mathcal{C}_b(\mathcal{X}) \times \mathcal{C}_b(\mathcal{Y}) \subset L^1(\mathcal{X}) \times L^1(\mathcal{Y})$. Esto es posible debido a que las funciones en $\mathcal{C}_b(\mathbb{R}^d)$ son capaces de aproximar suficientemente bien a las funciones en $L^1(\mathbb{R}^d)$ ¹². De todos modos, por simplicidad, se seguirán considerando que los potenciales de Kantorovich son funciones genéricas en $L^1(\mathcal{X})$ y $L^1(\mathcal{Y})$ respectivamente.

¹¹En rigor, esta igualdad es casi seguramente. Sin embargo, se puede demostrar que los potenciales se pueden elegir continuos, recuperando la igualdad puntual.

¹²Más precisamente, dada una medida de probabilidad $\mu \in \mathcal{M}_+^1(\mathbb{R}^d)$, el conjunto $\mathcal{C}_b(\mathbb{R}^d) \subset L^1(\mathbb{R}^d, \mu)$ es denso en $L^1(\mathbb{R}^d, \mu)$: para cualquier función $f \in L^1(\mathbb{R}^d, \mu)$ y error $\epsilon > 0$, existe una función $g \in \mathcal{C}_b(\mathbb{R}^d)$ tal que $\int_{\mathbb{R}^d} |f - g| \, d\mu < \epsilon$.

Al igual que para el caso discreto, el soporte para el plan de transporte óptimo está contenido en el subconjunto que activa la desigualdad dual en (2.2.9):

Corolario 2.2. Sea $(\phi^*, \psi^*) \in L^1(\mathcal{X}) \times L^1(\mathcal{Y})$ un par de potenciales de Kantorovich asociados al problema dual (2.2.9). Entonces, cualquier solución $\pi^* \in \Pi(\mu, \nu)$ para el problema primal (2.2.5) tiene su soporte en un subconjunto específico de $\mathcal{X} \times \mathcal{Y}$:

$$\text{Supp}(\pi^*) \subset \{(x, y) \in \mathcal{X} \times \mathcal{Y} : \phi^*(x) + \psi^*(y) = c(x, y)\}$$

El problema dual permitirá demostrar que el problema de Kantorovich es equivalente (en el sentido que lo indicará el Teorema 2.2) al problema de Monge en el caso particular $\mathcal{X} = \mathcal{Y} \subset \mathbb{R}^d$ cuando se considera un costo cuadrático, el cual es el escenario más importante debido a su conexión con el problema de Schrödinger con prior browniano en el Capítulo 3. Para mostrar este resultado (conocido como teorema de Brenier), se vuelve necesario revisar algunos conceptos de análisis convexo:

Definición 2.6 (subdiferencial). Dada una función convexa $f : \mathbb{R}^d \rightarrow \mathbb{R}$, su *subdiferencial* ∂f es la función (multivaluada) que asigna a cada $x \in \mathbb{R}^d$ un conjunto de subgradientes:

$$\partial f(x) := \{y \in \mathbb{R}^d : f(z) \geq \tilde{f}_y(z) := f(x) + \langle y, z - x \rangle, \forall z \in \mathbb{R}^d\}. \quad (2.2.10)$$

Este concepto permite generalizar el concepto de gradiente a funciones convexas que no necesariamente son diferenciables en todo su dominio. En efecto, la desigualdad en (2.2.10) indica que la recta tangente \tilde{f}_y (cuya pendiente es $y \in \partial f(x)$) está siempre por debajo del gráfico de f , lo cual es una propiedad que caracteriza a las funciones convexas diferenciables. En la Figura 2.9 se puede ver el subdiferencial de una función no diferenciable. Se puede probar lo siguiente (ver, por ejemplo, [VS03]):

Proposición 2.7. Sea $f : \mathbb{R}^d \rightarrow \mathbb{R}$ una función diferenciable en $x \in \mathbb{R}^d$. Entonces, $\partial f(x) = \{\nabla f(x)\}$.

Este resultado es el que permitirá posteriormente caracterizar el mapa de Monge $T : \mathcal{X} \rightarrow \mathcal{Y}$ como el gradiente de una función convexa en el Teorema 2.2.

Por otra parte, el concepto de transformada de Legendre permitirá reescribir la condición de factibilidad dual $\phi \oplus \psi \leq c$ utilizando el concepto de subdiferencial:

Definición 2.7 (transformada de Legendre). Dada una función $f : \mathbb{R}^d \rightarrow \mathbb{R}$, su *transformada de Legendre*¹³ es otra función $f_* : \mathbb{R}^d \rightarrow \mathbb{R}$ definida por:

$$f_*(y) := \sup_{x \in \mathbb{R}^d} \langle x, y \rangle - f(x).$$

La importancia de la transformada de Legendre, en este caso, es que permite caracterizar completamente el subdiferencial de una función convexa. En efecto, se tiene el siguiente resultado:

Proposición 2.8. Dada una función convexa y continua $f : \mathbb{R}^d \rightarrow \mathbb{R}$, se tiene que $\forall x, y \in \mathbb{R}^d$,

¹³En la literatura es usual usar la notación f^* en vez de f_* . Aquí se preferirá la segunda notación para no confundir con la notación usada para soluciones óptimas.

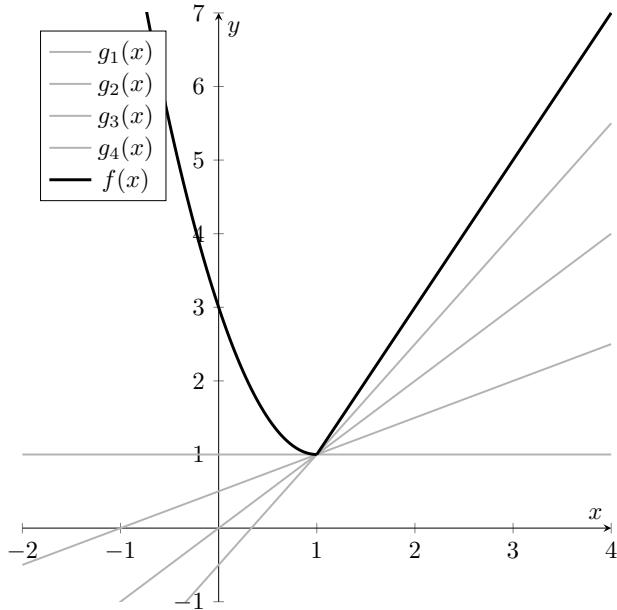


Figura 2.9: La función f no es diferenciable en $x = 1$, por lo que su conjunto de subgradientes no será un único punto. Las rectas g_i , $i \in \{1, 2, 3, 4\}$ son tangentes a f en $(1, 1)$, por lo que sus pendientes corresponden a subgradientes de f . Esta imagen se encuentra en el archivo `subdifferential.ipynb`.

$$y \in \partial f(x) \Leftrightarrow f(x) + f_*(y) = \langle x, y \rangle.$$

Con estos resultados auxiliares, se puede enunciar y demostrar el teorema de Brenier, el cual es un resultado clave que permite obtener soluciones del problema de Monge a partir del problema de Kantorovich, y viceversa (ver Figura 2.6):

Teorema 2.2 (Brenier). Sean $\mu \in \mathcal{M}_+^1(\mathcal{X})$ y $\nu \in \mathcal{M}_+^1(\mathcal{Y})$ medidas de probabilidad en $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ con segundo momento finito¹⁴:

$$\int_{\mathcal{X}} \|x\|^2 \, d\mu(x) < \infty, \quad \int_{\mathcal{Y}} \|y\|^2 \, d\nu(y) < \infty.$$

Si μ posee una función de densidad asociada¹⁵, entonces existe una única solución $\pi^* \in \Pi(\mu, \nu)$ para el problema de Kantorovich (2.2.5) con costo cuadrático, $c(x, y) = \frac{1}{2} \|x - y\|^2$. Además, esta solución es determinista (en el sentido que lo indica la Proposición 2.4) y de la forma

$$\pi^* = (\text{Id} \otimes \nabla f)_{\#} \mu,$$

para alguna función convexa $f : \mathbb{R}^d \rightarrow \mathbb{R}$.

Demostración. Para empezar, notar que la función objetivo en (2.2.5) con $c(x, y) = \frac{1}{2} \|x - y\|^2$ se puede descomponer como $c(x, y) = \frac{1}{2} (\|x\|^2 + \|y\|^2) + \langle x, y \rangle$, por lo tanto:

¹⁴Los momentos de una medida de probabilidad $\mu \in \mathcal{M}_+^1(\mathbb{R}^d)$ son los momentos de una variable aleatoria cuya ley es μ .

¹⁵Esta hipótesis se puede relajar a que $\mu \in \mathcal{M}_+^1(\mathcal{X})$ no le entregue masa a conjuntos medibles con dimensión de Hausdorff-Besicovitch menor que d .

$$\begin{aligned} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \, d\pi(x, y) &= \frac{1}{2} \int_{\mathcal{X} \times \mathcal{Y}} (\|x\|^2 + \|y\|^2) \, d\pi(x, y) - \int_{\mathcal{X} \times \mathcal{Y}} \langle x, y \rangle \, d\pi(x, y) \\ &= \underbrace{\frac{1}{2} \left(\int_{\mathcal{X}} \|x\|^2 \, d\mu(x) + \int_{\mathcal{Y}} \|y\|^2 \, d\nu(y) \right)}_{<\infty} - \int_{\mathcal{X} \times \mathcal{Y}} \langle x, y \rangle \, d\pi(x, y). \end{aligned}$$

Donde el primer sumando (finito) es constante con respecto a π , por lo que es suficiente resolver el siguiente problema de optimización:

$$\sup_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{Y}} \langle x, y \rangle \, d\pi(x, y).$$

Análogo a lo obtenido en el Teorema 2.1 (y recordando que los potenciales de Kantorovich se podían elegir continuos), el dual de este nuevo problema es:

$$\inf_{\substack{(f, g) \in \mathcal{C}_b(\mathcal{X}) \times \mathcal{C}_b(\mathcal{Y}) \\ f \oplus g \geq \langle \cdot, \cdot \rangle}} \int_{\mathcal{X}} f(x) \, d\mu(x) + \int_{\mathcal{Y}} g(y) \, d\nu(y). \quad (2.2.11)$$

Por otra parte, si se fija la primera variable dual, $f \in \mathcal{C}_b(\mathcal{X})$, la restricción $f \oplus g \geq \langle \cdot, \cdot \rangle$ se puede reescribir de la siguiente forma:

$$\begin{aligned} f(x) + g(y) &\geq \langle x, y \rangle, \quad \forall x \in \mathcal{X}, y \in \mathcal{Y} \\ \iff g(y) &\geq \langle x, y \rangle - f(x), \quad \forall x \in \mathcal{X}, y \in \mathcal{Y} \\ \iff g(y) &\geq \underbrace{\sup_{x \in \mathcal{X}} \langle x, y \rangle - f(x)}_{f_*(y)}, \quad \forall y \in \mathcal{Y}. \end{aligned}$$

Por lo tanto, para minimizar el funcional $g \mapsto \int_{\mathcal{Y}} g(y) \, d\nu(y)$ (recordar que f está fija) respetando la restricción $f \oplus g \geq \langle \cdot, \cdot \rangle$ basta tomar $g = f_*$. En consecuencia, el problema (2.2.11) se puede reescribir de forma irrestricta mediante la siguiente formulación semidual:

$$\inf_{f \in \mathcal{C}_b(\mathcal{X})} \int_{\mathcal{X}} f(x) \, d\mu(x) + \int_{\mathcal{Y}} f_*(y) \, d\nu(y).$$

Del mismo modo, fijando ahora la segunda variable dual, $g = f_*$, se muestra que la restricción $f \oplus g \geq \langle \cdot, \cdot \rangle$ es equivalente a la restricción $f(x) \geq f_{**}(x) = \sup_{y \in \mathcal{Y}} \langle x, y \rangle - f_*(x)$, $\forall x \in \mathcal{X}$, por lo que, nuevamente, se puede elegir a f_{**} como el primer potencial para minimizar ahora la primera integral.

En consecuencia, dado que la transformada de Legendre define una función convexa, se puede restringir la búsqueda a potenciales $f \in \mathcal{C}_b(\mathcal{X})$ convexos:

$$\inf_{f \in \mathcal{C}_b(\mathcal{X}) \text{ convexa}} \int_{\mathcal{X}} f(x) \, d\mu(x) + \int_{\mathcal{Y}} f_*(y) \, d\nu(y).$$

Por otra parte, si f^* es una variable dual óptima, por la condición de holgura complementaria (análogo a lo realizado en el Corolario 2.2):

$$\text{Supp}(\pi^*) \subset \{(x, y) \in \mathcal{X} \times \mathcal{Y} : f^*(x) + f_*^*(y) = \langle x, y \rangle\}.$$

Notar que, gracias a la Proposición 2.8, la igualdad $f^*(x) + f_*^*(y) = \langle x, y \rangle$ se alcanza si y solo si $y \in \partial f^*(x)$. Además, como la variable dual f^* es convexa y μ tiene función de densidad, resulta ser que f^* es diferenciable

en \mathcal{X}^{16} . De esta forma, $\partial f(x) = \{\nabla f(x)\}$, y así

$$\text{Supp}(\pi^*) \subset \{(x, \nabla f^*(x)) : x \in \mathcal{X}\} \subset \mathcal{X} \times \mathcal{Y}.$$

Por lo que π^* es una medida producto soportada en el grafo de la función ∇f^* y así, es un plan de transporte determinista con $\pi^* = (\text{Id} \otimes \nabla f^*)_\# \mu$. \square

Una consecuencia directa de este teorema es que permite obtener resultados de existencia y unicidad para el problema de Monge con costo cuadrático:

Corolario 2.3. Bajo las hipótesis del Teorema 2.2, el mapa de transporte $T = \nabla f$ es la única solución del problema de Monge con funcional de costo $c(x, y) = \frac{1}{2} \|x - y\|^2$.

Demostración. Notar que, por definición de $\pi^* = (\text{Id} \otimes \nabla f)_\# \mu$, el mapa $x \in \mathcal{X} \mapsto \nabla f(x) \in \mathcal{Y}$ cumple la propiedad $(\nabla f)_\# \mu = \nu$, por lo que $T = \nabla f$ es efectivamente un mapa de transporte. Por otra parte, dado que π^* es un plan de transporte determinista y óptimo, por la Proposición 2.4 se concluye que ∇f es solución del problema de Monge. Además, como el Teorema 2.2 garantiza unicidad para π^* , se obtiene la unicidad de la solución para el problema de Monge. \square

2.2.3. Casos particulares en \mathbb{R}^d

Si bien los problemas de transporte óptimo son complejos de analizar en general, existen algunos casos particulares en donde la solución se puede caracterizar fácilmente.

Transporte óptimo unidimensional

En el caso unidimensional $\mathcal{X} = \mathcal{Y} = \mathbb{R}$, la solución para los problemas de transporte óptimo se pueden definir haciendo uso de la función de distribución acumulada y su respectiva inversa, la función cuantil. La primera de estas funciones tiene la siguiente definición:

Definición 2.8 (función de distribución acumulada). Dada una medida de probabilidad $\mu \in \mathcal{M}_+^1(\mathbb{R})$, su función de distribución acumulada $F_\mu : \mathbb{R} \rightarrow [0, 1]$ se define como:

$$F_\mu(x) := \int_{-\infty}^x d\mu(s) = \int_{-\infty}^x p(s) ds,$$

donde la última igualdad solo tiene sentido si μ posee una función de densidad p .

Es importante recordar que la función de distribución acumulada F_μ identifica totalmente a la medida de probabilidad μ^{17} . El siguiente teorema entrega la solución para el problema de Kantorovich entre dos medidas de probabilidad en \mathbb{R} :

Teorema 2.3. Sean $\mu \in \mathcal{M}_+^1(\mathcal{X})$ y $\nu \in \mathcal{M}_+^1(\mathcal{Y})$ dos medidas de probabilidad en $\mathcal{X} = \mathcal{Y} = \mathbb{R}$ con funciones de distribución acumulada F_μ y F_ν respectivamente. Entonces, el plan de transporte óptimo, $\pi^* \in \Pi(\mu, \nu)$, para el problema de Kantorovich entre μ y ν con función de costo $c(x, y) = |x - y|$ tiene la siguiente función de distribución acumulada:

¹⁶Esto es consecuencia del teorema de Rademacher.

¹⁷Dado que $\mu((a, b]) = F_\mu(b) - F_\mu(a)$, se tiene que F_μ determina el valor de μ en una semi-álgebra de \mathbb{R} , por lo cual la extensión de Carathéodory es única.

$$F_{\pi^*}(x, y) = \inf\{F_\mu(x), F_\nu(y)\}.$$

Además, el valor óptimo del problema de Kantorovich es la distancia $L^1(\mathbb{R})$ entre las funciones de distribución acumulada de μ y ν :

$$\inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{Y}} |x - y| d\pi(x, y) = \|F_\mu - F_\nu\|_{L^1(\mathcal{X})} := \int_{-\infty}^{\infty} |F_\mu(z) - F_\nu(z)| dz.$$

Para extender el resultado a funcionales de costo de la forma $c(x, y) = |x - y|^p$ será necesario introducir la función cuantil:

Definición 2.9 (función cuantil). Dada una medida de probabilidad $\mu \in \mathcal{M}_+^1(\mathbb{R})$ con función de distribución acumulada F_μ , se define la *función cuantil* de μ , $Q_\mu : [0, 1] \rightarrow \mathbb{R}$, como:

$$Q_\mu(z) := \inf\{x \in \mathbb{R} : F_\mu(x) \geq z\}.$$

Esta función actúa como función inversa de la función de distribución acumulada ya que si F_μ es además estrictamente creciente se obtiene que $Q_\mu(z) = F_\mu^{-1}(z)$, mientras que en los casos donde no hay invertibilidad, se tiene que $F_\mu^{-1}(F_\mu(x)) \geq x$ y $F(F_\mu^{-1}(z)) \geq z$, por lo que se suele decir que es su inversa por la izquierda.

Con esta definición, se tiene el siguiente resultado:

Teorema 2.4. Sean $\mu \in \mathcal{M}_+^1(\mathcal{X})$ y $\nu \in \mathcal{M}_+^1(\mathcal{Y})$ dos medidas de probabilidad en $\mathcal{X} = \mathcal{Y} = \mathbb{R}$ con funciones de distribución acumulada F_μ y F_ν , respectivamente. Entonces, para el funcional de costo $c(x, y) = |x - y|^p$, $p \geq 1$, se tiene que un mapa de transporte óptimo para el problema de Monge (2.1.6) es

$$T(x) = Q_\nu(F_\mu(x)).$$

Mientras que el valor óptimo para el problema de Kantorovich es la distancia L^1 de sus funciones cuantil:

$$\|Q_\mu - Q_\nu\|_{L^1([0, 1])} := \int_0^1 |Q_\mu(z) - Q_\nu(z)|^p dz.$$

Las demostraciones de estos resultados se pueden encontrar en [Tho18].

Transporte óptimo entre distribuciones gaussianas

Otro caso donde se puede encontrar una solución de forma cerrada ocurre cuando se consideran las leyes de dos distribuciones gaussianas con un funcional de costo cuadrático:

Teorema 2.5. Dadas dos variables aleatorias gaussianas, $x \sim \mathcal{N}(\mu_1, \Sigma_1)$ e $y \sim \mathcal{N}(\mu_2, \Sigma_2)$, si se considera $\mu = \text{Ley}(x)$ y $\nu = \text{Ley}(y)$ las respectivas medidas asociadas a x e y respectivamente¹⁸, entonces, el valor óptimo para el problema de Kantorovich (2.2.5) entre μ y ν con $c(x, y) = \|x - y\|^2$ es

$$\|\mu_1 - \mu_2\|_2^2 + \text{Tr} \left(\Sigma_1 + \Sigma_2 - 2 \left(\Sigma_1^{\frac{1}{2}} \Sigma_2 \Sigma_1^{\frac{1}{2}} \right)^{\frac{1}{2}} \right).$$

Además, el mapa de Monge asociado $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$ viene dado por

$$T(x) = \mu_2 + A(x - \mu_1), \quad \text{donde } A = \Sigma_1^{-1/2} \left(\Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2} \right)^{1/2} \Sigma_1^{-1/2}.$$

¹⁸Para una variable aleatoria $x \sim p(x)$ (con $p : \mathcal{X} \rightarrow \mathbb{R}_+$ su función de densidad), su ley es la medida $\mu_x(A) = \int_A p(x) dx$

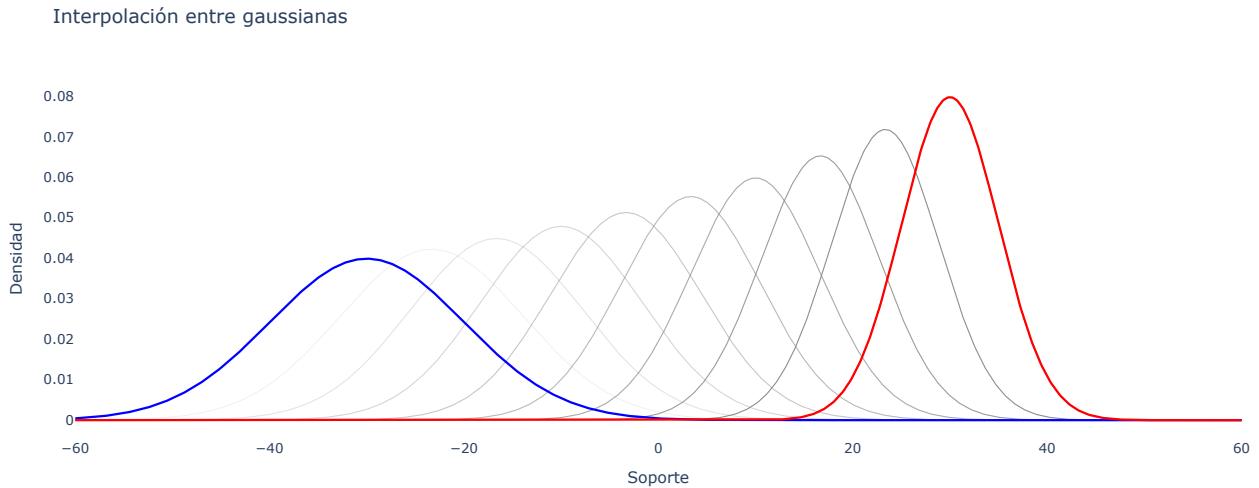


Figura 2.10: Interpolación entre dos curvas gaussianas utilizando la solución cerrada del Teorema 2.5. El código de esta simulación se encuentra en el archivo `distr_interpolation.ipynb`.

La demostración de esta propiedad se puede encontrar en [PC20]. Por otro lado, en la Figura 2.10 se observa el problema de transporte entre dos gaussianas de una forma dinámica, donde se ve la *trayectoria* que la distribución debe seguir durante el transporte. Esto será estudiado en la Sección 2.3.

El Teorema 2.5 entrega de forma explícita el mapa de Monge para el problema de transporte óptimo, el cual se sabe que existe gracias al Teorema 2.2. Además, es importante notar que Σ_1 y Σ_2 son matrices de covarianzas, por lo que deben ser simétricas y definidas positivas (asumiendo que x e y son variables aleatorias gaussianas no degeneradas). De este modo, la descomposición de Cholesky garantiza que sus raíces cuadradas están bien definidas y, por lo tanto, la matriz A está bien definida.

Por otra parte, la matriz A puede ser interpretada como una transformación lineal. De esta forma, el operador T es un operador afín que primero traslada las masas de $x \sim \mathcal{N}(\mu_1, \Sigma_1)$ al origen mediante $x - \mu_1$ y luego aplica la transformación lineal. Finalmente, T realiza una traslación al centro de $\mathcal{N}(\mu_2, \Sigma_2)$.

La transformación $x \mapsto Ax$ busca deformar la distribución $x \sim \mathcal{N}(\mu_1, \Sigma_1)$ para hacerla similar a la distribución $\mathcal{N}(\mu_2, \Sigma_2)$. Esto se vuelve evidente cuando se considera $d = 1$ ya que, en ese caso, $A = \frac{\sigma_2}{\sigma_1} \in \mathbb{R}_+$ es la raíz cuadrada del ratio entre ambas varianzas.

Por último, es importante destacar que el valor óptimo para el problema de Kantorovich entre gaussianas es utilizado en la métrica FID al momento de evaluar algunos modelos generativos como los modelos de difusión (ver, por ejemplo, [Bet+22]).

Mapa de Monge a partir de un potencial de Kantorovich

En la demostración del Teorema 2.2 se utilizó la transformada de Legendre (Definición 2.7) para reformular el problema dual (restringido) en un problema irrestricto, limitado únicamente a buscar sobre un tipo particular de funciones. En este apartado se extenderá esta idea al problema de Kantorovich con funcionales arbitrarios. Para un estudio más detallado, ver [VS03].

En el problema dual (2.2.9), los potenciales admisibles $(\phi, \psi) \in L^1(\mathcal{X}) \times L^1(\mathcal{Y})$ deben cumplir la restricción $\phi \oplus \psi \leq c$. Con esto, repitiendo la idea de la demostración del Teorema 2.2, si se fija el primer potencial $\phi \in L^1(\mathcal{X})$, el segundo potencial $\psi \in L^1(\mathcal{Y})$ debe ser tal que

$$\psi(y) \leq c(x, y) - \phi(x), \quad \forall x \in \mathcal{X}, y \in \mathcal{Y} \iff \psi(y) \leq \inf_{x \in \mathcal{X}} c(x, y) - \phi(x).$$

Por lo que se puede elegir el segundo potencial como la función $\phi^c(y) = \inf_{x \in \mathcal{X}} c(x, y) - \phi(x)$ para maximizar la función objetivo cuando el primer potencial está fijo. Considerando ahora este nuevo potencial $\phi^c \in L^1(\mathcal{Y})$ como fijo, se puede realizar un procedimiento análogo para actualizar el primer potencial:

$$\phi(x) \leq c(x, y) - \phi^c(y), \quad \forall x \in \mathcal{X}, y \in \mathcal{Y} \iff \phi(x) \leq \inf_{y \in \mathcal{Y}} c(x, y) - \phi^c(y).$$

Por lo que se puede actualizar el primer potencial a la función $\phi^{c\bar{c}}(x) := \inf_{y \in \mathcal{Y}} c(x, y) - \phi^c(y)$ para maximizar la función objetivo cuando el segundo potencial está fijo. Este procedimiento motiva las siguientes definiciones, las cuales se pueden considerar como una generalización de la transformada de Legendre:

Definición 2.10 (c -transformadas y concavidad). Sean $f : \mathcal{X} \rightarrow \mathbb{R}$ y $g : \mathcal{Y} \rightarrow \mathbb{R}$ funciones arbitrarias. Dado un funcional $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ acotado, se definen las siguientes transformaciones:

- c -transformada de f : función $f^c : \mathcal{Y} \rightarrow \mathbb{R}$ definida como $f^c(y) := \inf_{x \in \mathcal{X}} c(x, y) - f(x)$.
- \bar{c} -transformada de g : función $g^{\bar{c}} : \mathcal{X} \rightarrow \mathbb{R}$ definida como $g^{\bar{c}}(x) := \inf_{y \in \mathcal{Y}} c(x, y) - g(y)$.

Por otra parte, una función $f : \mathcal{X} \rightarrow \mathbb{R}$ se dirá c -cóncava si existe una función $g : \mathcal{Y} \rightarrow \mathbb{R}$ tal que $f = g^{\bar{c}}$. Análogamente, una función $g : \mathcal{Y} \rightarrow \mathbb{R}$ se dirá \bar{c} -cóncava si existe una función $f : \mathcal{X} \rightarrow \mathbb{R}$ tal que $g = f^c$.

Notar que la única diferencia entre la c -transformada y la \bar{c} -transformada es la variable sobre la que actúa el funcional c . Por lo tanto, si c es simétrico (i.e., $c(x, y) = c(y, x)$), ambas transformaciones son iguales.

Con estas definiciones, el procedimiento descrito anteriormente comienza con un primer potencial arbitrario, $\phi \in L^1(\mathcal{X})$, y se va actualizando alternadamente el otro potencial de tal forma que en cada actualización el valor del funcional de optimización dual en (2.2.9) no empeora (y, eventualmente, mejora). Por lo tanto, la cadena de actualizaciones de potenciales es la siguiente:

$$(\phi, \phi^c) \rightarrow (\phi^{c\bar{c}}, \phi^c) \rightarrow (\phi^{c\bar{c}}, \phi^{c\bar{c}c}). \quad (2.2.12)$$

Si bien se podrían realizar estas iteraciones indefinidamente, se puede demostrar que $\phi^{c\bar{c}c} = \phi^c$, por lo que la tercera actualización, $(\phi^{c\bar{c}}, \phi^{c\bar{c}c})$, es equivalente a $(\phi^{c\bar{c}}, \phi^c)$, lo cual corresponde a la segunda iteración. Por lo tanto, este procedimiento no permite mejorar la función objetivo más allá de dos iteraciones. De todos modos, este método sugiere que basta encontrar un único potencial $\phi \in L^1(\mathcal{X})$ para el problema dual (2.2.9). Observando que $f = \phi^{c\bar{c}}$ en (2.2.12) es una función c -cóncava, se puede probar el siguiente resultado:

Teorema 2.6 (formulación dual mediante c -transformadas). Sea $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ un funcional de costo continuo y acotado con $\mathcal{X} \subset \mathbb{R}^n$ e $\mathcal{Y} \subset \mathbb{R}^m$. Para dos medidas de probabilidad, $\mu \in \mathcal{M}_+^1(\mathcal{X})$ y $\nu \in \mathcal{M}_+^1(\mathcal{Y})$, el problema dual (2.2.9) es equivalente a:

$$\sup_{f \in L^1(\mathcal{X}) \text{ } c\text{-cóncava}} \int_{\mathcal{X}} f(x) d\mu(x) + \int_{\mathcal{Y}} f^c(y) d\nu(y).$$

En particular, se pueden elegir potenciales de Kantorovich de la forma $(\eta^{c\bar{c}}, \eta^c)$, para algún $\eta \in L^1(\mathcal{X})$.

Más aún, se tiene la siguiente caracterización para el mapa de transporte óptimo:

Teorema 2.7 (Gangbo-McCann). Sean $\mu \in \mathcal{M}_+^1(\mathcal{X})$ y $\nu \in \mathcal{M}_+^1(\mathcal{Y})$ dos medidas de probabilidad en $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$. Si f^* es el primer potencial óptimo c -convexo para el problema de Kantorovich entre μ y ν , entonces, el mapa de transporte óptimo para el problema de Monge, $T^* : \mathcal{X} \rightarrow \mathcal{Y}$ es:

$$T^*(x) = (\nabla_x c(x, \cdot)^{-1} \circ \nabla f^*)(x).$$

Notar que el Teorema 2.2 es un caso particular de este resultado. La relación entre T^* y f^* se simplifica aún más cuando el costo es invariante bajo traslaciones, es decir, si se puede escribir de la forma $c(x, y) = h(x - y)$, con $h : \mathbb{R}^d \rightarrow \mathbb{R}$ una función convexa. En este caso, a la función h se le denomina *potencial de costo*.

Corolario 2.4. Bajo las hipótesis del Teorema 2.7, si además el funcional de costo tiene un potencial de costo convexo, $h : \mathbb{R}^d \rightarrow \mathbb{R}$, entonces:

$$T^*(x) = x - (\nabla h_* \circ \nabla f^*)(x).$$

Una demostración de estos resultados se puede encontrar, por ejemplo, en [Bun23b].

2.3. Formulación dinámica

En esta sección se abordará una perspectiva dinámica del problema de transporte óptimo, que complementa y extiende la formulación estática vista anteriormente. La formulación dinámica permite interpretar el transporte de masa como un proceso continuo en el tiempo, permitiendo analizar escenarios más complejos como, por ejemplo, fenómenos dinámicos biológicos (ver [Bun23a]). Este nuevo enfoque permitirá establecer conexiones entre el transporte óptimo y otras áreas como el control óptimo en la Subsección 2.3.2 y la mecánica de fluidos en la Subsección 2.3.3. Además, dadas las propiedades métricas que se estudiarán en la Subsección 2.3.1, la formulación dinámica del transporte óptimo entregará un método de interpolación natural en el espacio de las medidas de probabilidad.

2.3.1. Distancias de Wasserstein

Para comenzar el estudio de la formulación dinámica del transporte óptimo, se revisarán algunas propiedades geométricas que induce el problema de Kantorovich en el conjunto de medidas de probabilidad $\mathcal{M}_+^1(\mathbb{R}^d)$. Por este motivo, a lo largo de toda esta subsección se considerará siempre que $\mathcal{X} = \mathcal{Y} \subset \mathbb{R}^d$. Por otra parte, las demostraciones de los resultados enunciados a lo largo de esta subsección serán omitidas para evitar detalles técnicos, pero se pueden encontrar en [VS03].

Una primera propiedad importante del problema de Kantorovich (2.2.5) es que para dos medidas $\mu, \nu \in \mathcal{M}_+^1(\mathcal{X})$, su valor óptimo define una noción de distancia entre μ y ν cuando se considera un funcional de costo $c(x, y) = \|x - y\|^p$, con $p > 1$. Más precisamente, el problema de Kantorovich induce una métrica en un subespacio específico de $\mathcal{M}_+^1(\mathcal{X})$, entregándole una estructura geométrica y una noción de convergencia al conjunto, inicialmente abstracto, $\mathcal{M}_+^1(\mathcal{X})$. Esta es una propiedad muy relevante del transporte óptimo, ya que permite comparar dos distribuciones de una forma matemáticamente robusta, pudiendo actuar como función de costo en problemas de aprendizaje automático.

Es importante notar que esta propiedad de ser una métrica en (un subconjunto de) $\mathcal{M}_+^1(\mathcal{X})$ no la posee la divergencia de Kullback-Leibler (definida en su versión absolutamente continua en la Definición 1.1 y extendida a su versión general en la Definición 3.4), ya que este operador no es simétrico ni cumple la desigualdad triangular. Si bien es posible corregir estos problemas definiendo la distancia de Jensen-Shannon, la distancia inducida por el problema de Kantorovich tendrá mejores propiedades permitiendo, por ejemplo,

interpolar de forma óptima (en el sentido de Monge-Kantorovich) entre dos distribuciones. Además, esta nueva distancia permitirá comparar distribuciones incluso si estas tienen soporte disjunto, lo cual indefine otros funcionales de discrepancia como la divergencia de Kullback-Leibler.

Definición 2.11 (métrica de Wasserstein). Dadas dos medidas $\mu, \nu \in \mathcal{M}_+^1(\mathcal{X})$ y el funcional de costo $c(x, y) = \|x - y\|^p$ con $p \in [1, \infty)$, se define la *p-distancia de Wasserstein* entre μ y ν como

$$\mathcal{W}_p(\mu, \nu) := \left(\min_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{X}} \|x - y\|^p d\pi(x, y) \right)^{1/p}.$$

En el caso $p = 2$, esta métrica se conoce como *distancia de Fréchet*.

El siguiente resultado afirma que \mathcal{W}_p es efectivamente una distancia en un sub-espacio específico de $\mathcal{M}_+^1(\mathcal{X})$:

Teorema 2.8 (espacio de Wasserstein). Sea $\mathcal{M}_+^{1,p}(\mathcal{X})$ el espacio de medidas de probabilidad en \mathcal{X} con el p -ésimo momento¹⁹ finito:

$$\mathcal{M}_+^{1,p}(\mathcal{X}) := \left\{ \mu \in \mathcal{M}_+^1(\mathcal{X}) : \int_{\mathcal{X}} \|x\|^p d\mu(x) < \infty \right\}.$$

Entonces, la función $\mathcal{W}_p : \mathcal{M}_+^{1,p}(\mathcal{X}) \times \mathcal{M}_+^{1,p}(\mathcal{X}) \rightarrow \mathbb{R}_+$ define una métrica en $\mathcal{M}_+^{1,p}(\mathcal{X})$ ²⁰ y el espacio métrico $(\mathcal{M}_+^{1,p}(\mathcal{X}), \mathcal{W}_p)$ se denomina *espacio de Wasserstein*.

Para otros funcionales de costo, la función \mathcal{W}_p^p no necesariamente es una distancia ya que, por lo general, no es posible demostrar la desigualdad triangular. Por otro lado, si bien se verá que la distancia de Wasserstein tiene muy buenas propiedades, el problema de Kantorovich sigue siendo un problema costoso de resolver, lo que justifica que se sigan usando funciones de discrepancias más sencillas como la divergencia de Kullback-Leibler que, si bien no son tan robustas con la distancia de Wasserstein, sí son fáciles de computar. En el Capítulo 3 se estudiará el algoritmo de Sinkhorn (ver Algoritmo 8), el cual permite aproximar una solución del problema de Kantorovich de forma eficiente, lo que ha aumentado aún más el interés en el transporte óptimo dentro de la comunidad del aprendizaje automático en los últimos años.

Por otra parte, el resultado anterior es generalizable a espacios métricos (\mathcal{X}, d) generales cuando se considera como función de costo a $c = d^p$, con $p > 1$. Es decir, si un conjunto \mathcal{X} tiene asociada una noción de cercanía (mediante una distancia d), entonces el conjunto de medidas $\mathcal{M}_+^{1,p}(\mathcal{X})$ puede dotarse de una noción de cercanía mediante la distancia de Wasserstein \mathcal{W}_p , la cual es obtenida a partir de la distancia d .

Interpolación y baricentros en $\mathcal{M}_+^1(\mathcal{X})$

Dadas dos medidas $\mu, \nu \in \mathcal{M}_+^1(\mathcal{X})$, una manera directa de interpolar entre estas dos medidas es considerar la medida $\tilde{\mu}_t = (1-t)\mu + t\nu$ que se obtiene al interpolar linealmente μ y ν . Sin embargo, muchas veces esta interpolación no es la natural cuando se considera un marco de trabajo dinámico, donde la interpolación representa el desplazamiento o transporte de una medida hacia otra. En esta sección se enunciará un resultado que afirma que el sub-espacio $\mathcal{M}_+^{1,p}(\mathcal{X})$ es un espacio geodésico cuando se dota de la métrica de Wasserstein $\mathcal{W}_p(\cdot, \cdot)$. Es decir, es posible trazar curvas de largo mínimo que conecten dos medidas de probabilidad $\mu, \nu \in \mathcal{M}_+^{1,p}(\mathcal{X})$ permitiendo, entre otras cosas, realizar una interpolación capaz de incorporar la geometría del problema, la cual viene codificada en la función de costo.

