Universidad de Valladolid

FACULTAD DE CIENCIAS

TRABAJO FIN DE MÁSTER

Máster en Matemáticas

Puentes de Schrödinger. Aplicaciones en métodos generativos.

Autor: Alejandro Martínez Álvarez
Tutor: Eustasio del Barrio Tellado
Año: 2025

Índice general

Resumen

El problema del Puente de Schrödinger es el de encontrar, de entre todas las distribuciones posibles en el espacio de funciones continuas de $[0, 1]$ en $\mathbb{R}^d$ con marginales inicial y final fijadas, la que minimice la divergencia de Kullback con respecto a la medida inducida por un movimiento Browniano reversible. Este problema está relacionado con el problema de transporte óptimo con regularización entrópica y permite, de hecho [25] y [26], reinterpretar este problema mediante una formulación dinámica similar a la introducida en Benamou y Brenier [2].

Recientemente este problema ha recibido mucha atención por su conexión con los métodos generativos basados en procesos de difusión. En estos métodos se busca aproximar una familia de transformaciones indexada por $t \in [0, 1]$ que convierta suavemente una distribución de referencia, frecuentemente Gaussiana, en otra distribución, idealmente la subyacente a la muestra de datos de interés.

Muchos de estos métodos recorren un doble camino progresivo-regresivo [33]. Primero se construye un proceso de difusión cuya distribución de equilibrio sea la distribución de referencia. La inversión temporal del proceso es, de nuevo, un proceso de difusión, que se podría emplear en la generación de nuevas observaciones de la distribución objetivo. El coeficiente de deriva de este proceso invertido depende de la función score de la densidad marginal del proceso de difusión, lo que obliga a emplear alguna técnica de estimación de tal score (procedimiento conocido como *score matching*).

En este contexto el recurso a los puentes de Schrödinger puede suponer cierta economía, puesto que la resolución numérica del problema mediante el algoritmo de Sinkhorn proporciona como subproducto la estimación necesaria de las funciones score [32].

En este TFM se pretende desarrollar la teoría necesaria para conectar las distintas formulaciones (estática y dinámica) de los problemas de transporte clásico y entrópico con el marco de los procesos de difusión. Adicionalmente, se explorará el rendimiento en la práctica del método basado en puentes de Schrödinger en el contexto de la generación de imágenes, comparando con otras alternativas actuales.

Palabras clave

Puentes de Schrödinger, Generación de imágenes, Transporte óptimo entrópico, Transporte óptimo clásico, Algoritmo de Sinkhorn.

Abstract

The Schrödinger Bridge problem is the problem of finding, among all possible distributions on the space of continuous functions from $[0, 1]$ to $\mathbb{R}^d$ with fixed initial and final marginals, the one that minimizes the Kullback divergence with respect to the measure induced by a reversible Brownian motion. This problem is related to the optimal transport problem with entropic regularization and, in fact, allowssee [25] and [26]a reinterpretation through a dynamic formulation similar to the one introduced by Benamou and Brenier [2].

Recently, this problem has gained considerable attention due to its connection with generative methods based on diffusion processes. These methods aim to approximate a family of transformations indexed by $t \in [0, 1]$ that smoothly maps a reference distribution, often Gaussian, into another distribution, ideally the one underlying the data of interest.

Many of these methods follow a bidirectional forward-reverse approach [33]. First, a diffusion process is constructed whose equilibrium distribution is the reference distribution. The time-reversed process is again a diffusion process, which can be used to generate new samples from the target distribution. The drift coefficient of this reversed process depends on the score function of the marginal density of the forward diffusion, which requires employing some score estimation technique (a procedure known as *score matching*).

In this context, Schrödinger bridges offer computational advantages, since solving the problem numerically using the Sinkhorn algorithm yields, as a byproduct, the necessary estimation of the score functions [32].

This work aims to develop the theoretical framework required to connect the different formulations (static and dynamic) of classical and entropic optimal transport problems with the setting of diffusion processes. Additionally, it explores the practical performance of the Schrödinger bridge-based method in the context of image generation, comparing it with other current alternatives.

Keywords

Schrödinger bridges, Image generation, Entropic optimal transport, Classical optimal transport, Sinkhorn algorithm.

Introducción

El *machine learning* es un campo de creciente interés con diversas aplicaciones que abarcan desde la automatización de procesos hasta la creación de contenido artístico y científico. Uno de los enfoques más destacados en este ámbito es la generación de imágenes. Los métodos más utilizados en este campo incluyen *Generative Adversarial Networks (GANs)* [16], *Variational Autoencoders (VAEs)* [23] y *Diffusion Models* [18], los cuales han demostrado ser efectivos para tareas como la creación o modificación de imágenes realistas.

Estudiaremos la extrapolación de un método alternativo basado en el concepto de transporte óptimo y *Puentes de Schrödinger*, adaptados específicamente para la generación de imágenes. Más detalladamente: el trabajo se centrará en la aplicación del *Puente de Schrödinger* a la generación de nuevas muestras. Para ello, se seguirán los resultados de Pooladian y Niles-Weed [32], quienes demuestran que los potenciales obtenidos al resolver el problema de transporte óptimo regularizado, vía algoritmo de Sinkhorn [24], pueden modificarse para construir un estimador natural denominado *Puente de Sinkhorn*. Asimismo, se establece que este puente converge hacia el *Puente de Schrödinger*, con una tasa de convergencia dependiente de la dimensión intrínseca de la medida objetivo y no de la dimensión ambiente.

En su estudio [32], Pooladian y Niles-Weed proporcionan un caso en el que se consideran muestras de puntos en un espacio bidimensional con distribuciones prefijadas, lo que permite comprobar empíricamente el comportamiento del *Puente de Sinkhorn* en un entorno controlado. Sin embargo, hasta donde alcanza el conocimiento del autor, la aplicación de este enfoque a la generación de imágenes no ha sido explorada en la literatura actual. Por ello, este trabajo tiene como objetivo adaptar el algoritmo propuesto por Pooladian y Niles-Weed a dicho contexto.

Para llegar a dichos conceptos será necesario introducir tanto el problema de transporte óptimo clásico como el problema de transporte óptimo entrópico.

El problema de transporte óptimo es una cuestión fundamental dentro de la probabilidad y estadística que busca determinar la manera más eficiente de transformar una distribución de probabilidad en otra, minimizando un cierto costo de transporte. Este problema tiene sus fundamentos en la obra de Gaspard Monge (1781) [29] y fue rigurosamente formalizado y extendido por Leonid Kantorovich (1942) [21] [20]. En la actualidad este problema ha vuelto a resurgir gracias a los avances computacionales y algoritmos eficientes. Como consecuencia, el transporte óptimo se utiliza cada vez más en problemas de ciencias de la imagen (como el procesamiento de color o textura), gráficos (para la manipulación de formas) o aprendizaje automático (para regresión, clasificación y modelado generativo).

En un contexto discreto, se consideran puntos $\{x_i\}_{i=1}^m$ de $\mathbb{R}^d$ con masas asociadas p_i y puntos $\{y_j\}_{j=1}^n$ de $\mathbb{R}^{d'}$ con masas q_j. Estas masas pueden normalizarse para obtener distribuciones de probabilidad en ambos espacios que llamaremos μ y ν. En este contexto el problema sería encontrar un plan de transporte

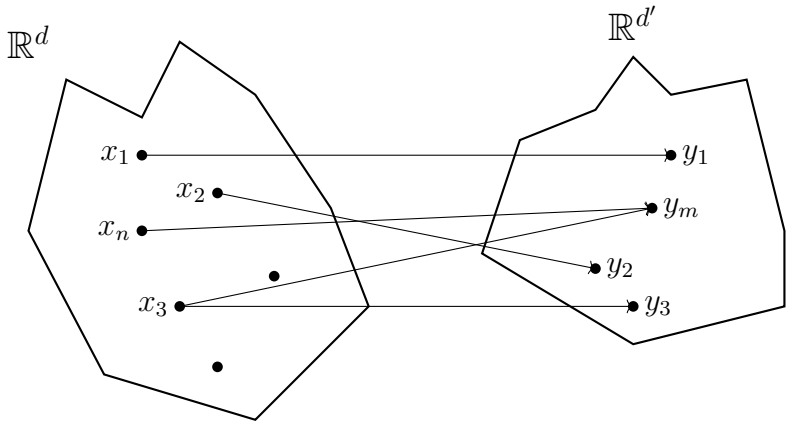
que minimice el coste cuadrático:

$$\min_{\pi \in \Pi(\mu,\nu)} \sum_{i=1}^{m} \sum_{j=1}^{n} \|x_i - y_j\|^2 \pi_{i,j},$$

sujeto a las restricciones

$$\begin{aligned}
&\sum_{j=1}^{n} \pi_{i,j} = p_i, && \text{para todo } i = 1, \ldots, m, \\
&\sum_{i=1}^{m} \pi_{i,j} = q_j, && \text{para todo } j = 1, \ldots, n, \\
&\pi_{i,j} \geq 0, && \text{para todo } i, j.
\end{aligned}$$

Donde $\Pi(\mu, \nu)$ denota el conjunto de planes de transporte con las marginales μ y ν.



- Toda la masa que hay en cada punto de origen se tiene que repartir entera entre los destinos. Es decir, si en el punto x_i hay masa p_i, se tiene que distribuir toda esa cantidad entre los distintos destinos y_j, sin perder ni ganar nada por el camino.
- Cada punto de destino debe recibir exactamente la cantidad de masa que le corresponde. Es decir, al punto y_j tiene que llegar una masa total de q_j, que puede venir repartida desde varios orígenes x_i, pero la suma tiene que ser exactamente esa.

Una parte de este estudio se basará en analizar las diversas formulaciones del problema, desde la formulación clásica de Monge hasta enfoques más recientes, como la formulación dinámica de Brenier [15], pasando por la formulación de Kantorovich y su correspondiente formulación dual.