¹⁹Recordar que los momentos de una medida de probabilidad $\mu \in \mathcal{M}_+^1(\mathcal{X})$ son los momentos de una variable aleatoria cuya ley es μ .

²⁰Fuera de este conjunto, la integral que define \mathcal{W}_p puede indefinirse, por lo que no puede ser una métrica.

En un espacio métrico general, (\mathcal{X}, d) , siempre es posible definir una nueva métrica (denominada *métrica intrínseca* inducida por d), la cual se define como el ínfimo de los largos de las distintas *curvas* que conectan dos puntos. Si para cualquier par de puntos este ínfimo es alcanzado por alguna curva (es decir, existe una curva que conecta los dos puntos cuyo largo es igual a $d(x, y)$), entonces el espacio se dice geodésico:

Definición 2.12 (espacio geodésico). Un espacio métrico (\mathcal{X}, d) es un *espacio geodésico* si para todo $x, y \in \mathcal{X}$,

$$d(x, y) = \min \{ \text{Largo}(w) : w : [0, 1] \rightarrow \mathcal{X} \text{ continua, con } w(0) = x \text{ y } w(1) = y \},$$

donde el concepto de *largo* tiene una definición precisa que se omite por simplicidad. En este caso, las curvas $w : [0, 1] \rightarrow \mathcal{X}$ que minimicen la expresión anterior se denominan *geodésicas* entre x e y . Estas geodésicas se dirán de *velocidad constante* si adicionalmente cumplen que

$$d(w(t_0), w(t_1)) = |t_1 - t_0| \cdot \underbrace{d(w(0), w(1))}_{d(x, y)}, \quad \forall t_0, t_1 \in [0, 1].$$

El siguiente resultado afirma que la distancia de Wasserstein permite trazar geodésicas:

Teorema 2.9 ($\mathcal{M}_+^{1,p}(\mathcal{X})$ es un espacio geodésico). El espacio de Wasserstein $(\mathcal{M}_+^{1,p}(\mathcal{X}), \mathcal{W}_p)$ es un espacio geodésico. Más aún, si $\mu, \nu \in \mathcal{M}_+^{1,p}(\mathcal{X})$ son medidas de probabilidad con $\pi^* \in \Pi(\mu, \nu)$ el plan óptimo de Kantorovich entre ellas, la *función de interpolación convexa* $P_t : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{X}$ definida como $P_t(x, y) := (1 - t)x + ty$ permite definir una curva geodésica a velocidad constante entre μ y ν mediante la medida push-forward inducida por P_t :

$$t \in [0, 1] \mapsto \mu_t = (P_t)_\# \pi^* \in \mathcal{M}_+^{1,p}(\mathcal{X}).$$

En particular, si T^* es mapa de Monge del problema (2.1.6), esta interpolación coincide con la *interpolación de McCann* entre μ y ν , la cual se define como

$$\mu_t = (T_t)_\# \mu, \tag{2.3.1}$$

donde $T_t = (1 - t)\text{Id} + tT^*$ es la interpolación convexa entre la función identidad $\text{Id}(x) = x$ y el mapa de Monge T^* . Esta interpolación ocurre en el espacio de las medidas de probabilidad y, por lo general, es una interpolación más realista que la interpolación euclíadiana $\tilde{\mu}_t = (1 - t)\mu + t\nu$ (ver Figura 2.11).

Por otra parte, dado que \mathcal{W}_p es una distancia, es posible generalizar la interpolación anterior a un conjunto de medidas mediante el cálculo del baricentro, el cual es un concepto que se puede aplicar a distancias en general. Dado un espacio métrico (\mathcal{X}, d) , un conjunto de puntos $\{x_i\}_{i=1}^k \subset \mathcal{X}$ y un vector de pesos $\lambda \in \Sigma_k$, se definen los *baricentros de orden p* de $\{x_i\}_{i=1}^k$ como los puntos que resuelven el problema

$$\min_{x \in \mathcal{X}} \sum_{i=1}^k \lambda_i d(x_i, x)^p. \tag{2.3.2}$$

Por ejemplo, si se considera $\mathcal{X} = \mathbb{R}^d$ (con la distancia usual) y $p = 2$, se puede demostrar que el baricentro de los puntos $\{x_i\}_{i=1}^k \subset \mathcal{X}$ existe, es único y viene dado por la media de los puntos:

$$x_B = \sum_{i=1}^k \lambda_i x_i. \tag{2.3.3}$$

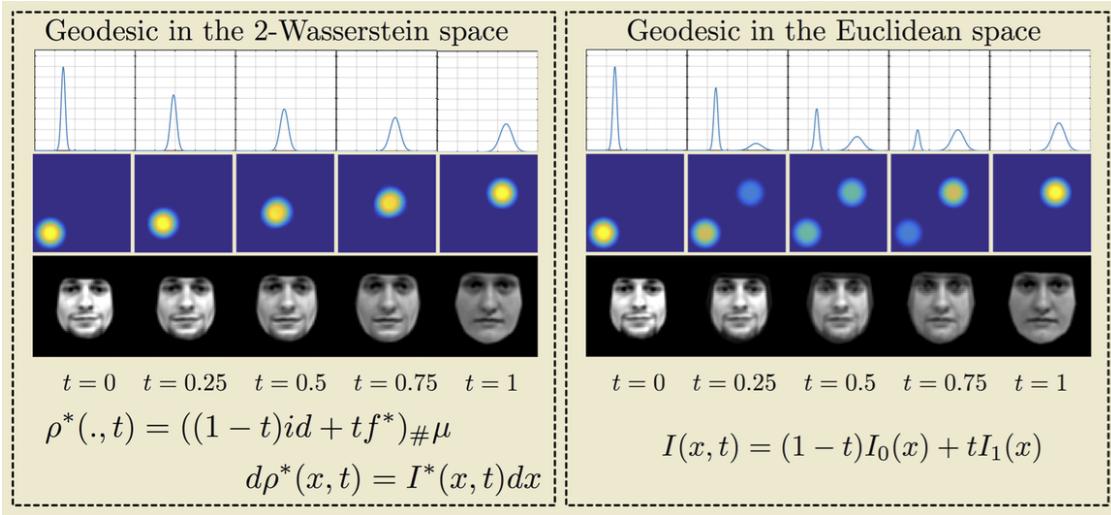


Figura 2.11: Interpolación en el espacio de Wasserstein (izquierda) y en el espacio de las muestras (derecha). Se observa interpolando en el espacio de Wasserstein se obtienen geodésicas más naturales que al hacerlo en el espacio observable. Imagen obtenida desde [Kol+17].

Cuando se considera el espacio de Wasserstein, $(\mathcal{M}_+^{1,p}(\mathcal{X}), \mathcal{W}_p)$, se tiene la siguiente definición:

Definición 2.13 (baricentro de Wasserstein). Dado un conjunto de medidas $\{\mu_i\}_{i=1}^k \subset \mathcal{M}_+^{1,p}(\mathcal{X})$ y un vector de pesos $\lambda \in \Sigma_k$, se dice que $\mu_B^* \in \mathcal{M}_+^{1,p}(\mathcal{X})$ es un *baricentro de Wasserstein de orden p* si es solución del problema

$$\min_{\mu \in \mathcal{M}_+^{1,p}(\mathcal{X})} \sum_{i=1}^k \lambda_i \mathcal{W}_p(\mu_i, \mu)^p,$$

donde $p \gg 1$ es un parámetro del problema.

Si bien el problema de los baricentros no es fácil de resolver en general, cuando se considera el espacio de Wasserstein con $p = 2$, el baricentro existe y es único:

Proposición 2.9. Dado un conjunto de medidas $\{\mu_i\}_{i=1}^k \subset \mathcal{M}_+^{1,p}(\mathcal{X})$ tales que alguna de ellas tenga función de densidad, y un vector de pesos $\lambda \in \Sigma_k$. Entonces, para $p = 2$ el baricentro $\mu_B \in \mathcal{M}_+^{1,p}(\mathcal{X})$ existe y es único. Más aún, se puede demostrar un resultado similar a la propiedad (2.3.3) en este espacio:

$$\mathbb{E}_{x_B \sim \mu_B} [x_B] = \sum_{i=1}^k \lambda_i \mathbb{E}_{x_i \sim \mu_i} [x_i].$$

Es decir, la esperanza de una variable aleatoria que distribuye μ_B es igual al baricentro de las esperanzas de variables aleatorias que distribuyen μ_i , $i \in \{1, \dots, k\}$.

En la Figura 2.13 se puede ver el cálculo de baricentros entre 4 elementos para distintos vectores de peso λ .

Por otra parte, en el caso discreto $\mathcal{X} = \{x_i\}_{i=1}^n$, el problema del baricentro de Wasserstein es un problema de programación lineal. En efecto, μ_B es la medida asociada al vector de probabilidad $a_B \in \Sigma_n$ que resuelve

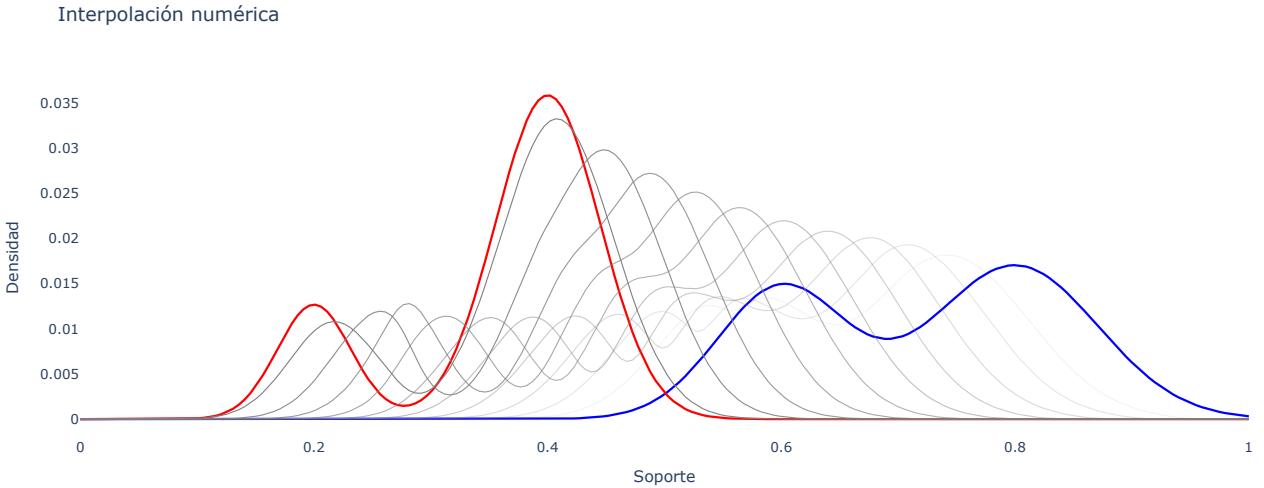


Figura 2.12: Interpolación entre dos mezclas gaussianas. En este caso, el mapa de transporte no tiene solución en forma cerrada y debe ser encontrado de forma numérica. El código se encuentra en el archivo `distr_interpolation.ipynb`.

el problema lineal

$$\arg \min_{\substack{\bar{a} \in \Sigma_n \\ (P_i)_{i=1}^k \subset \mathcal{M}_{nn,n}([0,1])}} \sum_{i=1}^k \lambda_i \langle C, P_i \rangle \quad \text{sujeto a} \quad P_i \mathbb{1}_n = a_i, P_i^\top \mathbb{1}_n = a_B, \forall i \in \{1, \dots, k\}, \quad (2.3.4)$$

donde $\{a_i\}_{i=1}^k$ son los vectores de probabilidad asociados a las medidas discretas $\{\mu_i\}_{i=1}^k \subset \mathcal{M}_+^1(\mathcal{X})$ y $C \in \mathcal{M}_{nn}(\mathbb{R}_+)$ es la matriz de distancias (elevada a p).

Convergencia débil en $\mathcal{M}_+^1(\mathcal{X})$

Para poder estudiar propiedades matemáticas del conjunto $\mathcal{M}_+^1(\mathcal{X})$ es necesario definir una estructura matemática sobre él. Dado que en este trabajo importan propiedades de continuidad (por ejemplo, para analizar la existencia de soluciones en los problemas de transporte) y de distancia (por ejemplo, para poder comparar distribuciones), el tipo de estructura matemática que se necesita sobre $\mathcal{M}_+^1(\mathcal{X})$ es una noción de convergencia. Una noción usual de convergencia sobre este espacio es la de *convergencia débil*:

Definición 2.14 (convergencia débil de medidas). Se dirá que una sucesión de medidas de probabilidad $(\mu_n)_{n \in \mathbb{N}} \subset \mathcal{M}_+(\mathcal{X})$ converge débilmente a la medida $\mu \in \mathcal{M}_+(\mathcal{X})$ si:

$$\int_{\mathcal{X}} f(x) d\mu_n(x) \rightarrow \int_{\mathcal{X}} f(x) d\mu(x), \quad \forall f \in \mathcal{C}_b(\mathcal{X}),$$

donde $\mathcal{C}_b(\mathcal{X})$ es el espacio de las funciones continuas y acotadas. Esta convergencia se denotará $\mu_n \rightharpoonup \mu$.

Para dar un ejemplo de esta noción de convergencia en $\mathcal{X} = \mathbb{R}$, se puede considerar una sucesión de medidas $(\delta_{x_n})_{n \in \mathbb{N}} \subset \mathcal{M}_+^1(\mathcal{X})$, donde cada medida está concentrada en un único punto $x \in \mathcal{X}$ y la secuencia de puntos $(x_n)_{n \in \mathbb{N}} \subset \mathcal{X}$ converge a un punto límite $\bar{x} \in \mathcal{X}$. En consecuencia, uno esperaría, naturalmente, que la sucesión de medidas $(\delta_{x_n})_{n \in \mathbb{N}}$ también converja a la medida límite $\delta_{\bar{x}}$. Se verá que esto sí ocurre bajo la noción de convergencia débil. En efecto, dada una función $f \in \mathcal{C}_b(\mathcal{X})$, se tiene que

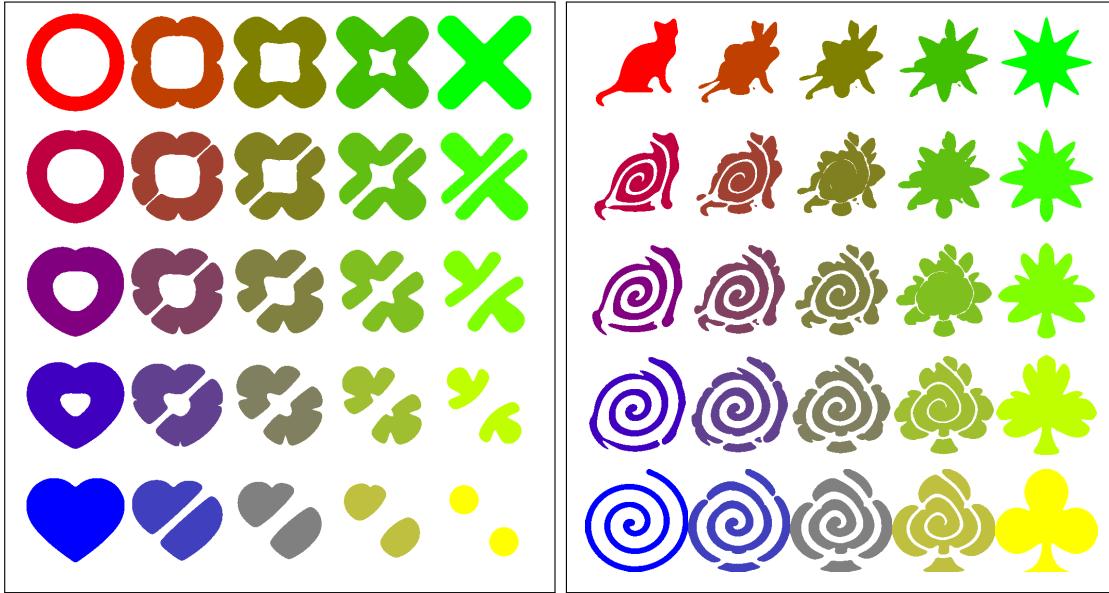


Figura 2.13: Interpolación baricéntrica entre 4 figuras en el plano (una figura en cada esquina). Cada imagen representa un vector de peso $\lambda \in \Sigma_4$ diferente, el cual indica el nivel de importancia de cada figura en la interpolación. Imagen obtenida desde [PC20].

$$\int_{\mathcal{X}} f(x) \, d\delta_{x_n}(x) = f(x_n) \quad \text{y} \quad \int_{\mathcal{X}} f(x) \, d\delta_{\bar{x}}(x) = f(\bar{x}).$$

Además, como $f : \mathcal{X} \rightarrow \mathbb{R}$ es continua, entonces $f(x_n) \rightarrow f(\bar{x})$, por lo que efectivamente $\delta_{x_n} \rightharpoonup \delta_{\bar{x}}$.

Por otra parte, si bien otras nociones de convergencia en $\mathcal{M}_+^1(\mathcal{X})$ pueden parecer más naturales, la convergencia débil es una buena noción de convergencia ya que no es muy rígida como para que haya pocas sucesiones convergentes, ni es muy débil como para no poder obtener buenos resultados matemáticos. Además, esta convergencia tiene orígenes matemáticos más profundos, los cuales le permiten heredar resultados importantes del análisis funcional. Por otro lado, si se consideran variables aleatorias asociadas a las medidas de la Definición 2.14, la convergencia en (2.14) se puede escribir como $\mathbb{E}_{x_n \sim \mu_n}[f(x_n)] \rightarrow \mathbb{E}_{x \sim \mu}[f(x)]$. Más aún, en el caso unidimensional $\mathcal{X} = \mathbb{R}$, esto es equivalente a la convergencia de las funciones de distribución acumulada de las variables aleatorias (en los puntos donde haya continuidad). Estas caracterizaciones forman parte del teorema de Portmanteau.

La noción de convergencia débil definida sobre todo $\mathcal{M}_+^1(\mathcal{X})$ se hereda directamente al subconjunto $\mathcal{M}_+^{1,p}(\mathcal{X})$ donde la distancia de Wasserstein define una métrica. El siguiente resultado afirma que converger en el sentido de convergencia débil es equivalente (salvo una condición extra de convergencia de momentos) a converger con la distancia de Wasserstein:

Teorema 2.10 (Convergencia con la distancia de Wasserstein). Dada una sucesión de medidas de probabilidad con p -ésimo momento finito, $(\mu_n)_{n \in \mathbb{N}} \subset \mathcal{M}_+^{1,p}(\mathcal{X})$, y un candidato a límite $\mu \in \mathcal{M}_+^{1,p}(\mathcal{X})$, se tiene que:

$$\mathcal{W}_p(\mu_k, \mu) \rightarrow 0 \iff (\mu_n \rightharpoonup \mu) \wedge \left(\int_{\mathcal{X}} \|x\|^p \, d\mu_n(x) \rightarrow \int_{\mathcal{X}} \|x\|^p \, d\mu(x) \right).$$

Del teorema anterior, es importante mencionar que la hipótesis adicional de la convergencia de los p -momentos es razonable. Por un lado, la convergencia débil de medidas, al estar integrando con respecto a funciones continuas y acotadas, no es capaz de capturar el comportamiento de las colas de las distribuciones. Por otra parte, si las colas de dos distribuciones difieren, el costo de transporte de masa en las colas (medido por $\|x - y\|^p$) será elevado, independientemente de si las distribuciones convergen débilmente. Al agregar la hipótesis de convergencia de momentos a la convergencia débil de medidas, se evita este problema y es posible garantizar la convergencia en el sentido del transporte óptimo.

Notar que si se considera $\mathcal{X} \subset \mathbb{R}^d$ compacto (i.e., cerrado y acotado), los p -momentos son siempre finitos. En efecto, si $\mu \in \mathcal{M}_+^1(\mathcal{X})$ con \mathcal{X} compacto:

$$\int_{\mathcal{X}} \|x\|^p d\mu(x) \leq \int_{\mathcal{X}} \left(\max_{x \in \mathcal{X}} \|x\|^p \right) d\mu(x) = \mu(\mathcal{X}) \cdot \max_{x \in \mathcal{X}} \|x\|^p < \infty,$$

donde $\max_{x \in \mathcal{X}} \|x\|^p < \infty$ ya que $\|\cdot\|^p$ es continua y \mathcal{X} es compacto. En consecuencia, $\mathcal{M}_+^{1,p}(\mathcal{X}) = \mathcal{M}_+^1(\mathcal{X})$. Más aún, se puede demostrar que los p -momentos también convergen. En consecuencia, bajo esta hipótesis adicional, la distancia de Wasserstein metriza totalmente la convergencia débil en $\mathcal{M}_+^1(\mathcal{X})$:

Corolario 2.5 (caracterización de la convergencia débil). Sea $\mathcal{X} \subset \mathbb{R}^d$ compacto. Dada una sucesión de medidas de probabilidad, $(\mu_n)_{n \in \mathbb{N}} \subset \mathcal{M}_+^1(\mathcal{X})$, y un candidato a límite $\mu \in \mathcal{M}_+^1(\mathcal{X})$, entonces:

$$\mathcal{W}_p(\mu_k, \mu) \rightarrow 0 \iff \mu_n \rightharpoonup \mu.$$

Por lo tanto, la distancia de Wasserstein (y por lo tanto, el problema de Kantorovich) es una noción de convergencia que no solo tiene la motivación geométrica del transporte óptimo, si no que también tiene propiedades prestadas desde el análisis funcional en espacios duales, lo que convierte al problema de transporte óptimo en método robusto para comparar dos medidas de probabilidad.

Comparación con otras funciones de pérdida

En este apartado se contrastará la distancia de Wasserstein que induce el problema de Kantorovich en $\mathcal{M}_+^1(\mathcal{X})$ con otras funciones de comparación de distribuciones utilizadas en el aprendizaje automático. En particular, se definirá el concepto de f -divergencia, el cual permite construir una familia de operadores para poder comparar fácilmente dos medidas, donde la divergencia de Kullback-Leibler es la f -divergencia más utilizada en el aprendizaje de máquinas.

Para hacer clara la definición de f -divergencia, se entregará primero su definición en el caso discreto:

Definición 2.15 (f -divergencia, caso discreto). Dado un conjunto finito $\mathcal{X} = \{x_i\}_{i=1}^n$ y $f : \mathbb{R}_+ \rightarrow \mathbb{R}$ una función continua²¹, convexa y tal que $f(1) = 0$. Entonces, para dos medidas $\mu, \nu \in \mathcal{M}_+^1(\mathcal{X})$ con vectores de probabilidad $a, b \in \Sigma_n$ respectivamente, se define su f -divergencia o divergencia de Csiszár como

$$D_f(\mu \| \nu) := \sum_{i: x_i \in \text{Supp}(\nu)} f\left(\frac{a_i}{b_i}\right) b_i + f'_\infty \sum_{i: x_i \notin \text{Supp}(\nu)} a_i,$$

donde $f'_\infty = \lim_{x \rightarrow \infty} \frac{f(x)}{x} \in [0, \infty]$ es la velocidad límite de f .

²¹En general, se suele definir como una función semi-continua inferior, pero se relajará la definición ya que no será necesario en este trabajo.

En esta definición, la exigencia $f(1) = 0$ muestra que si $\frac{a_i}{b_i} = 1, \forall i \in \{0, \dots, n\}$ (i.e., ambas medidas entregan la misma masa en los puntos de soporte común), entonces el primer sumando es nulo. Por otra parte, la segunda suma busca aumentar el valor de la divergencia cuando la parte singular de μ (i.e., los puntos donde no comparte soporte con ν) es muy grande, mientras que el factor f'_∞ (no necesariamente finito) está definido de esta forma para que la divergencia $D_f(\cdot \| \cdot)$ sea continua²² bajo la noción de convergencia débil de medidas (ver Definición 2.14).

Antes de extender esta definición al caso continuo, es necesario entregar el siguiente resultado, el cual afirma que una medida de probabilidad se puede descomponer, con respecto a otra medida, en una parte absolutamente continua (i.e., que posee función de densidad) y una parte singular (i.e., que no comparten soporte):

Teorema 2.11 (descomposición de Lebesgue). Dadas dos medidas de probabilidad, $\mu, \nu \in \mathcal{M}_+^1(\mathcal{X})$, entonces existen dos medidas $\mu^{\text{a.c.}(\nu)}, \mu^{\perp(\nu)} \in \mathcal{M}_+(\mathcal{X})$, tales que:

- $\mu = \mu^{\text{a.c.}(\nu)} + \mu^{\perp(\nu)}$.
- $\mu^{\text{a.c.}(\nu)} \ll \nu$ (i.e., $\mu^{\text{a.c.}(\nu)}$ tiene función de densidad $\frac{d\mu^{\text{a.c.}(\nu)}}{d\nu}$ con respecto a ν).
- $\mu^{\perp(\nu)} \perp \nu$ (i.e., $\mu^{\perp(\nu)}$ y ν tienen soporte disjunto).

Con esta descomposición, las definición anterior se extiende utilizando la función de densidad en la parte absolutamente continua y la medida singular en la parte singular del soporte:

Definición 2.16 (f -divergencia, caso continuo). Sea $f : \mathbb{R}_+ \rightarrow \mathbb{R}$ una función continua, convexa y tal que $f(1) = 0$. Entonces, para dos medidas $\mu, \nu \in \mathcal{M}_+^1(\mathcal{X})$ sobre $\mathcal{X} \subset \mathbb{R}^d$ se define su f -divergencia o divergencia de Csiszár como

$$D_f(\mu \| \nu) := \int_{\mathcal{X}} f\left(\frac{d\mu^{\text{a.c.}(\nu)}}{d\nu}(x)\right) d\nu(x) + f'_\infty \cdot \mu^{\perp(\nu)}(\mathcal{X}), \quad (2.3.5)$$

donde $\frac{d\mu^{\text{a.c.}(\nu)}}{d\nu}$ es la función de densidad (derivada de Radon-Nikodym) de $\mu^{\text{a.c.}(\nu)}$ con respecto a ν (ver Teorema A.1) y $f'_\infty = \lim_{x \rightarrow \infty} \frac{f(x)}{x} \in [0, \infty]$ es la velocidad límite de f .

Una f -divergencia natural²³ es la que está asociada a $f(x) = |x - 1|$. Esta divergencia resultará ser, al igual que el problema de Kantorovich, una métrica en $\mathcal{M}_+^1(\mathcal{X})$ denominada *distancia en variación total*:

Definición 2.17 (variación total). Dadas dos medidas de probabilidad $\mu, \nu \in \mathcal{M}_+^1(\mathcal{X})$, se define su (divergencia en) variación total como:

$$D_{\text{TV}}(\mu \| \nu) := \int_{\mathcal{X}} \left| \frac{d\mu^{\text{a.c.}(\nu)}}{d\nu}(x) - 1 \right| d\nu(x) + \mu^{\perp(\nu)}(\mathcal{X}).$$

A pesar de que su definición es poco clara, esta divergencia puede escribirse de forma sencilla en algunos casos. En efecto, cuando $\mathcal{X} = \{x_i\}_{i=1}^n$ es un espacio discreto y $\mu, \nu \in \mathcal{M}_+^1(\mathcal{X})$ son medidas con vectores de probabilidad $a, b \in \Sigma_n$ respectivamente:

²²Recordar que la continuidad es una propiedad deseable cuando se quiere maximizar o minimizar una función.

²³Esta divergencia se puede considerar natural ya que $f(x) = |x - 1|$ es la función más simple que cumple las propiedades pedidas.

$$\begin{aligned}
 D_{\text{TV}}(\mu \| \nu) &= \sum_{i: x_i \in \text{Supp}(\nu)} \left| \frac{a_i}{b_i} - 1 \right| b_i + \sum_{i: x_i \notin \text{Supp}(\nu)} a_i \\
 &= \sum_{i: x_i \in \text{Supp}(\nu)} |a_i - b_i| + \sum_{i: x_i \notin \text{Supp}(\nu)} |a_i - 0| \\
 &= \sum_{i=1}^n |a_i - b_i| \\
 &= \|a - b\|_{l^1},
 \end{aligned}$$

donde se identifica que $D_{\text{TV}}(\mu \| \nu)$ es precisamente la distancia l^1 de los vectores de probabilidad. Esta propiedad puede extenderse al continuo $\mathcal{X} \subset \mathbb{R}^d$ si las medidas μ y ν tienen funciones de densidad p_μ y p_ν respectivamente. En efecto, se demuestra de forma análoga que la distancia en variación total entre μ y ν corresponde a la distancia L^1 de sus funciones de densidad:

$$D_{\text{TV}}(\mu \| \nu) = \|p_\mu - p_\nu\|_{L^1(\mathcal{X})} := \int_{\mathcal{X}} |p_\mu(x) - p_\nu(x)| \, dx.$$

Esta última propiedad sugiere que la divergencia en variación total es una métrica en $\mathcal{M}_+^1(\mathcal{X})$. Esta afirmación es cierta. Más aún, notando que $D_{\text{TV}}(\mu \| \nu)$ se define como la norma L^1 (o l^1 en el caso discreto) de una diferencia, se puede demostrar que la métrica $d_{\text{TV}}(\mu, \nu) := D_{\text{TV}}(\mu \| \nu)$ proviene realmente de una norma mediante $d_{\text{TV}}(\mu, \nu) = \|\mu - \nu\|_{\text{TV}}$, donde $\|\mu\|_{\text{TV}} := |\mu|(\mathcal{X})$ es la norma de variación total en el espacio vectorial de todas las medidas finitas (no necesariamente de probabilidad), tanto positivas como negativas²⁴.

Sin embargo, si bien esta nueva noción de distancia en $\mathcal{M}_+^1(\mathcal{X})$ parece tener buenas propiedades, no es capaz de metrizar la convergencia débil de medidas como sí lo hace la distancia de Wasserstein:

Proposición 2.10. La distancia en variación total, $d_{\text{TV}}(\mu, \nu) := D_{\text{TV}}(\mu \| \nu)$, no es una distancia compatible con la convergencia débil en $\mathcal{M}_+^1(\mathcal{X})$ ²⁵. Más aún, si $\mathcal{X} \subset \mathbb{R}^d$ es acotado, esta distancia es más fuerte que la distancia de Wasserstein $\mathcal{W}_1(\mu, \nu)$:

$$\mathcal{W}_1(\mu, \nu) \leq k \cdot d_{\text{TV}}(\mu, \nu), \quad \forall \mu, \nu \in \mathcal{M}_+^1(\mathcal{X}),$$

donde $k > 0$ es una constante que depende únicamente de \mathcal{X} .

La proposición anterior permite concluir que la convergencia en variación total es más fuerte que la convergencia débil. Es decir, si una sucesión de medidas $(\mu_n)_{n \in \mathbb{N}} \subset \mathcal{M}_+^1(\mathcal{X})$ converge a $\mu \in \mathcal{M}_+^1(\mathcal{X})$ en variación total (i.e., $d_{\text{TV}}(\mu_n, \mu) \rightarrow 0$), entonces también converge débilmente (i.e., $\mathcal{W}_1(\mu_n, \mu) \rightarrow 0$). Para ver que el recíproco no es cierto (i.e., la distancia en variación total no es compatible con la convergencia débil), se puede considerar la sucesión de medidas $(\delta_{x_n})_{n \in \mathbb{N}} \subset \mathcal{M}_+^1(\mathbb{R})$, la cual se probó que converge débilmente a $\delta_{\bar{x}}$ cuando $x_n \rightarrow \bar{x}$ en \mathbb{R} . Sin embargo, esta convergencia no es cierta en variación total. En efecto, δ_{x_n} y $\delta_{\bar{x}}$ son medidas singulares (tienen soporte disjunto) por lo que la función de densidad en (2.3.5) es nula, mientras que la parte singular de δ_{x_n} es todo δ_{x_n} , luego:

$$d_{\text{TV}}(\delta_{x_n}, \delta_{\bar{x}}) = D_{\text{TV}}(\delta_{x_n} \| \delta_{\bar{x}}) = \int_{\mathbb{R}} |0 - 1| \, d\delta_{\bar{x}}(x) + \delta_{x_n}(\mathbb{R}) = 1 + 1 > 0, \quad \forall n \in \mathbb{N}.$$

²⁴ Sin embargo, esta norma no resulta útil en $\mathcal{M}_+^1(\mathcal{X})$ ya que, además de no ser un espacio vectorial, $\|\mu\|_{\text{TV}} = 1$, para toda medida de probabilidad $\mu \in \mathcal{M}_+^1(\mathcal{X})$.

²⁵ Es decir, no genera la misma topología, lo que se traduce en que no tiene los mismos conjuntos abiertos, funciones continuas ni conjuntos compactos.

Por lo que, efectivamente, δ_{x_n} no converge a $\delta_{\bar{x}}$ bajo esta noción de convergencia ya que no es posible que $d_{TV}(\delta_{x_n}, \delta_{\bar{x}}) \rightarrow 0$ (salvo que x_n sea constante a partir de un cierto momento).

El hecho que la convergencia en variación total sea más fuerte que la convergencia débil muchas veces es una mala propiedad ya que, bajo esta nueva noción de convergencia, son menos las sucesiones que convergen, lo cual suele ser necesario para probar resultados de existencia de soluciones en problemas de optimización. Desde otro punto de vista, la topología inducida por la distancia en variación total tiene más abiertos, lo que reduce la cantidad de conjuntos compactos.

Por otra parte, tanto en la Definición 2.16 como en la Definición 2.15 se puede observar la similitud de las f -divergencias con la divergencia de Kullback-Leibler usada en los modelos de difusión donde, por simplicidad, se asume que las medidas de probabilidad son absolutamente continuas con respecto a la medida de Lebesgue para poder trabajar directamente sobre funciones de densidad de probabilidad (ver Definición 1.1). Es directo ver que la divergencia de Kullback-Leibler es realmente una f -divergencia cuando se considera a f como la entropía de Shannon $f(x) = x \cdot \log(x)$. Notar que para este caso, $f'_\infty = \lim_{x \rightarrow \infty} \log(x) = \infty$, por lo que la divergencia de Kullback-Leibler siempre será ∞ si μ no tiene densidad con respecto a ν , lo cual es consistente con la Definición 3.4. En particular, $D_{KL}(\delta_x \| \delta_y) = \infty$ si $x \neq y$, por lo que esta divergencia tampoco es compatible con la convergencia débil de medidas. Más aún, se puede demostrar que ninguna f -divergencia es capaz de metrizar la convergencia débil debido a que, a diferencia de la distancia de Wasserstein, las f -divergencias no son capaces de acceder a la geometría del espacio subyacente, la cual viene codificada en la métrica del espacio \mathcal{X} y es informada a la distancia de Wasserstein mediante la función de costo.

\mathcal{W}_1 como función de pérdida en redes neuronales

Un enfoque alternativo (y quizás más robusto) al enfoque de máxima verosimilitud consiste en comparar directamente la distribución aprendida y la distribución real de los datos, donde el criterio de comparación más común es la divergencia de Kullback-Leibler la cual, cuando es minimizada en su forma *forward*, equivale al enfoque de máxima verosimilitud. Dadas las limitaciones que esta divergencia presenta, se necesita un criterio más general para poder comparar un modelo paramétrico μ_θ con una distribución de datos μ_{true} . Dado un modelo generativo de variable latente $x = g_\theta(z) \sim \mu_\theta$ (con $z \sim \mu_z$ la variable latente) que busca aprender la medida μ_{true} mediante el entrenamiento de una red neuronal g_θ , lo natural es considerar como función de pérdida a $d(\mu_\theta, \mu_{\text{true}})$, donde d es alguna distancia en $\mathcal{M}_+^1(\mathcal{X})$.

Para poder entrenar la red neuronal g_θ con descenso del gradiente, es necesario que el mapa

$$\begin{aligned}\Theta &\rightarrow \mathcal{M}_+^1(\mathcal{X}) \\ \theta &\mapsto \mu_\theta,\end{aligned}$$

sea continuo. De esta forma, la función que sea deseada minimizar, $\theta \mapsto d(\mu_\theta, \mu_{\text{true}})$, es continua y, eventualmente, derivable. Dado que la distancia de Wasserstein es más débil que la distancia en variación total, hay secuencias $(\theta_k)_{k \in \mathbb{N}} \subset \Theta$ con $\theta_k \rightarrow \theta^* \in \Theta$ tales que $\mu_{\theta_k} \rightarrow \mu_{\theta^*}$ con la distancia de Wasserstein pero no con distancia en variación total, lo cual indica que \mathcal{W}_p es una distancia más apropiada para usar como función de pérdida ya que hay más funciones que son continuas bajo esta topología²⁶. El siguiente resultado precisa esto:

Teorema 2.12. Sea $\mu_{\text{true}} \in \mathcal{M}_+^1(\mathcal{X})$ una distribución desconocida y $g_\theta(z) \sim \mu_\theta$ un modelo generativo de variable latente $z \sim \mu_z$, entonces:

- Si $\theta \rightarrow g_\theta$ es continua en θ , entonces también lo es $\mathcal{W}_1(\mu_\theta, \mu_{\text{true}})$.

²⁶Recordar que una función f es continua si y solo si $f(x_n) \rightarrow f(\bar{x})$, para toda sucesión $(x_n)_{n \in \mathbb{N}} \subset \mathcal{X}$ convergente a $\bar{x} \in \mathcal{X}$

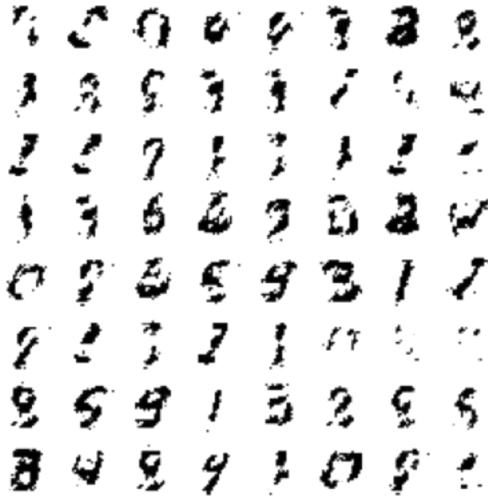


Figura 2.14: Imágenes generadas por una Wasserstein GAN en el dataset MNIST. La calidad de las imágenes es inferior a las generadas en el Capítulo 1 debido a que se entrenó el modelo neuronal menos tiempo. En el archivo `wgan.ipynb` hay una implementación minimal de este modelo generativo.