No obstante, la implementación computacional del problema de transporte óptimo clásico plantea desafíos significativos, particularmente en contextos de alta dimensión. Para abordar estas dificultades, se introduce la formulación entrópica, que incorpora un término de regularización con el propósito de suavizar la solución y mejorar la estabilidad numérica del problema. Esta reformulación no solo facilita la resolución computacional, sino que también aproxima el modelo a un problema de optimización convexa, permitiendo el uso de algoritmos más eficientes y escalables.

Para facilitar la comprensión, a continuación se presenta una estructura detallada del contenido de este trabajo, que aborda de manera progresiva los conceptos y métodos fundamentales hasta llegar a la aplicación práctica en generación de imágenes.

En el capítulo 1, se abordará tanto la formulación clásica del problema de transporte óptimo, incluyendo sus distintas versiones: el problema de Monge, la formulación de Kantorovich y su versión dinámica. Fundamentalmente seguiremos [36].

En el capítulo 2, se introducirá la formulación entrópica y se analizarán sus distintos enfoques: estático, dinámico y estocástico, haciendo especial hincapié en la interpretación de la formulación estocástica dentro del marco dinámico a través de la ecuación de Fokker-Planck. Principalmente nos basaremos en [36] y [31].

El capítulo 3 se dedicará al problema del *Puente de Schrödinger*. Este inicialmente formulado como un problema de mecánica estadística, puede reinterpretarse como un modelo de acoplamiento entre distribuciones de probabilidad condicionado a trayectorias estocásticas. Este modelo es equivalente al problema de transporte óptimo entrópico y desempeña un papel crucial en la estimación de transformaciones entre distribuciones. Su relevancia trasciende el ámbito teórico, resultando de particular interés en aplicaciones donde se requiere modelar la proximidad entre distribuciones y determinar trayectorias óptimas entre ellas. En este capítulo se profundizará en los aspectos técnicos que subyacen a este enfoque. Nos basaremos en [26] y [32]. Además introduciremos el estimador sugerido por [32], el *Puente de Sinkhorn*.

En el capítulo 4 estudiaremos el análisis estadístico del estimador, de hecho de la ley inducida por nuestro estimador. En primer lugar se estimará el error de una muestra sin discretización, cuando comparamos con la medida dada por el proceso óptimo, y luego se estudiará el error que se comete al discretizar. Con todo ello llegaremos al resultado principal de dicho capítulo, Error total del Puente de Sinkhorn.

El capítulo 5 se dedica a 2 experimentos. Por un lado, un ejemplo básico, en el que trabajamos con distribuciones, que tienen cierta forma en $\mathbb{R}^3$, para mostrar el funcionamiento del algoritmo y, así, observar cómo se transportan las distribuciones. Por otro lado, extrapolamos la idea del experimento al contexto de las imágenes. Además, comentaremos cómo afectan los resultados al variar los parámetros e interpretaremos los resultados.

Motivación

Este Trabajo de Fin de Máster nace del interés por aplicar herramientas matemáticas avanzadas a problemas reales. Se ha llevado a cabo en colaboración con el grupo de investigación en *Computer Vision* de la Fundación CIDAUT.

El planteamiento inicial consistía en generar escenarios sintéticos de conducción que pudieran emplearse en el desarrollo de vehículos autónomos. No obstante, la complejidad de esta tarea ha llevado a posponer dicho objetivo para futuras investigaciones. En su lugar, el trabajo se ha planteado como una primera aproximación al tema, con el propósito de asentar bases para una posible tesis doctoral o futuras líneas de investigación más amplias.

No obstante, esta experiencia ha brindado la oportunidad de poner en práctica conceptos de Teoría de la Probabilidad y Procesos Estocásticos adquiridos a lo largo del máster. De manera particular, se ha profundizado en el marco del transporte óptimo clásico, transporte óptimo entrópico y su relación con el puente de Schrödinger, estas formulaciones ofrecen posibilidades muy interesantes en ámbitos como la generación de datos.

A pesar del creciente interés que despiertan estas metodologías en la comunidad matemática y estadística, su aplicación práctica en contextos como la generación de imágenes sigue siendo escasa. Esta circunstancia es una oportunidad para la investigación en este ámbito.

Capítulo 1

Problema de transporte óptimo clásico

Después del planteamiento en el caso discreto de la introducción, nos situamos aquí en el caso continuo. Trabajaremos con medidas de probabilidad (utilizaremos indistintamente este término y probabilidades) π en el espacio producto $\mathbb{R}^d \times \mathbb{R}^d$ con momento de orden 2 finito. Al espacio de probabilidades que cumplen esta condición lo denotaremos $\mathcal{P}_2(\mathbb{R}^d)$.

Nuestro objetivo en este capítulo es presentar el problema de transporte óptimo clásico y varias formulaciones: Kantorovich, Monge y dinámica (o de Brenier).

En primer lugar se define el conjunto de transportes admisibles o conjunto de planes de transporte.

$$\Pi(\mu,\nu) = \{\pi \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d) : \pi[A \times \mathbb{R}^d] = \mu(A), \pi[\mathbb{R}^d \times B] = \nu(B)\}, \tag{1.1}$$

para todo A y B conjuntos medibles de $\mathbb{R}^d$. Es decir, el conjunto de probabilidades producto en $\mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$ tal que la primera marginal sea μ y la segunda ν. Las condiciones de las marginales se traducen en que toda la masa que cojamos desde un x coincida con $\mathrm{d}\mu(x)$, análogamente con y. Esto significa

$$\int_{\mathbb{R}^d} \mathrm{d}\pi(x,y) = \mathrm{d}\mu(x), \quad \int_{\mathbb{R}^d} \mathrm{d}\pi(x,y) = \mathrm{d}\nu(y).$$

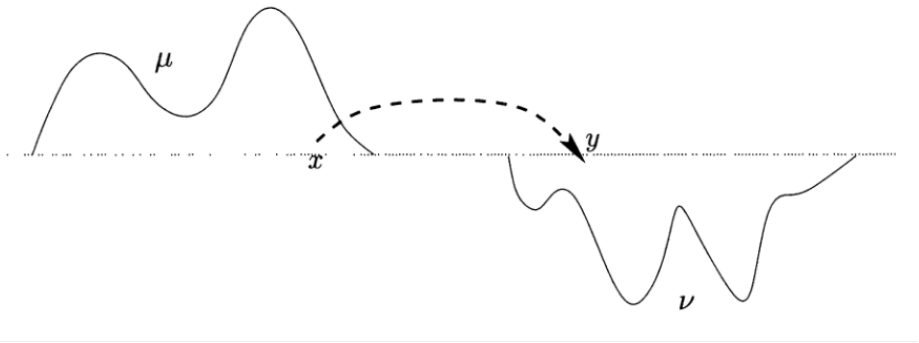


Figura 1.1: Problema de transporte de masa

Puesto que μ y ν están prefijadas, $\Pi(\mu,\nu)$ es no vacío. Es suficiente elegir $\mu \otimes \nu$ como probabilidad producto y obviamente cumple las condiciones de la definición.

Además, se define una función coste $c(x,y) : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}$. Coloquialmente, $c(x,y)$ es el coste de transportar una unidad de masa de $\mathbb{R}^d$ a $\mathbb{R}^d$.

Nota. Ejemplos de función coste podría ser cualquier $\|x-y\|^p$, con $1 \leq p \leq \infty$. El caso realmente interesante es $p = 2$, será con el que trabajaremos. En nuestro caso, se comentará de nuevo cuando sea pertinente, utilizaremos $c_2(x,y) := \frac{1}{2}\|x-y\|^2$.

Dado este contexto estamos en condiciones de enunciar el problema de transporte óptimo. Esta primera formulación se debe a Kantorovich que en la década de los 40 del siglo XX la presenta en dos artículos [21] y [20].

1.1. Formulación de Kantorovich

En la situación anterior planteamos el problema de Kantorovich, es decir, buscamos

$$\mathcal{T}(\mu,\nu) := \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c_2(x,y)\, \mathrm{d}\pi(x,y), \tag{1.2}$$

Dicha integral es el *coste de transporte*, para ahorrar notación la denotaremos como $I[\pi]$. Nuestro propósito aquí es presentar el teorema de dualidad de Kantorovich 1.5, antes de ello probaremos que el problema anterior admite minimizador.

Proposición 1.1 (Existencia de plan óptimo)**.** *El problema de minimización*

$$\min_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c_2(x,y)\, \mathrm{d}\pi(x,y).$$

admite un minimizador $\pi^\star$. Al plan de transporte óptimo lo denotaremos $\mathcal{T}(\mu,\nu)$

Demostración. Puesto que estamos trabajando con probabilidades de momento finito de orden 2, se tiene

$$M_2 := \int_{\mathbb{R}^d} \frac{\|x\|^2}{2}\, \mathrm{d}\mu(x) + \int_{\mathbb{R}^d} \frac{\|y\|^2}{2}\, \mathrm{d}\nu(y) < \infty. \tag{1.3}$$

Por tanto,

$$\pi \in \Pi(\mu,\nu) \Longrightarrow \int_{\mathbb{R}^d \times \mathbb{R}^d} \frac{\|x\|^2 + \|y\|^2}{2}\, \mathrm{d}\pi(x,y) \leq M_2.$$

Sea $\delta > 0$ y sean $K \subset \mathbb{R}^d, L \subset \mathbb{R}^d$ tal que

$$\mu[\mathbb{R}^d \backslash K] \leq \delta, \quad \nu[\mathbb{R}^d \backslash L] \leq \delta.$$

Para todo $\pi \in \Pi(\mu,\nu)$,

$$\pi[(\mathbb{R}^d \times \mathbb{R}^d)\backslash(K \times L)] \leq \pi[\mathbb{R}^d \times (\mathbb{R}^d \backslash L)] + \pi[(\mathbb{R}^d \backslash K) \times L] \leq 2\delta.$$