- Bajo ciertas hipótesis de regularidad sobre g_θ , $\mathcal{W}_1(\mu_\theta, \mu_{\text{true}})$ es continuo en todo Θ y diferenciable casi en todas partes.
- Los resultados anteriores son falsos si se considera la distancia en variación total.

En particular, las hipótesis de regularidad del teorema anterior se cumplen si g_θ es una red fully-connected (con función de activación ReLU, tanh o sigmoide) y μ_z es tal que $\mathbb{E}_{z \sim \mu_z} [\|z\|] < \infty$ (en particular, se puede considerar una variable latente gaussiana), por lo que es posible entrenar un modelo generativo neuronal (entrenado mediante algoritmos de gradiente) utilizando la distancia de Wasserstein pero no la distancia en variación total. La demostración de este resultado se puede encontrar en [ACB17], donde además resuelven el problema de transporte óptimo adaptando la función de costo de una GAN. Este modelo, conocido como *Wasserstein GAN* (WGAN), permitió solucionar muchos de los problemas expuestos en la Subsección 1.1.1 y es un ejemplo de uso del transporte óptimo en modelos generativos. En la Figura 2.14 se pueden ver muestras generadas a partir de este modelo.

2.3.2. Interpretación como problema de control óptimo estocástico

Si bien el problema de Kantorovich tiene buenas propiedades tanto matemáticas como prácticas, la naturaleza de su formulación es totalmente estática ya que los mapas de transporte $T : \mathcal{X} \rightarrow \mathcal{Y}$ solo indican una asignación de transporte pero no entregan la trayectoria que deben seguir los elementos de \mathcal{X} para llegar a \mathcal{Y} . Por otro lado, si bien la interpolación de McCann (2.3.1) permite este dinamismo gracias a las propiedades geodésicas del espacio de Wasserstein, es posible reformular el problema de transporte como un problema dinámico en el cual la masa *fluye* desde una medida a otra debido a la presencia un campo vectorial.

En esta subsección se escribirá el problema de Kantorovich con costo cuadrático en $\mathcal{X} = \mathcal{Y} \subset \mathbb{R}^d$ como otro problema de optimización equivalente que consistirá en guiar la trayectoria de un conjunto de partículas desde una distribución inicial hacia una distribución final mediante un control que dirigirá la trayectoria basado en el principio de mínima acción.

Para comenzar, notar que se puede escribir el costo cuadrático como un problema de minimización trivial:

$$\frac{1}{2} \|x - y\|^2 = \min_{r \in \mathcal{C}_{xy}} \int_0^1 \frac{1}{2} \|\dot{r}\|^2 dt,$$

donde el conjunto factible \mathcal{C}_{xy} es el conjunto de las trayectorias $r : [0, 1] \rightarrow \mathcal{X}$ (con derivada continua) que comienzan en x y terminan en y :

$$\mathcal{C}_{xy} = \{r \in C^1([0, 1], \mathcal{X}) : r(0) = x, r(1) = y\}.$$

Notar que el mínimo del problema anterior es alcanzado por el segmento de recta $r^*(t) = (1-t)x + ty$ ($t \in [0, 1]$), el cual es una geodésica en \mathcal{X} . Por lo tanto, cualquier otra trayectoria factible en \mathcal{C}_{xy} tendrá un valor mayor para el problema de optimización, por lo que la medida de probabilidad $\delta_{r^*} \in \mathcal{M}_+^1(C^1([0, 1], \mathcal{X}))$ que concentra toda su masa en la trayectoria r^* resolverá el siguiente problema de optimización:

$$\min_{P_{xy} \in \Gamma(\delta_x, \delta_y)} \mathbb{E}_{r \sim P_{xy}} \left[\int_0^1 \frac{1}{2} \|\dot{r}\|^2 dt \right], \quad (2.3.6)$$

donde el conjunto factible es el conjunto de medidas de probabilidad sobre el conjunto de trayectorias $C^1([0, 1], \mathcal{X})$ cuyas marginales en $t = 0$ y $t = 1$ son δ_x y δ_y respectivamente:

$$\Gamma(\delta_x, \delta_y) = \{P \in \mathcal{M}_+^1(C^1([0, 1], \mathcal{X})) : P_0 = \delta_x, P_1 = \delta_y\}.$$

En consecuencia, como el valor óptimo de (2.3.6) sigue siendo $\frac{1}{2} \|x - y\|^2$, se tiene la siguiente reformulación para el funcional de costo del problema de Kantorovich entre dos medidas $\mu_0, \mu_1 \in \mathcal{M}_+^1(\mathbb{R}^d)$:

$$\min_{\pi \in \Pi(\mu, \nu)} \mathbb{E}_{(x,y) \sim \pi} \left[\frac{1}{2} \|x - y\|^2 \right] = \min_{\pi \in \Pi(\mu, \nu)} \mathbb{E}_{(x,y) \sim \pi} \left[\min_{P_{xy} \in \Gamma(\delta_x, \delta_y)} \mathbb{E}_{r \sim P_{xy}} \left[\int_0^1 \frac{1}{2} \|\dot{r}\|^2 dt \right] \right], \quad (2.3.7)$$

donde se ha escrito el funcional de costo del problema de Kantorovich como una esperanza. Por otra parte, notar que si $\pi \in \Pi(\mu, \nu)$ y $P_{xy} \in \Gamma(\delta_x, \delta_y)$, entonces, la doble esperanza $\mathbb{E}_\pi [\mathbb{E}_{P_{xy}} [\cdot]]$ se puede escribir como una única esperanza $\mathbb{E}_P [\cdot]$ con $P \in \Gamma(\mu, \nu)$, donde $\Gamma(\mu, \nu)$ el conjunto de medidas de probabilidad en $C^1([0, 1], \mathbb{R}^d)$ cuyas marginales en $t = 0$ y $t = 1$ son $\mu, \nu \in \mathcal{M}_+^1(\mathcal{X})$ respectivamente:

$$\Gamma(\mu, \nu) = \{P \in \mathcal{M}_+^1(C^1([0, 1], \mathcal{X})) : P_0 = \mu, P_1 = \nu\}.$$

De esta forma, el problema (2.3.7) se puede escribir como

$$\min_{P \in \Gamma(\mu, \nu)} \mathbb{E}_{r \sim P} \left[\int_0^1 \frac{1}{2} \|\dot{r}\|^2 dt \right]. \quad (2.3.8)$$

Para escribir este nuevo problema como un problema de control óptimo, se puede considerar que la velocidad \dot{r} de la trayectoria es definida por un control admisible continuo $v : [0, 1] \times \mathcal{X} \rightarrow \mathcal{X}$ mediante la dinámica $\dot{r}(t) = v(t, r(t))$. Con esta observación, el problema (2.3.8) se puede escribir como la siguiente formulación de control óptimo:

$$\min_{v \in C([0,1] \times \mathcal{X})} \mathbb{E}_r \left[\int_0^1 \frac{1}{2} \|v(t, r(t))\|^2 dt \right] \quad \text{sujeto a} \quad \begin{cases} dr_t = v(t, r_t) dt \\ r_0 \sim \mu \\ r_1 \sim \nu. \end{cases} \quad (2.3.9)$$

En consecuencia, el problema de Kantorovich puede ser formulado dinámicamente como un problema de control óptimo estocástico, donde el control admisible v dirige la dinámica de las distribuciones marginales del proceso r . Esta reinterpretación permite generalizar el problema de Kantorovich a dinámicas más complejas. Por ejemplo, se puede considerar la siguiente generalización de (2.3.9):

$$\min_{v \in C([0,1] \times \mathbb{R}^d)} \mathbb{E}_r \left[\int_0^1 \frac{1}{2} \|v(t, r(t))\|^2 dt \right] \quad \text{sujeto a} \quad \begin{cases} dr_t = [f(t + r_t) + v(t, r_t)] dt \\ r_0 \sim \mu \\ r_1 \sim \nu, \end{cases} \quad (2.3.10)$$

donde f puede ser interpretado como un prior sobre el control admisible v . Notar que las restricciones de borde sobre r no cambian ya que lo único que se busca cambiar en este nuevo problema es el funcional de costo, el cual está *centrado* alrededor del prior f .

2.3.3. Formulación de Benamou-Brenier

El trabajo desarrollado en la subsección anterior permite interpretar el problema de transporte óptimo como un problema de fluidodinámica, donde una medida es transportada de un lugar a otro mediante un campo de velocidades que preserva la masa en todo instante del trayecto y que, además, al comienzo y al final del recorrido la masa está distribuida de la forma que exige el problema de transporte óptimo. Para construir esta formulación, se reescribirá la restricción de preservación de masa usando la ecuación de continuidad, la cual será crucial en la mayor parte de las formulaciones dinámicas del transporte óptimo y puede verse como una versión determinística de la ecuación de Fokker-Planck (ver Teorema A.4).

Ecuación de transporte

Dado un fluido inmerso en $\mathcal{X} \subset \mathbb{R}^d$ y guiado por un campo de velocidades dinámico en el tiempo, $v : \mathcal{X} \times [0, T] \rightarrow \mathcal{X}$, la ecuación de transporte es una ecuación en derivadas parciales (PDE) que codifica la conservación de la masa del fluido a lo largo de su recorrido a través del campo vectorial v . Para derivar esta ecuación, es necesario el siguiente resultado clásico de cálculo vectorial:

Teorema 2.13 (divergencia de Gauss). Sea $\Omega \subset \mathbb{R}^d$ un volumen cerrado y acotado con frontera $\partial\Omega$. Entonces, para un campo vectorial $\vec{F} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ continuamente diferenciable, se tiene que:

$$\oint_{\partial\Omega} \vec{F} \cdot d\vec{S} = \int_{\Omega} \nabla \cdot \vec{F} dV,$$

donde la primera integral es una integral de superficie mientras que la segunda integral es una integral estándar en \mathbb{R}^d . En esta última, $\nabla \cdot \vec{F}$ es la divergencia del campo $\vec{F} = (F_1, \dots, F_d)$:

$$\nabla \cdot \vec{F} := \sum_{i=1}^d \frac{\partial F_i}{\partial x_i}.$$

En la Figura 2.15 se puede obtener una intuición de este teorema.

Teniendo este resultado, la derivación de la ecuación de conservación de masa es directa. En efecto, si se considera un fluido en $\mathcal{X} \subset \mathbb{R}^d$ desplazándose en el espacio gracias a un campo de velocidades $v : \mathcal{X} \times [0, T] \rightarrow \mathcal{X}$, su densidad de masa $\rho : \mathcal{X} \times [0, T] \rightarrow \mathbb{R}_+$ dependerá del tiempo debido a los cambios locales de densidad provocados por el campo de velocidades. Es decir, la masa total del fluido en un volumen $\Omega \subset \mathcal{X}$ en el instante $t \in [0, T]$ fijo viene dada por:

$$\int_{\Omega} \rho(x, t) dx.$$

En consecuencia, si la masa total se conserva, la tasa de cambio de la masa dentro del volumen Ω debe ser igual a la masa que sale de Ω a través de su superficie $\partial\Omega$ ya que, de lo contrario, se estaría creando o perdiendo masa dentro de $\Omega \subset \mathcal{X}$. Por lo tanto:

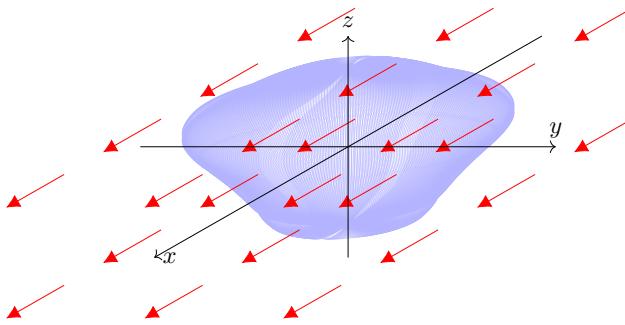


Figura 2.15: campo vectorial \vec{F} en \mathbb{R}^3 atravesando una superficie cerrada $\partial\Omega \subset \mathbb{R}^3$. El flujo total que pasa a través de $\partial\Omega$ (definido por la integral de superficie) es precisamente la cantidad de fluido se está *creando o perdiendo* en cada punto del espacio dentro de Ω , lo cual viene dado por el operador divergencia. De esta forma, sumando todas estas divergencias infinitesimales se obtiene el flujo neto a través de $\partial\Omega$. Esta imagen se encuentra en el archivo `surface_integral.ipynb`.

$$\frac{\partial}{\partial t} \left(\int_{\Omega} \rho(x, t) dx \right) = - \oint_{\partial\Omega} \rho(x, t) v(x, t) d\vec{S}.$$

Donde el signo negativo es debido a que se está considerando el flujo que sale de la superficie. Utilizando el Teorema 2.13 en el lado derecho con $F = \rho v$, se tiene que:

$$\frac{\partial}{\partial t} \left(\int_{\Omega} \rho(x, t) dx \right) = - \int_{\Omega} \nabla \cdot (\rho(x, t) v(x, t)) dx \implies \int_{\Omega} \left(\frac{\partial \rho}{\partial t}(x, t) + \nabla \cdot (\rho(x, t) v(x, t)) \right) dx = 0.$$

Donde se usó la regla de Leibniz para introducir la derivada en la integral. Luego, dado que el volumen de control $\Omega \subset \mathcal{X}$ es arbitrario, la igualdad anterior indica que

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho v) = 0. \quad (2.3.11)$$

Esta PDE se conoce como *ecuación de transporte* o *ecuación de continuidad* y es una ecuación fundamental tanto en la física como en la matemática ya que permite describir fenómenos que son conservados a lo largo del tiempo como la masa, la energía y la carga eléctrica. En este caso, la ecuación de transporte se está interpretando como una ecuación en fluidodinámica, donde forma parte de una de las ecuaciones de Euler para describir la dinámica que sigue un fluido a lo largo del tiempo. En particular, de (2.3.11) se obtiene directamente que si el fluido es incompresible (i.e. la densidad ρ es constante), la ecuación se reduce a $\nabla \cdot v = 0$, por lo que el campo de velocidades es solenoidal.

Es importante notar que esta es una ecuación de conservación fuerte en el sentido que no solo conserva la masa total del sistema si no que también la conserva de manera local, no permitiendo la *teleportación* de materia y forzando a que la trayectoria del fluido sea continua.

Interpretación fluidodinámica del transporte óptimo

Retomando la formulación de control óptimo (2.3.9), si las medidas μ y ν tienen densidades ρ_0 y ρ_1 respectivamente, las densidades marginales $\rho(t, \cdot)$ del sistema que evoluciona de acuerdo a la dinámica $dr_t = v(t, r_t) dt$ deben satisfacer (débilmente) la ecuación de continuidad (2.3.11), por lo que es posible reescribir el problema de transporte óptimo como la búsqueda de un campo de velocidades $v : \mathcal{X} \times [0, T] \rightarrow \mathcal{X}$ que haga evolucionar una función de densidad (desconocida) $\rho : \mathcal{X} \times [0, T] \rightarrow \mathbb{R}_+$ de tal forma que $\rho(\cdot, 0) = \rho_0$ y $\rho(\cdot, 1) = \rho_1$. Esto es lo que se conoce como la *formulación de Benamou-Brenier*:

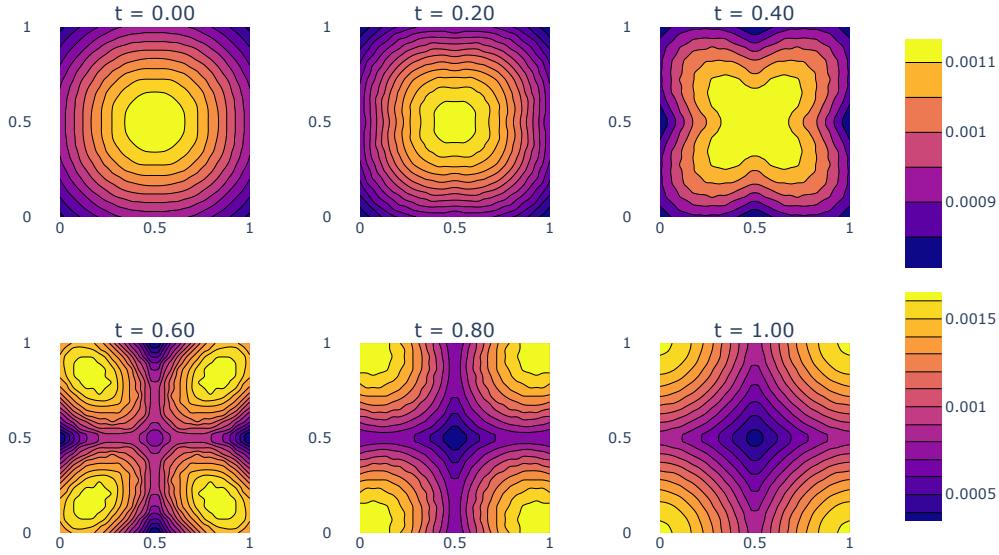


Figura 2.16: Simulación del problema de Benamou-Brenier para dos distribuciones. Se observa la interpolación entre ambas distribuciones para distintos tiempos. El código se encuentra en el archivo `benamou_brenier.ipynb`.

$$\inf_{(\rho, v)} \int_0^1 \int_{\mathcal{X}} \frac{1}{2} \|v(x, t)\|^2 \rho(x, t) \, dx \, dt \quad \text{sujeto a} \quad \begin{cases} \frac{\partial \rho}{\partial t} + \nabla \cdot (\rho v) = 0 \\ \rho(\cdot, 0) = \rho_0 \\ \rho(\cdot, 1) = \rho_1. \end{cases} \quad (2.3.12)$$

En la Figura 2.16 se puede observar una simulación de este problema.

Notar que el problema generalizado (2.3.10) también admite una formulación desde esta perspectiva. Si el prior f cumple una ecuación de continuidad (2.3.11) (sustituyendo v por f) cuya solución ρ no es consistente con las marginales μ y ν , esta ecuación de continuidad se puede interpretar como un prior sobre el proceso buscado, y así, el problema (2.3.10) busca una actualización del campo vectorial f que no difiera mucho de f pero que sí respete las distribuciones marginales:

$$\inf_{(\rho, v)} \int_0^1 \int_{\mathbb{R}^d} \frac{1}{2} \|v(x, t) - f(x, t)\|^2 \rho(x, t) \, dx \, dt \quad \text{sujeto a} \quad \begin{cases} \frac{\partial \rho}{\partial t} + \nabla \cdot (\rho v) = 0 \\ \rho(\cdot, 0) = \rho_0 \\ \rho(\cdot, 1) = \rho_1. \end{cases}$$

Notar que si el campo vectorial a priori f cumple las restricciones marginales, entonces $v = f$ resuelve el problema anterior. Además, si $f = 0$, se recupera el problema de transporte óptimo original.

Condiciones de optimalidad para Benamou-Brenier

En esta sección se encontrarán condiciones de optimalidad para el problema (2.3.12), las cuales serán posteriormente generalizadas para el problema del puente de Schrödinger en el Capítulo 3. Para no sobrecargar la notación, se omitirá la dependencia de las variables en las funciones.

Para comenzar, notar que el lagrangiano irrestringido del problema de Benamou-Brenier (2.3.12) sobre el conjunto factible²⁷ $\Gamma(\rho_0, \rho_1) \times C([0, 1] \times \mathcal{X})$ viene dado por:

²⁷Recordar que $\Gamma(\rho_0, \rho_1)$ es el conjunto de medidas en $C^1([0, 1], \mathcal{X})$ cuyas (densidades) marginales son ρ_0 y ρ_1 al comienzo y

$$L = \int_{\mathcal{X}} \int_0^1 \left[\frac{1}{2} \|v\|^2 \rho + \lambda(x, t) \left(\frac{\partial \rho}{\partial t} + \nabla \cdot (v\rho) \right) \right] dt dx,$$

donde

$$\int_0^1 \lambda \frac{\partial \rho}{\partial t} dt = (\lambda \rho)_{t=0}^{t=1} - \int_0^1 \frac{\partial \lambda}{\partial t} \rho dt \quad \text{y} \quad \int_{\mathcal{X}} \lambda \nabla \cdot (\rho v) dx = \underbrace{\oint_{\partial \mathcal{X}} \lambda(\rho v) d\vec{s}}_{=0} - \int_{\mathcal{X}} \nabla \lambda \cdot (\rho v) dx$$

Por lo tanto:

$$\begin{aligned} L &= \int_{\mathcal{X}} \int_0^1 \frac{1}{2} \|v\|^2 \rho dt dx + \int_{\mathcal{X}} \left((\lambda \rho)_{t=0}^{t=1} - \int_0^1 \frac{\partial \lambda}{\partial t} \rho dt \right) dx - \int_{\mathcal{X}} \int_0^1 \nabla \lambda \cdot (\rho v) dt dx \\ &= \int_{\mathcal{X}} \int_0^1 \left(\frac{1}{2} \|v\|^2 - \frac{\partial \lambda}{\partial t} - \nabla \lambda \cdot v \right) \rho dt dx + \underbrace{\int_{\mathcal{X}} (\lambda(x, 1)\rho_1(x) - \lambda(x, 0)\rho_0(x)) dx}_{\text{constante}} \end{aligned}$$

Para minimizar el lagrangiano sobre $\Gamma(\rho_0, \rho_1) \times C([0, 1] \times \mathcal{X})$ (con λ fijo), se comenzará fijando $\rho \in \Gamma(\rho_0, \rho_1)$ y se minimizará $v \in C([0, 1] \times \mathcal{X})$. Por condición de primer orden:

$$\frac{\partial L}{\partial v} = 0 \iff \frac{\partial}{\partial v} \left(\frac{1}{2} \|v\|^2 - \frac{\partial \lambda}{\partial t} - \nabla \lambda \cdot v \right) = v - \nabla \lambda = 0 \implies v^* = \nabla \lambda \text{ es control óptimo.}$$

Sustituyendo en el lagrangiano:

$$\begin{aligned} L &= \int_{\mathcal{X}} \int_0^1 \left(\frac{1}{2} \|\nabla \lambda\|^2 - \frac{\partial \lambda}{\partial t} - \|\nabla \lambda\|^2 \right) \rho dt dx + \text{constante} \\ &= - \int_{\mathcal{X}} \int_0^1 \left(\frac{\partial \lambda}{\partial t} + \frac{1}{2} \|\nabla \lambda\|^2 \right) \rho dt dx + \text{constante} \end{aligned}$$

Luego, si λ satisface la ecuación de Hamilton-Jacobi, $\frac{\partial \lambda}{\partial t} + \frac{1}{2} \|\nabla \lambda\|^2 = 0$, entonces $L = \text{constante}$ y cualquier $\rho \in \Gamma(\rho_0, \rho_1)$ minimiza el lagrangiano. En consecuencia, se probó lo siguiente:

Proposición 2.11. Si (ρ^*, λ^*) es solución del sistema acoplado

$$\begin{aligned} \frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \nabla \lambda) &= 0 \\ \frac{\partial \lambda}{\partial t} + \frac{1}{2} \|\nabla \lambda\|^2 &= 0 \end{aligned}$$

Con condiciones de borde $(\rho(\cdot, 0) = \rho_0) \wedge (\rho(\cdot, 1) = \rho_1)$, entonces, (ρ^*, v^*) es solución del problema de Benamou-Brenier (2.3.12) con $v^* = \nabla \lambda^*$.

Notar que la primera ecuación corresponde a la ecuación de transporte (conservación de masa), mientras que la segunda ecuación es la ecuación de Hamilton-Jacobi que codifica la optimalidad del lagrangiano. En el Capítulo 3 se verá una versión estocástica de esta formulación, donde la ecuación de transporte es cambiada por una ecuación tipo Fokker-Planck, mientras que la ecuación de Hamilton-Jacobi se cambia por una de Hamilton-Jacobi-Bellman.

al final del proceso. Además, el conjunto de controles admisibles se considera como todos los posibles controles continuos.

2.3.4. Modelos de difusión como mapa de transporte óptimo

Si bien los modelos de difusión a tiempo continuo estudiados en la Sección 1.4 tienen una naturaleza estocástica, la *probability flow ODE* (1.4.14) derivada a partir de la ecuación de Fokker-Planck del proceso de difusión induce un mapeo biunívoco entre el soporte de la distribución inicial p_{data} y el soporte de la distribución final p_{prior} . Más aún, este mapeo puede verse como un encoder de la distribución p_{data} donde p_{prior} se interpreta como el espacio latente asociado. Más específicamente, si $x \sim p_{\text{data}}$ es una muestra de la distribución inicial y $E_T(x)$ es la posición a la que llega x al final del proceso de difusión (simulado hasta tiempo T), entonces $E_T(x)$ puede verse como una codificación de x en el soporte de la distribución p_{prior} .

Con esta simplificación a un mapeo determinístico entre ambas distribuciones, es natural preguntarse si dicho mapa es un mapa de transporte óptimo entre las medidas asociadas a las densidades p_{data} y p_{prior} respectivamente. En [Khr+22] estudian esta pregunta obteniendo resultados parciales positivos.

En particular, para un modelo de difusión a tiempo continuo guiado por la VP-SDE (1.4.8) (i.e., la SDE asociada a la formulación continua del modelo DDPM propuesto por [HJA20]), los autores de [Khr+22] prueban empíricamente que cuando $T \rightarrow \infty$, el mapa $x \mapsto E_T(x)$ es (con un muy alto grado de precisión) similar al mapa de Monge²⁸ entre las distribuciones p_{data} y p_{prior} . Más aún, cuando p_{data} es una distribución gaussiana, los autores demuestran teóricamente que el mapa E_T es efectivamente el mapa de Monge entre ambas distribuciones cuando $T \rightarrow \infty$.

Para mostrar este resultado de forma empírica, en el estudio se consideraron diferentes distribuciones de datos, incluyendo el dataset ImageNet, el cual se considera de alta dimensión desde la perspectiva del transporte óptimo. En todos los casos, los autores de [Khr+22] lograron alcanzar una precisión a nivel de máquina, con un error relativo máximo del orden de 10^{-15} , donde compararon el mapa de transporte inducido por el proceso de difusión con el mapa de Monge obtenido al resolver la ecuación de Fokker-Planck asociada al problema.

Si bien este es un resultado al menos admirable, un trabajo posterior [LS22] construyó un contraejemplo donde la igualdad entre ambos mapas de transporte no se cumple. Más específicamente, construyeron un problema de transporte donde el mapa inducido por el modelo de difusión determinista difiere del mapa de Monge en un punto, mostrando que, o bien los modelos de difusión son solo una muy buena aproximación de los mapas de Monge, o bien se debe reducir el conjunto de distribuciones donde la igualdad es cierta. De todos modos, aunque los modelos de difusión (en su versión determinista) converjan al mapa de Monge, se sigue teniendo la limitación de tener que considerar un horizonte de tiempo infinito para obtener dicho mapa de transporte óptimo.

Por otra parte, es importante aclarar que [Khr+22] propone únicamente que el mapa que induce el modelo de difusión aproxima al mapa de Monge entre dos distribuciones (donde la distribución de llegada es gaussiana), pero no afirma que toda la trayectoria del proceso determinista (1.4.14) sea la solución de la formulación de Benamou-Brenier (ver Subsección 2.3.3). Esto puede verse claramente en la Figura 2.17 donde se observa que las trayectorias seguidas por las partículas en un modelo de difusión son curvas y no rectas como es el caso de la formulación fluidodinámica de Benamou-Brenier.

En esta línea de trabajo, propuestas recientes como *flow matching* (ver [Lip+23]) y *rectified flows* (ver [LGL22]) han ganado popularidad dentro de los modelos fundacionales.

²⁸Notar que para este problema el mapa de Monge existe y está bien definido gracias al teorema de Brenier Teorema 2.2.

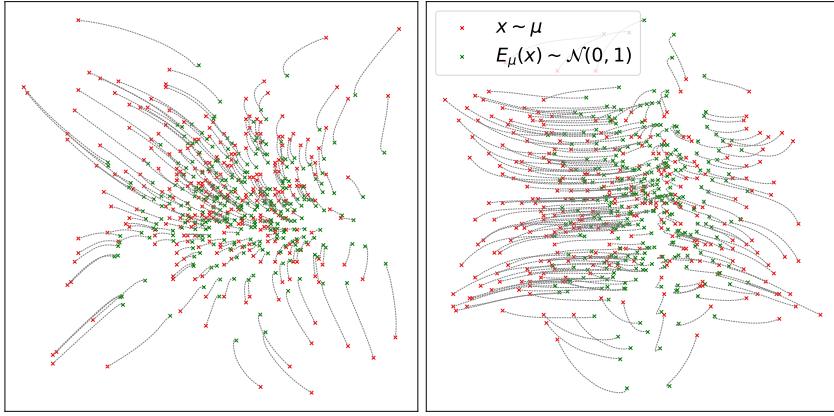


Figura 2.17: Trayectorias seguidas por las muestras de dos distribuciones iniciales $x \sim \mu$ (una en cada gráfico) cuando la distribución de llegada $E_\mu(x)$ es una variable aleatoria gaussiana estándar. Imagen obtenida desde [Khr+22].

2.3.5. Limitaciones del transporte óptimo dinámico

Si bien la distancia que induce el problema de Kantorovich en $\mathcal{M}_+^1(\mathcal{X})$ tiene buenas propiedades, esta métrica no ha sido muy utilizada en la comunidad del aprendizaje de máquinas ya que es computacionalmente costosa y sufre de la maldición de la dimensionalidad: cuando se busca aplicar el transporte óptimo a problemas de aprendizaje automático (p.g. como una función de pérdida), por lo general las medidas de probabilidad $\mu \in \mathcal{M}_+^1(\mathcal{X})$ y $\nu \in \mathcal{M}_+^1(\mathcal{Y})$ no son conocidas y solo se tiene acceso a una cantidad finita de muestras provenientes de estas medidas, por lo que se vuelve necesario aproximar la solución del problema de transporte óptimo. Sin embargo, en [Dud69] muestran que esta aproximación se deteriora exponencialmente a medida que aumenta la dimensión de los datos:

Teorema 2.14 (Dudley). Sea $\mathcal{X} \subset \mathbb{R}^d$ compacto y $\mu, \nu \in \mathcal{M}_+^1(\mathcal{X})$ dos medidas de probabilidad en este espacio. Dados dos conjuntos de muestras independientes $\{x_i\}_{i=1}^n, \{y_j\}_{j=1}^n \subset \mathbb{R}^d$ con $x_i \sim \mu$ e $y_j \sim \nu$ para $i, j \in \{1, \dots, n\}$, se pueden aproximar las medidas μ y ν mediante sus medidas empíricas

$$\hat{\mu} := \frac{1}{n} \sum_{i=1}^n \delta_{x_i} \quad \hat{\nu} := \frac{1}{n} \sum_{j=1}^n \delta_{y_j}.$$

Bajo esta aproximación, para $d \gg 1$ y $p \in [1, \infty)$ se tiene el siguiente error para el estimador $\mathcal{W}_p(\hat{\mu}, \hat{\nu})$ de $\mathcal{W}_p(\mu, \nu)$:

$$\mathbb{E}_{(x_i, y_i)_{i=1}^n \sim (\mu \otimes \nu)^n} [\|\mathcal{W}_p(\hat{\mu}, \hat{\nu}) - \mathcal{W}_p(\mu, \nu)\|] = \mathcal{O}(n^{-\frac{1}{d}}).$$

Notar que en el teorema anterior, el estimador $\mathcal{W}_p(\hat{\mu}, \hat{\nu})$ es un estimador consistente ya que $\hat{\mu} \rightharpoonup \mu$ y $\hat{\nu} \rightharpoonup \nu$, por lo que $\mathcal{W}_p(\hat{\mu}, \hat{\nu}) \rightarrow \mathcal{W}_p(\mu, \nu)$ (recordar que la distancia de Wasserstein es compatible con la convergencia débil).

En consecuencia, si bien el problema de Kantorovich entrega buenas propiedades métricas que pueden resultar en una buena función de pérdida para usar en el aprendizaje automático, el hecho de que sufra de la maldición de la dimensionalidad ha provocado que se prefieran usar f -divergencias que, si bien no son capaces de acercarse a la geometría del problema, sí son muy eficientes de computar.

Con el fin de aminorar los problemas encontrados en el transporte óptimo, en el siguiente capítulo se estudiará una versión regularizada del problema de Kantorovich la cual entregará un problema estrictamente convexo, podrá ser resuelto de manera eficiente y, como se verá en la Sección 3.2, resultará ser equivalente al problema del puente de Schrödinger estático. Además, la versión dinámica del problema de Schrödinger podrá ser vista como una versión ruidosa de la formulación de Benamou-Brenier donde la ecuación de continuidad (2.3.11) será sustituida por una ecuación de Fokker-Planck (ver Sección A.2).

Capítulo 3

Regularización y problema de Schrödinger

Si bien el problema de Kantorovich discreto (2.2.1) es un problema de programación lineal (en particular, es convexo), es costoso de resolver por su cantidad cuadrática de incógnitas y nada garantiza la unicidad de la solución¹, la cual en general no se tiene cuando el problema presenta cierto tipo de simetrías. Además, aun cuando la formulación dual (2.2.7) logra disminuir el número de incógnitas a una cantidad lineal, no soluciona el problema de la no unicidad de la solución, lo cual puede generar problemas al momento de resolver numéricamente el problema de optimización, ya sea el primal o el dual. Por otra parte, tal como se mencionó al final del Capítulo 2 el problema de transporte óptimo sufre de la maldición de la dimensionalidad al intentar aproximar la distancia de Wasserstein mediante una aproximación empírica, lo que limita fuertemente su uso en altas dimensiones.

Motivado por estas limitaciones, en [Cut13] se propuso una técnica de regularización para el problema de Kantorovich, la cual posteriormente resultó ser equivalente a la formulación estática del puente de Schrödinger. La técnica de regularización entrópica consiste en agregar un término adicional a la función objetivo del problema de Kantorovich con el fin de poder obtener un problema dual que se pueda resolver eficientemente y que, además, caracterice la solución del problema primal.

3.1. Transporte óptimo entrópico

Una forma de reparar el problema de la no unicidad de la solución es regularizar el problema de Kantorovich agregando un término conveniente que transforme la función objetivo en una función estrictamente convexa, permitiendo garantizar la unicidad del minimizador. Es decir, el problema de optimización regularizado es de la forma

$$\inf_{P \in \Pi_d(\mu, \nu)} \sum_{i=1}^n \sum_{j=1}^m C_{ij} P_{ij} + \epsilon \cdot \text{Regularizador}(P), \quad (3.1.1)$$

en el caso discreto y

¹Si bien se puede tener múltiples minimizadores, el valor mínimo es único, por lo que la no unicidad no influye sobre el valor de la distancia de Wasserstein entre dos distribuciones. Sin embargo, sí puede generar problemas al intentar aproximarla numéricamente.

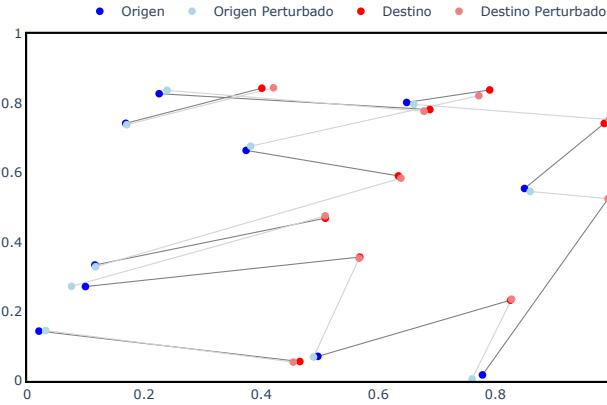


Figura 3.1: Inestabilidad del mapa de Kantorovich bajo pequeñas perturbaciones en los parámetros del problema. En este caso, todos los elementos de \mathcal{X} e \mathcal{Y} fueron perturbados con un ruido gaussiano de baja varianza. La imagen se encuentra en el archivo `ot_instability.ipynb`.

$$\inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) \, d\pi(x, y) + \epsilon \cdot \text{Regularizador}(\pi), \quad (3.1.2)$$

en su versión continua. En ambos casos, $\epsilon > 0$ es un ponderador que indica cuánta importancia se le da al regularizador durante la optimización. Recordando que el problema de Kantorovich define una distancia en (un subconjunto de) $\mathcal{M}_+^1(\mathcal{X})$ cuando $\mathcal{X} = \mathcal{Y} \subset \mathbb{R}^d$ y $c(x, y) = \|x - y\|^p$, este nuevo problema de optimización permite obtener una versión suavizada de la distancia de Wasserstein.

Para la elección de la función de regularización, es útil observar que las soluciones del problema de Kantorovich por lo general pueden ser muy inestables con respecto a pequeñas variaciones en las medidas μ y ν o en el funcional de costo c , lo cual se puede ver en la Figura 3.1. Dado que esto no es una propiedad deseable, se puede elegir un regularizador que además de ser estrictamente convexo, promueva soluciones estables, pudiendo interpretar los problemas regularizados (3.1.1) y (3.1.2) como un trade-off entre tener un plan de Kantorovich inestable y tener una plan de transporte sub-óptimo pero más estable a cambios en los parámetros del problema. En este caso, el regularizador también se puede entender como un regularizador sobre el modelo ya que evita el sobreajuste a los datos de entrenamiento y le entrega robustez al problema con respecto a la función objetivo, donde el término adicional puede ser considerado como una incertidumbre sobre el funcional de costo real. Por otra parte, en muchos casos donde se utiliza la distancia de Wasserstein (p.g. como función de pérdida en un modelo de aprendizaje automático) es suficiente tener una aproximación razonable, por lo que el sesgo inducido por el regularizador no afecta considerablemente a la utilidad de la métrica de Wasserstein.