Esto prueba que $\Pi(\mu,\nu)$ es ajustado, en virtud del teorema de Prohorov [5] es relativamente compacto con respecto a la topología débil. Además, veamos que dicho conjunto es débilmente cerrado:

Sea $\{\pi_n\}_n \in \Pi(\mu,\nu)$ y tal que $\{\pi_n\} \to \pi \in \overline{\Pi}(\mu,\nu)$. Veamos que $\pi \in \Pi(\mu,\nu)$. Ahora bien, si elegimos $h(x) = f(x,y)$ se tiene

$$\int_{\mathbb{R}^d} h(x)\,\mathrm{d}\mu(x) = \int_{\mathbb{R}^d\times\mathbb{R}^d} f(x,y)\,\mathrm{d}\pi_n(x,y) \to^{w.}$$
$$\int_{\mathbb{R}^d\times\mathbb{R}^d} f(x,y)\,\mathrm{d}\pi(x,y) = \int_{\mathbb{R}^d} h(x)\,\mathrm{d}\tilde{\mu}(x).$$

Y como $\mathcal{C}_b$ es una clase determinante de la probabilidad entonces $\mu = \tilde{\mu}$. En conclusión, $\Pi(\mu,\nu)$ es compacto (con respecto a la topología débil). Esto garantiza la existencia del mínimo. □

1.1.1. Formulación dual

En virtud de la linealidad de la esperanza y de la condición (1.1) se deduce que $\pi \in \Pi(\mu,\nu)$ si y sólo si existe una medida no negativa en $\mathbb{R}^d \times \mathbb{R}^d$ tal que, para todo par $(f,g) \in L^1(\nu) \times L^1(\mu)$, se tiene

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} [f(x)+g(y)]\,\mathrm{d}\pi(x,y) = \int_{\mathbb{R}^d} f(x)\,\mathrm{d}\mu(x) + \int_{\mathbb{R}^d} g(y)\,\mathrm{d}\nu(y). \tag{1.4}$$

Con intención de hacer más liviana la notación denotaremos a la suma de la derecha $J(f,g)$ y Φ al conjunto de funciones de $L^1(\mu) \times L^1(\nu)$ que cumplen

$$f(x) + g(y) \leq c_2(x,y), \quad \mathrm{d}\mu - \text{p.c.t. } x \in \mathbb{R}^d,\ \mathrm{d}\nu - \text{p.c.t. } y \in \mathbb{R}^d.$$

Nota. En este trabajo presentamos el caso de función de coste cuadrático y $\mathbb{R}^d$, estos resultados se pueden extender a funciones de coste inferiormente semicontinuas y espacios polacos (métricos, separables y completos).

Es conocido que un problema de minimización lineal convexo como (1.2) admite una formulación dual, en nuestro contexto de transporte de masas la introduce Kantorovich en [20]. Como se ha comentado en la nota anterior, pese a que el teorema de dualidad se presenta en espacios Polacos y función coste inferiormente semicontinua (se puede generalizar, aun, a espacios más inusuales) la formulación aquí expuesta se reduce a utilizar la función de coste cuadrático y $\mathbb{R}^d$.

Teorema 1.2 (Teorema de dualidad de Kantorovich)**.** *Sean $\mathbb{R}^d$, $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ y $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ probabilidades, y consideremos $c_2(x,y) := \frac{1}{2}\|x-y\|^2$ la función de coste cuadrático. Con la notación introducida se tiene:*

$$\min_{\pi\in\Pi(\mu,\nu)} I[\pi] = \max_{(f,g)\in\Phi} J(f,g). \tag{1.5}$$

En el caso de que restrinjamos el conjunto Φ a funciones continuas y acotadas el valor del supremo del lado derecho se mantiene.

Desarrollando la expresión de $I[\pi]$ y utilizando la hipótesis de que las probabilidades tienen momento de orden 2 finito se tiene

$$I[\pi] = \int_{\mathbb{R}^d} \frac{\|x\|^2}{2}\,\mathrm{d}\mu(x) + \int_{\mathbb{R}^d} \frac{\|y\|^2}{2}\,\mathrm{d}\nu(y) - \int_{\mathbb{R}^d\times\mathbb{R}^d} [x\cdot y]\,\mathrm{d}\pi(x,y).$$

Por otro lado, la condición de $(f,g) \in \Phi$ es

$$f(x) + g(y) \leq \frac{\|x-y\|^2}{2},$$

después de reorganizar términos al lado derecho de la desigualdad se tiene

$$x \cdot y \leq \left[\frac{\|x\|^2}{2} - f(x)\right] + \left[\frac{\|y\|^2}{2} - g(y)\right].$$

Definimos las funciones

$$\varphi(x) = \left[\frac{\|x\|^2}{2} - f(x)\right], \quad \psi(y) = \left[\frac{\|y\|^2}{2} - g(y)\right].$$

En definitiva, el problema de transporte óptimo (1.2) se escribe del siguiente modo

$$\min_{\pi \in \Pi(\mu,\nu)} I[\pi] = M_2 - \max_{\pi \in \Pi(\mu,\nu)} \left\{ \int_{\mathbb{R}^d \times \mathbb{R}^d} [x \cdot y] \, \mathrm{d}\pi(x,y) \right\}, \tag{1.6}$$

y

$$\sup_{(f,g) \in \Phi} J = M_2 - \min_{(\varphi,\psi) \in \widetilde{\Phi}} \{J(\varphi,\psi)\}, \tag{1.7}$$

donde $\widetilde{\Phi}$ es el conjunto de los pares $(\varphi, \psi) \in L^1(\mu) \times L^1(\nu)$, tal que para casi todo x, y,

$$x \cdot y \leq \varphi(x) + \psi(y). \tag{1.8}$$

La formulación dual también cambia

$$\max_{\pi \in \Pi(\mu,\nu)} \left\{ \int_{\mathbb{R}^d \times \mathbb{R}^d} [x \cdot y] \, \mathrm{d}\pi(x,y) \right\} = \min_{(\varphi,\psi) \in \widetilde{\Phi}} \{J(\varphi,\psi)\}. \tag{1.9}$$

Introducimos ahora la doble convexificación vía la transformada de Legendre en la última igualdad. Para la definición y propiedades de esta transformada y resultados sobre análisis convexo ver Apéndice A. A través de la transformada de Legendre, la ecuación (1.8) se traduce en

$$\psi(y) \geq \varphi^*(y). \tag{1.10}$$

y, en consecuencia, $J(\varphi, \psi) \geq J(\varphi, \varphi^*)$. Si repetimos el razonamiento fijando la otra componente se deduce que $J(\varphi, \varphi^*) \geq J(\varphi^{**}, \varphi^*)$. De esas dos desigualdades se sigue que

$$\inf_{(\varphi,\psi) \in \widetilde{\Phi}} J(\varphi,\psi) \geq \inf_{\varphi \in L^1(\mu)} J(\varphi^{**}, \varphi^*).$$

Si admitimos que $(\varphi^{**}, \varphi^*) \in L^1(\mu) \times L^1(\nu)$ entonces $(\varphi^{**}, \varphi^*) \in \widetilde{\Phi}$ y se deduce de la anterior desigualdad que el inferior de J en $\widetilde{\Phi}$ no cambia si se restringe J a un conjunto más pequeño de $\widetilde{\Phi}$ formado por $(\varphi^{**}, \varphi^*)$. Pero esas funciones son funciones convexas inferiormente semicontinuas, ya que están definidas como el supremo de familias de funciones lineales.

En virtud del [36, Lema 2.10] la desigualdad anterior es, de hecho, una igualdad. La idea de la demostración es tomar una sucesión que minimiza el funcional dual y modificarla de forma que siga siendo mínima pero que esté acotada por funciones afines y que sea integrable respecto a las medidas consideradas. Esto permite asegurar que la sucesión pertenece al dominio del funcional y que efectivamente converge al mínimo, completando así la justificación de la dualidad.

Teorema 1.3 (Existencia de un par óptimo de funciones convexas conjugadas)**.** *Sean μ, ν en $\mathcal{P}_2(\mathbb{R}^d)$. Entonces existe un par (φ, φ^*) de funciones convexas conjugadas inferiormente semicontinuas en $\mathbb{R}^d$, tal que*

$$\inf_{\widetilde{\Phi}} J = J(\varphi, \varphi^*)$$

Nota. La prueba se encuentra en [36, Teorema 2.9]. La idea de la prueba es intercambiar compacidad fuerte por compacidad débil más monotonía.

Demostración. Sustituimos la función $x \cdot y$ por una función no negativa; para ello, añadimos algunas funciones integrables bien elegidas a ambos lados de la desigualdad que define $\widetilde{\Phi}$. Por ejemplo,

$$(\varphi, \psi) \in \widetilde{\Phi} \iff \left[\varphi(x) + \frac{\|x\|^2}{2}\right] + \left[\psi(y) + \frac{\|y\|^2}{2}\right] \geq \frac{\|x\|^2}{2} + \frac{\|y\|^2}{2} + x \cdot y = \frac{\|x+y\|^2}{2}.$$

Sea $(\varphi_k, \psi_k)_{k\in\mathbb{N}}$ una sucesión mínima para J. En vista del comentario anterior, podemos asumir que

$$0 \leq \varphi_k(x) + \frac{\|x\|^2}{2}, \quad 0 \leq \psi_k(y) + \frac{\|y\|^2}{2},$$

$$\left[\varphi_k(x) + \frac{\|x\|^2}{2}\right] + \left[\psi_k(y) + \frac{\|y\|^2}{2}\right] \geq \frac{\|x+y\|^2}{2} \geq 0.$$

El punto importante aquí es que todo es no negativo.

Aunque $\varphi_k(x) + \frac{\|x\|^2}{2}$ esté acotado en $L^1(\mu)$, esto no es suficiente para garantizar la compacidad débil en L^1: se necesita una estimación adicional de equi-integrabilidad. Por eso pasaremos por un procedimiento de truncamiento. Así, para cada $\ell \in \mathbb{N}$, definimos $\varphi_k^{(\ell)}$, $\psi_k^{(\ell)}$ por truncamiento:

$$\varphi_k^{(\ell)}(x) + \frac{\|x\|^2}{2} = \text{mín}\left(\varphi_k(x) + \frac{\|x\|^2}{2}, \ell\right),$$

$$\psi_k^{(\ell)}(y) + \frac{\|y\|^2}{2} = \text{mín}\left(\psi_k(y) + \frac{\|y\|^2}{2}, \ell\right).$$

Es fácil verificar las siguientes propiedades:

$$\begin{cases} 0 \leq \varphi_k^{(\ell)}(x) + \dfrac{\|x\|^2}{2} \leq \ell, \\ 0 \leq \psi_k^{(\ell)}(y) + \dfrac{\|y\|^2}{2} \leq \ell, \end{cases} \tag{1.11}$$

$$\begin{cases} \varphi_k^{(1)} \leq \varphi_k^{(2)} \leq \cdots \leq \varphi_k^{(\ell)} \leq \ldots, \\ \psi_k^{(1)} \leq \psi_k^{(2)} \leq \cdots \leq \psi_k^{(\ell)} \leq \ldots, \end{cases} \tag{1.12}$$

$$J(\varphi_k^{(\ell)}, \psi_k^{(\ell)}) \leq J(\varphi_k, \psi_k), \tag{1.13}$$

$$\varphi_k^{(\ell)}(x) + \psi_k^{(\ell)}(y) \geq \text{mín}\left(\frac{\|x+y\|^2}{2}, \ell\right) - \left(\frac{\|x\|^2}{2} + \frac{\|y\|^2}{2}\right). \tag{1.14}$$

Por (1.11), sabemos que para cada ℓ,

$$\varphi_k^{(\ell)}(x) = -\frac{\|x\|^2}{2} + O(\ell).$$

Dado que $-\|x\|^2/2$ es una función fija en $L^1(\mu)$, deducimos que la sucesión $(\varphi_k^{(\ell)})_{k\in\mathbb{N}}$ define una sucesión débilmente compacta en $L^1(\,\mathrm{d}\mu)$. Por lo tanto, extrayendo una subsucesión,

$$\varphi_k^{(\ell)} \to \varphi^{(\ell)}, \quad \text{débilmente en } L^1(\,\mathrm{d}\mu), \text{ cuando } k \to \infty.$$

De modo similar, para cada ℓ, la sucesión $(\psi_k^{(\ell)})_{k\in\mathbb{N}}$ converge débilmente en $L^1(\nu)$, tras extraer una subsucesión, hacia alguna $\psi^{(\ell)} \in L^1(\nu)$. Por una *extracción diagonal*, podemos extraer una subsucesión de $k \in \mathbb{N}$ para la cual la convergencia vale para todo ℓ. Entonces, como la convergencia débil preserva el orden, deducimos de (1.12), (1.13), (1.14) que

$$\begin{cases} \varphi^{(1)} \le \varphi^{(2)} \le \cdots \le \varphi^{(\ell)} \le \ldots, \\ \psi^{(1)} \le \psi^{(2)} \le \cdots \le \psi^{(\ell)} \le \ldots, \end{cases} \tag{1.15}$$

$$J(\varphi^{(\ell)}, \psi^{(\ell)}) = \lim_{k\to\infty} J(\varphi_k^{(\ell)}, \psi_k^{(\ell)}) \le \liminf_{k\to\infty} J(\varphi_k, \psi_k) = \inf_{\widetilde{\Phi}} J, \tag{1.16}$$

$$\varphi^{(\ell)}(x) + \psi^{(\ell)}(y) \ge \min\left(\frac{\|x+y\|^2}{2}, \ell\right) - \left(\frac{\|x\|^2}{2} + \frac{\|y\|^2}{2}\right). \tag{1.17}$$

Las sucesiones $(\varphi^{(\ell)})$ y $(\psi^{(\ell)})$ están acotadas en L^1, son no decrecientes y están acotadas inferiormente por una función fija en L^1. Por lo tanto, podemos aplicar el *teorema de convergencia monótona*, y deducir la existencia de límites en L^1 de φ y ψ, definidos casi en todo punto:

$$\varphi = \lim_{\ell\to\infty} \varphi^{(\ell)}, \qquad \psi = \lim_{\ell\to\infty} \psi^{(\ell)},$$

los cuales satisfacen

$$J(\varphi, \psi) = \lim_{\ell\to\infty} J(\varphi^{(\ell)}, \psi^{(\ell)}) \le \inf_{\widetilde{\Phi}} J.$$

Al pasar al límite en (1.17), vemos que $(\varphi, \psi) \in \widetilde{\Phi}$. Por lo tanto, es un par óptimo.

Aplicamos la doble convexificación a (φ, ψ) para obtener un par óptimo formado por funciones convexas conjugadas propias. □

Teorema 1.4 (Teorema de transporte óptimo para coste cuadrático)**.** *Sean μ, ν dos probabilidades en $\mathcal{P}_2(\mathbb{R}^d)$. Si consideramos el problema de transporte óptimo con función de coste cuadrático, entonces*

1. *$\pi \in \Pi(\mu, \nu)$ es óptimo si y solo si existe una función convexa φ tal que $y \in \partial\varphi(x)$, para $\mathrm{d}\pi$ -casi todo (x, y).*

   En ese caso el par (φ, φ^*) ha de ser el mínimo en el problema

$$\inf_{(f,g)\in\widetilde{\Phi}} \left\{ \int_{\mathbb{R}^d} f(x)\,\mathrm{d}\mu(x) + \int_{\mathbb{R}^d} g(y)\,\mathrm{d}\nu(y) \right\}.$$

2. *Si μ tiene densidad entonces existe un único transporte óptimo π tal que*

$$\pi = (Id \times \nabla\varphi)\#\mu, \tag{1.18}$$

donde $\nabla\varphi$ es el único gradiente de una función convexa tal que $\nabla\varphi\#\mu = \nu$.

Nota. 1. A $\nabla\varphi$ la denominaremos *potencial de Brenier* que transporta μ a ν.

2. Notemos que los potenciales son únicos salvo traslaciones por constante $a \to (\varphi + a, \psi - a)$.

Demostración. 1. Por la proposición 1.1 existe un plan óptimo $\pi^\star$ y por el resultado 1.3 existe un par de funciones convexas conjugadas (φ, φ^*) óptimas. Por la formulación equivalente que se ha visto en el teorema 1.5

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} [\varphi(x) + \varphi^*(y) - x \cdot y]\, \mathrm{d}\pi^\star(x, y) = 0.$$

El integrando es una función no negativa luego se tiene que anular casi siempre. En virtud de la caracterización de la subdiferencial (ver apéndice A) se verifica la implicación hacia la derecha.

Para ver la otra, es dar las igualdades en sentido inverso.

2. En primer lugar notemos que la integral del borde del dominio de φ es 0 puesto que μ tiene densidad. Si $\pi \in \Pi(\mu, \nu)$ y $T\#\mu = \nu$ entonces

$$\mathrm{d}\pi\text{-casi siempre }, y = T(x) \iff \pi = (\mathrm{Id} \times T)\#\mu.$$

Por ser φ una función convexa entonces la subdiferencial en x es $\nabla\varphi$ $\mathrm{d}\mu$ casi siempre(ver apéndice A). Además, notemos que una propiedad que se cumple $\mathrm{d}\mu$ casi siempre también se cumple $\mathrm{d}\pi$ casi siempre, luego se tiene $y = \nabla\varphi(x)$ $\mathrm{d}\pi$-casi siempre .

Hasta aquí hemos probado que cualquier plan de transporte óptimo toma la forma

$$(\mathrm{Id} \times \nabla\varphi)\#\mu,$$

para alguna función convexa φ tal que $\nabla\varphi\#\mu = \nu$, y hemos probado que al menos hay un plan de transferencia que lo cumple. Centrémonos en la unicidad. Sea $\bar\varphi$ otra función convexa tal que

$$\nabla\bar\varphi\#\mu = \nu.$$

Queremos probar que $\nabla\varphi = \nabla\bar\varphi$, μ casi siempre. Por el apartado 1, dado que $\bar\varphi$ es convexa y μ tiene densidad, se sigue del teorema de Rademacher que $\bar\varphi$ es diferenciable μ-casi en todas partes, y por tanto $\nabla\bar\varphi$ está bien definida μ-casi siempre. Definimos entonces $\bar\pi := (\mathrm{Id} \times \nabla\bar\varphi)\#\mu$, que pertenece a $\Pi(\mu, \nu)$. Como $\nabla\bar\varphi(x) \in \partial\bar\varphi(x)$ en los puntos de diferenciabilidad, se sigue que $y \in \partial\bar\varphi(x)$. Por tanto, en virtud del apartado 1 del teorema, $\bar\pi$ es un plan óptimo y el par $(\bar\varphi, \bar\varphi^*)$ es óptimo del dual. Por tanto,

$$\int_{\mathbb{R}^d} \bar\varphi \, \mathrm{d}\mu + \int_{\mathbb{R}^d} \bar\varphi^* \, \mathrm{d}\nu = \int_{\mathbb{R}^d} \varphi \, \mathrm{d}\mu + \int_{\mathbb{R}^d} \varphi^* \, \mathrm{d}\nu$$

Sea π el plan de transporte óptimo asociado a φ. A partir de esta igualdad tenemos:

$$\begin{aligned}\int_{\mathbb{R}^d\times\mathbb{R}^d} [\bar\varphi(x) + \bar\varphi^*(y)] \, \mathrm{d}\pi(x, y) &= \int_{\mathbb{R}^d\times\mathbb{R}^d} [\varphi(x) + \varphi^*(y)] \, \mathrm{d}\pi(x, y) \\ &= \int_{\mathbb{R}^d\times\mathbb{R}^d} (x \cdot y) \, \mathrm{d}\pi(x, y)\end{aligned}$$

Como $\pi = (\mathrm{Id} \times \nabla\varphi)\#\mu$, esto se puede reescribir como

$$\int_{\mathbb{R}^d} [\bar{\varphi}(x) + \bar{\varphi}^*(\nabla\varphi(x))] \ \mathrm{d}\mu(x) = \int_{\mathbb{R}^d} x \cdot \nabla\varphi(x) \ \mathrm{d}\mu(x)$$

Por lo tanto,

$$\int_{\mathbb{R}^d} [\bar{\varphi}(x) + \bar{\varphi}^*(\nabla\varphi(x)) - x \cdot \nabla\varphi(x)] \ \mathrm{d}\mu(x) = 0$$

Como el integrando es no negativo, debe anularse μ-casi en todas partes. Usando la proposición A.3 del apéndice A, concluimos que

$$\nabla\varphi(x) \in \partial\bar{\varphi}(x) \quad \text{para } \mu\text{-casi todo } x.$$

Como $\bar{\varphi}$ es diferenciable μ-casi siempre, concluimos que

$$\nabla\varphi(x) = \nabla\bar{\varphi}(x), \quad \text{para } \mu\text{-casi todo } x.$$

□

La ecuación (1.18) enlaza perfectamente con el problema de Monge, en el sentido de que $\nabla\varphi$ es la única solución del problema de transporte de Monge. Esta formulación del problema lo introducimos a continuación.