Como se verá en la siguiente subsección, una buena elección para el regularizador es la función de información media, la cual es estrictamente convexa y, además, promueve soluciones más suaves (en algún sentido a precisar). En particular, en el caso discreto, este regularizador forzará a que el plan de transporte entrópico óptimo sea denso, en el sentido de que todos los puntos en $\mathcal{X} = \{x_i\}_{i=1}^n$ envían masa a todos los puntos en $\mathcal{Y} = \{y_j\}_{j=1}^m$, a diferencia de lo que se observa en muchos casos donde los planes de Kantorovich son *deterministas* al estar asociados a mapas de Monge donde no se permite división de masa por parte de los elementos de \mathcal{X} . Además, como se verá en la Subsección 3.2.2, este problema se puede resolver de forma eficiente, siendo esto último uno de los mayores beneficios del transporte óptimo entrópico (EOT).

3.1.1. Regularización entrópica

Para el problema de Kantorovich discreto (2.2.1) se utilizará como regularizador (el negativo de) la función de entropía con el fin de promover una alta entropía del plan de transporte. Por otra parte, se verá que utilizar este regularizador es equivalente (salvo constantes aditivas) a regularizar el problema utilizando la divergencia de Kullback-Leibler (reversa) con respecto a la medida producto $\mu \otimes \nu$. Este último punto permitirá escribir el problema de transporte óptimo entrópico como un caso particular del problema del puente de Schrödinger en la Sección 3.2.

Problema entrópico discreto

Para estudiar el problema de transporte óptimo entrópico se comenzará dando su formulación en un escenario discreto, donde $\mathcal{X} = \{x_i\}_{i=1}^n$ e $\mathcal{Y} = \{y_j\}_{j=1}^m$ son conjuntos finitos que serán dotados de medidas discretas.

Se comenzará definiendo la entropía de una medida de probabilidad discreta:

Definición 3.1 (entropía discreta). Para una medida discreta $\mu \in \mathcal{M}_+^1(\mathcal{X})$ con vector de probabilidad $a \in \Sigma_n$, se define la entropía discreta de μ como:

$$\mathcal{H}(\mu) := - \sum_{i=1}^n a_i \log(a_i), \quad (3.1.3)$$

donde, por continuidad, se define $0 \cdot \log(0) = 0$ en el caso que algún $a_i \in [0, 1]$ sea nulo.

Dado que la medida discreta $\mu \in \mathcal{M}_+^1(\mathcal{X})$ se identifica con el vector de probabilidad $a \in \Sigma_n$, es usual escribir $\mathcal{H}(a)$ en vez de $\mathcal{H}(\mu)$. Del mismo modo, dado que las medidas medidas $\mu \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})$ se identifican con su matriz de probabilidades² $P \in \mathcal{M}_{n,m}([0, 1])$ mediante $P_{ij} = \mu(\{(x_i, y_j)\})$ (ver (2.2.4)), se suele escribir $\mathcal{H}(P)$ en vez de $\mathcal{H}(\mu)$.

Se verá que esta función cumple las propiedades deseadas:

Proposición 3.1. Las siguientes propiedades son ciertas:

1. La función de entropía discreta $\mu \mapsto \mathcal{H}(\mu)$ es estrictamente cóncava y, en consecuencia, el regularizador $\text{Regularizador}(P) = -\mathcal{H}(P)$ es estrictamente convexo y el problema (3.1.1) tiene solución única.
2. Dadas dos medidas discretas, $\mu \in \mathcal{M}_+^1(\mathcal{X})$ y $\nu \in \mathcal{M}_+^1(\mathcal{Y})$, con vectores de probabilidad $a \in \Sigma_n$ y $b \in \Sigma_m$ respectivamente, el problema de optimización

$$\max_{P \in \Pi_d(\mu, \nu)} \mathcal{H}(P), \quad (3.1.4)$$

es maximizado por la matriz $ab^\top \in \mathcal{M}_{n,m}([0, 1])$, la cual es la matriz de probabilidades asociada a la medida producto $\mu \otimes \nu \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})$.

Demostración. Para la primera proposición, se verá que la función de entropía discreta definida en (3.1.3) tiene matriz hessiana definida negativa para toda medida $\mu \in \mathcal{M}_+^1(\mathcal{X})$:

$$\frac{\partial \mathcal{H}}{\partial a_i}(\mu) = -(\log(a_i) + 1) \implies \frac{\partial^2 \mathcal{H}}{\partial a_i a_j}(\mu) = \begin{cases} \frac{-1}{a_i} & \text{si } i = j \\ 0 & \text{si } i \neq j. \end{cases}$$

²En este caso, la suma en (3.1.3) es sobre todos los elementos de la matriz de probabilidades P .

Por lo tanto, $D^2(\mathcal{H}(\mu)) = \text{diag}\left(\frac{-1}{a_1}, \dots, \frac{-1}{a_n}\right)$ y, así, todos los valores propios de $D^2(\mathcal{H}(\mu))$ son negativos³, por lo que la matriz es definida negativa para toda medida $\mu \in \mathcal{M}_+^1(\mathcal{X})$.

Para la segunda proposición, el problema de optimización con sus restricciones escritas de forma estándar es:

$$\begin{aligned} & \max_{P \in \mathcal{M}_{n,m}(\mathbb{R})} - \sum_{i=1}^n \sum_{j=1}^m P_{ij} \log(P_{ij}) \\ & \text{s.a: } a_i - \sum_{j=1}^m P_{ij} = 0, \quad \forall i \in \{1, \dots, n\} \\ & \qquad b_j - \sum_{i=1}^n P_{ij} = 0, \quad \forall j \in \{1, \dots, m\} \\ & \qquad -P_{ij} \leq 0, \quad \forall i \in \{1, \dots, n\}, j \in \{1, \dots, m\}. \end{aligned}$$

Dado que el problema es convexo, se pueden utilizar las condiciones de Karush–Kuhn–Tucker (KKT) para identificar a un máximo. El lagrangiano del problema $L : \mathcal{M}_{n,m}(\mathbb{R}) \times \mathcal{M}_{n,m}(\mathbb{R}) \times \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$ es:

$$L(P, \lambda, \kappa, \eta) = - \sum_{i=1}^n \sum_{j=1}^m P_{ij} \log(P_{ij}) + \kappa_i \left(a_i - \sum_{j=1}^m P_{ij} \right) + \eta_j \left(b_j - \sum_{i=1}^n P_{ij} \right) + \lambda_{ij}(-P_{ij}).$$

Las condiciones de KKT para este problema son:

- Factibilidad: $P \in \Pi_d(\mu, \nu)$
- Holgura complementaria: $\lambda_{ij}(-P_{ij}) = 0, \forall i \in \{1, \dots, n\}, j \in \{1, \dots, m\}$
- $\lambda_{ij} \geq 0, \forall i \in \{1, \dots, n\}, j \in \{1, \dots, m\}$.
- Punto crítico para el lagrangiano: $\nabla_P L = 0$.

Para que la última condición se cumpla es necesario que:

$$\frac{\partial L}{\partial P_{ij}}(P, \lambda, \kappa, \eta) = -(\log(P_{ij}) + 1) - \kappa_i - \eta_j - \lambda_{ij} = 0 \implies P_{ij} = e^{-1-\kappa_i-\eta_j-\lambda_{ij}}.$$

Se verá que es posible encontrar una tripleta (P, κ, η) que cumpla las condiciones de KKT cuando $\lambda_{ij} = 0$. Bajo esta consideración, la última igualdad permite escribir P_{ij} como un producto de la forma $P_{ij} = p_i q_j$ y así, para que la condición de factibilidad se cumpla es necesario que:

$$\begin{aligned} \sum_{j=1}^m P_{ij} &= a_i \implies p_i \sum_{j=1}^m q_j = a_i \implies p_i = a_i \\ \sum_{i=1}^n P_{ij} &= b_j \implies q_j \sum_{i=1}^n p_i = b_j \implies q_j = b_j, \end{aligned}$$

y, por lo tanto, con $P_{ij} = a_i b_j$ se cumplen todas las condiciones de KKT, por lo que $P = ab^\top$ es un máximo del problema. \square

Por la proposición anterior, el mapa de transporte que maximiza la entropía en (3.1.4) es la medida producto $\mu \otimes \nu$ definida para cada par $(x_i, y_j) \in \mathcal{X} \times \mathcal{Y}$ como

³Recordar que los valores propios de una matriz diagonal son precisamente los valores de su diagonal.

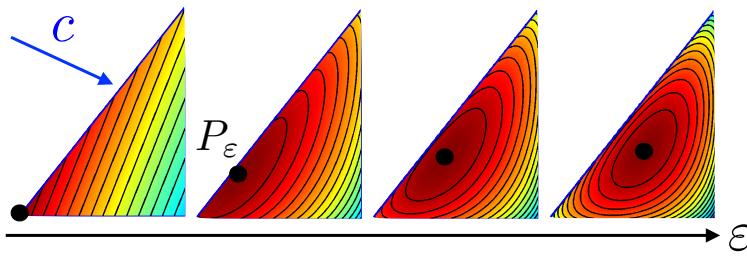


Figura 3.2: Evolución de la solución óptima $P_\epsilon \in \Pi_d(\mu, \nu)$ del problema regularizado (3.1.6) para distintos valores de ϵ . Dado que el problema de Kantorovich discreto no regularizado ($\epsilon = 0$) es un problema de programación lineal, el argumento minimizante se encontrará siempre en un vértice del polítopo $\Pi_d(\mu, \nu)$. Al aumentar el valor de ϵ , la solución óptima para el problema entrópico tiende a alejarse de los vértices y converger hacia un centro de alta entropía. Más aún, se puede demostrar que la matriz de transporte óptima para este problema regularizado tiene todas sus coordenadas positivas (dado que no tiene ninguna restricción de desigualdad activa, lo cual ocurre en los vértices de $\Pi_d(\mu, \nu)$), por lo que todos los puntos de \mathcal{X} transportan masa a todos los elementos de \mathcal{Y} . Imagen obtenida desde [PC20].

$$(\mu \otimes \nu)(\{x_i\} \times \{y_j\}) = \mu(\{x_i\}) \nu(\{y_j\}) = (ab^\top)_{ij} = a_i b_j.$$

Este plan de transporte distribuye la masa de cada $x_i \in \mathcal{X}$ en todos los elementos de \mathcal{Y} , donde cada $y_j \in \mathcal{Y}$ recibe una cantidad proporcional a su masa $\nu(\{y_j\}) = b_j$, por lo que la cantidad de masa transferida desde x_i hacia y_j (según $\mu \otimes \nu$) corresponde a $a_i b_j$. Recordando que en el caso discreto siempre se puede considerar que $a_i, b_j > 0$, el plan de transporte $\mu \otimes \nu$ envía masa a todos los elementos de \mathcal{Y} desde todos los elementos de \mathcal{X} , siendo este el extremo opuesto a lo que realiza el mapa de Monge. Por otra parte, desde una perspectiva probabilística, la medida producto $\mu \otimes \nu$ puede interpretarse como la distribución conjunta de dos variables aleatorias independientes, mientras que un coupling general $\pi \in \Pi(\mu, \nu)$ puede interrelacionar ambas componentes.

Con estas propiedades demostradas, el problema de transporte óptimo entrópico con Regularizador(P) = $-\mathcal{H}(P)$ se puede escribir explícitamente como

$$\inf_{P \in \Pi_d(\mu, \nu)} \sum_{i=1}^n \sum_{j=1}^m C_{ij} P_{ij} + \epsilon \cdot \sum_{i=1}^n \sum_{j=1}^m P_{ij} \log(P_{ij}). \quad (3.1.5)$$

En la Figura 3.2 se puede observar cómo se traslada la medida óptima dentro del conjunto factible $\Pi_d(\mu, \nu)$ al aumentar el valor de ϵ .

Por otra parte, debido a que la medida producto maximiza la función de entropía discreta en (3.1.4), el problema entrópico (3.1.5) entregará una solución que se podrá interpretar como una interpolación entre el plan de Kantorovich y el plan de transporte trivial $\mu \otimes \nu$. Esto sugiere que, posiblemente, el regularizador $-\mathcal{H}(H)$ sea equivalente a otro regularizador que promueva soluciones similares a la medida producto $\mu \otimes \nu$. Para estudiar esto, se dará una versión discreta de la divergencia de Kullback-Leibler utilizada para medir discrepancias entre medidas en el Capítulo 1:

Definición 3.2 (divergencia de Kullback-Leibler, caso discreto). Dadas dos medidas discretas $\mu, \nu \in \mathcal{M}_+^1(\mathcal{X})$ con vectores de probabilidad $a, b \in \Sigma_n$ respectivamente, se define la divergencia de Kullback-Leibler entre μ y ν como:

$$D_{KL}(\mu \| \nu) := \sum_{i=1}^n a_i \log \left(\frac{a_i}{b_i} \right), \quad (3.1.6)$$

donde, por continuidad, se define $0 \cdot \log(0) = 0$ en el caso que algún $a_i \in [0, 1]$ sea nulo. Además, en línea con la Definición 2.15, si para algún índice $i \in \{0, \dots, n\}$ se tiene que $a_i > b_i = 0$, se define $D_{KL}(\mu \| \nu) = \infty$.

Notar que, al igual que para la entropía discreta (3.1.3), para calcular la divergencia de Kullback-Leibler entre medidas en un espacio producto $\mathcal{X} \times \mathcal{Y}$, la suma en (3.1.6) debe ser sobre todos los elementos de las matrices de probabilidad asociada a cada medida.

La siguiente propiedad relaciona el concepto de entropía de un coupling con la divergencia de Kullback-Leibler:

Proposición 3.2. Dadas dos medidas discretas, $\mu \in \mathcal{M}_+^1(\mathcal{X})$ y $\nu \in \mathcal{M}_+^1(\mathcal{Y})$, con vectores de probabilidad $a \in \Sigma_n$ y $b \in \Sigma_m$ respectivamente, entonces, para un coupling $\pi \in \Pi(\mu, \nu)$ se tiene la siguiente descomposición:

$$D_{KL}(\pi \| \mu \otimes \nu) = -H(\pi) + H(\mu) + H(\nu),$$

donde $\mu \otimes \nu \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})$ es la medida producto, cuya matriz de probabilidades es $ab^\top \in \mathcal{M}_{n,m}([0, 1])$.

Demuestração. Sea $P \in \mathcal{M}_{n,m}([0, 1])$ la matriz de probabilidades asociada a π . Notando que $(ab^\top)_{ij} = a_i b_j$, se tiene lo siguiente:

$$\begin{aligned} D_{KL}(\pi \| \mu \otimes \nu) &= \sum_{i=1}^n \sum_{j=1}^m P_{ij} \log \left(\frac{P_{ij}}{a_i b_j} \right) \\ &= \sum_{i=1}^n \sum_{j=1}^m P_{ij} \log(P_{ij}) - \sum_{i=1}^n \sum_{j=1}^m P_{ij} \log(a_i) - \sum_{i=1}^n \sum_{j=1}^m P_{ij} \log(b_j) \\ &= -H(\pi) - \sum_{i=1}^n \left(\log(a_i) \sum_{j=1}^m P_{ij} \right) - \sum_{j=1}^m \left(\log(b_j) \sum_{i=1}^n P_{ij} \right) \\ &= -H(\pi) + H(\mu) + H(\nu), \end{aligned}$$

donde en la penúltima igualdad se usó que $\sum_{j=1}^m P_{ij} = a_i$ y $\sum_{i=1}^n P_{ij} = b_j$ ya que la medida π tiene marginales μ y ν . \square

Notar que esta propiedad permite concluir directamente que la función $\pi \mapsto D_{KL}(\pi \| \mu \otimes \nu)$ hereda las buenas propiedades de la función de entropía (3.1.3) vistas en la Proposición 3.1: es estrictamente convexa y, bajo la restricción $\pi \in \Pi(\mu, \nu)$, es minimizada por la medida producto $\mu \otimes \nu \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})$. Con esto, es razonable considerar Regularizador(P) = $D_{KL}(\pi \| \mu \otimes \nu)$ en (3.1.1), lo cual es equivalente a utilizar el regularizador $-\mathcal{H}(\pi)$:

$$\begin{aligned} \min_{P \in \Pi_d(\mu, \nu)} \langle C, P \rangle_F + \epsilon \cdot D_{KL}(P \| \mu \otimes \nu) &= \min_{P \in \Pi_d(\mu, \nu)} \langle C, P \rangle_F + \epsilon \cdot (H(\mu) + H(\nu) - H(P)) \\ &= \epsilon \cdot (H(\mu) + H(\nu)) + \left(\min_{P \in \Pi_d(\mu, \nu)} \langle C, P \rangle_F - \epsilon \cdot H(P) \right). \end{aligned}$$

Por lo que el minimizador es el mismo en ambos casos y los respectivos valores óptimos solo difieren por una constante que depende de μ , ν y ϵ . Dado que el problema del puente de Schrödinger es formulado mediante

la minimización de la divergencia de Kullback-Leibler, se preferirá definir el problema de transporte óptimo entrópico utilizando este regularizador, aunque algunas veces se utilizará el problema equivalente (3.1.5) para mostrar o ilustrar propiedades del problema entrópico.

Definición 3.3 (problema entrópico, versión discreta). Sean $\mu \in \mathcal{M}_+^1(\mathcal{X})$ y $\nu \in \mathcal{M}_+^1(\mathcal{Y})$ dos medidas de probabilidad discretas en $\mathcal{X} = \{x_i\}_{i=1}^n$ e $\mathcal{Y} = \{y_j\}_{j=1}^m$, con vectores de probabilidad $a \in \Sigma_n$ y $b \in \Sigma_m$ respectivamente. Para $\epsilon > 0$ y una matriz de costo $C \in \mathcal{M}_{n,m}(\mathbb{R})$, el problema de transporte óptimo entrópico entre μ y ν es:

$$\min_{P \in \Pi_d(\mu, \nu)} \langle C, P \rangle_F + \epsilon \cdot D_{KL}(P \| \mu \otimes \nu) = \sum_{i=1}^n \sum_{j=1}^m C_{ij} P_{ij} + \epsilon \sum_{i=1}^n \sum_{j=1}^m P_{ij} \log \left(\frac{P_{ij}}{a_i b_j} \right). \quad (3.1.7)$$

Este problema tiene una formulación similar al problema del puente de Schrödinger estático estudiado en la Sección 3.2, donde el término $\langle C, P \rangle_F$ en (3.1.7) será absorbido por la divergencia $D_{KL}(P \| \mu \otimes \nu)$ para obtener un caso particular del problema de Schrödinger estático cuando se considera una distribución de referencia específica conocida como *kernel de Gibbs*.

Para concluir la formulación discreta, se estudiará la relación del problema entrópico (3.1.7) con el problema de Kantorovich original. En particular, se mostrará que la solución entrópica converge a el plan de Kantorovich de máxima entropía cuando $\epsilon \rightarrow 0^+$, mientras que la solución del problema entrópico tiende a la medida producto cuando $\epsilon \rightarrow \infty$.

Proposición 3.3. Sean $\mu \in \mathcal{M}_+^1(\mathcal{X})$ y $\nu \in \mathcal{M}_+^1(\mathcal{Y})$ dos medidas de probabilidad discreta con vectores de medida $a \in \Sigma_n$ y $b \in \Sigma_m$ respectivamente. Si $P_\epsilon \in \mathcal{M}_{n,m}([0, 1])$ es la (única) solución del problema entrópico discreto (3.1.7) para $\epsilon > 0$, entonces la función $\epsilon \mapsto P_\epsilon$ tiene los siguientes límites:

$$\begin{aligned} \lim_{\epsilon \rightarrow 0^+} P_\epsilon &= \arg \max_{P \in \Pi_d(\mu, \nu)} \left\{ H(P) : P \in \arg \min_{P \in \Pi_d(\mu, \nu)} \langle C, P \rangle_F \right\} \\ \lim_{\epsilon \rightarrow \infty} P_\epsilon &= ab^\top, \end{aligned}$$

donde $ab^\top \in \Pi_d(\mu, \nu)$ es la matriz de probabilidad asociada a la medida producto $\mu \otimes \nu \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})$.

Demostración. Para el primer límite, es necesario probar que $\lim_{\epsilon \rightarrow 0^+} P_\epsilon$ es un plan de transporte factible, minimiza el problema de Kantorovich (2.2.1) y que, de todas las soluciones del problema de Kantorovich, esta es la de mayor entropía.

Sea $(\epsilon_n)_{n \in \mathbb{N}} \subset \mathbb{R}_{++}$ una sucesión positiva con $\epsilon \rightarrow 0^+$. Como el conjunto de matrices $\Pi_d(\mu, \nu)$ es un conjunto acotado, la sucesión de soluciones $(P_{\epsilon_n})_{n \in \mathbb{N}} \subset \Pi_d(\mu, \nu)$ posee una subsucesión $(P_{\epsilon_{y_n}})_{n \in \mathbb{N}} \subset \Pi_d(\mu, \nu)$ convergente a una matriz $P_0 \in \mathcal{M}_{n,m}([0, 1])$. Además, como $\Pi_d(\mu, \nu)$ es un conjunto cerrado, se tiene además que $P_0 \in \Pi_d(\mu, \nu)$, por lo que la matriz de probabilidad límite, P_0 , es efectivamente un plan de transporte.

Para ver que P_0 es solución del problema no regularizado (2.2.1), se mostrará que $\langle C, P_0 \rangle$ toma el mismo valor que $\langle C, P^* \rangle$ para una solución P^* del problema de Kantorovich. Para esto, si P^* es solución del problema (2.2.1), entonces, para todo $n \in \mathbb{N}$:

$$\langle C, P^* \rangle \leq \langle C, P_{y_n} \rangle \implies 0 \leq \langle C, P_{y_n} \rangle_F - \langle C, P^* \rangle_F.$$

Por otra parte, como P_{y_n} es solución para el problema entrópico (3.1.5) (el cual es equivalente a (3.1.7)), se tiene la desigualdad:

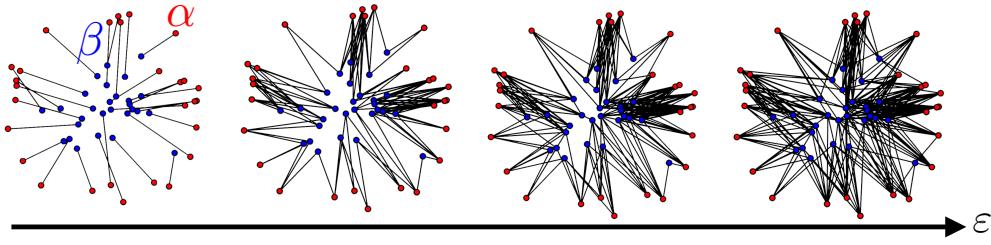


Figura 3.3: (Izquierda) plan de Kantorovich entre dos medidas μ y ν con vectores de probabilidad α y β respectivamente. (Derecha) evolución de la solución entrópica $P_\epsilon \in \Pi_d(\mu, \nu)$ para distintos valores de ϵ . Solo se muestran los arcos para valores $(P_\epsilon)_{ij} \in [0, 1]$ mayores a un cierto umbral. Imagen obtenida desde [PC20].

$$\langle C, P_{y_n} \rangle_F - y_n \cdot \mathcal{H}(P_{y_n}) \leq \langle C, P^* \rangle_F - y_n \cdot \mathcal{H}(P^*).$$

Luego, por ambas desigualdades:

$$0 \leq \langle C, P_{y_n} \rangle_F - \langle C, P^* \rangle_F \leq y_n (\mathcal{H}(P_{y_n}) - \mathcal{H}(P^*)). \quad (3.1.8)$$

Como la función $P \mapsto \mathcal{H}(P)$ es continua, $\mathcal{H}(P_{y_n}) \rightarrow \mathcal{H}(P_0)$. Del mismo modo, como $P \mapsto \langle C, P \rangle_F$ es continua (ya que es lineal), $\langle C, P_{y_n} \rangle_F \rightarrow \langle C, P_0 \rangle_F$. En conclusión, tomando $y_n \rightarrow 0^+$ se obtiene que $\langle C, P_0 \rangle_F = \langle C, P^* \rangle_F$, por lo que P_0 es efectivamente una solución para el problema de Kantorovich.

Por último, la desigualdad (3.1.8) muestra que $\mathcal{H}(P^*) \leq \mathcal{H}(P_{y_n})$, para todo $n \in \mathbb{N}$, por lo que tomando límite y usando nuevamente que la función de entropía es continua, se concluye que $\mathcal{H}(P^*) \leq \mathcal{H}(P_0)$.

Para el segundo límite basta notar que el funcional de costo $\langle C, P \rangle_F$ es acotado cuando $P \in \Pi_d(\mu, \nu)$ ya es un funcional continuo y $P \in \Pi_d(\mu, \nu)$ es un conjunto cerrado y acotado. De este modo, si $\epsilon > 0$ crece indefinidamente en (3.1.7), el máximo convergerá al mínimo de $D_{KL}(P \| \mu \times \nu)$, el cual es el máximo de $\mathcal{H}(P)$ que, según la Proposición 3.1, es la medida producto $\mu \otimes \nu$. \square

Esta propiedad puede observarse en la Figura 3.3. Por otro lado, es importante mencionar que es posible regularizar el problema de Kantorovich utilizando otra f -divergencia entre π y $\mu \otimes \nu$, y no necesariamente la divergencia de Kullback-Leibler. Sin embargo, se suele preferir este regularizador ya que el problema dual se puede resolver eficientemente.

Problema entrópico continuo

En este apartado se extenderá el análisis anterior al caso continuo $\mathcal{X} \subset \mathbb{R}^n$, $\mathcal{Y} \subset \mathbb{R}^m$. Dado que la entropía y la divergencia de Kullback-Leibler son regularizadores equivalentes, basta extender solo uno a su versión continua. Se elegirá la divergencia de Kullback-Leibler debido a que la Definición 2.16 permite extender f -divergencias al continuo, siendo consistente con la Definición 3.3:

Definición 3.4 (divergencia de Kullback-Leibler, caso continuo). Dadas dos medidas de probabilidad $\mu, \nu \in \mathcal{M}_+^1(\mathcal{X})$, se define la divergencia de Kullback entre μ y ν como

$$D_{KL}(\mu \| \nu) := \begin{cases} \int_{\mathcal{X}} \log\left(\frac{d\mu}{d\nu}(x)\right) d\mu(x) & \text{si } \mu \ll \nu \\ \infty & \text{si } \mu \not\ll \nu, \end{cases}$$

donde $\frac{d\mu}{d\nu} : \mathcal{X} \rightarrow \mathbb{R}_+$ es la derivada de Radon-Nikodym de μ con respecto a ν (ver Teorema A.1) y la relación de orden \ll es la de continuidad absoluta entre medidas (ver Definición A.3).

El valor para $D_{KL}(\mu \| \nu)$ cuando $\mu \not\ll \nu$ se debe a que la velocidad límite para esta f -divergencia (con f la información de Shannon) es no acotada (ver Definición 2.15). En la literatura, esta función muchas veces es conocida como *entropía relativa* en su formulación continua, pero en este trabajo se seguirá utilizando el nombre *divergencia de Kullback-Leibler* para mantener la consistencia entre las formulaciones discretas y continuas.

Notar que si las medidas $\mu, \nu \in \mathcal{M}_+^1(\mathcal{X})$ tienen funciones de densidad p_μ y p_ν respectivamente, se recupera la definición estándar usada en la Definición 1.1. De manera informal:

$$\frac{d\mu}{d\nu}(x) = \frac{d\mu}{dx}(x) : \frac{d\nu}{dx}(x) = \frac{p_\mu(x)}{p_\nu(x)} \implies D_{KL}(\mu \| \nu) = \int_{\mathcal{X}} \log\left(\frac{p_\mu(x)}{p_\nu(x)}\right) p_\mu(x) dx,$$

donde dx es la medida de Lebesgue en \mathcal{X} . En estos casos, al igual que en el Capítulo 1, se preferirá la notación $D_{KL}(p_\mu \| p_\nu)$ para hacer explícita la dependencia de las densidades.

Con esta extensión del término regularizador al caso continuo, la formulación continua del problema (3.1.7) es la siguiente:

Definición 3.5 (problema entrópico, versión continua). Sean $\mu \in \mathcal{M}_+^1(\mathcal{X})$ y $\nu \in \mathcal{M}_+^1(\mathcal{Y})$ medidas de probabilidad sobre $\mathcal{X} \subset \mathbb{R}^n$, $\mathcal{Y} \subset \mathbb{R}^m$. El problema de transporte óptimo entrópico entre estas medidas es el siguiente:

$$\inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y) + \epsilon \cdot D_{KL}(\pi \| \mu \otimes \nu). \quad (3.1.9)$$

Donde la divergencia $D_{KL}(\pi \| \mu \otimes \nu)$ es la integral⁴

$$\int_{\mathcal{X} \times \mathcal{Y}} \log\left(\frac{d\pi}{d(\mu \otimes \nu)}(x, y)\right) d\pi(x, y).$$

Por otra parte, la medida de referencia $\mu \otimes \nu$ usada en la divergencia, tanto en el caso discreto como en el caso continuo, solo es importante para indicar el soporte de integración. En efecto, la solución no cambia si se utiliza otra medida producto como referencia, siempre y cuando comparten el soporte:

Proposición 3.4. Sea $\pi \in \Pi(\mu, \nu)$ y $\mu' \in \mathcal{M}_+^1(\mathcal{X})$, $\nu' \in \mathcal{M}_+^1(\mathcal{Y})$ otras medidas de probabilidad tales que $\text{Supp}(\mu') = \text{Supp}(\mu)$ y $\text{Supp}(\nu') = \text{Supp}(\nu)$ ⁵, entonces:

$$D_{KL}(\pi \| \mu \otimes \nu) = D_{KL}(\pi \| \mu' \otimes \nu') - D_{KL}(\mu \otimes \nu \| \mu' \otimes \nu'),$$

con $D_{KL}(\mu \otimes \nu \| \mu' \otimes \nu') = D_{KL}(\mu \| \mu') + D_{KL}(\nu \| \nu')$ una constante independiente de π . En particular, el problema de optimización (3.1.9) no cambia su solución si se utiliza $D_{KL}(\pi \| \mu' \otimes \nu')$ en vez de $D_{KL}(\pi \| \mu \otimes \nu)$.

Demostración. Por simplicidad, se probará la proposición en el caso discreto. Para esto, sean $\mu, \mu' \in \mathcal{M}_+^1(\mathcal{X})$ dos medidas con el mismo soporte $\mathcal{X} = \{x_i\}_{i=1}^n$ y vectores de probabilidad $a, a' \in \Sigma_n$ respectivamente, y sean $\nu, \nu' \in \mathcal{M}_+^1(\mathcal{Y})$ otras dos medidas con el mismo soporte $\mathcal{Y} = \{y_j\}_{j=1}^m$ y vectores de probabilidad $b, b' \in \Sigma_m$ respectivamente. Entonces, para una medida $\pi \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})$ con matriz de probabilidad $P \in \Pi_d(\mu, \nu)$ se tiene que:

⁴Notar que si $\pi \in \Pi(\mu, \nu)$, entonces $\pi \ll \mu \otimes \nu$.

⁵Notar que esto implica que $\mu \sim \mu'$ y $\nu \sim \nu'$, donde la relación de equivalencia $\mu_1 \sim \mu_2$ significa que $\mu_1 \ll \mu_2$ y $\mu_2 \ll \mu_1$.

$$\begin{aligned}
 D_{\text{KL}}(\pi \| \mu \otimes \nu) &= \sum_{i=1}^n \sum_{j=1}^m \log \left(\frac{P_{ij}}{a_i b_j} \right) P_{ij} \\
 &= \sum_{i=1}^n \sum_{j=1}^m \left[\log \left(\frac{P_{ij}}{a'_i b'_j} \right) + \log \left(\frac{a'_i b'_j}{a_i b_j} \right) \right] P_{ij} \\
 &= \sum_{i=1}^n \sum_{j=1}^m \log \left(\frac{P_{ij}}{a'_i b'_j} \right) P_{ij} + \sum_{i=1}^n \sum_{j=1}^m \left[\log \left(\frac{a'_i}{a_i} \right) P_{ij} + \log \left(\frac{b'_j}{b_j} \right) P_{ij} \right] \\
 &= D_{\text{KL}}(\pi \| \mu' \otimes \nu') + \sum_{i=1}^n a_i \log \left(\frac{a'_i}{a_i} \right) + \sum_{j=1}^m b_j \log \left(\frac{b'_j}{b_j} \right) \\
 &= D_{\text{KL}}(\pi \| \mu' \otimes \nu') - D_{\text{KL}}(\mu \| \mu') - D_{\text{KL}}(\nu \| \nu'),
 \end{aligned}$$

donde en la penúltima igualdad se usó que $\sum_{j=1}^m P_{ij} = a_i$ y $\sum_{i=1}^n P_{ij} = b_j$. Además, las dos divergencias individuales se pueden unir en una única divergencia producto:

$$\begin{aligned}
 D_{\text{KL}}(\mu \| \mu') + D_{\text{KL}}(\nu \| \nu') &= \sum_{i=1}^n \sum_{j=1}^m a_i b_j \left[\log \left(\frac{a_i}{a'_i} \right) + \log \left(\frac{b_j}{b'_j} \right) \right] \\
 &= \sum_{i=1}^n \sum_{j=1}^m a_i b_j \log \left(\frac{a_i b_j}{a'_i b'_j} \right) \\
 &= D_{\text{KL}}(\mu \otimes \nu \| \mu' \otimes \nu'),
 \end{aligned}$$

donde en la primera igualdad se usó que $\sum_{i=1}^n a_i = \sum_{j=1}^m b_j = 1$ ya que μ y ν son medidas de probabilidad. \square

Por otra parte, esta formulación continua también hereda la propiedad vista en la Proposición 3.3, donde para valores muy grandes de ϵ , el plan de transporte entrópico se asemeja a la distribución conjunta de dos variables aleatorias independientes con leyes μ y ν respectivamente. Esto puede ser observado en la Figura 3.4.

3.1.2. Problema entrópico dual

En esta subsección se estudiará el problema dual del problema de Kantorovich regularizado, tanto en su versión discreta como en su versión continua. En ambos casos, a diferencia de lo que ocurre en el problema de Kantorovich dual estudiado en la Subsección 2.2.2, aquí se podrá obtener la (única) solución del problema primal a partir de una solución del problema dual. Más aún, este problema permitirá desarrollar el algoritmo de Sinkhorn en la Subsección 3.2.2, el cual es el algoritmo estándar para resolver el problema del puente de Schrödinger.

Problema dual discreto

Como se mencionó en la subsección anterior, usar un regularizador en la función objetivo del problema entrópico (3.1.1) transforma el problema de optimización en uno estrictamente convexo, garantizando la unicidad de la solución. Sin embargo, al igual como ocurre en el problema de Kantorovich no regularizado, este problema sigue siendo costoso de resolver ya que tiene la misma cantidad de incógnitas que el problema de Kantorovich original (2.2.1), lo cual motiva a volver a estudiar la formulación dual del problema. Para el caso discreto, se tiene el siguiente resultado de dualidad:

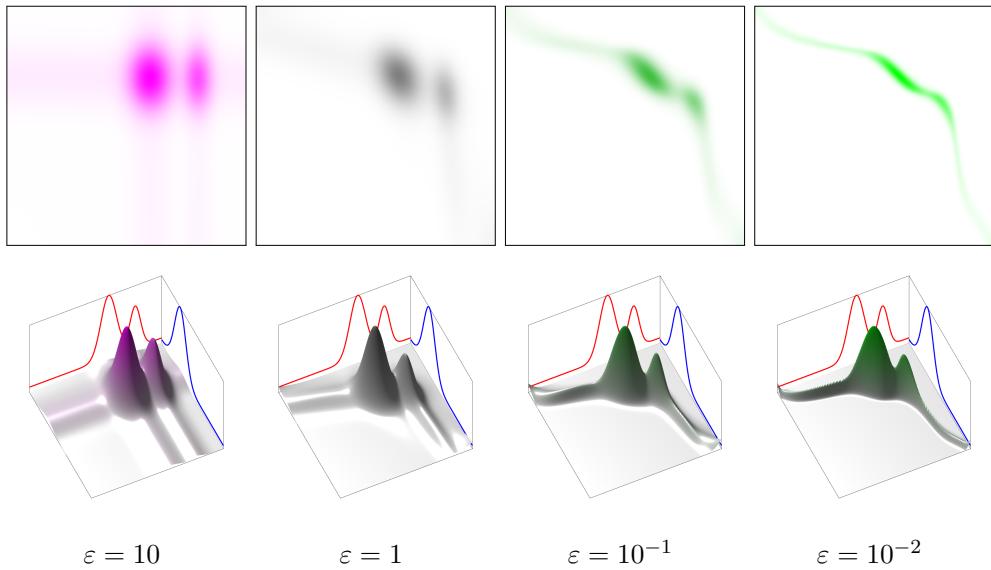


Figura 3.4: Evolución de la solución $\pi_\epsilon \in \Pi(\mu, \nu)$ para distintos valores de $\epsilon > 0$ en el problema entrópico con costo cuadrático entre dos medidas de probabilidad continuas $\mu, \nu \in \mathcal{M}_+^1(\mathbb{R})$. Imagen obtenida desde [PC20].

Proposición 3.5 (problema entrópico dual, caso discreto). El problema dual del problema entrópico discreto (3.1.5) es el problema de optimización irrestringido

$$\sup_{(\phi, \psi) \in \mathbb{R}^n \times \mathbb{R}^m} \langle C, P \rangle_F - \epsilon \cdot \mathcal{H}(P) = \sup_{(\phi, \psi) \in \mathbb{R}^n \times \mathbb{R}^m} \langle \phi, a \rangle + \langle \psi, b \rangle - \epsilon \sum_{i=1}^n \sum_{j=1}^m \exp\left(\frac{\phi_i + \psi_j - C_{ij} - \epsilon}{\epsilon}\right). \quad (3.1.10)$$

Además, hay dualidad fuerte:

$$\inf_{P \in \Pi_d(\mu, \nu)} \langle C, P \rangle_F - \epsilon \cdot \mathcal{H}(P) = \sup_{(\phi, \psi) \in \mathbb{R}^n \times \mathbb{R}^m} \langle \phi, a \rangle + \langle \psi, b \rangle - \epsilon \sum_{i=1}^n \sum_{j=1}^m \exp\left(\frac{\phi_i + \psi_j - C_{ij} - \epsilon}{\epsilon}\right).$$

Por otra parte, la (única) solución del problema primal (3.1.5) es

$$P_{ij}^* = \exp\left(\frac{\phi_i^* + \psi_j^* - C_{ij} - \epsilon}{\epsilon}\right),$$

donde $(\phi^*, \psi^*) \in \mathbb{R}^n \times \mathbb{R}^m$ son multiplicadores de Lagrange óptimos para el problema (3.1.12) (conocidos como potenciales entrópicos), los cuales no deben ser confundidos con los potenciales de Kantorovich en (2.2.7).

Demostración. El problema entrópico discreto (3.1.5) escrito con todas sus restricciones de forma explícita es el siguiente:

$$\begin{aligned}
 & \min_{P \in \mathcal{M}_{n,m}([0,1])} \sum_{i=1}^n \sum_{j=1}^m P_{ij} C_{ij} + \epsilon \sum_{i=1}^n \sum_{j=1}^m P_{ij} \log(P_{ij}) \\
 \text{s.a.: } & \sum_{j=1}^m P_{ij} = a_i, \quad \forall i \in \{1, \dots, n\}, \\
 & \sum_{i=1}^n P_{ij} = b_j, \quad \forall j \in \{1, \dots, m\}.
 \end{aligned}$$

Notar que este problema ya no es de programación lineal debido a los términos $P_{ij} \log(P_{ij})$ dentro de la función objetivo. Por lo tanto, el problema dual debe ser determinado utilizando dualidad lagrangiana.

Asociando un multiplicador de Lagrange $\phi \in \mathbb{R}^n$ a las restricciones asociadas a la marginal μ y un multiplicador $\psi \in \mathbb{R}^m$ a las restricciones asociadas a la marginal ν , el lagrangiano del problema entrópico discreto, $L : \mathcal{M}_{n,m}([0,1]) \times \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$, es:

$$L(P, \phi, \psi) = \sum_{i=1}^n \sum_{j=1}^m P_{ij} C_{ij} + \epsilon \sum_{i=1}^n \sum_{j=1}^m P_{ij} \log(P_{ij}) + \sum_{i=1}^n \phi_i \left(a_i - \sum_{j=1}^m P_{ij} \right) + \sum_{j=1}^m \psi_j \left(b_j - \sum_{i=1}^n P_{ij} \right),$$

y por lo tanto, para obtener la función dual $\theta : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$ definida por

$$\theta(\phi, \psi) = \min_{P \in \mathcal{M}_{n,m}([0,1])} L(P, \phi, \psi), \quad (3.1.11)$$

basta aplicar la condición de primer orden sobre $P \mapsto L(P, \alpha, \beta)$. De este modo, considerando a ϕ y ψ fijos, la solución óptima P^* para el problema (3.1.11) verifica:

$$\frac{\partial L}{\partial P_{ij}}(P^*, \phi, \psi) = C_{ij} + \epsilon(\log(P_{ij}^*) + 1) - \phi_i - \psi_j = 0 \implies P_{ij}^* = \exp\left(\frac{\phi_i + \psi_j - C_{ij} - \epsilon}{\epsilon}\right).$$

Luego, la función dual es:

$$\begin{aligned}
 \theta(\phi, \psi) &= L(P^*, \phi, \psi) \\
 &= \sum_{i=1}^n \sum_{j=1}^m P_{ij}^* (C_{ij} + \epsilon \cdot \log(P_{ij}^*) - \phi_i - \psi_j) + \sum_{i=1}^n \phi_i a_i + \sum_{j=1}^m \psi_j b_j \\
 &= \sum_{i=1}^n \sum_{j=1}^m P_{ij}^* (C_{ij} + [\phi_i + \psi_j - C_{ij} - \epsilon] - \phi_i - \psi_j) + \sum_{i=1}^n \phi_i a_i + \sum_{j=1}^m \psi_j b_j \\
 &= -\epsilon \sum_{i=1}^n \sum_{j=1}^m \exp\left(\frac{\phi_i + \psi_j - C_{ij} - \epsilon}{\epsilon}\right) + \sum_{i=1}^n \phi_i a_i + \sum_{j=1}^m \psi_j b_j.
 \end{aligned}$$

Concluyendo lo que se quería demostrar. \square

Es importante mencionar que este problema dual no posee solución única. En efecto, si $(\phi^*, \psi^*) \in \mathbb{R}^n \times \mathbb{R}^m$ es solución de (3.1.10), entonces $(\tilde{\phi}, \tilde{\psi}) = (\phi + \eta \mathbf{1}_n, \psi - \eta \mathbf{1}_m)$ con $\eta \in \mathbb{R}$ también es solución. En efecto, el término exponencial en (3.1.10) no cambia, mientras que para la otra parte:

$$\langle \phi + \eta \mathbb{1}_n, a \rangle + \langle \psi - \eta \mathbb{1}_m, b \rangle = \langle \phi, a \rangle + \eta \langle \mathbb{1}_n, a \rangle + \langle \psi, b \rangle - \eta \langle \mathbb{1}_m, b \rangle = \langle \phi, a \rangle + \langle \psi, b \rangle.$$

Donde se usó que $\langle \mathbb{1}_n, a \rangle = \langle \mathbb{1}_m, b \rangle = 1$ ya que $a \in \Sigma_n$ y $b \in \Sigma_m$ son vectores de probabilidad. Sin embargo, se puede demostrar que todas las soluciones son de este tipo, por lo que la solución al problema (3.1.10) es única salvo constantes aditivas.

Por otra parte, al comparar este problema dual con la formulación dual del problema de Kantorovich no regularizado (2.2.7) se observa que la regularización entrópica relaja la restricción $\phi_i + \psi_j \leq C_{ij}$ para los potenciales de Kantorovich (ϕ, ψ) en (2.2.7) agregándola a la función objetivo como un penalizador exponencial en el problema (3.1.10). Esta relajación permite pasar de trabajar una formulación dual con restricciones en el problema de Kantorovich a una irrestricta en el problema regularizado.

Por último, en la Subsección 3.2.2, se estudiará el algoritmo de Sinkhorn, el cual permite resolver este problema de manera eficiente, pudiendo computar una buena aproximación de la distancia de Wasserstein para problemas de transporte óptimo de alta dimensionalidad y a gran escala.

Problema dual continuo

Las propiedades demostradas para el problema dual discreto se siguen cumpliendo en el caso continuo. En particular, se sigue teniendo dualidad fuerte:

Proposición 3.6 (problema entrópico dual, caso continuo). El problema dual del problema entrópico continuo (3.1.9) es el problema de optimización irrestricto

$$\sup_{(\phi, \psi) \in \mathcal{C}_b(\mathcal{X}) \times \mathcal{C}_b(\mathcal{Y})} \int_{\mathcal{X}} \phi(x) d\mu(x) + \int_{\mathcal{Y}} \psi(y) d\nu(y) - \epsilon \int_{\mathcal{X} \times \mathcal{Y}} \exp\left(\frac{(\phi \oplus \psi)(x, y) - c(x, y) - \epsilon}{\epsilon}\right) d(\mu \otimes \nu)(x, y). \quad (3.1.12)$$

Además, hay dualidad fuerte:

$$\begin{aligned} & \sup_{(\phi, \psi) \in \mathcal{C}_b(\mathcal{X}) \times \mathcal{C}_b(\mathcal{Y})} \int_{\mathcal{X}} \phi(x) d\mu(x) + \int_{\mathcal{Y}} \psi(y) d\nu(y) - \epsilon \int_{\mathcal{X} \times \mathcal{Y}} \exp\left(\frac{(\phi \oplus \psi)(x, y) - c(x, y) - \epsilon}{\epsilon}\right) d(\mu \otimes \nu)(x, y) \\ &= \inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y) + \epsilon \cdot D_{KL}(\pi \| \mu \otimes \nu). \end{aligned}$$

Por otra parte, la (única) solución del problema primal (3.1.9) es

$$d\pi(x, y) = \exp\left(\frac{\phi^* \oplus \psi^* - c - \epsilon}{\epsilon}\right) (x, y) d\mu(x) d\nu(y),$$

donde $(\phi^*, \psi^*) \in \mathcal{C}_b(\mathcal{X}) \times \mathcal{C}_b(\mathcal{Y})$ son multiplicadores de Lagrange óptimos para el problema (3.1.12) (conocidos como potenciales entrópicos), los cuales no deben ser confundidos con los potenciales de Kantorovich en (2.2.9).

Al igual que para el problema dual entrópico discreto (3.1.10), las soluciones son únicas salvo constantes aditivas⁶. Además, esta formulación también se puede interpretar como una relajación de la restricción $\phi \oplus \psi \leq c$ del problema de Kantorovich dual (2.2.9).

⁶Sin embargo, es posible modificar los valores que toman las funciones ϕ^* y ψ^* fuera del soporte de $\mu \otimes \nu$, por lo que la unicidad es en el sentido de las medidas μ y ν .

Notar que el hecho de no tener restricciones en este problema dual permite escribir el problema (3.1.12) como la maximización de una esperanza:

$$\sup_{(\phi, \psi) \in \mathcal{C}_b(\mathcal{X}) \times \mathcal{C}_b(\mathcal{Y})} \mathbb{E}_{(x,y) \sim \mu \otimes \nu} \left[(\phi \oplus \psi)(x,y) - \epsilon \cdot \exp \left(\frac{(\phi \oplus \psi)(x,y) - c(x,y) - \epsilon}{\epsilon} \right) \right].$$

Por lo que es posible utilizar cualquier algoritmo de maximización de esperanzas para resolver el problema. Por otra parte, si bien el problema de Kantorovich dual (2.2.9) también se puede escribir de esta forma, la restricción $\phi \oplus \psi \leq c$ se suele introducir en la esperanza mediante otro problema de optimización, por lo que en la práctica no es posible maximizar esta reformulación de forma eficiente.

3.2. Problema de Schrödinger estático

Como se vio en la sección anterior, la regularización entrópica soluciona varias de las limitaciones presentes en el problema de Kantorovich, como suavizar el plan de transporte óptimo y garantizar su unicidad. En esta sección se verá que el problema de transporte entrópico es equivalente al problema de Schrödinger estático, lo cual permite heredar toda la teoría estudiada hasta el momento a este nuevo problema.

Para motivar el problema del puente de Schrödinger, se intentará resolver el problema de Kantorovich discreto (2.2.1) comenzando con un candidato a solución. Para esto, notar que un plan de transporte razonable entre μ y ν es aquel que evita transportar masa entre pares de puntos $(x, y) \in \mathcal{X} \times \mathcal{Y}$ cuyo costo de transporte C_{ij} sea elevado. Con esta idea, resulta natural proponer como plan de transporte al kernel de Gibbs, el cual se define de la siguiente forma:

Definición 3.6 (kernel de Gibbs). Dados dos conjuntos finitos $\mathcal{X} = \{x_i\}_{i=1}^n$ e $\mathcal{Y} = \{y_j\}_{j=1}^m$, una matriz de costo $C \in \mathcal{M}_{n,m}(\mathbb{R})$ y un parámetro de temperatura $T > 0$, se define el *kernel de Gibbs* $\mathcal{K} \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y})$ como la medida (no necesariamente de probabilidad) cuya matriz de masa es

$$\mathcal{K}_{ij} := \exp \left(-\frac{C_{ij}}{T} \right), \quad (3.2.1)$$

donde se ha abusado de notación para representar a la medida y la matriz de probabilidades con el mismo símbolo \mathcal{K} .

Es importante destacar la estrecha similitud entre el kernel de Gibbs y la distribución de Boltzmann definida en (1.2.5)⁷. Sin embargo, aunque ambas son medidas con la misma función de masa (salvo constante multiplicativa de normalización), el kernel de Gibbs es una medida en un espacio producto mientras que la distribución de Boltzmann es una medida de probabilidad en un espacio simple. Esta diferencia se debe a que la distribución de Boltzmann está pensada como una distribución sobre un espacio de estados mientras que el kernel de Gibbs está pensado como un kernel de transición entre dos estados dentro de un sistema dinámico. Por otro lado, al igual que para la distribución de Boltzmann, la función exponencial en (3.2.1) es usada para obtener una medida positiva mientras que el parámetro de temperatura T cumple un rol similar al parámetro de temperatura (inversa) γ usado para guiar la generación de los modelos de difusión en (1.3.14).

Desde la perspectiva del transporte óptimo, el kernel de Gibbs puede interpretarse como un pseudo plan de transporte razonable pero que no necesariamente respeta las distribuciones marginales μ y ν y, más aún, no es necesariamente una medida de probabilidad. Por lo tanto, para obtener un plan de transporte factible a partir del kernel de Gibbs es necesario proyectar (en algún sentido) la medida positiva $\mathcal{K} \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y})$ sobre

⁷Recordar que esta distribución permitió interpretar la maximización de la ELBO de un modelo de difusión (1.3.8) como la minimización de la energía libre de Helmholtz de un sistema termodinámico.

el conjunto factible $\Pi_d(\mu, \nu) \subset \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y}) \subset \mathcal{M}_+(\mathcal{X} \times \mathcal{Y})$. Dado que la medida de discrepancia estándar entre medidas es la divergencia de Kullback-Leibler⁸, resulta natural estudiar el siguiente problema:

$$\arg \min_{P \in \Pi_d(\mu, \nu)} D_{\text{KL}}(P \| \mathcal{K}).$$

Sorprendentemente, la solución a este nuevo problema, si bien no es un plan de Kantorovich, sí es la solución al problema regularizado (3.1.5). Más aún, el parámetro de temperatura T en el kernel de Gibbs (3.2.1) es precisamente el ponderador de regularización:

$$\begin{aligned} D_{\text{KL}}(P \| \mathcal{K}) &= \sum_{i=1}^n \sum_{j=1}^m \log \left(\frac{P_{ij}}{\mathcal{K}_{ij}} \right) P_{ij} \\ &= \sum_{i=1}^n \sum_{j=1}^m \left(\log(P_{ij}) + \frac{C_{ij}}{T} \right) P_{ij} \\ &= \frac{1}{T} \left(T \cdot \log(P_{ij}) P_{ij} + \sum_{i=1}^n \sum_{j=1}^m C_{ij} P_{ij} \right) \\ &= \frac{1}{T} (-T \cdot \mathcal{H}(P) + \langle C, P \rangle_F). \end{aligned}$$

En consecuencia:

$$\arg \min_{P \in \Pi_d(\mu, \nu)} D_{\text{KL}}(P \| \mathcal{K}) = \arg \min_{P \in \Pi_d(\mu, \nu)} \langle C, P \rangle_F - \epsilon \mathcal{H}(P). \quad (3.2.2)$$

El lado izquierdo de (3.2.2) es un caso particular de lo que se conoce como el *problema del puente de Schrödinger* (SBP), el cual consiste en buscar una medida lo más similar a otra medida de referencia (en este caso, \mathcal{K}) que tenga distribuciones marginales específicas:

Definición 3.7 (problema del puente de Schrödinger, versión estática). Sean \mathcal{X} e \mathcal{Y} dos espacios medibles (discretos o continuos). Entonces, el puente de Schrödinger entre las medidas $\mu \in \mathcal{M}_+^1(\mathcal{X})$ y $\nu \in \mathcal{M}_+^1(\mathcal{Y})$ dada una medida de referencia $R \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})$ es la (única) solución del problema

$$\arg \min_{\pi \in \Pi(\mu, \nu)} D_{\text{KL}}(\pi \| R). \quad (3.2.3)$$

Dada la igualdad en (3.2.2), se observa que el problema de Kantorovich entrópico, al menos en su formulación discreta, es un caso particular del problema del puente de Schrödinger cuando se considera $R = \mathcal{K}$. Sin embargo, en la siguiente subsección se verá que ambos problemas son realmente equivalentes, por lo que todo lo estudiado hasta el momento se hereda directamente al problema (estático) de Schrödinger.

Posteriormente, en la Sección 3.3, se revisará la formulación dinámica de este problema, la cual resultará ser una versión estocástica de la formulación dinámica estudiada en la Sección 2.3. Más aún, esta formulación dinámica podrá verse como una generalización directa de los modelos de difusión estudiados en el Capítulo 1.

Por otra parte, es importante mencionar que la divergencia de Kullback-Leibler que se busca minimizar en (3.2.3) induce una noción de convergencia en $\mathcal{M}_+^1(\mathcal{X})$ de forma similar a la que induce el problema de Kantorovich mediante la distancia de Wasserstein. Para ver esto, se tiene el siguiente resultado previo:

⁸Si bien la divergencia de Kullback-Leibler se definió para comparar distribuciones de probabilidad, su definición se puede extender a medidas arbitrarias (no necesariamente acotadas). En [Léo13] extienden esta definición con el rigor necesario.

Proposición 3.7. Dada una medida de probabilidad $\nu \in \mathcal{M}_+^1(\mathcal{X})$, el operador $\mu \in \mathcal{M}_+^1(\mathcal{X}) \mapsto D_{\text{KL}}(\mu \| \nu)$ es convexo y no negativo. Más aún, es estrictamente convexo en el subconjunto donde $D_{\text{KL}}(\mu \| \nu)$ es finito y es nulo si y solo si $\mu = \nu$.

Demostración. Notando que $d\mu = \frac{d\mu}{d\nu} d\nu$ para $\mu \ll \nu$, la entropía relativa se puede escribir como

$$D_{\text{KL}}(\mu \| \nu) = \int_{\mathcal{X}} \log \left(\frac{d\mu}{d\nu}(z) \right) \frac{d\mu}{d\nu}(z) d\nu(z) = \mathbb{E}_{z \sim \nu} \left[h \left(\frac{d\mu}{d\nu}(z) \right) \right],$$

donde la función $h : [0, \infty] \rightarrow [-e^{-1}, \infty]$ definida como $h(z) = z \log z$ es convexa y estrictamente convexa en el subconjunto donde es finita. Por lo tanto, como la esperanza es un operador lineal, se tiene la convexidad de $D_{\text{KL}}(\cdot \| \nu)$.

Para la positividad basta utilizar la desigualdad de Jensen:

$$D_{\text{KL}}(\mu \| \nu) = \mathbb{E}_{z \sim \nu} \left[h \left(\frac{d\mu}{d\nu}(z) \right) \right] \geq h \left(\mathbb{E}_{z \sim \nu} \left[\frac{d\mu}{d\nu}(z) \right] \right) = h(\mathbb{E}_{z \sim \mu}[1]) = h(1) = 0.$$

Para la reflexividad, notar que:

$$D_{\text{KL}}(\mu \| \nu) = 0 \iff \frac{d\mu}{d\nu} = 1 \iff \mu = \nu.$$

Lo que concluye la demostración. □

En consecuencia, si bien la divergencia de Kullback-Leibler no es una distancia como sí lo es la distancia de Wasserstein, la propiedad $D_{\text{KL}}(\mu \| \nu) = 0 \iff \mu = \nu$ motiva a definir la siguiente noción de convergencia de medidas:

Definición 3.8 (convergencia en entropía relativa). Una secuencia de medidas de probabilidad $(\mu_n)_{n \in \mathbb{N}} \subset \mathcal{M}_+^1(\mathcal{X})$ converge en entropía relativa a $\mu \in \mathcal{M}_+^1(\mathcal{X})$ si

$$\lim_{n \rightarrow \infty} D_{\text{KL}}(\mu_n \| \mu) = 0.$$

Con esta nueva topología, resulta natural preguntarse qué relación tiene con las otras nociones de convergencia estudiadas en la Subsección 2.3.1. La siguiente desigualdad afirma que la convergencia en entropía relativa es más fuerte que la convergencia en variación total (ver Definición 2.17) y, en consecuencia, es más fuerte que la convergencia débil si \mathcal{X} es compacto:

Teorema 3.1 (desigualdad de Pinsker). Dadas dos medidas de probabilidad $\mu, \nu \in \mathcal{M}_+^1(\mathcal{X})$, se tiene que:

$$D_{\text{TV}}(\mu \| \nu) \leq \sqrt{2 D_{\text{KL}}(\mu \| \nu)}.$$

En particular, si una sucesión $(\mu_n)_{n \in \mathbb{N}} \subset \mathcal{M}_+^1(\mathcal{X})$ converge en entropía relativa a $\mu \in \mathcal{M}_+^1(\mathcal{X})$, también lo hará en variación total.

La demostración de esta propiedad se puede encontrar en [Nut22].

3.2.1. Equivalencia con el transporte óptimo entrópico

En esta subsección se verá la equivalencia entre el problema del puente de Schrödinger y el problema de transporte óptimo entrópico, donde la solución de un problema será solución del otro. Esta equivalencia es importante ya que indica que el problema del puente de Schrödinger, el cual se motivó únicamente como una forma de eludir los problemas de los modelos de difusión, tiene interpretaciones y propiedades importantes, las cuales se heredan directamente del transporte óptimo. Los detalles técnicos de esta equivalencia, los cuales son sutiles pero necesarios para mostrar la integrabilidad de las funciones involucradas, se pueden revisar en [Nut22].

Problema de Schrödinger como problema entrópico

Para transformar el problema del puente de Schrödinger en un problema de Kantorovich regularizado, se mostrará que un funcional de costo específico, dependiente de la medida de referencia R , permite transformar la divergencia de Kullback-Leibler en un funcional de minimización con la misma forma que el funcional del problema entrópico.

Proposición 3.8 (SBP como EOT). Dada una medida de referencia $R \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})$ con $R \ll \mu \otimes \nu$, entonces, toda solución del problema de Schrödinger (3.2.3) es solución del problema de transporte óptimo entrópico (3.1.9) con el funcional de costo $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ definido como

$$c(x, y) = -\epsilon \cdot \log \left(\frac{dR}{d(\mu \otimes \nu)}(x, y) \right).$$

Demostración. Notar que $dR(x, y) = e^{-\frac{c(x, y)}{\epsilon}} d(\mu \otimes \nu)(x, y)$, luego⁹:

$$\frac{d\pi}{dR}(x, y) = e^{\frac{c(x, y)}{\epsilon}} \frac{d\pi}{d(\mu \otimes \nu)}(x, y).$$

Sustituyendo esta función de densidad:

$$\begin{aligned} D_{KL}(\pi \| R) &= \int_{\mathcal{X} \times \mathcal{Y}} \log \left(\frac{d\pi}{dR}(x, y) \right) d\pi(x, y) \\ &= \int_{\mathcal{X} \times \mathcal{Y}} \log \left(e^{\frac{c(x, y)}{\epsilon}} \frac{d\pi}{d(\mu \otimes \nu)}(x, y) \right) d\pi(x, y) \\ &= \int_{\mathcal{X} \times \mathcal{Y}} \frac{c(x, y)}{\epsilon} d\pi(x, y) + \int_{\mathcal{X} \times \mathcal{Y}} \log \left(\frac{d\pi}{d(\mu \otimes \nu)}(x, y) \right) d\pi(x, y) \\ &= \frac{1}{\epsilon} \left(\int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y) + \epsilon \cdot D_{KL}(\pi \| \mu \otimes \nu) \right). \end{aligned}$$

En consecuencia, ambos funcionales de optimización son equivalentes, salvo constantes positivas de multiplicación.

□

Esta interpretación del problema de Schrödinger como problema de transporte óptimo es la que permite heredar la teoría estudiada a este nuevo problema, otorgándole una robustez que actualmente no se sabe si poseen los modelos de difusión estudiados en el Capítulo 1.

⁹La notación diferencial indica que $R(C) = \int_C e^{-\frac{c}{\epsilon}} d(\mu \otimes \nu)$ para $C \in \mathcal{B}(\mathcal{X} \times \mathcal{Y})$.

Problema entrópico como problema de Schrödinger

De forma análoga a lo desarrollado en el resultado anterior, se verá que una medida de referencia específica, dependiente del funcional de costo c , permite transformar el funcional de minimización del problema entrópico en una divergencia de Kullback-Leibler.

Proposición 3.9 (EOT como SBP). Dado un funcional de costo $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$, toda solución del problema de transporte entrópico (3.1.9) es solución del problema de Schrödinger (3.2.3) cuando se considera la medida de referencia $R \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})$ dada por $dR \propto e^{-\frac{c(x,y)}{\epsilon}} d(\mu \otimes \nu)$, es decir:

$$R(C) = \frac{1}{a} \int_C e^{-\frac{c(x,y)}{\epsilon}} d(\mu \otimes \nu)(x, y), \quad \forall C \in \mathcal{B}(\mathcal{X} \times \mathcal{Y}),$$

donde $a = \int_{\mathcal{X} \times \mathcal{Y}} e^{-\frac{c(x,y)}{\epsilon}} d(\mu \otimes \nu)(x, y)$ es la constante de normalización.

Demostración. Basta reordenar los términos para formar la medida de referencia R . Denotando por $a = \int_{\mathcal{X} \times \mathcal{Y}} e^{-\frac{c(x,y)}{\epsilon}} d(\mu \otimes \nu)(x, y)$ la constante de normalización:

$$\begin{aligned} & \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y) + \epsilon \cdot D_{KL}(\pi \| \mu \otimes \nu) \\ &= \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\pi(x, y) + \epsilon \int_{\mathcal{X} \times \mathcal{Y}} \log \left(\frac{d\pi}{d(\mu \otimes \nu)}(x, y) \right) d\pi(x, y) \\ &= \epsilon \int_{\mathcal{X} \times \mathcal{Y}} \left[\log \left(e^{\frac{c(x,y)}{\epsilon}} \right) + \log \left(\frac{d\pi}{d(\mu \otimes \nu)}(x, y) \right) \right] d\pi(x, y) \\ &= \epsilon \int_{\mathcal{X} \times \mathcal{Y}} \log \left(\frac{1}{e^{-\frac{c(x,y)}{\epsilon}}} \cdot \frac{d\pi}{d(\mu \otimes \nu)}(x, y) \right) d\pi(x, y) \\ &= \epsilon \int_{\mathcal{X} \times \mathcal{Y}} \log \left(\frac{1}{\frac{1}{a} e^{-\frac{c(x,y)}{\epsilon}}} \cdot \frac{d\pi}{d(\mu \otimes \nu)}(x, y) \right) d\pi(x, y) - \epsilon \int_{\mathcal{X} \times \mathcal{Y}} \log(a) d\pi(x, y) \\ &= \epsilon \int_{\mathcal{X} \times \mathcal{Y}} \log \left(\frac{d\pi}{dR}(x, y) \right) d\pi(x, y) - \epsilon \cdot \log(a) \\ &= \epsilon \cdot D_{KL}(\pi \| R) - \epsilon \cdot \log(a). \end{aligned}$$

En consecuencia, ambos funcionales de optimización son equivalentes, salvo constantes de multiplicación y traslación.

□

Es importante destacar que, si bien la medida de referencia en la Proposición 3.9 parece poco natural, se observa que al considerar el caso usual $c(x, y) = \frac{1}{2} \|x - y\|^2$, R se reduce a una distribución gaussiana. Este resultado será claro en la Sección 3.3, donde se mostrará que el problema de Schrödinger con referencia browniana (lo que se podría considerar como el caso estándar de este problema) es equivalente a realizar transporte óptimo con costo cuadrático.

3.2.2. Algoritmo de Sinkhorn

En esta subsección se estudiará el algoritmo de Sinkhorn, también conocido como IPFP (*iterative proportional fitting procedure*), el cual es un método eficiente para resolver el problema del puente de Schrödinger estático o, equivalentemente, el problema de transporte entrópico. Para esto, es necesario recordar que una solución

$(\phi^*, \psi^*) \in \mathbb{R}^n \times \mathbb{R}^m$ del problema entrópico dual (3.1.10) permite obtener una solución $P^* \in \Pi_d(\mu, \nu)$ del problema entrópico primal (3.1.5), de la forma

$$P_{ij}^* = \exp\left(\frac{\phi_i^* + \psi_j^* - C_{ij} - \epsilon}{\epsilon}\right) = \underbrace{\exp\left(\frac{\phi_i^* - \frac{\epsilon}{2}}{\epsilon}\right)}_{u_i^*} \exp\left(-\frac{C_{ij}}{\epsilon}\right) \underbrace{\exp\left(\frac{\psi_j^* - \frac{\epsilon}{2}}{\epsilon}\right)}_{v_j^*}. \quad (3.2.4)$$

Es decir, la solución óptima para el problema entrópico, $P^* \in \mathcal{M}_{n,m}([0, 1])$, se puede factorizar usando dos vectores positivos $u^* \in \mathbb{R}_{++}^n$, $v^* \in \mathbb{R}_{++}^m$ y (la matriz asociada a) el Kernel de Gibbs con $T = \epsilon$ mediante

$$P_{ij}^* = u_i^* \mathcal{K}_{ij} v_j^* \implies P^* = \text{diag}(u^*) \mathcal{K} \text{diag}(v^*).$$

Por lo tanto, el problema entrópico (3.1.5) se reduce a buscar vectores $(u^*, v^*) \in \mathbb{R}_{++}^n \times \mathbb{R}_{++}^m$ tales que $P^* = \text{diag}(u^*) \mathcal{K} \text{diag}(v^*) \in \Pi_d(\mu, \nu)$. Notar que estos vectores son únicos salvo constantes multiplicativas positivas. En efecto, si (u^*, v^*) es solución, entonces $(ru^*, \frac{1}{r}v^*)$ con $r > 0$ también es solución¹⁰. Esto indica que solo es importante la dirección de estos vectores y no su magnitud, lo cual motivará a identificar todos los múltiplos de un vector con un único representante en la Definición 3.9.

Para resolver este problema eficientemente, es conveniente recordar de (2.2.2) que para $P \in \mathcal{M}_{n,m}([0, 1])$, P está en $\Pi_d(\mu, \nu)$ si y solo si $P\mathbf{1}_m = a$ y $P^\top\mathbf{1}_n = b$, es decir, el par (u^*, v^*) debe cumplir que

$$\text{diag}(u^*) \underbrace{\mathcal{K} \text{diag}(v^*) \mathbf{1}_m}_{v^*} = u^* \odot (\mathcal{K}v^*) = a \quad \text{diag}(v^*) \underbrace{\mathcal{K}^\top \text{diag}(u^*) \mathbf{1}_n}_{u^*} = v^* \odot (\mathcal{K}^\top u^*) = b,$$

donde \odot es el producto de Hadamard (coordenada a coordenada). Para encontrar vectores $u^* \in \mathbb{R}_{++}^n$, $v^* \in \mathbb{R}_{++}^m$ que cumplan estas condiciones, el algoritmo de Sinkhorn comienza fijando $v^{(0)} = \mathbf{1}_m$ y luego actualiza los vectores mediante iteraciones que fuerzan que se cumplan cada una de las condiciones de forma individual:

$$u^{(l+1)} = a \oslash (\mathcal{K}v^{(l)}) \quad (3.2.5)$$

$$v^{(l+1)} = b \oslash (\mathcal{K}^\top u^{(l+1)}), \quad (3.2.6)$$

donde \oslash es la división de Hadamard (coordenada a coordenada). Estas iteraciones permiten construir una aproximación $P^{(l)}$ de P^* para cada $l \geq 1$ mediante

$$P^{(l)} = \text{diag}(u^{(l)}) \mathcal{K} \text{diag}(v^{(l)}).$$

Si bien este algoritmo va actualizando los potenciales u y v forzando una restricción a la vez, en el Teorema 3.3 se prueba la convergencia a la solución real del problema entrópico (3.1.7). Además, esta convergencia entrega un criterio de parada natural al observar la diferencia entre las medidas μ y ν , y las distribuciones marginales de P . En el Algoritmo 8 se detalla este procedimiento, mientras que en la Figura 3.5 se observan soluciones obtenidas para el problema entrópico discreto utilizando este procedimiento.

Por último, es importante mencionar que las inicializaciones $v^{(0)} = \mathbf{1}_m$ y $P^{(0)} = \mathcal{K}$ se puede sustituir por otras, pudiendo cambiando únicamente los valores finales de los vectores $u^{(l)}$ y $v^{(l)}$ pero no la convergencia de $P^{(l)}$ a P^* . Además, este algoritmo se puede extender a su versión continua, donde el kernel de Gibbs toma la forma

¹⁰Recordar que los vectores (u^*, v^*) se definen a partir de los potenciales (ϕ^*, ψ^*) que resuelven el problema entrópico (3.1.5), los cuales son únicos salvo constantes aditivas.

Algoritmo 8 Algoritmo de Sinkhorn

Require: Vectores de probabilidad $a \in \Sigma_n$, $b \in \Sigma_m$.
Require: Matriz de costos $C \in \mathcal{M}_{n,m}(\mathbb{R})$, parámetro de regularización $\epsilon > 0$.
Require: Tolerancia de error $\text{tol} > 0$.

- 1: Calcular $\mathcal{K} = \exp(-C/\epsilon)$ (elemento a elemento).
- 2: Inicializar $v^{(0)} = \mathbf{1}_m$ y $P^{(0)} = \mathcal{K}$.
- 3: $l \leftarrow 0$
- 4: **while** $\|P^{(l)}\mathbf{1}_m - a\|_1 + \|(P^{(l)})^\top \mathbf{1}_n - b\|_1 > \text{tol}$ **do**
- 5: $u^{(l+1)} \leftarrow a \oslash (\mathcal{K}v^{(l)})$
- 6: $v^{(l+1)} \leftarrow b \oslash (\mathcal{K}^\top u^{(l+1)})$
- 7: $P^{(l+1)} \leftarrow \text{diag}(u^{(l+1)}) \mathcal{K} \text{diag}(v^{(l+1)})$
- 8: $l \leftarrow l + 1$
- 9: **end while**
- 10: **return** $P^{(l+1)}$

$$d\mathcal{K}(x, y) = \exp\left(-\frac{c(x, y)}{\epsilon}\right) d(\mu \otimes \nu)(x, y).$$

Esta extensión se puede revisar en [Nut22]. Por otra parte, el Algoritmo 8 puede ser modificado para resolver el problema del baricentro de Wasserstein (2.3.4) cuando se aplica una regularización entrópica. En este caso, el problema del baricentro regularizado toma la forma

$$\arg \min_{\substack{a_B \in \Sigma_n \\ (P_i)_{i=1}^k \subset \mathcal{M}_{n,n}([0,1])}} \sum_{i=1}^k \lambda_i D_{\text{KL}}(P_i \| \mathcal{K}) \quad \text{sujeto a} \quad P_i \mathbf{1}_n = a_i, P_i^\top \mathbf{1}_n = a_B, \forall i \in \{1, \dots, k\},$$

donde $\mathcal{K} = \exp\left(-\frac{\epsilon}{c}\right)$ es el kernel de Gibbs.

Convergencia bajo la métrica de Hilbert

En este apartado se estudiará la velocidad de convergencia del algoritmo de Sinkhorn a la solución óptima P^* del problema entrópico discreto. Para esto, será necesario definir un espacio adecuado donde estudiar la noción de cercanía entre los elementos involucrados en el algoritmo de Sinkhorn. Dado que todas las soluciones del problema son de la forma $(ru^*, \frac{1}{r}v^*)$ con $r > 0$ y (u^*, v^*) una solución arbitraria, resulta útil la siguiente definición, la cual permite tratar a todos los vectores con la misma dirección como iguales:

Definición 3.9 (cono proyectivo real). Dados dos vectores positivos $u, v \in \mathbb{R}_{++}^d$, se dirá que $u \sim v$ si existe un $r > 0$ tal que $u = rv$. Con esto, se define el *cono proyectivo real* como el espacio cociente inducido por esta relación de equivalencia:

$$\mathbb{R}_{++}^d / \sim = \{[u] : u \in \mathbb{R}_{++}^d\},$$

donde $[u] = \{v \in \mathbb{R}_{++}^d : u \sim v\}$ es la clase de equivalencia de $u \in \mathbb{R}_{++}^d$, la cual representa el rayo con la dirección de u en el ortante positivo.

En este espacio, es usual definir la métrica proyectiva de Hilbert, la cual será usada para medir la cercanía entre las iteraciones del algoritmo de Sinkhorn, $(u^{(l)}, v^{(l)})$ y la solución (u^*, v^*) , la cual es única vista como un elemento de \mathbb{R}_{++}^d / \sim . Por simplicidad, se denotarán los elementos de \mathbb{R}_{++}^d / \sim como vectores usuales en \mathbb{R}_{++}^d , teniendo siempre presente que solo importa la dirección de los vectores y no su magnitud.

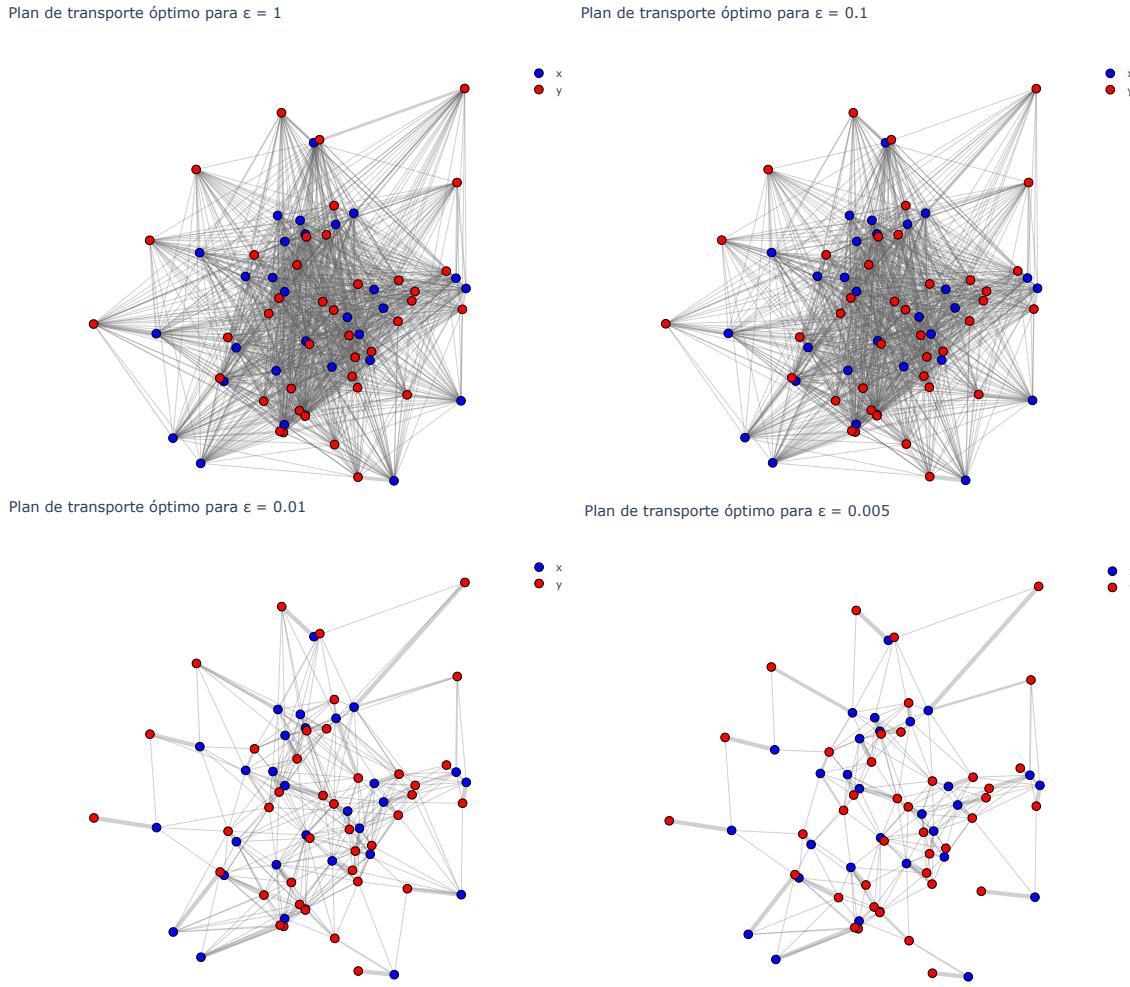


Figura 3.5: Soluciones para el problema entrópico utilizando el Algoritmo 8. El grueso de los arcos es proporcional a la cantidad de masa transferida según el plan de transporte. Se observa como el plan de transporte entrópico se va volviendo determinista a medida que disminuye ϵ . Esta simulación se encuentra en el archivo `sinkhorn.ipynb`.