1.2. Problema de Monge

Vamos a formular una versión más fuerte que la de Kantorovich. Esto es lo que se llama el problema de Monge. Se enuncia en el año 1781 en [29].

Dados dos espacios de probabilidad $(\Omega_1, \sigma_1, \nu)$ y $(\Omega_2, \sigma_2, \mu)$ (usualmente serán $\mathbb{R}^d$ y la σ-álgebra de Borel), el problema de Monge del transporte óptimo propone encontrar el mínimo valor de

$$\int_{\Omega_1} c(x, T(x)) \, \mathrm{d}\mu(x), \tag{1.19}$$

en el conjunto de aplicaciones medibles $T : \Omega_1 \longrightarrow \Omega_2$ que verifican $T\#\mu = \nu$. Como corolario del teorema 1.4 (cuando nos situamos en esas hipótesis) la aplicación de transporte óptimo es $\nabla\varphi$, es decir,

$$\int_{\mathbb{R}^d} \|x - \nabla\varphi(x)\|^2 \, \mathrm{d}\mu(x) = \inf_{T\#\mu=\nu} \int_{\mathbb{R}^d} \|x - T(x)\|^2 \, \mathrm{d}\mu(x). \tag{1.20}$$

Sin embargo, calcular φ es muy complicado y hay que buscar alternativas; es por ello que surgen formulaciones como la entrópica, que trataremos en un capítulo posterior.

1.3. Formulación dinámica

Hasta ahora nos hemos centrado en formulaciones que son independientes del tiempo, es decir, el coste de transportar una unidad de masa dependen únicamente del momento inicial y el momento final. Vamos a introducir una formulación, que llamaremos dinámica, en la que intervienen los tiempos intermedios. Para ello nos apoyaremos en el lenguaje de la mecánica de fluidos y las ecuaciones de Hamilton-Jacobi.

Este punto de vista permite interpretar el problema de transporte de masa en un problema de *distancia* (encontrar la distancia entre una probabilidad μ y ν), entonces el problema de minimización se traduce en un problema de geodésicas. Recordemos que estamos trabajando con la función coste cuadrático $c_2(z) = \|z\|^2$, para facilitar el desarrollo se ha suprimido el factor $\frac{1}{2}$. Esto se puede extender a variedades de Riemmann y otras funciones de coste, pero insistimos en que para nuestro objetivo esto no es necesario.

En este nuevo planteamiento estudiaremos el proceso de transporte vía la familia de trayectorias: para cada x asociamos una trayectoria $(T_t(x))_{0\leq t\leq 1}$, que denotaremos $(T_t x)$, y $C_2[(T_t x)]$ corresponde al coste de *desplazamiento*. Necesitaremos exigir que $(T_t x)$, vista como aplicación en t, sea $\mathcal{C}^1$ y continua a trozos $\mathrm{d}\mu$- casi siempre. El problema dinámico luce así

$$\inf_{T_0=\mathrm{Id}, T_1\#\mu=\nu} \left\{ \int_{\mathbb{R}^d} C_2[(T_t x)]\, \mathrm{d}\mu(x) \right\}. \tag{1.21}$$

Una condición suficiente para que se tenga la compatibilidad entre (1.21) y la formulación estática es que para todo x e y se cumpla

$$c_2(x,y) = \inf_{z_0=x, z_1=y} \{C_2[(z_t)_{0\leq t\leq 1}]\}. \tag{1.22}$$

Por otro lado, notemos que se tiene

$$C_2[(z_t)] = \int_0^1 \|\dot{z}_t\|^2 \,\mathrm{d}t,$$

donde el punto denota la derivada con respecto de t. Esta propiedad nos resultará útil para la prueba del siguiente resultado:

Proposición 1.5 (Las trayectorias son lineas rectas)**.** *Consideremos c_2 en $\mathbb{R}^d$ entonces*

$$\inf_{z_0=x, z_1=y} \left\{ \int_0^1 \|\dot{z}_t\|^2 \,\mathrm{d}t \right\} = \|x-y\|^2. \tag{1.23}$$

Además, el ínfimo es un mínimo y de hecho se alcanza de manera única en

$$z_t = x + t(y-x).$$

Demostración. Observamos que la función $v \mapsto \|v\|^2$ es estrictamente convexa en $\mathbb{R}^d$, por lo que podemos aplicar la desigualdad de Jensen. Como $\dot{z}_t \in \mathbb{R}^d$ para cada t, se tiene

$$\int_0^1 \|\dot{z}_t\|^2 \,\mathrm{d}t \geq \left\| \int_0^1 \dot{z}_t \,\mathrm{d}t \right\|^2.$$

Pero por el teorema fundamental del cálculo:

$$\int_0^1 \dot{z}_t \, \mathrm{d}t = z_1 - z_0 = y - x,$$

por lo que

$$\int_0^1 \|\dot{z}_t\|^2 \, \mathrm{d}t \geq \|y - x\|^2.$$

Esto prueba que

$$\inf_{z_0=x, z_1=y} \int_0^1 \|\dot{z}_t\|^2 \, \mathrm{d}t \geq \|x - y\|^2.$$

Para ver que la cota inferior se alcanza, consideramos la recta $z_t = x + t(y - x)$. Entonces $\dot{z}_t = y - x$ para todo $t \in [0, 1]$, y se tiene

$$\int_0^1 \|\dot{z}_t\|^2 \, \mathrm{d}t = \int_0^1 \|y - x\|^2 \, \mathrm{d}t = \|y - x\|^2.$$

Así, el ínfimo es en realidad un mínimo, y se alcanza en la curva $z_t = x + t(y - x)$.

Finalmente, la igualdad en la desigualdad de Jensen solo se da si $\dot{z}_t$ es constante casi en todas partes, lo cual implica que la única curva que minimiza la energía es aquella con derivada constante, es decir, la recta entre x y y. Por lo tanto, el minimizador es único. □

Teorema 1.6 (Teorema de transporte óptimo dependiente del tiempo)**.** *Consideramos la función coste cuadrático $c_2(x, y)$ en $\mathbb{R}^d$. Sean μ y ν probabilidades en $\mathbb{R}^d$, absolutamente continuas con respecto a la medida de Lebesgue, y sea $C_2[(z_t)] = \int_0^1 c_2(\dot{z}_t) \, \mathrm{d}t$. Sea ∇f el único ($\mathrm{d}\mu$-casi siempre) gradiente de una función convexa f tal que $[Id - \nabla f]\#\mu = \nu$. Entonces la solución del problema* (1.21) *viene determinada por*

$$T_t(x) = x - t\nabla f(x), \quad 0 \leq t \leq 1. \tag{1.24}$$

Demostración. Si existe, un optimizador ha de verificar que $T_0(x) = x$ y $T_1(x) = x - \nabla f(x)$,y $T_t(x)$ ha de interpolar linealmente $T_0(x)$ y $T_1(x)$. Esto prueba la ecuación de arriba.

Veamos ahora que la familia $T_t(x) = x - t\nabla f(x)$ minimiza el funcional dinámico

$$C_2[(z_t)] = \int_0^1 \|\dot{z}_t\|^2 \, \mathrm{d}t,$$

donde, para cada x, $z_t(x) := T_t(x)$ define una curva en $\mathbb{R}^d$ que va de x a $x - \nabla f(x)$.

Por la proposición previa, sabemos que para cada $x \in \mathbb{R}^d$, la trayectoria que minimiza la acción

$$\int_0^1 \|\dot{z}_t\|^2 \, \mathrm{d}t$$

entre los puntos x y $x - \nabla f(x)$ es la línea recta

$$z_t(x) = x - t\nabla f(x).$$

Lo que finaliza la prueba. □

En este caso la solución de esta formulación coincide con el *interpolante de McCann*:

$$\gamma(t) = \left[(1 - t)\mathrm{Id} + t\nabla f)\#\mu\right], \quad \gamma(0) = \mu, \quad \gamma(1) = \nu.$$

1.3.1. Ecuación de continuidad

Sean μ, ν dos probabilidades absolutamente continuas en $\mathbb{R}^d$, en virtud del teorema 1.4, sabemos que existe un único (casi siempre) gradiente de una función convexa ∇f tal que $\nabla f \# \mu = \nu$. Se define

$$\mathsf{p}_t = [\mu, \nu]_t := [(1-t)\mathrm{Id} + t\nabla f] \, \# \mu.$$

Obviamente se tiene

$$\mathsf{p}_0 = [\mu, \nu]_0 := \mu, \quad \mathsf{p}_1 = [\mu, \nu]_1 := \nu$$

Es decir, la familia $(\mathsf{p}_t)_{0 \le t \le 1}$ interpola de forma lineal μ y ν. De hecho, del teorema 1.4 se sigue que el coste de transporte óptimo de μ a p_t es:

$$\begin{aligned} \mathcal{T}(\mu, \mathsf{p}_t) &= \int_{\mathbb{R}^d} \|x - [(1-t)x + t\nabla f(x))]\|^2 \, \mathrm{d}\mu(x) \\ &= t^2 \int_{\mathbb{R}^d} \|x - \nabla f(x))\|^2 \, \mathrm{d}\mu(x) = t^2 \mathcal{T}(\mu, \nu). \end{aligned}$$

En términos de la distancia 2 de Wasserstein $\mathcal{W}_2 = \sqrt{\mathcal{T}}$ se tiene

$$\mathcal{W}_2(\mu, \mathsf{p}_t) = t\mathcal{W}_2(\mu, \nu).$$

Por otro lado, notemos que no necesariamente ha de existir la densidad de p_t, sin embargo tenemos la condición suficiente de que si μ y ν tienen densidad entonces p_t también la tiene. Recogemos esta y otra propiedad en el siguiente resultado.