Proposición 3.10 (métrica de Hilbert). Dados dos vectores $u, v \in \mathbb{R}_{++}^d$, se define la *métrica proyectiva de Hilbert* como

$$d_{\mathcal{H}}(u, v) = \log \left(\max_{1 \leq i, j \leq d} \frac{u_i v_j}{v_i u_j} \right). \quad (3.2.7)$$

En \mathbb{R}_{++}^d / \sim , $d_{\mathcal{H}}(\cdot, \cdot)$ define una métrica y, más aún, esta métrica es completa¹¹.

Notar que $d_{\mathcal{H}}(\cdot, \cdot)$ está bien definida para elementos de \mathbb{R}_{++}^d / \sim . En efecto, si $p \in [u]$ y $q \in [v]$ (i.e., existen $r, s > 0$ tales que $p = ru$ y $q = sv$), entonces:

$$d_{\mathcal{H}}(p, q) = \log \left(\max_{1 \leq i, j \leq d} \frac{(ru_i)(sv_j)}{(sv_i)(ru_j)} \right) = \log \left(\max_{1 \leq i, j \leq d} \frac{u_i v_j}{v_i u_j} \right) = d_{\mathcal{H}}(u, v).$$

¹¹Si bien esta es una propiedad técnica, es esencial para usar resultados como el teorema del punto fijo de Banach, el cual permite concluir la convergencia del algoritmo de Sinkhorn.

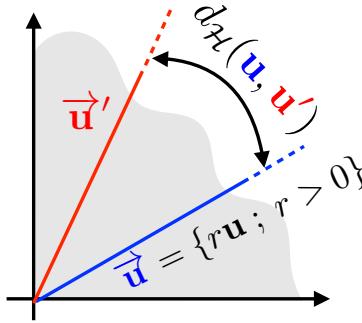


Figura 3.6: Distancia de Hilbert entre dos rayos $u, u' \in \mathbb{R}_{++}/\sim$, donde la distancia depende únicamente del ángulo entre ambos rayos. Imagen obtenida desde [PC20].

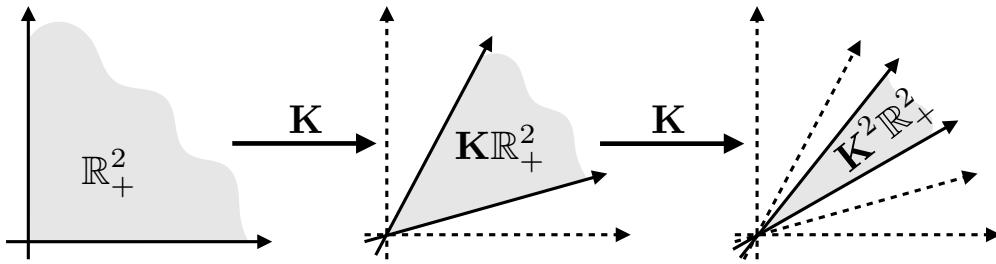


Figura 3.7: Contracción del primer cuadrante en \mathbb{R}^2 provocada por la multiplicación consecutiva de una matriz positiva K . Se observa que a medida que se va multiplicando por K , los rayos están cada vez más cerca, convergiendo a un único rayo en el cuadrante positivo. Imagen obtenida desde [PC20].

Por lo que la distancia no depende del representante de $[u]$ o $[v]$. En la Figura 3.6 se puede observar la distancia entre dos elementos del cono proyectivo \mathbb{R}^d_{++}/\sim .

El siguiente resultado afirma que si M es una matriz de entradas positivas, la distancia entre los rayos Mu y Mv es menor que la distancia entre los rayos u y v . Es decir, al multiplicar rayos por matrices positivas, estos se vuelven cada vez más cercanos. Esto puede verse en la Figura 3.7.

Teorema 3.2 (contracciones en \mathbb{R}^d_{++}/\sim). Dada una matriz con entradas positivas, $M \in \mathcal{M}_{n,m}(\mathbb{R}_{++})$, entonces el operador $x \in \mathbb{R}^d_{++}/\sim \mapsto Mx \in \mathbb{R}^d_{++}/\sim$ es una contracción, es decir:

$$d_H(Mu, Mv) \leq \lambda(M) d_H(u, v) \quad \forall u, v \in \mathbb{R}^d_{++}/\sim,$$

donde $\lambda(M) < 1$ es una constante que depende únicamente de M .

Este resultado permite mostrar que las iteraciones $(u^{(l)}, v^{(l)})$ se van acercando cada vez más a la solución real (u^*, v^*) , la cual es única salvo ponderaciones positivas (i.e., es única vista como elemento de \mathbb{R}^n_{++}/\sim y \mathbb{R}^m_{++}/\sim respectivamente). En efecto, de (3.2.5) se tiene que:

$$u^{(l+1)} = a \odot (\mathcal{K}v^{(l)}) \quad \text{y} \quad u^* = a \odot (\mathcal{K}v^*).$$

Luego, notando que la distancia en (3.2.7) no cambia si las coordenadas de u y v se ponderan por un mismo vector, se tiene lo siguiente:

$$\begin{aligned}
 d_{\mathcal{H}}(u^{(l+1)}, u^*) &= d_{\mathcal{H}}(a \oslash (\mathcal{K}v^{(l)}), a \oslash (\mathcal{K}v^*)) \\
 &= d_{\mathcal{H}}(\mathbb{1}_n \oslash (\mathcal{K}v^{(l)}), \mathbb{1}_n \oslash (\mathcal{K}v^*)) \\
 &= d_{\mathcal{H}}((\mathcal{K}v^*) \oslash (\mathcal{K}v^{(l)}), \mathbb{1}_n) \\
 &= d_{\mathcal{H}}(\mathcal{K}v^*, \mathcal{K}v^{(l)}) \\
 &\leq \lambda(\mathcal{K}) d_{\mathcal{H}}(v^*, v^{(l)}).
 \end{aligned}$$

Del mismo modo, de (3.2.6) se obtiene que $d_{\mathcal{H}}(v^{(l+1)}, v^*) \leq \lambda(\mathcal{K}) d_{\mathcal{H}}(u^*, u^{(l+1)})$. Por lo tanto, sustituyendo esta expresión y por inducción se concluye que:

$$d_{\mathcal{H}}(u^{(l)}, u^*) \leq \lambda(\mathcal{K})^2 d_{\mathcal{H}}(u^{(l-1)}, u^*) \leq \lambda(\mathcal{K})^{2l} d_{\mathcal{H}}(u^{(0)}, u^*).$$

De forma análoga se demuestra que $d_{\mathcal{H}}(v^{(l)}, v^*) \leq \lambda(\mathcal{K})^{2l} d_{\mathcal{H}}(v^{(0)}, v^*)$.

Por otra parte, estas desigualdades permiten acotar la distancia a las soluciones óptimas u^* y v^* mediante la discrepancia entre las distibuciones marginales de $P^{(l)}$ y las marginales esperadas μ y ν . En efecto, usando la desigualdad triangular:

$$d_{\mathcal{H}}(u^{(l)}, u^*) \leq d_{\mathcal{H}}(u^{(l)}, u^{(l+1)}) + d_{\mathcal{H}}(u^{(l+1)}, u^*) \leq d_{\mathcal{H}}(u^{(l)}, a \oslash (\mathcal{K}v^{(l)})) + \lambda(\mathcal{K})^2 d_{\mathcal{H}}(u^{(l)}, u^*).$$

Por lo tanto:

$$d_{\mathcal{H}}(u^{(l)}, u^*) \leq \frac{d_{\mathcal{H}}(u^{(l)}, a \oslash (\mathcal{K}v^{(l)}))}{1 - \lambda(\mathcal{K})^2} = \frac{d_{\mathcal{H}}(u^{(l)} \odot (\mathcal{K}v^{(l)}), a)}{1 - \lambda(\mathcal{K})^2} = \frac{d_{\mathcal{H}}(P^{(l)}\mathbb{1}_m, a)}{1 - \lambda(\mathcal{K})^2}.$$

Mientras que para la segunda marginal se demuestra de forma equivalente que

$$d_{\mathcal{H}}(v^{(l)}, v^*) \leq \frac{d_{\mathcal{H}}((P^{(l)})^\top \mathbb{1}_n, b)}{1 - \lambda(\mathcal{K})^2}.$$

Estas propiedades se resumen en el siguiente teorema, el cual, además, incluye una cota para la distancia uniforme entre P^* y $P^{(l)}$:

Teorema 3.3 (convergencia del algoritmo de Sinkhorn). Dadas las iteraciones (3.2.5) y (3.2.6) del algoritmo de Sinkhorn, entonces:

$$(u^{(l)}, v^{(l)}) \rightarrow (u^*, v^*) \quad \text{en } \mathbb{R}_{++}^d / \sim.$$

Además, si P^* es la (única) solución del problema entrópico (3.1.7) y $P^{(l)} = \text{diag}(u^{(l)}) \mathcal{K} \text{diag}(v^{(l)})$ es su aproximación dada por el Algoritmo 8, entonces:

$$\|\log P^{(l)} - \log P^*\|_\infty \leq d_{\mathcal{H}}(u^{(l)}, u^*) + d_{\mathcal{H}}(v^{(l)}, v^*), \tag{3.2.8}$$

donde

$$d_{\mathcal{H}}(u^{(l)}, u^*) \leq \frac{d_{\mathcal{H}}(P^{(l)}\mathbb{1}_m, a)}{1 - \lambda(\mathcal{K})^2} \quad \text{y} \quad d_{\mathcal{H}}(v^{(l)}, v^*) \leq \frac{d_{\mathcal{H}}((P^{(l)})^\top \mathbb{1}_n, b)}{1 - \lambda(\mathcal{K})^2}, \tag{3.2.9}$$

más aún, estas distancias decaen de forma exponencial:

$$d_{\mathcal{H}} \left(u^{(l)}, u^* \right) = \mathcal{O}(\lambda(\mathcal{K})^{2l}) \quad \text{y} \quad d_{\mathcal{H}} \left(v^{(l)}, v^* \right) = \mathcal{O}(\lambda(\mathcal{K})^{2l}).$$

El resultado anterior muestra una convergencia lineal para el algoritmo de Sinkhorn. Además, las desigualdades (3.2.8) y (3.2.9) indican que las discrepancias entre las marginales de $P^{(l)}$ y los vectores de probabilidad $a \in \Sigma_n$ y $b \in \Sigma_m$ son un buen criterio de parada para este algoritmo. Esto puede observarse en la Figura 3.8 para el caso discreto y en la Figura 3.9 para el caso continuo, donde además se grafica la proyección báricéntrica del plan de transporte entrópico, la cual permite obtener, de manera forzada, un mapa $T : \mathcal{X} \rightarrow \mathcal{Y}$ mediante el promedio en \mathcal{Y} del plan de Kantorovich π :

$$T(x) = \int_{\mathcal{Y}} y \frac{d\pi}{d(\mu \otimes \nu)}(x, y) d\nu(y).$$

3.3. Formulación dinámica

De forma similar a lo realizado para el problema de Kantorovich no regularizado, es posible formular el problema de Schrödinger (3.2.3) como un problema dinámico. En esta marco de trabajo, el objetivo es encontrar un proceso estocástico $x = (x_t)_{t \in [0,1]}$ que tenga ciertas distribuciones marginales al comienzo y al final del proceso, y que además esté lo más cerca posible (en el sentido de la entropía relativa) a un proceso estocástico de referencia:

Definición 3.10 (problema del puente de Schrödinger, versión dinámica). Considerando $\mathcal{X} = \mathcal{Y} \subset \mathbb{R}^d$, el puente de Schrödinger entre las medidas $\mu \in \mathcal{M}_+^1(\mathcal{X})$ y $\nu \in \mathcal{M}_+^1(\mathcal{Y})$ dada una medida de referencia $R \in \mathcal{M}_+^1(C([0,1], \mathcal{X}))$ es la (única) solución del problema

$$\arg \min_{P \in \Gamma(\mu, \nu)} D_{\text{KL}}(P \| R), \quad (3.3.1)$$

donde, al igual que en el Capítulo 1, $\Gamma(\mu, \nu)$ el conjunto de medidas de probabilidad en $C^1([0, 1], \mathcal{X})$ cuyas marginales en $t = 0$ y $t = 1$ son μ y ν respectivamente:

$$\Gamma(\mu, \nu) = \{P \in \mathcal{M}_+^1(C([0, 1], \mathcal{X})) : P_0 = \mu, P_1 = \nu\}.$$

Sorprendentemente, este problema puede ser resuelto fácilmente si se conoce una solución para el problema de Schrödinger estático (3.2.3). En efecto, la divergencia $D_{\text{KL}}(P \| R)$ entre dos procesos estocásticos se puede descomponer en la divergencia entre su marginales en tiempo $t \in \{0, 1\}$ y el resto del proceso en tiempo $t \in (0, 1)$. Para ver esto, se tiene la siguiente descomposición de la entropía relativa cuando está definida sobre medidas en un espacio producto:

Proposición 3.11 (regla de la cadena para la entropía relativa). Sean $\mu, \nu \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})$ dos medidas de probabilidad definidas en un espacio producto, entonces:

$$D_{\text{KL}}(\mu \| \nu) = D_{\text{KL}}(\mu_0 \| \nu_0) + \mathbb{E}_{x \sim \mu_0} [D_{\text{KL}}(\mu_{|x} \| \nu_{|x})],$$

donde μ_0 y ν_0 son la primera marginal de μ y ν respectivamente, mientras que $\mu_{|x}$ y $\nu_{|x}$ corresponden a la segunda marginal condicionada a que la primera componente sea x .

Demostración. Por simplicidad en la notación, se asumirá que μ y ν poseen función de densidad $p_\nu(x, y)$ y $p_\nu(x, y)$ respectivamente. De esta forma:

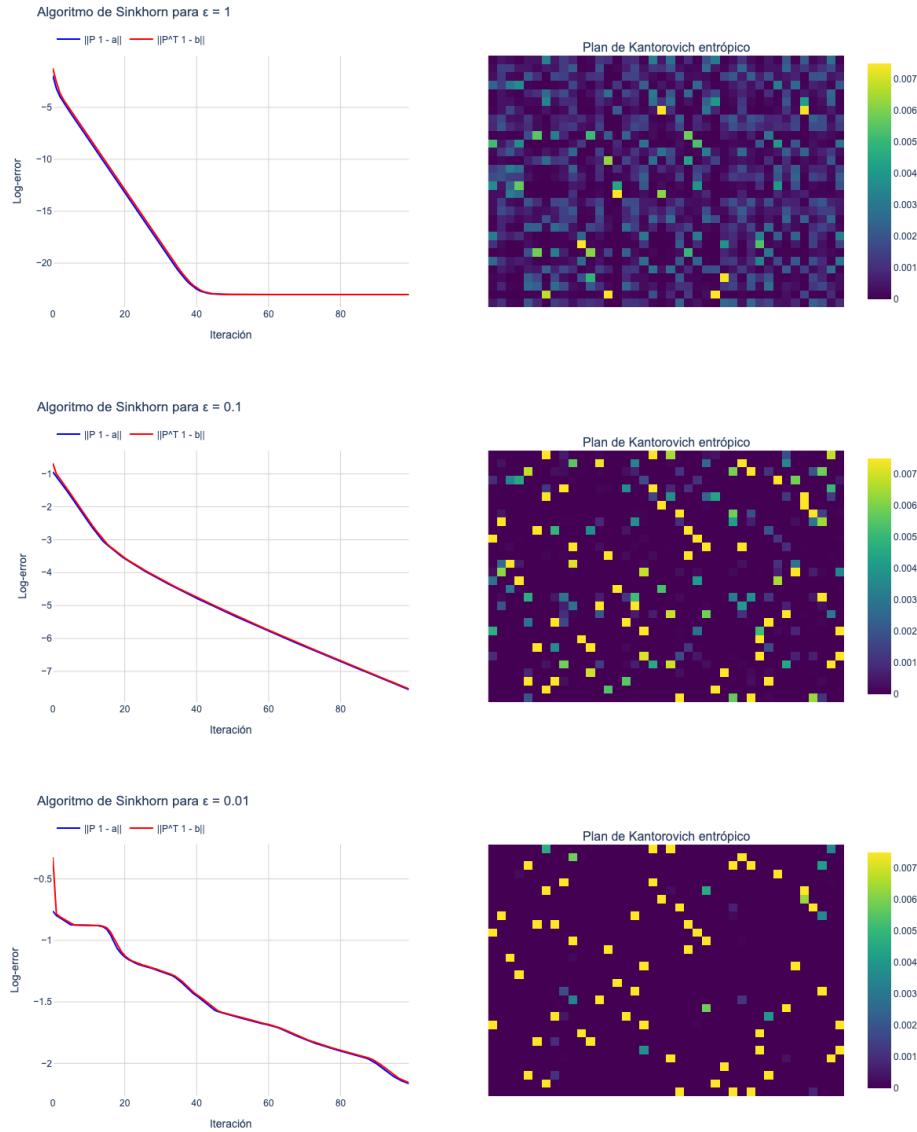


Figura 3.8: Iteraciones del algoritmo de Sinkhorn entre dos distribuciones discretas. Se observa que a medida se disminuye ϵ , el error disminuye más lentamente, mientras que el plan de transporte se vuelve más determinista. La simulación se encuentra en el archivo `sinkhorn.ipynb`.

$$\begin{aligned}
 D_{KL}(\mu \| \nu) &= \int_{\mathcal{X} \times \mathcal{Y}} \log \left(\frac{p_\mu(x, y)}{p_\nu(x, y)} \right) p_\mu(x, y) d(x, y) \\
 &= \int_{\mathcal{X} \times \mathcal{Y}} \log \left(\frac{p_\mu(x)}{p_\nu(x)} \right) p_\mu(x) p_\mu(y|x) d(x, y) + \int_{\mathcal{X} \times \mathcal{Y}} \log \left(\frac{p_\mu(y|x)}{p_\nu(y|x)} \right) p_\mu(x) p_\mu(y|x) d(x, y) \\
 &= \int_{\mathcal{X}} \log \left(\frac{p_\mu(x)}{p_\nu(x)} \right) p_\mu(x) \left(\int_{\mathcal{Y}} p_\mu(y|x) dy \right) dx + \int_{\mathcal{X}} \left[\log \left(\frac{p_\mu(y|x)}{p_\nu(y|x)} \right) p_\mu(y|x) dy \right] p_\mu(x) dx \\
 &= D_{KL}(p_\mu(x) \| p_\nu(x)) + \mathbb{E}_{x \sim p_\mu(x)} [D_{KL}(p_\mu(y|x) \| p_\nu(y|x))].
 \end{aligned}$$

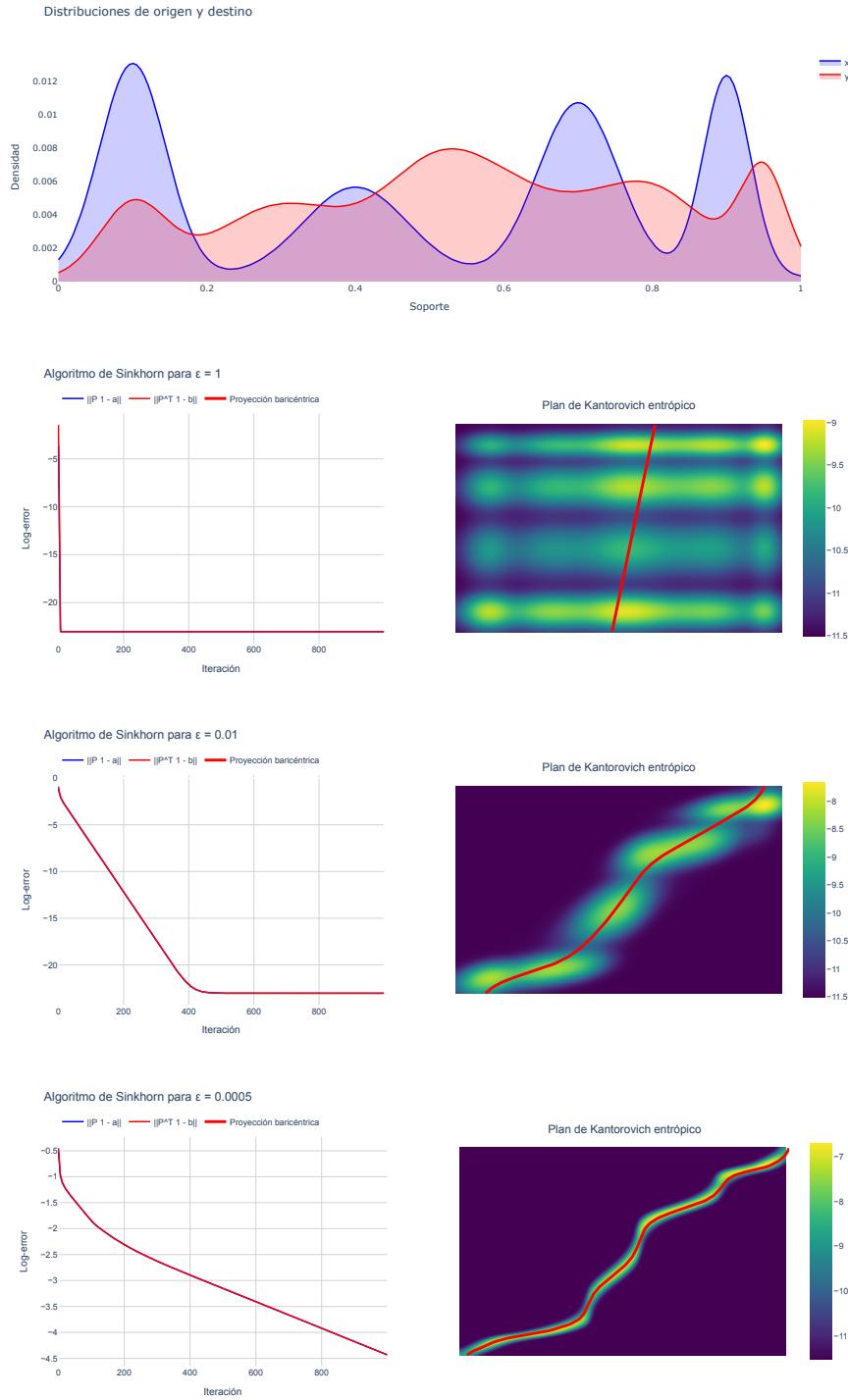


Figura 3.9: Iteraciones del algoritmo de Sinkhorn entre dos distribuciones continuas. Se observa que la proyección baricéntrica del plan de transporte óptimo converge al mapa de Monge del problema cuando $\epsilon \rightarrow 0$. Además, se observa que para $\epsilon = 1$ la solución (similar a la medida producto) se encuentra en muy pocas iteraciones. La simulación se encuentra en el archivo `sinkhorn.ipynb`.

□

En consecuencia, $D_{KL}(P \| R)$ admite la siguiente descomposición:

$$D_{KL}(P \| R) = D_{KL}(P_{01} \| R_{01}) + \mathbb{E}_{(x,y) \sim P_{01}} [D_{KL}(P_{|xy} \| R_{|xy})],$$

donde $P_{01}, R_{01} \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})$ son las medidas marginales de los procesos en tiempo $t = 0$ y $t = 1$, mientras que $P_{|xy}, R_{|xy}$ indican las medidas de los procesos condicionados a empezar en x y terminar en y .

Por otra parte, notar que para el problema dinámico (3.3.1) se puede elegir $P_{|xy} = R_{|xy}$ para hacer el segundo sumando nulo. De esta forma:

$$\min_{P \in \Gamma(\mu, \nu)} D_{KL}(P \| R) = \min_{P_{01} \in \Pi(\mu, \nu)} D_{KL}(P_{01} \| R_{01}).$$

Es decir, el problema dinámico tiene el mismo valor óptimo que el problema estático. Más aún, si P_{01}^* resuelve el problema estático, entonces la medida

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

resuelve el problema dinámico, mostrando que ambos problemas son equivalentes.

Por otro lado, cuando se considera como proceso de referencia a un movimiento browniano de difusividad ϵ (ver Definición A.6), el problema de Schrödinger se reduce al problema de transporte óptimo entrópico con costo cuadrático

Proposición 3.12 (SBP browniano estático). Dado $P \in \Gamma(\mu, \nu)$ y W^ϵ un movimiento browniano con difusividad ϵ , entonces:

$$D_{KL}(P_{01} \| W_{01}^\epsilon) = \frac{1}{\epsilon} \left[\int_{\mathcal{X} \times \mathcal{Y}} \frac{1}{2} \|x - y\|^2 dP_{01}(x, y) - \epsilon \cdot \mathcal{H}(P_{01}) + \text{constante} \right].$$

Demostración. Dado que W^ϵ es un movimiento browniano de difusividad ϵ , $\frac{dW_{1|0}^\epsilon}{dy}(y|x)$ es la función de densidad de una variable aleatoria gaussiana, por lo que resulta conveniente realizar la siguiente descomposición:

$$\frac{dW_{01}^\epsilon}{dx dy}(x, y) = \frac{dW_{1|0}^\epsilon}{dy}(y|x) \cdot \underbrace{\frac{dW_0^\epsilon}{dx}(x)}_{\frac{d\mu}{dx}(x)}, \quad dP_{01}(x, y) = dP_{1|0}(y|x) \underbrace{dP_0(x)}_{d\mu(x)}.$$

Con esto:

Simulación de Puente Browniano



Figura 3.10: Puente browniano entre los puntos $x_0 = 2$ y $x_1 = 5$. Esta simulación se encuentra en el archivo `sdes.ipynb`.

$$\begin{aligned}
 D_{\text{KL}}(P_{01} \| W_{01}^\epsilon) &= \int_{\mathcal{X} \times \mathcal{Y}} \log \left(\frac{dP_{01}}{dW_{01}^\epsilon} \right) dP_{01}(x, y) \\
 &= \int_{\mathcal{X} \times \mathcal{Y}} \log \left(\frac{dP_{01}}{dx \ dy} \right) dP_{01}(x, y) - \int_{\mathcal{X} \times \mathcal{Y}} \log \left(\frac{dW_{01}^\epsilon}{dx \ dy} \right) dP_{01}(x, y) \\
 &= -H(P_{01}) - \int_{\mathcal{X} \times \mathcal{Y}} \log \left[\frac{1}{(2\pi\epsilon)^{\frac{d}{2}}} \exp \left(\frac{-1}{2\epsilon} \|x - y\|^2 \right) \cdot \frac{d\mu}{dx}(x) \right] dP_{01}(x, y) \\
 &= -H(P_{01}) + \frac{d}{2} \log(2\pi\epsilon) + \int_{\mathcal{X} \times \mathcal{Y}} \frac{1}{2\epsilon} \|x - y\|^2 dP_{01}(x, y) - \int_{\mathcal{X} \times \mathcal{Y}} \log \left(\frac{d\mu}{dx}(x) \right) dP_{01}(x, y) \\
 &= \int_{\mathcal{X} \times \mathcal{Y}} \frac{1}{2\epsilon} \|x - y\|^2 dP_{01}(x, y) - H(P_{01}) - \underbrace{\int_{\mathcal{X}} \log \left(\frac{d\mu}{dx}(x) \right) \left(\int_{\mathcal{Y}} dP_{1|0}(y|x) \right) d\mu(x)}_{-\mathcal{H}(\mu)} + \text{constante} \\
 &= \frac{1}{\epsilon} \left[\int_{\mathcal{X} \times \mathcal{Y}} \frac{1}{2} \|x - y\|^2 dP_{01}(x, y) - \epsilon \cdot H(P_{01}) + \text{constante} \right].
 \end{aligned}$$

□

A modo de ejemplo, en la Figura 3.11 se muestran algunos puentes de Schrödinger trazados entre dos distribuciones discretas. Es importante destacar que lo único necesario para realizar esto es resolver el problema entrópico con costo cuadrático, ya que la distribución condicional $W_{|xy}$ del movimiento browiano empezando en x y terminando en y es totalmente conocida y se denomina *puente browniano*, el cual tiene la siguiente SDE:

$$dx_t = \frac{y - x_t}{1 - t} dt + dw_t$$

donde se consideró, por simplicidad, que la difusión es constante y unitaria. En la Figura 3.10 se puede ver un puente browiano unidimensional.

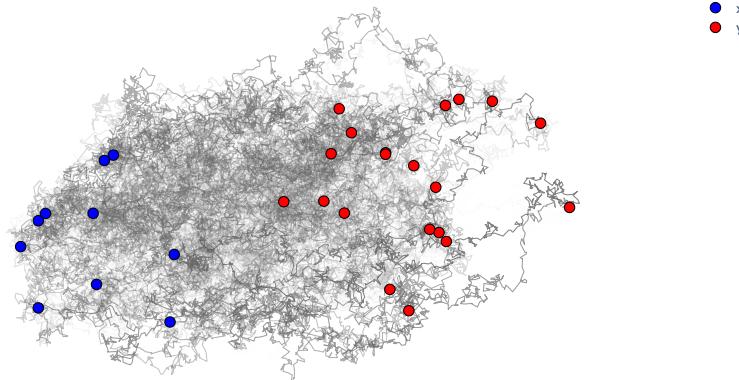
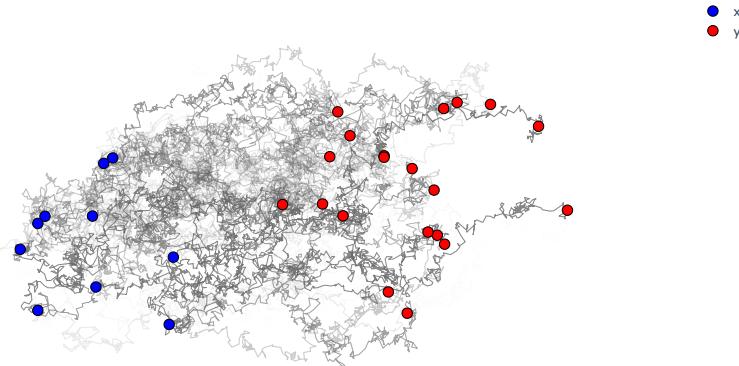
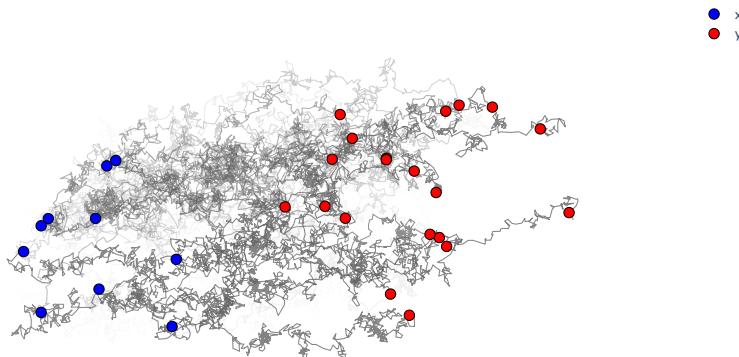
Puentes de Schrödinger para $\epsilon = 1$ Puentes de Schrödinger para $\epsilon = 0.1$ Puentes de Schrödinger para $\epsilon = 0.05$ 

Figura 3.11: Puentes de Schrödinger para distintos valores de ϵ . Se observa como disminuye la aleatoriedad del plan de transporte estocástico cuando disminuye ϵ . El código que genera estas imágenes se encuentra en el archivo `sbp.ipynb`.

3.3.1. Formulación de Brenamou-Brenier para el problema de Schrödinger

La interpretación fluidodinámica que se le dio al problema de transporte óptimo en el Capítulo 2 puede ser generalizada al problema del puente de Schrödinger de forma natural. Recordando que la dinámica que se optimiza en la formulación de Benamou-Brenier (2.3.12) es determinista, en esta nueva formulación se

considera un proceso estocástico de la forma

$$dx_t = v(x_t, t) dt + \sigma dw_t, \quad \sigma > 0 \text{ constante},$$

donde este nuevo sistema dinámico puede interpretarse como el sistema determinista de Benamou-Brenier con un término de ruido adicional. De esta forma, este nuevo proceso tendrá asociada la siguiente ecuación de Fokker-Planck:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (v\rho) - \frac{\sigma^2}{2} \Delta \rho = 0.$$

Esta ecuación corresponde a la ecuación de transporte con un término adicional que indica la estocasticidad del modelo. Por lo tanto, la formulación estocástica para Benamou-Brenier es la siguiente:

$$\inf_{(\rho, v)} \int_0^1 \int_{\mathcal{X}} \|v(x, t)\|^2 \rho(x, t) dx dt \quad \text{sujeto a} \quad \begin{cases} \frac{\partial \rho}{\partial t} + \nabla \cdot (\rho v) - \frac{\sigma^2}{2} \Delta \rho = 0 \\ \rho(\cdot, 0) = \rho_0 \\ \rho(\cdot, 1) = \rho_1, \end{cases} \quad (3.3.2)$$

donde el único cambio fue cambiar la ecuación de transporte por la ecuación de Fokker-Planck. Notar que cuando $\sigma \rightarrow 0$ se recupera la formulación de Benamou-Brenier original.

Para encontrar las condiciones de optimalidad de este problema de control, se puede repetir el mismo procedimiento hecho en la Subsección 2.3.3, donde la única diferencia es que se añade un término adicional al lagrangiano del problema, el cual puede ser trabajado mediante la segunda identidad de Green:

$$-\frac{\sigma^2}{2} \int_{\mathcal{X}} \lambda \Delta \rho dx = -\frac{\sigma^2}{2} \cdot 0 - \frac{\sigma^2}{2} \int_{\mathcal{X}} \rho \Delta \lambda dx.$$

Por lo tanto, el lagrangiano toma la siguiente forma para esta formulación:

$$L = \int_{\mathcal{X}} \int_0^1 \left(\frac{1}{2} \|v\|^2 - \frac{\partial \lambda}{\partial t} - \nabla \lambda \cdot v - \frac{\sigma^2}{2} \Delta \lambda \right) \rho dt dx + \text{constante}.$$

Nuevamente, por condición de primer orden, se tiene que $v^* = \nabla \lambda$ es control óptimo, por lo que

$$L = - \int_{\mathcal{X}} \int_0^1 \left(\frac{\partial \lambda}{\partial t} + \frac{1}{2} \|\nabla \lambda\|^2 + \frac{\sigma^2}{2} \Delta \lambda \right) \rho dt dx + \text{constante}.$$

Luego, si λ satisface la ecuación de Hamilton-Jacobi-Bellman, $\frac{\partial \lambda}{\partial t} + \frac{1}{2} \|\nabla \lambda\|^2 + \frac{\sigma^2}{2} \Delta \lambda = 0$, entonces $L = \text{constante}$ y cualquier $\rho \in \Gamma(\rho_0, \rho_1)$ minimiza el lagrangiano. En consecuencia, se probó el siguiente para la versión estocástica de Benamou-Brenier:

Proposición 3.13. Si (ρ^*, λ^*) es solución del sistema acoplado

$$\begin{aligned} \frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \nabla \lambda) - \frac{\sigma^2}{2} \Delta \rho &= 0 \\ \frac{\partial \lambda}{\partial t} + \frac{1}{2} \|\nabla \lambda\|^2 + \frac{\sigma^2}{2} \Delta \lambda &= 0, \end{aligned}$$

con condiciones de borde $(\rho(\cdot, 0) = \rho_0) \wedge (\rho(\cdot, 1) = \rho_1)$, entonces, (ρ^*, v^*) es solución del problema de Benamou-Brenier estocástico (3.3.2) con $v^* = \nabla \lambda^*$.

Notar que la primera ecuación del sistema acoplado es la ecuación de Fokker-Planck de la SDE $dx_t = v(x_t, t) dt + \sigma dw_t$ asociada al transporte, mientras que la segunda ecuación (Hamilton-Jacobi-Bellman) es la que codifica la optimalidad del control. Además, dado que esta formulación de Benamou-Brenier equivale al problema de Schrödinger, se puede considerar que el resultado anterior es un criterio de optimalidad para el problema del puente de Schrödinger dinámico.

Sistema de Schrödinger

El sistema acoplado de la Proposición 3.13 para el problema de Schrödinger puede ser transformado en otro sistema acoplado equivalente mediante las transformaciones de Hopf-Cole. Aplicando el cambio de variable $(\lambda, \rho) \rightarrow (\phi, \hat{\phi})$ definido como

$$\phi = \exp\left(\frac{\lambda}{\sigma^2}\right) \quad \hat{\phi} = \rho \exp\left(\frac{-\lambda}{\sigma^2}\right),$$

el sistema de la Proposición 3.13 se reduce a:

$$\frac{\partial \phi}{\partial t} + \frac{\sigma^2}{2} \Delta \phi = 0 \tag{3.3.3}$$

$$\frac{\partial \hat{\phi}}{\partial t} - \frac{\sigma^2}{2} \Delta \hat{\phi} = 0, \tag{3.3.4}$$

con condiciones de borde $\phi(\cdot, 0)\hat{\phi}(\cdot, 0) = \rho_0$ y $\phi(\cdot, 1)\hat{\phi}(\cdot, 1) = \rho_1$. Este sistema es conocido como *sistema de Schrödinger* y será el que permitirá ver que los modelos de difusión estudiados en el Capítulo 1 son un caso particular del problema del puente de Schrödinger. Para esto, es necesario generalizar el resultado anterior a una familia más amplia de procesos estocásticos. Si se fija como medida de referencia al proceso $dx_t = f(x_t, t) dt + \sigma g(t) dw_t$, se puede probar la siguiente extensión del resultado anterior (ver [CGP20] para una demostración):

Proposición 3.14. Dado el problema de Schrödinger $\min_{P \in \Gamma(\rho_0, \rho_1)} D_{KL}(P \| R)$ con R la medida de un proceso que resuelve la SDE $dx_t = f(x_t, t) dt + \sigma g(t) dw_t$ (con $\sigma > 0$ una constante), entonces

$$\inf_{(\rho, v)} \int_0^1 \int_{\mathcal{X}} \|v(x, t)\|^2 \rho(x, t) dx dt \quad \text{sujeto a} \quad \begin{cases} dx_t = (f(x_t, t) + g(t)v(x_t, t)) dt + \sigma g(t) dw_t \\ \rho(\cdot, 0) = \rho_0, \\ \rho(\cdot, 1) = \rho_1, \end{cases}$$

es la formulación de Benamou-Brenier para este problema de Schrödinger. Más aún, las condiciones de optimalidad para este problema son las siguientes:

$$\begin{aligned} \frac{\partial \phi}{\partial t} + \frac{\sigma^2}{2} \text{Tr}(g^2 \Delta \phi) + \nabla \phi^\top f &= 0 \\ \frac{\partial \hat{\phi}}{\partial t} - \frac{\sigma^2}{2} \text{Tr}(g^2 \Delta \hat{\phi}) + \nabla \cdot (\hat{\phi} f) &= 0, \end{aligned}$$

donde el control óptimo es $v(x_t, t) = \sigma^2 g(t) \nabla \log \phi(x_t, t)$. Así, fijando $\sigma = 1$, la solución P^* del problema de Schrödinger viene dada por la ley de cualquiera de las siguientes SDEs:

$$\begin{aligned} dx_t &= (f + g^2 \nabla \log \phi(x_t, t)) dt + g dw_t, \quad x_0 \sim \rho_0 \\ dx_t &= (f - g^2 \nabla \log \hat{\phi}(x_t, t)) dt + g dw_t, \quad x_1 \sim \rho_1 \end{aligned}$$

En consecuencia, para resolver el problema del puente de Schrödinger para una SDE de referencia usual, se debe resolver el sistema acoplado de la Proposición 3.14 y luego simular el proceso forward (usando ϕ) o el proceso backward (usando $\hat{\phi}$) para tener una simulación del proceso P^* que resuelve el problema de Schrödinger.

Problema de Schrödinger como generalización de los modelos de difusión

Para el entrenamiento de un modelo de difusión a tiempo continuo, el proceso de difusión por lo general debe tener una forma muy simple (p.g. lineal) para poder computar la función de score $\nabla_{x_t} \log p(x_t | x_0)$ de forma cerrada, lo cual muchas veces limita la aplicabilidad de este tipo de modelos. Para concluir este trabajo, se verá un último resultado que permite entrenar un modelo para el problema de Schrödinger mediante verosimilitud. Este método, al igual que los modelos de difusión, permite realizar transporte óptimo en alta dimensión y de manera eficiente. Además, al ser un resultado teórico en el continuo, no se ve afectado por la elección del solver usado.

Se comenzará con el siguiente resultado, el cual transforma el sistema de PDEs acoplado de la Proposición 3.14 en un sistema acoplado de SDEs:

Teorema 3.4 (Feynman-Kac para el problema de Schrödinger). Sea

$$\begin{aligned} dx_t &= (f + gz_t) dt + g dw_t \\ dy_t &= \frac{1}{2} \hat{z}_t^\top z_t dt + z_t^\top dw_t \\ d\hat{y}_t &= \left(\frac{1}{2} \hat{z}_t^\top \hat{z}_t^\top \nabla \cdot (g\hat{z}_t + f) + \hat{z}_t^\top z_t \right) dt + \hat{z}_t^\top dw_t, \end{aligned}$$

un sistema acoplado con condiciones de borde $x_0 \sim \delta_{x_0}$ e $y_1 + \hat{y}_1 = \log \rho_1(x_1)$. Entonces, las relaciones de Feynman-Kac entre el sistema de PDEs de la Proposición 3.14 y este nuevo sistema de SDEs son las siguientes:

$$\begin{aligned} y_t &= \log \phi(x_t, t), \quad z_t = g \nabla \log \phi(x_t, t) \\ \hat{y}_t &= \log \hat{\phi}(x_t, t), \quad \hat{z}_t = g \nabla \log \hat{\phi}(x_t, t). \end{aligned}$$

Además, se tiene que $y_t + \hat{y}_t = \log p(x_t)$, $\forall t \in [0, 1]$, donde p es la densidad de la solución del problema de Schrödinger de la Proposición 3.14.

En consecuencia, es posible recuperar los procesos forward y backward de la Proposición 3.14 (que corresponden a la solución del problema de Schrödinger hacia adelante y hacia atrás en el tiempo) a partir de z_t y \hat{z}_t . Para entrenar estos modelos, se utilizará el siguiente resultado, cuya demostración se puede encontrar en [CLT23]:

Teorema 3.5 (verosimilitud para el problema de Schrödinger). Para un modelo (z_t, \hat{z}_t) del problema de Schrödinger (de acuerdo a la Teorema 3.4), la verosimilitud de un punto $x_0 \in \mathcal{X}$ puede ser escrita como:



Figura 3.12: Imágenes generadas utilizando el enfoque del Teorema 3.5 sobre los datasets MNIST, CelebA y CIFAR-10. Imagen obtenida desde [CLT23].

$$\log p(x_0) = \mathbb{E} [\log p(x_1)] - \int_0^1 \mathbb{E} \left[\frac{1}{2} \|z_t + \hat{z}_t\|^2 + \nabla \cdot (\sigma_t \hat{z}_t - f) \right] dt \quad (3.3.5)$$

Por lo tanto, entrenando un modelo paramétrico según (3.3.5) para z_t y otro para \hat{z}_t , es posible entrenar un modelo neuronal de forma eficiente para el problema de Schrödinger. Para las divergencias dentro de la función objetivo se puede usar un estimador eficiente como el estimador de Hutchinson. En la Figura 3.12 se pueden ver muestras generadas utilizando este enfoque.

Notar que la función objetivo (3.3.5) colapsa a la función objetivo de un modelo de difusión cuando $(z_t, \hat{z}_t) = (0, gs_\theta)$ (ver Teorema 1.5). Esto ocurre cuando la medida de referencia R por si sola es capaz de alcanzar la distribución ρ_1 ya que, en este caso, no hay que hacer un esfuerzo adicional para alcanzar ρ_1 , por lo que $z_t = 0$ y \hat{z}_t colapsa a gs_θ .

Conclusión

En este trabajo se presentó de manera autocontenido y unificada tres tópicos que han ganado gran relevancia en la comunidad de la inteligencia artificial generativa: los modelos de difusión, el transporte óptimo y el problema del puente de Schrödinger. Para ello, se comenzó el estudio con una descripción general de algunos modelos generativos neuronales clásicos, lo cual permitió introducir conceptos clave como variable latente e inferencia aproximada. Posteriormente, se realizó un análisis exhaustivo de los modelos generativos basados en difusión, colocando énfasis en realizar un desarrollo *natural*, tanto en su formulación como en su implementación. Esto permitió identificar algunas limitaciones intrínsecas de este tipo de modelos, lo que motivó el estudio del transporte óptimo para luego concluir con una equivalencia entre el transporte óptimo regularizado y el problema del puente de Schrödinger.

En particular, los temas cubiertos para cada tópico fueron los siguientes:

- Modelos de difusión.
 - Modelos generativos alternativos: redes generativas adversarias (Subsección 1.1.1), modelos basados en energía (Subsección 1.1.2) y autoencoders variacionales (Sección 1.2). En particular, esta última familia de modelos pudo verse como una formulación previa a los modelos de difusión, permitiendo introducir el concepto de ELBO y su interpretación física.
 - Formulación discreta: se definieron los procesos forward (1.3.1) y backward (1.3.2) de un modelo de difusión como una cadena de Markov discreta, lo que permitió obtener fácilmente las distribuciones $q(x_t|x_0)$ y $q(x_{t-1}|x_t, x_0)$ (en la Proposición 1.10 y en la Proposición 1.11 respectivamente), ambas usadas durante el entrenamiento. Además, se mostró que este tipo de modelos admiten diferentes reparametrizaciones.
 - Detalles prácticos: se implementaron dos arquitecturas neuronales importantes (U-Net y DiT) en la Subsección 1.3.2, ambas usadas en los modelos de difusión. Además, en la Subsección 1.3.3 se estudiaron algunos aspectos útiles al momento de entrenar y usar estos modelos, siendo DDIM y guidance los temas más importantes de esta parte.
 - Formulación a tiempo continuo: en la Subsección 1.4.1 se comenzó estudiando algunos modelos que buscan aprender la función de score (score matching y denoising score matching) para luego extender los procesos forward y backward de un modelo de difusión al continuo mediante el uso de ecuaciones diferenciales estocásticas en la Subsección 1.4.2.
 - Limitaciones de los modelos de difusión: en la Subsección 1.4.4 se estudiaron las limitaciones que motivaron el estudio del transporte óptimo.
- Transporte óptimo.
 - Se comenzó estudiando los problemas de Monge (Sección 2.1) y Kantorovich (Sección 2.2), ambos

tanto en su formulación discreta como continua, colocando énfasis en una introducción natural al tema en vez de priorizar un desarrollo completamente riguroso.

- En la Subsección 2.2.2 se estudió la formulación dual, la cual es necesaria para obtener resultados como el Teorema 2.2. Además, esta formulación dual es *relajada* en la Subsección 3.1.2, permitiendo obtener una solución primal a partir de una solución dual.
 - Formulación dinámica: en la Subsección 2.3.1 se mostró que el problema de Kantorovich induce una distancia en (un subconjunto de) $\mathcal{M}_+^1(\mathcal{X})$, la cual permite interpolar entre distribuciones de probabilidad y caracteriza la convergencia débil de medidas. Luego, en la Subsección 2.3.2 y en la Subsección 2.3.3 se vieron dos reformulaciones dinámicas del problema, las cuales fueron generalizadas posteriormente en el Capítulo 3.
- Problema de Schrödinger.
- Regularización entrópica: algunas limitaciones del problema de Kantorovich sugieren sumar un término regularizador al problema original, lo cual es hecho en la Subsección 3.1.1. Las ventajas de esto son estudiadas en la Subsección 3.1.2.
 - SBP estático: en la Sección 3.2 se ve la equivalencia entre este nuevo problema regularizado y una versión estática del problema de Schrödinger, obteniendo un algoritmo eficiente para resolver ambos problemas.
 - SBP dinámico: por último, en la Sección 3.3 se extienden los resultados anteriores de forma análoga a lo realizado en la Sección 2.3.

Impacto práctico y metodológico

Uno de los aportes más destacables de este trabajo es la integración de conceptos teóricos complejos en un marco accesible y aplicable. Este enfoque permite a la comunidad investigadora, especialmente a quienes se centran en modelos generativos, entender y aprovechar herramientas matemáticas avanzadas como la teoría del transporte óptimo y su conexión con procesos estocásticos. En particular, la equivalencia entre el transporte óptimo regularizado y los puentes de Schrödinger ofrece una perspectiva unificadora que facilita el desarrollo de algoritmos más eficientes y versátiles (como el Algoritmo 8).

En términos prácticos, las implementaciones realizadas en esta tesis proporcionan una base sólida para futuras investigaciones. Los modelos de difusión desarrollados, junto con las arquitecturas neuronales implementadas (como U-Net y DiT), representan herramientas de referencia al momento de implementar nuevos modelos. Por otro lado, los métodos propuestos para resolver el problema del transporte óptimo y su versión regularizada han demostrado ser computacionalmente eficientes, especialmente en dominios donde los datos están estructurados de forma no trivial.

Asimismo, los puentes de Schrödinger, al ser formulados como una extensión estocástica del transporte óptimo, abren nuevas posibilidades para modelar sistemas donde la incertidumbre y el ruido juegan un papel fundamental. Esto es particularmente relevante en áreas como la modelización climática, la física computacional y las ciencias biomédicas, donde las distribuciones de probabilidad de interés no siempre son fácilmente parametrizables.

En conclusión, esta tesis no solo aporta al entendimiento teórico de modelos generativos avanzados, sino que también ofrece herramientas prácticas y un marco conceptual que pueden ser utilizados y extendidos por la comunidad científica y tecnológica para abordar problemas actuales y futuros en inteligencia artificial generativa.

Trabajo futuro

El objetivo general de este trabajo fue estudiar el transporte óptimo y el problema de Schrödinger como una generalización natural de los modelos generativos basados en difusión. En esa misma línea, una familia alternativa de modelos generativos que ha ganado popularidad en los últimos meses son los modelos basados en flujos normalizantes a tiempo continuo (propuestos inicialmente en [Che+19]), los cuales consisten en transformar una distribución en otra guiando su trayectoria mediante un campo de velocidades. Los modelos de difusión pueden verse como un caso particular de este tipo de modelos mediante la probability flow ODE (1.4.14), mientras que la formulación de Benamou-Brenier en la Subsección 2.3.3 busca un campo de velocidades que realice la interpolación de McCann asociada al transporte óptimo entre las distribuciones. Por otra parte, los puentes de Schrödinger pueden verse como una versión estocástica de esta familia de modelos.

Dentro de la categoría de modelos basados en flujos continuos, las propuestas más populares son los *rectified flows* (propuesto en [LGL22]) y *flow matching* (propuesto en [Lip+23]). Estas metodologías no solo representan otra generalización elegante de los modelos de difusión, sino que también comparten muchas de las características positivas de los puentes de Schrödinger estudiadas en este trabajo.

Apéndice

A.1. Medidas de probabilidad

La teoría de la medida es una rama del análisis matemático que permite tener un marco teórico robusto para el estudio de las probabilidades, el cual muchas veces se vuelve necesario al trabajar con variables aleatorias continuas. En esta sección se entregará la intuición de algunos resultados necesarios para este trabajo, sin entrar en detalles técnicos como la construcción de la integral de Lebesgue o la medibilidad de las funciones, lo que extendería considerablemente la sección y desviaría el foco del trabajo. En consecuencia, muchos de los resultados entregados no son del todo riguroso pero sí son suficientes para el desarrollo de la teoría del transporte óptimo.

En la teoría de la medida se definen de forma precisa dos conjuntos los cuales se suelen ver como un único conjunto en la interpretación clásica de las probabilidades. Estos conjuntos corresponden al conjunto muestral \mathcal{X} y el espacio de eventos \mathcal{S} . Usando como ejemplo el experimento de lanzar dos monedas, el espacio muestral corresponde a todos los posibles resultados del experimento, es decir $\mathcal{X} = \{(C, C), (C, S), (S, C), (S, S)\}$, mientras que el espacio de eventos corresponde a diferentes subconjuntos del espacio muestral \mathcal{X} que puedan ser de interés para calcularles una probabilidad. Por ejemplo, un evento A podría ser que la primera moneda sea C (i.e., $A = \{(C, C), (C, S)\} \subset \mathcal{P}(\mathcal{X})^1$) mientras que otro evento B puede ser que ambas monedas sean S (i.e., $B = \{(S, S)\} \subset \mathcal{P}(\mathcal{X})$). De este modo, para este caso particular, los posibles eventos de este experimento son las diferentes combinaciones de resultados en \mathcal{X} que se pueden obtener:

$$\begin{aligned}\mathcal{S} = & \{ \\ & \{\}, \\ & \{(C, C)\}, \{(C, S)\}, \{(S, C)\}, \{(S, S)\}, \\ & \{(C, C), (C, S)\}, \{(C, C), (S, C)\}, \{(C, C), (S, S)\}, \{(C, S), (S, C)\}, \{(C, S), (S, S)\}, \{(S, C), (S, S)\}, \\ & \{(C, C), (C, S), (S, C)\}, \{(C, C), (C, S), (S, S)\}, \{(C, C), (S, C), (S, S)\}, \{(C, S), (S, C), (S, S)\}, \\ & \{(C, C), (C, S), (S, C), (S, S)\}\end{aligned}$$

Notar que, el espacio muestral completo \mathcal{X} también es por si solo un evento, al igual que su complemento $\{\}$ (evento vacío). Es lógico esperar que para alguna noción de probabilidad, la probabilidad del evento $\mathcal{X} \in \mathcal{S}$ sea 1, mientras que la probabilidad del evento vacío $\emptyset \in \mathcal{S}$ se espera que sea 0. Además, en este espacio de eventos \mathcal{S} hay otros eventos que también tiene sentido asignarles probabilidad 0 como por ejemplo el evento $\{(C, C), (S, S)\}$. En teoría de la medida, al conjunto de eventos \mathcal{S} se le denomina σ -álgebra y es sobre los

¹Para un conjunto arbitrario \mathcal{X} , el conjunto potencia (también llamado conjunto de las partes) corresponde al conjunto de todos los subconjuntos posibles de \mathcal{X} : $\mathcal{P}(\mathcal{X}) := \{A : A \subset \mathcal{X}\}$. En particular, $\mathcal{X} \in \mathcal{P}(\mathcal{X})$ y $\emptyset \in \mathcal{P}(\mathcal{X})$.

elementos de este conjunto donde define una noción de probabilidad, lo cual es más razonable que asignarle probabilidades a los elementos del espacio muestral \mathcal{X} directamente. Por otra parte, la noción de probabilidad que se asigna sobre los elementos de \mathcal{S} debe ser consistente con lo que uno esperaría de la interpretación clásica de las probabilidades. En particular, la probabilidad del evento formado por dos eventos disjuntos (i.e., cuya intersección es vacía) debe ser la suma de la probabilidad de los eventos individuales.

Con las observaciones anteriores es natural construir las siguientes definiciones:

Definición A.1 (σ -álgebra). Dado un conjunto \mathcal{X} no necesariamente finito, otro conjunto $\mathcal{S} \subset \mathcal{P}(\mathcal{X})$ se dirá σ -álgebra (en \mathcal{X}) si:

- $\mathcal{X} \in \mathcal{S}$.
- Si $A \in \mathcal{S}$ entonces $A^c := \{w \in \mathcal{X} : w \notin A\} \in \mathcal{S}$.
- Si $(A_n)_{n \in \mathbb{N}} \subset \mathcal{S}$ entonces $\bigcup_{n \in \mathbb{N}} A_n \in \mathcal{S}$.

Por último, a los elementos $A \in \mathcal{S}$ se les denomina *conjuntos medibles* y al par $(\mathcal{X}, \mathcal{S})$ se le denomina *espacio medible*.

En el marco probabilístico, la primera condición indica que el espacio muestral \mathcal{X} es un evento por si solo. La segunda condición exige que para todo evento, el evento complementario (que se puede entender como la negación del evento original) también es efectivamente un evento. Por otra parte, la última condición indica que unir una cantidad numerable de eventos es también un evento.

A.1.1. Definición de medida

Para asignar una probabilidad a los eventos de una σ -álgebra se utiliza la noción de medida, la cual se extiende más allá del marco probabilístico:

Definición A.2 (medida). Dado un espacio medible $(\mathcal{X}, \mathcal{S})$, una medida sobre este espacio es una función $\mu : \mathcal{S} \rightarrow \mathbb{R} \cup \infty$ que además cumple los axiomas de Kolmogorov:

- Positividad: $\mu(A) \geq 0^2$ para todo conjunto medible $A \in \mathcal{S}$.
- Anulación en el vacío: $\mu(\emptyset) = 0$.
- σ -aditividad: si $(A_n)_{n \in \mathbb{N}} \subset \mathcal{S}$ es una familia de conjuntos medibles disjuntos de a pares³, entonces:

$$\mu \left(\bigcup_{n \in \mathbb{N}} A_n \right) = \sum_{n \in \mathbb{N}} \mu(A_n)$$

A la tupla $(\mathcal{X}, \mathcal{S}, \mu)$ se le denomina *espacio de medida*. Si $\mu(\mathcal{X}) = 1$ se dice que μ es una *medida de probabilidad* y $(\mathcal{X}, \mathcal{S}, \mu)$ se denomina *espacio de probabilidad*.

A lo largo de este trabajo, $\mathcal{M}_+(\mathcal{X})$ representará el conjunto de medidas definidas sobre el espacio medible $(\mathcal{X}, \mathcal{S})$. Del mismo modo, $\mathcal{M}_+^1(\mathcal{X}) \subset \mathcal{M}_+(\mathcal{X})$ representará el conjunto de medidas de probabilidad definidas en $(\mathcal{X}, \mathcal{S})$. Por otra parte, este capítulo está enfocado en formular la teoría únicamente para medidas de probabilidad ya que este es un marco de trabajo lo suficientemente general y está en línea con lo realizado en el Capítulo 1. Debido a esto, al hablar de medidas siempre se estará haciendo referencia a medidas (positivas)

²Si bien es posible trabajar con valores negativos (medidas con signo), en este trabajo se asumirá siempre la positividad.

³Es decir, $A_i \cap A_j = \emptyset, \forall i \neq j$.

de probabilidad. Sin embargo, la mayor parte de los resultados se puede extender sin mayores cambios al conjunto de medidas positivas $\mathcal{M}_+(\mathcal{X})$ e incluso al conjunto de medidas con signo. Por último, es importante tener en cuenta que dada una medida positiva finita $\mu \in \mathcal{M}_+(\mathcal{X})$ (i.e., $\mu(\mathcal{X}) = M < \infty$), siempre es posible construir una medida de probabilidad mediante normalización (i.e., $\mu_{\text{probabilidad}} = \frac{1}{M}\mu$).

Medidas en espacios discretos

Si el conjunto muestral \mathcal{X} es finito, entonces se dirá que se está trabajando sobre un *espacio discreto*. En este caso, es usual considerar la σ -álgebra de las partes, por lo que el espacio medible en el caso discreto es $(\mathcal{X}, \mathcal{P}(\mathcal{X}))$. Por lo tanto, si $\mathcal{X} = \{x_1, \dots, x_n\}$, las medidas (discretas) de probabilidad sobre este espacio son siempre de la forma

$$\mu = \sum_{i=1}^n \lambda_i \delta_{x_i} \quad (\text{A.1.1})$$

donde δ_{x_i} es un delta de Dirac⁴ y $\lambda_i \geq 0$ indica la probabilidad (también llamada *masa*) que la medida μ le entrega al elemento $x_i \in \mathcal{X}$. Dado que μ debe ser medida de probabilidad, necesariamente $\sum_{i=1}^n \lambda_i = 1$. Al vector $\lambda \in \mathbb{R}_+^n$ se le denomina *vector de probabilidad asociado a μ* y determina totalmente la medida $\mu \in \mathcal{M}_+^1(\mathcal{X})$.

El conjunto de todos los vectores de probabilidad se denomina *simplex de probabilidad* y cada vector de este conjunto determina de manera biunívoca una medida de probabilidad en el espacio medible discreto mediante (A.1.1):

$$\Sigma_n := \left\{ \lambda \in \mathbb{R}_+^n : \|\lambda\|_1 = \sum_{i=1}^n \lambda_i = 1 \right\} \quad (\text{A.1.2})$$

Medidas en espacios continuos

De acuerdo a la Definición A.2, una medida μ no es necesariamente finita, lo que permite extender la teoría a otros tópicos no relacionados directamente con la teoría de probabilidades. Por ejemplo, en el plano \mathbb{R}^2 se puede definir, naturalmente, la medida de Lebesgue, la cual asigna a cada figura en el plano su respectiva área. Esta medida no es finita (se pueden hacer figuras tan grandes como se quiera) ni tiene interpretación probabilística.

Para poder definir la medida estándar en $\mathcal{X} = \mathbb{R}^d$ es necesario primero indicar cuál es la σ -álgebra adecuada en este espacio. Si bien en el caso discreto es usual considerar la σ -álgebra de las partes (donde todos los subconjuntos de \mathcal{X} son conjuntos medibles), en el caso continuo esto genera ciertos problemas patológicos que sugieren definir una σ -álgebra más apropiada. La σ -álgebra usual en estos casos, denotada por $\mathcal{B}(\mathbb{R}^d)$, es la menor σ -álgebra que contiene todos los conjuntos abiertos de \mathbb{R}^d y se denomina *σ -álgebra de los borelianos de \mathbb{R}^d* . Si bien $\mathcal{B}(\mathbb{R}^d) \subsetneq \mathcal{P}(\mathbb{R}^d)$, esta σ -álgebra es lo suficientemente grande y general como para contener todos los subconjuntos de \mathbb{R}^d considerados en la práctica. Más aún, encontrar un subconjunto de \mathbb{R}^d que no pertenezca a $\mathcal{B}(\mathbb{R}^d)$ (i.e., que no sea medible), es un problema no trivial.

Con respecto a la medida que se considera en el escenario continuo, para el caso $d = 1$ es usual definir, bajo la σ -álgebra $\mathcal{B}(\mathbb{R})$, la medida $\mu \in \mathcal{M}_+(\mathbb{R})$ definida en cada intervalo $(a, b] \in \mathcal{B}(\mathbb{R})$ como su largo (i.e., $\mu((a, b]) = b - a$ ⁵). Esta medida es conocida como medida de Lebesgue y se considera la medida por defecto

⁴Es decir, $\delta_{x_i}(A)$ vale 1 si $x_i \in A$ y 0 si no.

⁵El teorema de Hahn-Carathéodory garantiza que es suficiente definir una medida en intervalos de este tipo para poder extenderla de forma única a todos los elementos de $\mathcal{B}(\mathbb{R})$.

en \mathbb{R} . De forma equivalente, se puede construir la medida de Lebesgue en \mathbb{R}^d (bajo la σ -álgebra $\mathcal{B}(\mathbb{R}^d)$), donde la medida de un conjunto medible $A \in \mathcal{B}(\mathbb{R}^d)$ es precisamente su volumen en el espacio.

Medidas en espacios más generales

La teoría del transporte óptimo y del puente de Schrödinger en el Capítulo 3 se puede extender sin mayores cambios a espacios más generales que el caso discreto y continuo. Por lo tanto, por completitud, se entregarán las definiciones precisas para estos casos.

Un espacio topológico (\mathcal{X}, τ) se dirá *separable* si posee un subconjunto denso numerable. Si existe una métrica $d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$ tal que su topología inducida es igual a τ , el espacio topológico se dirá *metrizable*. Si d además es una distancia completa (en el sentido de Cauchy), (\mathcal{X}, τ) se dirá *completamente metrizable*. Con esto, un *espacio polaco* es un espacio topológico completamente metrizable y separable.

Al trabajar un espacio topológico (\mathcal{X}, τ) como un espacio medible, siempre se considerará, implícitamente, la σ -álgebra de sus *borelianos*, $\mathcal{B}(\mathcal{X})$, la cual se define como la menor σ -álgebra que contiene a τ . Así, una medida sobre esta σ -álgebra se denominará *medida de Borel*.

En transporte óptimo y en teoría de probabilidades en general, es usual enunciar los teoremas asumiendo que los espacios involucrados son, al menos, espacios polacos, ya que este tipo de espacios es lo suficientemente general como para cubrir un amplio espectro de problemas, y evita algunos casos patológicos encontrados en formulaciones más generales. En particular, tanto el caso discreto $\mathcal{X} = \{x_1, \dots, x_n\}$ como el caso continuo $\mathcal{X} = \mathbb{R}^d$ son espacios polacos.

A.1.2. Existencia de funciones de densidad

Cuando se consideran variables aleatorias con valores en \mathbb{R}^d (como las estudiadas en el Capítulo 1), estas inducen naturalmente una medida de probabilidad en \mathbb{R}^d , donde la masa que asigna dicha medida a cada conjunto medible $A \in \mathcal{B}(\mathbb{R}^d)$ corresponde a la probabilidad de que la variable aleatoria tome ese valor. Por ejemplo, dada una variable aleatoria gaussiana $z \sim \mathcal{N}(0, 1)$, la medida que esta induce sobre \mathbb{R} viene dada por:

$$\mu_z(A) = \int_A p_z(x) \, dx \quad (\text{A.1.3})$$

donde $A \in \mathcal{B}(\mathbb{R}^d)$ es un conjunto medible de \mathbb{R} (por ejemplo, un intervalo) y p_z es la función de densidad de una variable aleatoria gaussiana estándar. En este caso, se dice que la medida $\mu_z \in \mathcal{M}_+^1(\mathbb{R})$ es *absolutamente continua* con respecto a la medida de Lebesgue dx ya que existe una función de densidad p_z que permite escribir μ_z como una integral con respecto a la medida de Lebesgue.

Sin embargo, esto no es siempre posible. Por ejemplo, si la variable aleatoria z concentrarse toda su masa en un único punto $x_0 \in \mathbb{R}$, no sería posible encontrar una función $p_z : \mathbb{R} \rightarrow \mathbb{R}$ para obtener la igualdad dada en (A.1.3) debido a que la integral de Riemann (vinculada a la medida de Lebesgue) no es capaz de asignarle masa a puntos individuales ya que, la medida de Lebesgue del intervalo formado por un único punto tiene medida cero (debido a que tiene longitud cero).

Este problema de la no existencia de una función de densidad se tendrá siempre que la medida a la que se le busca densidad le asigne masa a elementos demasiado pequeños con respecto a la medida con la que se está comparando. Esta observación motiva la siguiente definición:

Definición A.3 (continuidad absoluta de medidas). Dado un espacio medible $(\mathcal{X}, \mathcal{S})$ y dos medidas de probabilidad $\mu, \nu \in \mathcal{M}_+^1(\mathcal{X})$, se dirá que μ es *absolutamente continua (a.c.)* con respecto a ν si

$$\mu(A) = 0 \implies \nu(A) = 0, \quad \forall A \in \mathcal{S}$$

Esta propiedad se denotará como $\mu \ll \nu$.

En los ejemplos anteriores, la medida asociada a la variable aleatoria gaussiana es absolutamente continua con respecto a la medida de Lebesgue, mientras que la medida que concentra toda su masa en un punto no lo es.

El siguiente resultado indica que si $\mu \ll \nu$, entonces μ posee una función de densidad con respecto a ν :

Teorema A.1 (Radon-Nikodym). Sea $(\mathcal{X}, \mathcal{S})$ un espacio medible y $\mu, \nu \in \mathcal{M}_+^1(\mathcal{X})$ dos medidas de probabilidad con $\mu \ll \nu$. Entonces, existe una función $f : \mathcal{X} \rightarrow \mathbb{R}_+$ denominada *función de densidad*, tal que:

$$\mu(A) = \int_{\mathcal{X}} f \, d\nu, \quad \forall A \in \mathcal{S}$$

La función f es única y es denominada *derivada de Radon-Nikodym de μ con respecto a ν* , por lo que se suele denotar como $\frac{d\mu}{d\nu}$.

A.1.3. Medida push-forward

Para concluir esta sección, se revisará el concepto de *medida push-forward*, el cual consiste en traspasar la medida desde un espacio de probabilidad a un espacio medible cualquiera mediante una función entre ambos espacios. Esta definición será crucial al momento de definir el problema de Monge en la próxima sección.

Definición A.4 (Medida push-forward). Sea $T : \mathcal{X} \rightarrow \mathcal{Y}$ una función entre el espacio de probabilidad $(\mathcal{X}, \mathcal{S}_{\mathcal{X}}, \mu)$ y el espacio medible $(\mathcal{Y}, \mathcal{S}_{\mathcal{Y}})$. La medida push-forward $T_{\#}\mu \in \mathcal{M}_+^1(\mathcal{Y})$ que induce μ sobre \mathcal{Y} mediante T se define como:

$$T_{\#}\mu(B) := (T_{\#}\mu)(B) = \mu(T^{-1}(B)) = \mu(\{x \in \mathcal{X} : T(x) \in B\}), \quad \forall B \in \mathcal{S}_{\mathcal{Y}}$$

Es decir, la función $T : \mathcal{X} \rightarrow \mathcal{Y}$ construye una medida en el espacio de llegada a partir de la medida en el espacio de origen mediante preimágenes. Notar que si $T : \mathcal{X} \rightarrow \mathcal{Y}$ es inyectiva, la definición anterior es equivalente a que $\mu(A) = \nu(T(A)), \forall A \in \mathcal{S}_{\mathcal{X}}$.

Se tiene la siguiente caracterización para saber si una medida es efectivamente la medida push-forward:

Proposición A.1. Sea $T : \mathcal{X} \rightarrow \mathcal{Y}$ una función entre el espacio de probabilidad $(\mathcal{X}, \mathcal{S}_{\mathcal{X}}, \mu)$ y el espacio medible $(\mathcal{Y}, \mathcal{S}_{\mathcal{Y}})$. Entonces, $\nu \in \mathcal{M}_+^1(\mathcal{Y})$ es la medida push-forward inducida por μ mediante T si y solo si

$$\int_{\mathcal{Y}} h(y) \, d\nu = \int_{\mathcal{X}} h(T(x)) \, d\mu, \quad \forall h \in \mathcal{C}(\mathcal{Y})$$

donde $\mathcal{C}(\mathcal{Y})$ representa el conjunto de funciones $h : \mathcal{Y} \rightarrow \mathbb{R}$ continuas en \mathcal{Y} .

Una forma de interpretar la medida push-forward es notando que el mapa $T : \mathcal{X} \rightarrow \mathcal{Y}$ es una función que traslada puntos entre dos espacios de medida, mientras que $T_{\#} : \mathcal{M}_+(\mathcal{X}) \rightarrow \mathcal{M}_+(\mathcal{Y})$ es una extensión de T que traslada medidas en \mathcal{X} a medidas en \mathcal{Y} .

Se tienen las siguientes propiedades de las medidas push-forward:

Proposición A.2. Sean \mathcal{X} e \mathcal{Y} espacios medibles y $\mu \in \mathcal{M}_+^1(\mathcal{X})$. Las siguientes afirmaciones son ciertas:

- El operador $T_\# : \mathcal{M}_+^1(\mathcal{X}) \rightarrow \mathcal{M}_+^1(\mathcal{Y})$ es lineal:

$$T_\#(\mu_1 + \mu_2) = T_\#\mu_1 + T_\#\mu_2$$

donde $\mu_1, \mu_2 \in \mathcal{M}_+^1(\mathcal{X})$.

- Fórmula de cambio de variable:

$$\int_{\mathcal{Y}} f(y) \, d(T_\#\mu)(y) = \int_{\mathcal{X}} f(T(x)) \, d\mu(x)$$

- Composición de mapas de transporte:

$$(S \circ T)_\# \mu = S_\#(T_\#\mu)$$

La demostración de estas propiedades se puede encontrar en [Tho18].

A.2. Procesos de $\hat{\text{ITO}}$

En esta sección se dará una introducción informal a las ecuaciones diferenciales estocásticas, mostrando por qué es necesario construir una nueva teoría para su desarrollo. Un desarrollo más riguroso se puede encontrar en [SS19], [Eva13] y [KS88]. Para comenzar, se entrega la siguiente jerarquía para los procesos estocásticos, la cual es importante para tener en cuenta las propiedades de los procesos que se están utilizando tanto en los modelos de difusión como en la formulación continua del transporte óptimo y el problema de Schrödinger en los Capítulo 2 y Capítulo 3 respectivamente:

- Un proceso estocástico es cualquier proceso aleatorio $x = (x_t)_{t \geq 0}$ indexado por el tiempo. Entre estos procesos están los procesos de Markov, los cuales son procesos cuyo futuro solo depende del presente y pueden ser procesos a tiempo discreto o a tiempo continuo. Debido a esto último, no todos los procesos de Markov tienen necesariamente una función de densidad marginal (ver Subsección A.1.2).
- Dentro de los procesos de Markov están los procesos de $\hat{\text{ITO}}$, los cuales son los procesos de Markov a tiempo continuo que se pueden representar mediante una ecuación diferencial estocástica (SDE). Por otra parte, no todos los procesos de Markov a tiempo continuo se pueden representar mediante una SDE (por ejemplo, un proceso de Poisson es un proceso de Markov pero no un proceso de $\hat{\text{ITO}}$).
- Un subconjunto particular de los procesos de $\hat{\text{ITO}}$ son los procesos de difusión, los cuales son procesos de $\hat{\text{ITO}}$ guiados por un movimiento browiano (i.e. poseen una SDE de la forma $\mu(x_t, t)dt + \sigma(x_t, t)dW$, ver Definición A.6). En particular, estos procesos tienen trayectorias continuas. Un proceso de $\hat{\text{ITO}}$ que no es un proceso de difusión es, por ejemplo, un proceso de Lévy, los cuales tienen SDEs de la forma $dx_t = (\mu dt + \sigma dW_t) + dJ_t$ donde el primer sumando es la parte difusiva y el segundo sumando es el que genera las discontinuidades en las trayectorias. En este trabajo, dado que se consideran únicamente trayectorias continuas, se utilizarán equivalentemente los términos *proceso de $\hat{\text{ITO}}$* y *proceso de difusión*.

Las SDEs son una herramienta poderosa para modelar sistemas dinámicos influenciados por componentes aleatorias. Estos modelos son esenciales en diversas áreas como física, biología, economía e ingeniería, donde los sistemas estudiados no pueden ser descritos adecuadamente solo por ecuaciones diferenciales ordinarias (ODEs) debido a la presencia de ruido o incertidumbre.