Proposición 1.7. *Con la notación de arriba, se tienen las siguientes propiedades*

1. $[\mu, \nu]_t = [\nu, \mu]_{1-t}$.
2. *Si μ o ν son absolutamente continuas (con respecto a la medida de Lebesgue), entonces p_t también lo es para todo $t \in [0, 1]$.*

Demostración. 1. Notemos que

$$\begin{aligned} [\mu, \nu]_t &= ((1-t)\mathrm{Id} + t\nabla f)\#\mu \\ &= ((1-t)\mathrm{Id} + t\nabla f)\#(\nabla f^* \# \nu) \\ &= [((1-t)\mathrm{Id} + t\nabla f) \circ \nabla f^*] \, \# \nu \\ &= ((1-t)f^* + t\mathrm{Id})\#\nu. \end{aligned}$$

2. En virtud de lo anterior basta considerar el caso en que μ sea absolutamente continua. Definimos

$$f_t(x) = tf(x) + (1-t)\frac{\|x\|^2}{2},$$

y notemos que

$$\langle \nabla f_t(x) - \nabla f_t(y), x - y \rangle \ge (1-t)\|x-y\|^2.$$

En particular,

$$\|\nabla f_t(x) - \nabla f_t(y)\| \ge (1-t)\|x-y\|.$$

f_t es uniformemente convexa y por ello f_t^* es diferenciable en todo punto, y por la ecuación anterior se deduce que $\nabla f_t^* = (\nabla f)^{-1}$ es Lipschitz. En particular, si tenemos un conjunto de medida de Lebesgue nula A entonces $\nabla f_t^*(A)$ es también cero y se puede escribir (ver lema C.2 del apéndice C)

$$\mathsf{p}_t[A] = \mu[\partial f_t^*(A)] = \mu[\nabla f_t^*(A)] = 0.$$

□

Estamos en condiciones de plantear la *ecuación de continuidad*. Sean $(T_t)_{0\leq t\leq 1}$ las soluciones de (1.21), que sabemos que son rectas; consideramos la medida de probabilidad en tiempos intermedios $\mathsf{p}_t = T_t\#\mu$. Nos planteamos cuál es la evolución natural de p_t, es por ello que surge el siguiente teorema.

Teorema 1.8 (Ecuación de continuidad)**.** *Consideramos $\mathbb{R}^d$ y sea $(T_t)_{0\leq t\leq 1}$ una familia localmente Lipschitz de difeomorfimos en $\mathbb{R}^d$, con $T_0 = Id$. Sea, además, $v_t = v(t,x)$ un campo de vectores con trayectorias (T_t). Sea μ una probabilidad en $\mathbb{R}^d$ y $\mathsf{p}_t = T_t\#\mu$. Entonces, $\mathsf{p}_t = \mathsf{p}(t,\cdot)$ es la única solución de la ecuación de transporte*

$$\frac{\partial \mathsf{p}_t}{\partial t} + \nabla\cdot(\mathsf{p}_t v_t) = 0, \quad 0 < t < 1, \tag{1.25}$$

con $\mathsf{p}_0 = \mu$, en $\mathcal{C}([0,1);\mathcal{P}(\mathbb{R}^d))$, donde en $\mathcal{P}(\mathbb{R}^d)$ se considera la topología débil.

Nota. Cuando decimos que $(T_t)_{0\leq t\leq 1}$ es una familia de localmente Lipschitz nos referimos a:

- T_t es una biyección $\mathbb{R}^d \to \mathbb{R}^d$ para todo t.
- Para todo $T < 1$ y para todo compacto $K \subset \mathbb{R}^d$, las aplicaciones $(t,x) \to T_t(x)$ y T_t^{-1} son Lipschitz en $[0,T]\times K$.

Demostración. En primer lugar veamos que $\mathsf{p}_t = T_t\#\mu$ verifica la ecuación de continuidad (1.25). Como hemos advertido la igualdad es en sentido débil: sea $g \in \mathcal{D}(\mathbb{R}^d)$ una función test (función $\mathcal{C}^\infty$ con soporte compacto) y $T \in (0,1)$, la aplicación $t \longrightarrow \int g\,\mathrm{d}\mathsf{p}_t$ es Lipschitz en $(0,T)$: sea, $t_1, t_2 \in (0,T)$

$$\begin{aligned}\left\|\int g\,\mathrm{d}\mathsf{p}_{t_1} - \int g\,\mathrm{d}\mathsf{p}_{t_2}\right\| &= \left\|\int (g\circ T_{t_1})\,\mathrm{d}\mu - \int (g\circ T_{t_2})\,\mathrm{d}\mu\right\| \\ &\leq \int \|[(g\circ T_{t_1}) - (g\circ T_{t_2})]\|\,\mathrm{d}\mu.\end{aligned}$$

Dado que g es una función test y T_t^{-1} es continua entonces la función $f\circ T_t$ tiene soporte compacto. Además, de nuevo, por ser g función test se tiene $\|\mathbf{D}g(x)\| \leq M$ entonces:

$$\int M\|T_{t_1} - T_{t_2}\|\,\mathrm{d}\mu \leq ML|t_1 - t_2|.$$

Por otro lado, para casi todo x, t

$$\frac{\partial}{\partial t}(g\circ T) = (\nabla\circ T)\cdot\frac{\partial T_t}{\partial t} = (\nabla\circ T)\cdot(v_t\circ T_t).$$

La primera igualdad se debe a la regla de la cadena mientras que la segunda a la definición de v_t. Entonces, para $h > 0$ se puede escribir

$$\frac{1}{h}\left(\int g\,\mathrm{d}\mathsf{p}_{t+h} - \int g\,\mathrm{d}\mathsf{p}_t\right) = \int\left(\frac{g\circ T_{t+h}(x) - g\circ T_t(x)}{h}\right)\mathrm{d}\mu.$$

El integrando está uniformemente acotado en $[0, T-h] \times \mathbb{R}^d$, y para casi todo t converge a $(\nabla g \circ T_t) \cdot v_t$ cuando $h \to 0$, para casi todo x.

En virtud del teorema de la convergencia dominada, se tiene que $t \longrightarrow \int g \, \mathrm{d}\mathsf{p}_t$ es diferenciable para casi todo t, y

Además, $t \longrightarrow \int g \, \mathrm{d}\mathsf{p}_t$ tiene derivada

$$\frac{\mathrm{d}}{\mathrm{d}t} \int g \, \mathrm{d}\mathsf{p}_t = \int (\nabla g \circ T_t) \cdot (v_t \circ T_t) \, \mathrm{d}\mu = \int (\nabla g \cdot v_t) \, \mathrm{d}\mathsf{p}_t,$$

que es lo que queríamos.

Ahora, veamos la unicidad. Por linealidad, si existiesen dos soluciones $\tilde{\mathsf{p}}_t$ y p'_t con el mismo dato inicial, entonces su diferencia $\mathsf{p}_t = \tilde{\mathsf{p}}_t - \mathsf{p}'_t$ también satisfaría la ecuación de continuidad con condición inicial $\mathsf{p}_0 = 0$. Veamos que $\mathsf{p}_T = 0$ para todo $T < 1$.

Supongamos que podemos construir una función Lipschitz $g(t, x)$ definida en $[0, T]$, de soporte compacto y que resuelve

$$\begin{cases} \frac{\partial g}{\partial t} = -v \cdot \nabla g, \\ \\ g|_{t=T} = g_T, \end{cases}$$

donde g_T es una función test arbitraria. Entonces, razonando de manera análoga a la parte de arriba se tiene que $t \longrightarrow \int g_t \, \mathrm{d}\mathsf{p}_t$ es Lipschitz y satisface (haciendo uso de la continuidad en $t = 0$).

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} \int g_t \, \mathrm{d}\mathsf{p}_t &= \int \frac{\partial g_t}{\partial t} \, \mathrm{d}\mathsf{p}_t + \int g_t \, \mathrm{d}(\frac{\partial \mathsf{p}_t}{\partial t}) \\ &= -\int v_t \cdot \nabla g_t \ \mathrm{d}\mathsf{p}_t + \int g_t \ \mathrm{d}[\nabla \cdot (v_t \mathsf{p}_t)] = 0, \end{aligned}$$

para casi todo t; luego

$$\int g_T \, \mathrm{d}\mathsf{p}_T = \int g_0 \, \mathrm{d}\mathsf{p}_0 = 0.$$

Como la función g_T es arbitraria se tiene $\mathsf{p}_T = 0$. Esto significa que $\tilde{\mathsf{p}}_t = \mathsf{p}'_t$, probando la unicidad.

Solamente queda construir la función $g(t, x)$. La ecuación que ha de verificar se puede reescribirse como

$$\frac{\partial g}{\partial t} + v \cdot \nabla g = 0.$$

Dicha solución debe satisfacer

$$g_t(T_t x) = g_T(T_T x),$$

o, equivalentemente

$$g_t = g_T \circ T_T \circ T_t^{-1}.$$

Puesto que (T_t) es una familia locamente Lipschitz esta fórmula define una función que cumple con lo que queremos. □

Nota. ▪ Insistimos en que la ecuación anterior se verifica en *sentido débil*. Al conjunto de pares (p_t, v_t) que cumplen esta ecuación lo denotaremos por C.

- La *ecuación de continuidad* es el equivalente a que se cumplan las restricciones marginales (es decir, el conjunto $\Pi(\mu,\nu)$ mencionado antes). Expresa las condiciones necesarias en los extremos de la trayectoria $(\mathsf{p}_t)_{t\in[0,1]}$.

La ecuación (1.25) es una forma particular de la ecuación de Fokker-Planck (ver apéndice C) en donde no aparece el término de difusión. En este caso, la evolución de $\mathsf{p}(x,t)$ está determinada por $v(x,t)$, y no por ruido. Esta ecuación está asociada con la ecuación diferencial estocástica siguiente

$$\mathrm{d}X_t = v_t(X_t)\,\mathrm{d}t.$$

En este caso el transporte es completamente determinista.

Esta dinámica es central en el estudio del transporte óptimo, donde la *ecuación de continuidad* describe la evolución de la masa de probabilidad bajo el flujo $v(x,t)$.

Desde este punto de vista se llega a la formulación del mínimo camino de $\mathcal{W}_2$ introducida por Brenier en [2].

Teorema 1.9 (Teorema de Benamou-Brenier)**.** *En las condiciones anteriores se tiene*

$$\mathcal{W}_2^2(\mu,\nu) = \min_{(\mathsf{p}_t,v_t)\in\mathsf{C}} \int_0^1 \int_{\mathbb{R}^d} \|v_t(x)\|^2 \,\mathrm{d}\mathsf{p}_t(x)\,\mathrm{d}t. \tag{1.26}$$

Demostración. En virtud del teorema 1.6, sabemos que el transporte óptimo dinámico entre μ y ν está dado por la familia de aplicaciones $T_t(x) = x - t\nabla f(x)$, donde ∇f es el gradiente de una función convexa tal que $T_1\#\mu = \nu$.

Por la proposición 1.5, las curvas $z_t(x) := T_t(x)$ minimizan el coste de desplazamiento

$$C_2[(z_t(x))] = \int_0^1 \|\dot{z}_t(x)\|^2 \,\mathrm{d}t = \|\nabla f(x)\|^2,$$

y por tanto el coste total mínimo viene dado por

$$\int_{\mathbb{R}^d} \|\nabla f(x)\|^2 \,\mathrm{d}\mu(x) = \mathcal{W}_2^2(\mu,\nu).$$

Además, según el teorema 1.8, la familia de medidas interpoladas $\mathsf{p}_t = T_t\#\mu$ y el campo de velocidades constante $v_t(x) = -\nabla f(x)$ satisfacen la ecuación de continuidad.

Esto demuestra que dicho par (p_t, v_t) es admisible y alcanza el mínimo en la expresión dinámica. Para cualquier otro par (p_t, v_t) admisible, el coste asociado siempre es mayor o igual, por lo que se concluye que:

$$\mathcal{W}_2^2(\mu,\nu) = \min_{(\mathsf{p}_t,v_t)\in\mathsf{C}} \int_0^1 \int_{\mathbb{R}^d} \|v_t(x)\|^2 \,\mathrm{d}\mathsf{p}_t(x)\,\mathrm{d}t.$$

□

1.4. Consideraciones computacionales

El desarrollo hasta aquí descrito se basa en el caso de $\mathbb{R}^d$, función de coste cuadrático y medidas de $\mathcal{P}_2(\mathbb{R}^d)$. En este trabajo también estamos interesados en aplicación computacional, en el que se pasará al caso discreto. Un libro de referencia en estos aspectos es [31].

En primer lugar notemos que todos los resultados de existencia siguen siendo válidos pues basta elegir medidas discretas. Estas cumplen las hipótesis con las que estamos trabajando.

Se consideran puntos $\{x_i\}_{i=1}^m$ de $\mathbb{R}^d$ con masas asociadas p_i y puntos $\{y_j\}_{j=1}^n$ de $\mathbb{R}^{d'}$ con masas q_j (realmente no tienen por qué ser de la misma dimensión). Estas masas pueden normalizarse para obtener distribuciones de probabilidad en ambos espacios que llamaremos de nuevo μ y ν. En este contexto el problema sería encontrar un plan de transporte $\gamma \in \mathbb{R}^{d\times d'}$ que minimice el coste cuadrático total:

$$\min_{\pi\in\Pi(\mu,\nu)} \sum_{i=1}^{m}\sum_{j=1}^{n} \|x_i - y_j\|^2 \pi_{i,j},$$

sujeto a las siguientes restricciones:

$$\begin{aligned}
&\sum_{j=1}^{n} \pi_{i,j} = p_i, && \text{para todo } i = 1,\dots,m,\\
&\sum_{i=1}^{m} \pi_{i,j} = q_j, && \text{para todo } j = 1,\dots,n,\\
&\pi_{i,j} \geq 0, && \text{para todo } i,j.
\end{aligned}$$

donde $\Pi(\mu,\nu)$ denota el conjunto de planes de transporte con las marginales μ y ν.

En este caso el problema a minimizar es resolver un problema de programación lineal. Por ello también hay una formulación dual asociada: introducimos variables duales $\phi_i \in \mathbb{R}^d$ y $\psi_j \in \mathbb{R}^{d'}$ asociadas a las masas p_i y q_j. El problema dual es:

$$\max_{\phi,g} \left\{ \sum_{i=1}^{m} p_i\phi_i + \sum_{j=1}^{n} q_j\psi_j \right\},$$

sujeto a:

$$\phi_i + \psi_j \leq \|x_i - y_j\|^2 \quad \text{para todo } i = 1,\dots,m,\ j = 1,\dots,n.$$

En dicho libro [31] se expone que el problema de transporte clásico tiene una complejidad computacional alta, ya que los métodos tradicionales como la programación lineal pueden requerir un tiempo de ejecución de $O(n^3)$ en el peor de los casos. Esto lo hace ineficiente para problemas de gran tamaño. Por otro lado es muy sensible a perturbaciones. Pequeñas variaciones en los datos pueden provocar cambios bruscos en la solución, lo que afecta su estabilidad.

Es por este tipo de motivos por los que se introduce la formulación entrópica, que añade un parámetro de regularización. La veremos a continuación.

Capítulo 2

Problema de transporte óptimo entrópico

La formulación entrópica surge al introducir un término de regularización, por ello mejora el problema de transporte óptimo clásico al suavizar la solución, evitando asignaciones exactas y generando correspondencias más distribuidas y estables. Además, facilita la resolución computacional del problema, ya que convierte el modelo en uno más cercano a problemas de optimización convexa, mejorando su eficiencia y estabilidad. Por último, en problemas de alta dimensión o grandes volúmenes de datos, la entropía favorece una mejor convergencia.

En este capítulo veremos varias formulaciones: la estática y la dinámica. Dentro de la segunda entraremos en la formulación estocástica, vía la *ecuación de Fokker-Planck*. También introduciremos el *algoritmo de Sinkhorn*.

La perturbación que añadiremos será la *divergencia de Kullback-Leibler* por un factor de escala ε.

Definición 2.1. La entropía relativa (entre medidas de probabilidad) o divergencia de Kullback-Leibler (entre dos medidas), dependerá del contexto, se define del siguiente modo:

$$\mathrm{KL}(\mu|\nu) := \begin{cases} \int_{\mathbb{R}^d} \log(\frac{\mathrm{d}\mu}{\mathrm{d}\nu})\,\mathrm{d}\mu, & \text{si } \mu \ll \nu, \\ \\ +\infty, & \text{en otro caso.} \end{cases}$$

En virtud de la regla de la cadena y la *desigualdad de Jensen* se prueba que $\mathrm{KL}(\mu|\nu) \geq 0$ y es $0 \iff \mu = \nu$.

Tal como hemos comentado añadiremos una perturbación y entonces el problema a minimizar será

$$\mathcal{T}_\varepsilon(\mu,\nu) := \min_{\pi\in\Pi(\mu,\nu)} \left\{ \int_{\mathbb{R}^d\times\mathbb{R}^d} \frac{1}{2}\|x-y\|^2\,\mathrm{d}\pi(x,y) \;+\; \varepsilon\,\mathrm{KL}(\pi|\mu\otimes\nu) \right\}. \tag{2.1}$$

Al igual que en el caso clásico llamaremos $I_\varepsilon[\pi]$ a lo que queremos minimizar, ponemos ε para hacer notar la dependencia de ese factor. En general no haremos referencia a esa dependencia, pero en el caso del enunciado del problema entrópico sí.

En general, se cumple que este problema es estrictamente convexo. En primer lugar, porque el primer término es un funcional lineal y en segundo lugar porque $\mathrm{KL}(P|Q)$ es estrictamente convexa con respecto a P:

Sean P_1 y P_2 probabilidades con densidades f_1 y f_2, respectivamente. Y Q con densidad g. Sea $t \in (0,1)$. Entonces

$$\begin{aligned}&\mathrm{KL}(tP_1 + (1-t)P_2|Q)\\&= \int \log\left(\frac{tf_1(x)+(1-t)f_2(x)}{g(x)}\right)\frac{tf_1(x)+(1-t)f_2(x)}{g(x)}g(x)\ \mathrm{d}x,\end{aligned}$$

es conocido que la función $x\log(x)$ es estrictamente convexa, luego

$$\begin{aligned}&\int \log\left(\frac{tf_1(x)+(1-t)f_2(x)}{g(x)}\right)\frac{tf_1(x)+(1-t)f_2(x)}{g(x)}g(x)\ \mathrm{d}x\\&< \int tf_1(x)\log\left(\frac{f_1(x)}{g(x)}\right)+(1-t)f_2(x)\log\left(\frac{f_2(x)}{g(x)}\right)\ \mathrm{d}x\\&= t\,\mathrm{KL}(P_1|Q)+(1-t)\,\mathrm{KL}(P_2|Q)\end{aligned}$$

y el conjunto $\Pi(\mu,\nu)$ es un subconjunto convexo del espacio de las medidas acotadas (se deduce de la linealidad de la integral y de la definición $\Pi(\mu,\nu)$). De esta forma pasamos de un problema lineal a una función objetivo estrictamente convexa, la cual tiene un solo mínimo que llamaremos *plan de transporte entrópico óptimo* (en caso de existir).