Para motivar su estudio, se puede considerar un sistema físico en \mathbb{R}^d que evoluciona durante un intervalo de tiempo $[t_0, T]$ de acuerdo a la siguiente dinámica:

$$\frac{dx(t)}{dt} = \mu(x(t), t) + \sigma(x(t), t)w(t) \quad (\text{A.2.1})$$

Donde $\mu : \mathbb{R}^d \times [t_0, T] \rightarrow \mathbb{R}^d$ es un forzante o estímulo determinista y $\sigma : \mathbb{R}^d \times [t_0, T] \rightarrow \mathcal{M}_{d,d}(\mathbb{R})$ es un factor matricial del forzante σw , el cual depende de una cantidad aleatoria $w : [0, T] \rightarrow \mathbb{R}^d$. De este modo, el sistema x tiene una componente evolutiva fija y otra estocástica, la cual puede representar incertidumbre sobre el forzante total o estímulos externos que no están considerados en el forzante μ .

En el campo de las SDEs, la función μ es denominada *drift*, mientras que la función σ se conoce como *dispersión* o *volatilidad*. El factor estocástico w es un término que representa la incertidumbre y es comúnmente considerado como *ruido blanco*. Para hacer más preciso este último término, se definirá de la siguiente forma:

Definición A.5 (ruido blanco). El ruido blanco es un proceso estocástico $w = (w_t)_{t \geq 0}$ en \mathbb{R}^d con las siguientes propiedades:

- Para dos tiempos distintos $t_1, t_2 > 0$, las variables aleatorias w_{t_1} y w_{t_2} son independientes.
- El mapeo $t \mapsto w(t)$ es un proceso gaussiano con media nula y correlación $\mathbb{E}[w(t_1)w(t_2)^\top] = \delta(t_1 - t_2)Q$, donde Q se conoce como la densidad espectral del proceso.

De la definición se obtiene que las trayectorias $t \mapsto w(t)$ son discontinuas casi seguramente (c.s.) y que el ruido blanco es no acotado. Estas propiedades no permiten analizar la ecuación diferencial definida en (A.2.1) de una forma usual ya que no cumple los requisitos del teorema clásico de existencia y unicidad para ODEs:

Teorema A.2 (Cauchy-Picard-Lindelöf). Sea $I \subset \mathbb{R}$ un intervalo cerrado y acotado y $f : I \times \mathbb{R}^d \rightarrow \mathbb{R}$ una función continua y Lipschitz con respecto a la segunda variable. Entonces, para cada $t_0 \in I$ y cada $x_0 \in \mathbb{R}^d$ existe una única solución $x : I \rightarrow \mathbb{R}^d$ de clase \mathcal{C}^1 del problema de Cauchy

$$\begin{cases} x'(t) = f(t, x(t)), & \forall t \in I \\ x(t_0) = x_0 \end{cases}$$

Dada la discontinuidad casi segura del ruido blanco, no se cumple la continuidad del lado derecho de (A.2.1) requerida en el Teorema A.2. Sin embargo, es posible realizar iteraciones de Picard análogas a la forma clásica para obtener la existencia y unicidad de la solución en el caso estocástico.

Como se verá en la siguiente subsección, este tipo de ecuaciones tampoco permitirán utilizar las integrales clásicas para este tipo de ecuaciones, por lo que se necesitará definir un nuevo concepto de integral.

A.2.1. Integral de Itô

La integral de Itô es un concepto fundamental en el cálculo estocástico, particularmente en el estudio de SDEs. A diferencia de las integrales de Riemann o de Lebesgue, la integral de Itô está diseñada para integrar con respecto a procesos estocásticos, especialmente con respecto al movimiento browniano, el cual se define a continuación.

Definición A.6 (movimiento browniano estándar). El movimiento browniano o proceso de Wiener es un proceso estocástico $W = (W_t)_{t \geq 0}$ en \mathbb{R}^d que está caracterizado⁶ por 4 propiedades:

- Inicio determinista: $W_0 = 0$ casi seguramente.

⁶Existen otras caracterizaciones equivalentes. Por ejemplo, se puede definir como el límite escalado de un paseo aleatorio (teorema de Donkster) o como una martingala continua con variación cuadrática (caracterización de Lévy).

- Incrementos independientes: para tiempos $t_1 \leq t_2 < t_3 \leq t_4$, los incrementos $W_{t_2} - W_{t_1}$ y $W_{t_4} - W_{t_3}$ son independientes.
- Incrementos gaussianos: para tiempos $t_1 < t_2$, $W_{t_2} - W_{t_1} \sim \mathcal{N}(0, (t_2 - t_1)I_d)$.
- Trayectorias continuas: el mapa $t \mapsto W_t$ es continuo casi seguramente.

De esta definición, se observa que un movimiento browniano es un proceso gaussiano con función de media idénticamente nula y función de covarianza $K(t_1, t_2) = \min(t_1, t_2)I_d$.

Para resolver la ecuación diferencial definida en (A.2.1), se podría tratar de integrar, al menos formalmente, dicha expresión en un intervalo de tiempo $[t_0, T]$. De esta forma:

$$x_T - x_{t_0} = \int_{t_0}^T \mu(x_t, t) dt + \int_{t_0}^T \sigma(x_t, t) w_t dt \quad (\text{A.2.2})$$

La primera integral es una integral clásica y puede ser resuelta en el sentido de Riemann o en el sentido de Lebesgue. Por otra parte, la segunda no permite ninguna de las dos interpretaciones anteriores, en efecto:

Como integral de Riemann-Darboux Para una partición $\mathcal{P} = (t_k)_{k=0}^n$ del intervalo $[t_0, T]$, se pueden estudiar las sumas de Riemann asociadas al integrando $\sigma(x_t, t)w_t$, las cuales son de la forma

$$\sum_{k=1}^n \sigma(x(\tilde{t}_k), \tilde{t}_k) w(\tilde{t}_k) (t_k - t_{k-1})$$

donde $\tilde{t}_k = \sup_{t \in [t_{k-1}, t_k]} \sigma(x_t, t)w_t$ en el caso de la suma superior y $\tilde{t}_k = \inf_{t \in [t_{k-1}, t_k]} \sigma(x_t, t)w_t$ en el caso de la suma inferior. Dado que el ruido blanco es no acotado en cualquier intervalo, los tiempos \tilde{t}_k no están definidos para ninguna partición \mathcal{P} independientemente de que $\|\mathcal{P}\| := \max_{1 \leq k \leq n} t_k - t_{k-1}$ tienda a cero. De esta forma, no es posible definir la segunda integral en (A.2.2) en el sentido de Riemann.

Como integral de Lebesgue El incremento $w_t dt$ en (A.2.2) corresponde precisamente al incremento de un movimiento browniano W . Tratando a este proceso como una medida de Lebesgue-Stieltjes⁷ del tipo $\mu_W((t_{k-1}, t_k]) = W_{t_k} - W_{t_{k-1}}$, la segunda integral en (A.2.2) vista en el sentido de Lebesgue toma la forma

$$\int_{t_0}^T \sigma(x_t, t) w_t dt = \int_{t_0}^T \sigma(x_t, t) dW_t = \lim_{\|\mathcal{P}\| \rightarrow 0} \sum_{k=1}^n \sigma(x(\tilde{t}_k), \tilde{t}_k) (W_{t_k} - W_{t_{k-1}}) \quad (\text{A.2.3})$$

Interpretar esta integral en el sentido de Lebesgue tampoco es posible ya que dicho límite no es único (el movimiento browniano no es de variación acotada). Por lo tanto, no es posible definir la integral estocástica en (A.2.2) en el sentido de Lebesgue. En particular, tampoco se podría definir como una integral de Riemann-Stieltjes. Sin embargo, de este análisis se rescata que $w_t dt = dW_t$, lo cual muchas veces es interpretado diciendo que el ruido blanco es la *derivada* del movimiento browniano. Sin embargo, esto es únicamente a modo de interpretación ya que el movimiento browniano es no diferenciable casi en todas partes.

En consecuencia, para poder darle sentido a la segunda integral en (A.2.2) es necesario poder definir una integral que permita diferenciales estocásticos. Para esto, se puede fijar la elección $\tilde{t}_k = t_k$ para lograr que el límite en (A.2.3) sí sea único (en media cuadrática). La construcción de esta nueva integral pasa a llamarse *integral de Itô*⁸, la cual toma la siguiente forma:

⁷Formalmente esto no tiene sentido ya que las medidas son funciones deterministas.

⁸También es posible definir la integral de Stratonovich considerando a \tilde{t}_k como el punto medio del intervalo $[t_{k-1}, t_k]$, la cual permite que las reglas del cálculo clásico sigan siendo válidas. Sin embargo, esta integral dificulta el análisis teórico ya que no

$$\int_{t_0}^T \sigma(x_t, t) dW_t = \lim_{\|\mathcal{P}\| \rightarrow 0} \sum_{k=1}^n \sigma(x(t_k), t_k) (W_{t_k} - W_{t_{k-1}})$$

Bajo esta nueva integral, la ecuación diferencial inicial (A.2.1) se denotará como

$$dx_t = \mu(x_t, t) dt + \sigma(x_t, t) dW_t \quad (\text{A.2.4})$$

Y pasará a llamarse *ecuación diferencial estocástica*, cuya solución corresponde al proceso estocástico dado por

$$x_t = x_{t_0} + \int_{t_0}^t \mu(x_s, s) ds + \int_{t_0}^t \sigma(x_s, s) dW_s$$

De aquí en adelante, cuando se hable de una SDE genérica se estará asumiendo que tiene la forma dada en (A.2.4). Notar que la solución de una SDE corresponde a un proceso aleatorio, por lo que para obtener una trayectoria específica es necesario poder generar muestras a partir de su distribución. En la figura A.1 se observan algunas trayectorias generadas a partir de la solución de una SDE específica.

Una herramienta fundamental para trabajar con SDEs es el lema o fórmula de Itô, el cual proporciona una manera de diferenciar funciones de procesos estocásticos. Esta fórmula es el análogo estocástico a la regla de la cadena en el cálculo clásico, y es crucial para manipular ecuaciones diferenciales estocásticas. En consecuencia, este resultado permite demostrar todos los resultados sobre SDEs utilizados a lo largo de este trabajo.

A modo de completitud, se enunciará este resultado en el caso unidimensional ($d = 1$):

Teorema A.3 (lema de Itô, caso escalar). Dado un proceso de Itô en \mathbb{R} , solución de la SDE $dx_t = \mu(x_t, t) dt + \sigma(x_t, t) dW_t$ en el intervalo temporal $[t_0, T]$ y una función $\phi : \mathbb{R} \times [t_0, T]$ dos veces diferenciables, entonces, el proceso dado por $\phi_t = \phi(x_t, t)$ también es un proceso de Itô asociado a la SDE

$$d\phi_t = \left(\frac{\partial \phi}{\partial t} + \mu_t \frac{\partial \phi}{\partial x} + \frac{1}{2} \sigma_t^2 \frac{\partial^2 \phi}{\partial x^2} \right) dt + \sigma_t \frac{\partial \phi}{\partial x} dW_t. \quad (\text{A.2.5})$$

donde $\mu_t = \mu(x_t, t)$ y $\sigma_t = \sigma(x_t, t)$, y las derivadas son evaluadas en (x_t, t) .

Notar que la presencia del término con la segunda derivada no tiene un análogo directo en el cálculo clásico, siendo un reflejo de la naturaleza estocástica del proceso $x = (x_t)_{t \in [0, T]}$. Por otra parte, este lema se puede extender a procesos de Itô en \mathbb{R}^d pero se omite su formulación por simplicidad.

A.2.2. Resultados necesarios

En esta subsección se entregarán sin demostración algunos resultados necesarios para el desarrollo de la teoría de los modelos de difusión a tiempo continuo. Además, estos resultados serán utilizados en la formulación dinámica del transporte óptimo en la Sección 2.3, y en la formulación dinámica del problema del puente de Schrödinger en la Sección 3.3.

Ecuación de Fokker-Planck

La ecuación de Fokker-Planck, también conocida como *ecuación forward de Kolmogorov*, es una ecuación en derivadas parciales (PDE) que describe la evolución temporal de la función de densidad de un proceso

cumple algunas propiedades esenciales que sí cumple la integral de Itô.

estocástico asociado a una SDE de la forma (A.2.4). Por simplicidad, solo se entregará el resultado en el caso unidimensional.

Teorema A.4 (ecuación de Fokker-Planck, caso escalar). Dada una SDE escalar ($d = 1$) de la forma

$$dx_t = \mu(x_t, t) dt + \sigma(x_t, t) dW_t$$

Entonces, la evolución de la función de densidad de probabilidad de x_t , $p(x, t) := p(x_t)$, viene dada por la solución de la PDE

$$\frac{\partial p(x, t)}{\partial t} = -\frac{\partial}{\partial x}(\mu(x, t)p(x, t)) + \frac{1}{2}\frac{\partial^2}{\partial x^2}[\sigma^2(x, t)p(x, t)]. \quad (\text{A.2.6})$$

Al igual que para la fórmula de Itô en (A.3), este resultado puede ser extendido para procesos de Itô en \mathbb{R}^d . Es importante destacar que si el término de difusión $\sigma(x_t, t)$ es nulo, la ecuación de Fokker-Planck se reduce a la ecuación de transporte (2.3.11) utilizada en la formulación dinámica del transporte óptimo (ver formulación de Benamou-Brenier en la Subsección 2.3.3). Posteriormente, en el Capítulo 3, la ecuación de Fokker-Planck será crucial para definir el sistema de Schrödinger, el cual caracterizará la solución al problema del puente de Schrödinger dinámico entre dos medidas de probabilidad.

A modo de ejemplo, se aplicará el Teorema A.4 sobre el movimiento browniano estándar, cuya SDE es $dx_t = dW_t$. En este caso, la ecuación de Fokker-Planck se reduce a

$$\frac{\partial p(x, t)}{\partial t} = \frac{1}{2}\frac{\partial^2 p(x, t)}{\partial x^2}.$$

La cual es una ecuación de difusión (en su forma más simple), cuya solución se conoce en forma cerrada. En efecto, si se tiene una condición inicial $p(x, 0) = \delta(x)$ (i.e., la el proceso estocástico $(x_t)_{t>0}$ comienza de forma determinista en el punto $0 \in \mathbb{R}$), la solución de esta PDE es:

$$p(x, t) = \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{x^2}{2t}\right)$$

Por lo que la densidad marginal $p(x_t)$ del movimiento browniano $dx_t = dW_t$ (comenzando en $x_0 = 0$) es precisamente la densidad de una variable aleatoria gaussiana $\mathcal{N}(0, t)$, lo cual es lo esperado de acuerdo a la Definición A.6.

Probability flow ODE

El siguiente teorema enuncia la ecuación dada en Ecuación 1.4.14 con más generalidad:

Teorema A.5 (probability flow ODE). Dado un proceso estocástico solución de la SDE

$$dx_t = \mu(x_t, t) dt + \sigma(x_t, t) dw_t, \quad x_0 \sim p_0(x_0)$$

donde $\mu : \mathbb{R}^d \times [0, 1] \rightarrow \mathbb{R}$ y $\sigma : \mathbb{R}^d \times [0, 1] \rightarrow \mathcal{M}_{d,d}(\mathbb{R})$, el siguiente proceso determinista tiene las mismas densidades marginales $(p_t)_{t \leq 0}$ que el proceso estocástico asociado a la SDE (A.5):

$$\frac{dx_t}{dt} = \tilde{\mu}(x_t, t), \quad x_0 \sim p_0(x_0)$$

donde

$$\tilde{\mu}(x_t, t) = \mu(x, t) - \frac{1}{2}\nabla \cdot (\sigma(x, t)\sigma(x, t)^\top) - \frac{1}{2}\sigma(x, t)\sigma(x, t)^\top \nabla_x \log p_t(x)$$

Teorema de inversión

En algunos casos, como en modelos de difusión basados en SDEs, es necesario poder invertir temporalmente un proceso estocástico. El siguiente resultado indica que, bajo condiciones razonables, esto es posible y más aún, entrega la forma que tendrá la SDE asociada al proceso inverso.

Teorema A.6 (inversión de Anderson). Dado un proceso estocástico $(x_t)_{t \in [0, T]}$ asociado a la SDE $dx_t = \mu(x_t, t)dt + \sigma(x_t, t)dW_t$ tal que μ y σ son lo suficientemente regulares como para que la densidad marginal $p_t(x_t)$ sea la única solución suave de la ecuación de Fokker-Planck asociada. Entonces, existe un único proceso reverso $(Y_t)_{t \in [0, T]} = (X_{T-t})_{t \in [0, T]}$ cuya SDE viene dada por:

$$dy_t = [f(y_t, t) - \nabla_{y_t} \cdot (\sigma(x_t, t)\sigma(x_t, t)^\top) - \sigma(x_t, t)\sigma(x_t, t)^\top \nabla_{y_t} \log p_{T-t}(y_t)] dt + \sigma(y_t, t)d\bar{w}_t$$

donde \bar{w} es un movimiento browniano fluyendo hacia atrás en el tiempo.

La demostración de este teorema puede ser encontrada en [And82].

A.2.3. Procesos de Ornstein–Uhlenbeck

Los procesos de Ornstein–Uhlenbeck son procesos estocásticos cruciales en diversas áreas de la ciencia y la economía. Estos procesos son fundamentales para modelar fenómenos que tienden a regresar a un valor medio o de equilibrio a lo largo del tiempo. Si bien forman una familia simple de proceso de estocástico, se les dedica una subsección aparte ya que son ampliamente utilizados en modelos generativos basados en difusión.

Un proceso de Ornstein–Uhlenbeck (centrado en 0) está asociado a una SDE lineal de la forma

$$dx_t = -\alpha x_t dt + \sigma dW_t$$

donde $\alpha, \sigma > 0$ son constantes. Como se verá, el parámetro α controla la tasa de reversión hacia la media, mientras que σ determina la volatilidad del proceso. Para esta SDE, la ecuación de Fokker-Planck (A.2.6) correspondiente es:

$$\frac{\partial p(x, t)}{\partial t} = a \frac{\partial}{\partial x} (x p(x, t)) + \frac{\sigma^2}{2} \frac{\partial^2 p(x, t)}{\partial x^2}$$

Además, considerando $\partial_t p = 0$ es posible encontrar la distribución estacionaria de este tipo de procesos:

$$0 = a \frac{\partial}{\partial x} (x p(x, t)) + \frac{\sigma^2}{2} \frac{\partial^2 p(x, t)}{\partial x^2} \implies p_{\text{est}}(x) = \sqrt{\frac{a}{\pi \sigma^2}} \exp\left(-\frac{ax^2}{\sigma^2}\right)$$

En caso de necesitar la densidad marginal $p(x_t)$ de forma cerrada para este tipo de procesos, la fórmula de Itô (A.2.5) permite obtener la siguiente solución:

$$x_t = e^{-\alpha t} x_0 + \sigma \int_0^t e^{-\alpha(t-s)} dW_s$$

donde x_0 es la condición inicial y la integral se debe interpretar en el sentido de Itô.

Propiedades Una propiedad importante de este tipo de procesos es la reversión a la media, lo que significa que a tiempos largos, el proceso se estabilizará en 0, independientemente de la condición inicial. Más generalmente, es posible estabilizar el proceso a otro valor μ considerando la SDE

$$dx_t = \alpha(\mu - x_t)dt + \sigma dW_t$$

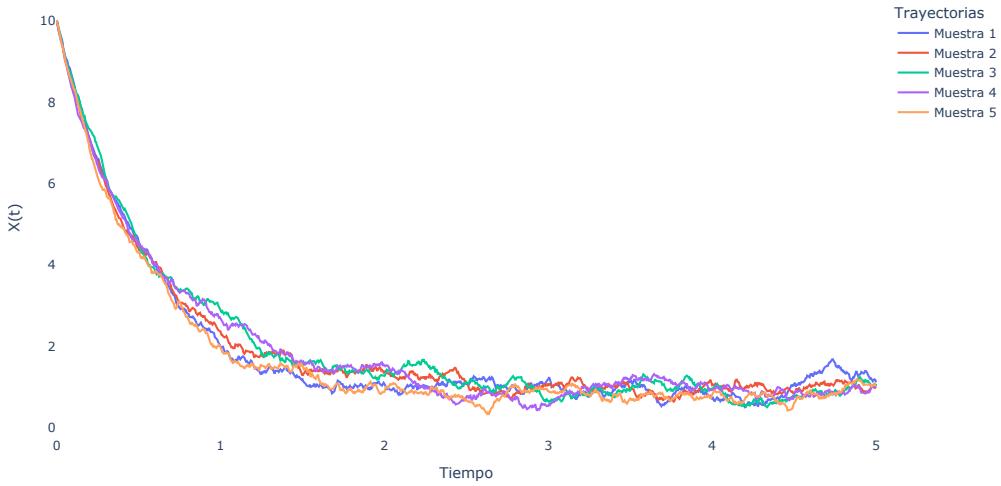


Figura A.1: Simulación mediante el algoritmo de Euler-Maruyama (ver Algoritmo 7) de 5 trayectorias para el proceso de Ornstein–Uhlenbeck $dx_t = 2(1 - x_t)dt + \frac{1}{2}dW_t$ en el intervalo temporal $[0, 5]$ y con condición inicial $x_0 = 10$. Esta simulación se encuentra en el archivo `sdes.ipynb`.

Los procesos de Ornstein–Uhlenbeck son unos de los procesos más simples que poseen esta propiedad, por lo que son ampliamente utilizados para modelar fenómenos en distintas áreas. En la Figura A.1 se realizó una simulación de este tipo de procesos utilizando el método de Euler-Maruyama dado en el Algoritmo 7, aunque es posible utilizar otros métodos de simulación más eficientes.

Por otra parte, los procesos de Ornstein–Uhlenbeck cumplen que, dada la condición inicial, sus distribuciones marginales $x_t|x_0$ distribuyen normalmente, por lo que el proceso completo es un proceso gaussiano. Además, este tipo de procesos poseen la propiedad de Markov (el futuro del proceso solo depende de su estado actual y no del pasado). Se puede probar que los procesos de Ornstein–Uhlenbeck son los únicos que cumplen las 3 propiedades nombradas hasta ahora.

Es importante destacar que la propiedad de reversión a la media junto a la de ser un proceso gaussiano es más fuerte aún ya que este tipo de procesos es ergódico y tiene una distribución invariante gaussiana.

Una observación adicional es que el análogo en tiempo discreto del proceso de Ornstein–Uhlenbeck es el proceso autorregresivo de primer orden AR(1), el cual viene dado por

$$x_{t+1} = c + \phi x_t + \epsilon_t$$

donde $|\phi| < 1$ es el parámetro de reversión a la media y ϵ_t es un término de error con distribución normal. Este paralelismo se aprovecha para la estimación y simulación de procesos de Ornstein–Uhlenbeck en situaciones donde solo se dispone de datos en intervalos discretos de tiempo, como es común en series temporales financieras.

Aplicaciones Dado lo simple de este tipo de modelos, es común utilizarlos en distintas ciencias con el fin de poder analizar fenómenos aleatorios. A continuación se enumeran algunas aplicaciones de los procesos de Ornstein–Uhlenbeck:

- **Economía:** bajo la premisa de que las tasas de interés no pueden crecer indefinidamente y que a largo plazo se deben estabilizar en un valor medio, es posible modelar su evolución mediante este tipo de procesos (modelo de Vasicek). Bajo el mismo argumento se pueden utilizar procesos de Ornstein–Uhlenbeck

para modelar el PIB de un país o tasas de cambio entre dos monedas.

- **Biología:** se puede usar este tipo de procesos para modelar la evolución de una población, donde se asume que hay un valor de estabilización que se alcanzará en el largo plazo.
- **Física:** originalmente este tipo de procesos se usó para modelar el movimiento aparentemente aleatorio de una partícula (modelo de Einstein) bajo un proceso de disipación de energía. El ejemplo más típico es el de un resorte, el cual se rige por la ley de Hooke, sometido a una fuerza de roce. En el largo plazo, se espera que el resorte vuelva a su condición de equilibrio.

Bibliografía

- [AB09] Sanjeev Arora y Boaz Barak. *Computational Complexity: A Modern Approach*. Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo: Cambridge University Press, 2009. ISBN: 978-0-521-42426-4. URL: <http://www.cambridge.org/9780521424264>.
- [ACB17] Martin Arjovsky, Soumith Chintala y Léon Bottou. *Wasserstein GAN*. 2017. arXiv: 1701.07875 [stat.ML]. URL: <https://arxiv.org/abs/1701.07875>.
- [And82] B. D. O. Anderson. «Reverse-time diffusion equation models». En: *Stochastic Process. Appl.* 12.3 (1982).
- [BDS19] Andrew Brock, Jeff Donahue y Karen Simonyan. *Large Scale GAN Training for High Fidelity Natural Image Synthesis*. 2019. arXiv: 1809.11096 [cs.LG]. URL: <https://arxiv.org/abs/1809.11096>.
- [Bet+22] Eyal Begale et al. *A Study on the Evaluation of Generative Models*. 2022. arXiv: 2206.10935 [cs.LG]. URL: <https://arxiv.org/abs/2206.10935>.
- [Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
- [BKM17] David M. Blei, Alp Kucukelbir y Jon D. McAuliffe. «Variational Inference: A Review for Statisticians». En: *Journal of the American Statistical Association* 112.518 (abr. de 2017), págs. 859-877. ISSN: 1537-274X. DOI: 10.1080/01621459.2017.1285773. URL: <http://dx.doi.org/10.1080/01621459.2017.1285773>.
- [Bor+23a] Valentin De Bortoli et al. *Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modeling*. 2023. arXiv: 2106.01357 [stat.ML].
- [Bor+23b] Valentin De Bortoli et al. *Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modeling*. 2023. arXiv: 2106.01357 [stat.ML]. URL: <https://arxiv.org/abs/2106.01357>.
- [Bor23] Valentin De Bortoli. *Convergence of denoising diffusion models under the manifold hypothesis*. 2023. arXiv: 2208.05314 [stat.ML]. URL: <https://arxiv.org/abs/2208.05314>.
- [Bun23a] Charlotte Bunne. «Neural Optimal Transport for Dynamical Systems: Methods and Applications in Biomedicine». PhD Thesis. Zurich, Switzerland: ETH Zurich, 2023.
- [Bun23b] Charlotte Bunne. *Optimal Transport in Learning, Control, and Dynamical Systems*. ICML Tutorial. 2023.
- [BV04] Stephen Boyd y Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
- [CGP20] Yongxin Chen, Tryphon T. Georgiou y Michele Pavon. *Stochastic control liaisons: Richard Sinkhorn meets Gaspard Monge on a Schroedinger bridge*. 2020. arXiv: 2005.10963 [math.OC]. URL: <https://arxiv.org/abs/2005.10963>.
- [Che+19] Ricky T. Q. Chen et al. *Neural Ordinary Differential Equations*. 2019. arXiv: 1806.07366 [cs.LG]. URL: <https://arxiv.org/abs/1806.07366>.
- [Che23] Ting Chen. *On the Importance of Noise Scheduling for Diffusion Models*. 2023. arXiv: 2301.10972 [cs.CV]. URL: <https://arxiv.org/abs/2301.10972>.

- [CLT23] Tianrong Chen, Guan-Horng Liu y Evangelos A. Theodorou. *Likelihood Training of Schrödinger Bridge using Forward-Backward SDEs Theory*. 2023. arXiv: 2110.11291 [stat.ML]. URL: <https://arxiv.org/abs/2110.11291>.
- [Cut+17] Marco Cuturi et al. «Learning with Regularized Distances: Optimal Transport and Dynamic Time Warping». En: ENS et. al. 2017.
- [Cut13] Marco Cuturi. *Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances*. 2013. arXiv: 1306.0895 [stat.ML].
- [Dev+19] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL]. URL: <https://arxiv.org/abs/1810.04805>.
- [DG03] Sanjoy Dasgupta y Anupam Gupta. «An Elementary Proof of a Theorem of Johnson and Lindenstrauss». En: *Random Structures & Algorithms* 22.1 (2003). Received 16 December 2001; accepted 11 July 2002, págs. 60-65. DOI: 10.1002/rsa.10073.
- [DN21] Prafulla Dhariwal y Alex Nichol. *Diffusion Models Beat GANs on Image Synthesis*. 2021. arXiv: 2105.05233 [cs.LG].
- [Dos+21] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV].
- [Dud69] R. M. Dudley. «The Speed of Mean Glivenko-Cantelli Convergence». En: *Annals of Mathematical Statistics* 40.1 (1969), págs. 40-50. DOI: 10.1214/aoms/1177697802. URL: <https://doi.org/10.1214/aoms/1177697802>.
- [Erd23] Kemal Erdem. «Step by Step visual introduction to Diffusion Models». En: <https://erdem.pl> (2023). URL: <https://erdem.pl/2023/11/step-by-step-visual-introduction-to-diffusion-models>.
- [EUD17] Stefan Elfwing, Eiji Uchibe y Kenji Doya. *Sigmoid-Weighted Linear Units for Neural Network Function Approximation in Reinforcement Learning*. 2017. arXiv: 1702.03118 [cs.LG].
- [Eva13] Lawrence C. Evans. *An introduction to stochastic differential equations*. Providence, Rhode Island: American Mathematical Society, 2013. ISBN: 978-1-4704-1054-4.
- [FKW22] Ghazal Farhani, Alexander Kazachek y Boyu Wang. *Momentum Diminishes the Effect of Spectral Bias in Physics-Informed Neural Networks*. 2022. arXiv: 2206.14862 [cs.LG].
- [GGR22] Albert Gu, Karan Goel y Christopher Ré. *Efficiently Modeling Long Sequences with Structured State Spaces*. 2022. arXiv: 2111.00396 [cs.LG]. URL: <https://arxiv.org/abs/2111.00396>.
- [Goo+14] Ian J. Goodfellow et al. *Generative Adversarial Networks*. 2014. arXiv: 1406.2661 [stat.ML].
- [Haf18] Danijar Hafner. *Building Variational Auto-Encoders in TensorFlow*. Blog post. 2018. URL: <https://danijar.com/building-variational-auto-encoders-in-tensorflow/>.
- [HG23] Dan Hendrycks y Kevin Gimpel. *Gaussian Error Linear Units (GELUs)*. 2023. arXiv: 1606.08415 [cs.LG].
- [HJA20] Jonathan Ho, Ajay Jain y Pieter Abbeel. *Denoising Diffusion Probabilistic Models*. 2020. arXiv: 2006.11239 [cs.LG].
- [HS22] Jonathan Ho y Tim Salimans. *Classifier-Free Diffusion Guidance*. 2022. arXiv: 2207.12598 [cs.LG]. URL: <https://arxiv.org/abs/2207.12598>.
- [Hyv05] Aapo Hyvärinen. «Estimation of Non-Normalized Statistical Models by Score Matching». En: *Helsinki Institute for Information Technology (HIIT) Department of Computer Science* (2005).
- [IS15] Sergey Ioffe y Christian Szegedy. *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. 2015. arXiv: 1502.03167 [cs.LG].
- [Kap+20] Jared Kaplan et al. *Scaling Laws for Neural Language Models*. 2020. arXiv: 2001.08361 [cs.LG].
- [Kar+22] Tero Karras et al. *Elucidating the Design Space of Diffusion-Based Generative Models*. 2022. arXiv: 2206.00364 [cs.CV]. URL: <https://arxiv.org/abs/2206.00364>.

- [Khr+22] Valentin Khrulkov et al. *Understanding DDPM Latent Codes Through Optimal Transport*. 2022. arXiv: 2202.07477 [stat.ML]. URL: <https://arxiv.org/abs/2202.07477>.
- [Kol+17] Soheil Kolouri et al. «Optimal Mass Transport: Signal processing and machine-learning applications». En: *IEEE Signal Processing Magazine* 34.4 (2017), págs. 43-59.
- [KS88] I. Karatzas y S.E. Shreve. *Brownian Motion and Stochastic Calculus*. New York: Springer-Verlag Inc., 1988. DOI: 10.1007/978-1-4684-0302-2.
- [KW22] Diederik P Kingma y Max Welling. *Auto-Encoding Variational Bayes*. 2022. arXiv: 1312.6114 [stat.ML].
- [LC19] Gabriel Loaiza-Ganem y John P. Cunningham. *The continuous Bernoulli: fixing a pervasive error in variational autoencoders*. 2019. arXiv: 1907.06845 [stat.ML].
- [Léo13] Christian Léonard. *A survey of the Schrödinger problem and some of its connections with optimal transport*. 2013.
- [LGL22] Xingchao Liu, Chengyue Gong y Qiang Liu. *Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow*. 2022. arXiv: 2209.03003 [cs.LG]. URL: <https://arxiv.org/abs/2209.03003>.
- [Lip+23] Yaron Lipman et al. *Flow Matching for Generative Modeling*. 2023. arXiv: 2210.02747 [cs.LG]. URL: <https://arxiv.org/abs/2210.02747>.
- [LS22] Hugo Lavenant y Filippo Santambrogio. «The flow map of the Fokker–Planck equation does not provide optimal transport». En: *Applied Mathematics Letters* 133 (2022), pág. 108225. ISSN: 0893-9659. DOI: <https://doi.org/10.1016/j.aml.2022.108225>. URL: <https://www.sciencedirect.com/science/article/pii/S089396592200180X>.
- [Lu+22] Cheng Lu et al. *DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps*. 2022. arXiv: 2206.00927 [cs.LG]. URL: <https://arxiv.org/abs/2206.00927>.
- [Lu+23] Cheng Lu et al. *DPM-Solver++: Fast Solver for Guided Sampling of Diffusion Probabilistic Models*. 2023. arXiv: 2211.01095 [cs.LG]. URL: <https://arxiv.org/abs/2211.01095>.
- [MO14] Mehdi Mirza y Simon Osindero. *Conditional Generative Adversarial Nets*. 2014. arXiv: 1411.1784 [cs.LG].
- [Mur23] Kevin P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023. URL: <http://probml.github.io/book2>.
- [ND21] Alex Nichol y Prafulla Dhariwal. *Improved Denoising Diffusion Probabilistic Models*. 2021. arXiv: 2102.09672 [cs.LG].
- [Nic+22] Alex Nichol et al. *GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models*. 2022. arXiv: 2112.10741 [cs.CV]. URL: <https://arxiv.org/abs/2112.10741>.
- [Nut22] Marcel Nutz. «Introduction to Entropic Optimal Transport». Unpublished lecture notes. 2022.
- [PC20] Gabriel Peyré y Marco Cuturi. *Computational Optimal Transport*. 2020. arXiv: 1803.00567 [stat.ML].
- [PX23] William Peebles y Saining Xie. *Scalable Diffusion Models with Transformers*. 2023. arXiv: 2212.09748 [cs.CV]. URL: <https://arxiv.org/abs/2212.09748>.
- [Rad+21] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: 2103.00020 [cs.CV]. URL: <https://arxiv.org/abs/2103.00020>.
- [Raz+19] Ali Razavi et al. *Preventing Posterior Collapse with delta-VAEs*. 2019. arXiv: 1901.03416 [cs.LG].
- [RFB15] Olaf Ronneberger, Philipp Fischer y Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597 [cs.CV].

- [RMC16] Alec Radford, Luke Metz y Soumith Chintala. *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks*. 2016. arXiv: 1511.06434 [cs.LG].
- [Rom+22] Robin Rombach et al. *High-Resolution Image Synthesis with Latent Diffusion Models*. 2022. arXiv: 2112.10752 [cs.CV].
- [Sam22] Aleksandr Samarin. *Power of Diffusion Models*. 2022. URL: <https://astralord.github.io/posts/power-of-diffusion-models/>.
- [SE20] Yang Song y Stefano Ermon. *Generative Modeling by Estimating Gradients of the Data Distribution*. 2020. arXiv: 1907.05600 [cs.LG].
- [SH22] Tim Salimans y Jonathan Ho. *Progressive Distillation for Fast Sampling of Diffusion Models*. 2022. arXiv: 2202.00512 [cs.LG]. URL: <https://arxiv.org/abs/2202.00512>.
- [SK21] Yang Song y Diederik P. Kingma. *How to Train Your Energy-Based Models*. 2021. arXiv: 2101.03288 [cs.LG]. URL: <https://arxiv.org/abs/2101.03288>.
- [SME22] Jiaming Song, Chenlin Meng y Stefano Ermon. *Denoising Diffusion Implicit Models*. 2022. arXiv: 2010.02502 [cs.LG].
- [Soh+15] Jascha Sohl-Dickstein et al. *Deep Unsupervised Learning using Nonequilibrium Thermodynamics*. 2015. arXiv: 1503.03585 [cs.LG].
- [Son+19] Yang Song et al. *Sliced Score Matching: A Scalable Approach to Density and Score Estimation*. 2019. arXiv: 1905.07088 [cs.LG].
- [Son+21a] Yang Song et al. *Maximum Likelihood Training of Score-Based Diffusion Models*. 2021. arXiv: 2101.09258 [stat.ML]. URL: <https://arxiv.org/abs/2101.09258>.
- [Son+21b] Yang Song et al. *Score-Based Generative Modeling through Stochastic Differential Equations*. 2021. arXiv: 2011.13456 [cs.LG].
- [Son+23] Yang Song et al. *Consistency Models*. 2023. arXiv: 2303.01469 [cs.LG]. URL: <https://arxiv.org/abs/2303.01469>.
- [SS19] Simo Särkkä y Arno Solin. *Applied Stochastic Differential Equations*. Cambridge University Press, 2019.
- [Tho18] Matthew Thorpe. *Introduction to Optimal Transport*. Current Version: Thursday 8th March, 2018. F2.08, Centre for Mathematical Sciences, University of Cambridge, 2018.
- [Tur21] Angus Turner. *Diffusion Models as a kind of VAE*. https://angusturner.github.io/generative_models/2021/06/29/diffusion-probabilistic-models-I.html. 2021.
- [Vas+23] Ashish Vaswani et al. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL].
- [Vin11] Pascal Vincent. «A Connection Between Score Matching and Denoising Autoencoders». En: *Neural Computation* 23.7 (2011), págs. 1661-1674. DOI: 10.1162/NECO_a_00142.
- [VS03] C. Villani y American Mathematical Society. *Topics in Optimal Transportation*. Graduate studies in mathematics. American Mathematical Society, 2003. ISBN: 9781470418045. URL: <https://books.google.cl/books?id=MyPjjgEACAAJ>.
- [Zhu+20] Jun-Yan Zhu et al. *Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks*. 2020. arXiv: 1703.10593 [cs.CV]. URL: <https://arxiv.org/abs/1703.10593>.