Nos fijamos en $\pi \ll \mu\otimes\nu$ (el otro caso al ser $+\infty$ no interesa en nuestro problema de minimización). Reescribimos $\mathrm{d}\pi$ del siguiente modo:

$$\mathrm{d}\pi(x,y) = \frac{\mathrm{d}\pi(x,y)}{\mathrm{d}(\mu\otimes\nu)}\,\mathrm{d}(\mu\otimes\nu) = \frac{\mathrm{d}\pi(x,y)}{\mathrm{d}(\mu\otimes\nu)}\,\mathrm{d}\mu(x)\,\mathrm{d}\nu(y).$$

Llamamos $r(x,y) = \frac{\mathrm{d}\pi(x,y)}{\mathrm{d}(\mu\otimes\nu)}$. Luego,

$$\mathrm{d}\pi(x,y) = r(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\nu(y).$$

Esa función $r(x,y)$ será de alta importancia en la formulación dual, veremos su forma explícita a continuación.

2.1. Formulación estática

Al igual que en el problema clásico nos interesa conocer una formulación dual del problema que sea equivalente, para ello manipulemos la expresión (2.1). Se tiene

Teorema 2.2 (Teorema de dualidad)**.** *Sean $\mathbb{R}^d$, $\varepsilon > 0$, $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ y $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ probabilidades. Consideremos $c_2(x,y)$ la función de coste cuadrático. Con la notación introducida hasta aquí se tiene:*

$$\min_{\pi\in\Pi(\mu,\nu)} I_\varepsilon[\pi] = \max_{(f,g)\in L^1(\mu)\times L^1(\nu)} J_\varepsilon(f,g), \tag{2.2}$$

donde

$$\begin{aligned}J_\varepsilon(f,g) := &\int_{\mathbb{R}^d} f(x)\ \mathrm{d}\mu(x) + \int_{\mathbb{R}^d} g(y)\ \mathrm{d}\nu(y)\\&+\varepsilon\int_{\mathbb{R}^d\times\mathbb{R}^d}\left[\exp\left(\frac{f(x)+g(y)-c_2(x,y)}{\varepsilon}\right)-1\right]\ \mathrm{d}\mu(x)\ \mathrm{d}\nu(y).\end{aligned}$$

Manipulando la expresión $I_\varepsilon[\pi]$

$$I_\varepsilon[\pi] = \int_{\mathbb{R}^d} c_2(x,y)r(x,y) + \varepsilon \log\left(r(x,y)\right) r(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\nu(y). \tag{2.3}$$

Sean $(f,g) \in L^1(\mu) \times L^1(\nu)$, restamos el siguiente término a ambos lados de la igualdad

$$\int_{\mathbb{R}^d} f(x)\,\mathrm{d}\mu(x) + \int_{\mathbb{R}^d} g(y)\,\mathrm{d}\nu(y) \equiv \int_{\mathbb{R}^d\times\mathbb{R}^d} (f(x)+g(y))r(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\nu(y).$$

Si reorganizamos la parte de la derecha (con el término ya restado) se llega a

$$\begin{aligned} I_\varepsilon[\pi] = &\int_{\mathbb{R}^d} f(x)\ \mathrm{d}\mu(x) + \int_{\mathbb{R}^d} g(y)\ \mathrm{d}\nu(y) \\ &+ \varepsilon \int_{\mathbb{R}^d\times\mathbb{R}^d} r(x,y) \log\left(\frac{r(x,y)}{\exp\left(\frac{f(x)+g(y)-c_2(x,y)}{\varepsilon}\right)}\right)\ \mathrm{d}\mu(x)\ \mathrm{d}\nu(y). \end{aligned}$$

Queremos probar que el mínimo de esa expresión es igual al superior de $J_\varepsilon(f,g)$.

- En primer lugar veremos que la primera expresión es mayor o igual que la segunda (se utilizará $s\log(s) \geq s-1, s \geq 0$). Obviamos las dos primeras integrales ya que se van a mantener a ambos lados de la igualdad:

$$\begin{aligned} &\varepsilon \int_{\mathbb{R}^d\times\mathbb{R}^d} r(x,y) \log\left(\frac{r(x,y)}{\exp\left(\frac{f(x)+g(y)-\frac{1}{2}\|x-y\|^2}{\varepsilon}\right)}\right)\ \mathrm{d}\mu(x)\ \mathrm{d}\nu(y) \geq \\ &\varepsilon\left(\int r(x,y)\,\mathrm{d}\mu(x)\ \mathrm{d}\nu(y) - \int_{\mathbb{R}^d\times\mathbb{R}^d} \exp\left(\frac{f(x)+g(y)-\frac{1}{2}\|x-y\|^2}{\varepsilon}\right)\ \mathrm{d}\mu(x)\ \mathrm{d}\nu(y)\right). \end{aligned}$$

  La integral de la izquierda, si imponemos que $r(x,y)$ sea una densidad, es 1 y llegamos a lo que queremos.

- Además $s\log(s) = s-1 \iff s = 1$: si encontrásemos f y g de manera que se cumpla esa equivalencia tendríamos la igualdad de las formulaciones y los óptimos. Es decir, si existen $(f^\star, g^\star)$ tal que

$$r(x,y) = \exp\left(\frac{f^\star(x) + g^\star(y) - 1/2\|x-y\|^2}{\varepsilon}\right),$$

  es una densidad con respecto a $\mu\otimes\nu$ y

$$\int_{\mathbb{R}^d} r(x,y)\,\mathrm{d}\nu(y) = 1, \tag{2.4}$$

$$\int_{\mathbb{R}^d} r(x,y)\,\mathrm{d}\mu(x) = 1. \tag{2.5}$$

  Puesto que estamos trabajando con función de coste cuadrático y con medidas de momento de orden 2 finito entonces siempre existen esas funciones y se pueden obtener mediante el *algoritmo de Sinkhorn*. Citamos [24] para encontrar esos resultados aplicados a espacios compactos, los veremos en la siguiente sección y los extenderemos a $\mathbb{R}^d$.

Entonces $\pi^\star$ y $(f^\star, g^\star)$ son óptimas. El término del logaritmo se anula y

$$\mathcal{T}_\varepsilon(\mu,\nu) = I_\varepsilon[\pi^\star] = \int_{\mathbb{R}^d} f^\star(x)\ \mathrm{d}\mu(x) + \int_{\mathbb{R}^d} g^\star(y)\ \mathrm{d}\nu(y).$$

En conclusión, se llega a la formulación dual del problema entrópico.

En virtud de las condiciones que hay que aplicar para las funciones en el óptimo se deduce que ha de cumplirse

$$\mathrm{d}\pi^\star(x,y) = \exp\left(\frac{f^\star(x) + g^\star(y) - 1/2\|x-y\|^2}{\varepsilon}\right)\mathrm{d}\mu(x)\,\mathrm{d}\nu(y). \tag{2.6}$$

Esas funciones que cumplen la igualdad se conocen como *potenciales entrópicos óptimos*.

Siguiendo con las ecuaciones (2.4), se tiene (razonamos con una de ellas, la otra es análoga)

$$\int_{\mathbb{R}^d} r(x,y)\,\mathrm{d}\mu(x) = 1 \tag{2.7}$$

$$\Longrightarrow \int_{\mathbb{R}^d} \exp\left(\frac{f^\star(x) + g^\star(y) - 1/2\|x-y\|^2}{\varepsilon}\right)\mathrm{d}\mu(x) = 1 \tag{2.8}$$

$$\Longrightarrow \int_{\mathbb{R}^d} \exp\left(\frac{f^\star(x) - 1/2\|x-y\|^2}{\varepsilon}\right)\mathrm{d}\mu(x) = \exp\left(\frac{g^\star(y)}{\varepsilon}\right). \tag{2.9}$$

Tomamos logaritmos y resulta

$$f^\star(x) = -\varepsilon\log\left(\int_{\mathbb{R}^d} \exp\left(\frac{g^\star(y) - 1/2\|x-y\|^2}{\varepsilon}\right)\mathrm{d}\nu(y)\right).$$

En conclusión, si un par de funciones cumplen las siguientes condiciones entonces necesariamente maximizan el dual.

$$f^\star(x) = -\varepsilon\log\left(\int_{\mathbb{R}^d} \exp\left(\frac{g^\star(y) - 1/2\|x-y\|^2}{\varepsilon}\right)\mathrm{d}\nu(y)\right), \tag{2.10}$$

$$g^\star(y) = -\varepsilon\log\left(\int_{\mathbb{R}^d} \exp\left(\frac{f^\star(x) - 1/2\|x-y\|^2}{\varepsilon}\right)\mathrm{d}\mu(x)\right). \tag{2.11}$$

Por otro lado, veamos un nuevo punto de vista que se obtiene aplicando definiciones: si tenemos un vector aleatorio $(X,Y) \sim \pi^\star$ llamamos *proyección baricéntrica* a $T_\varepsilon(X) = \mathbb{E}(Y|X=x)$. Si suponemos que μ tiene densidad p y ν tiene densidad q entonces

$$\begin{aligned} T_\varepsilon(X) &= \int_{\mathbb{R}^d} y \exp\left(\frac{f^\star(x) + g^\star(y) - 1/2\|x-y\|^2}{\varepsilon}\right) q(y)\,\mathrm{d}y \\ &= \int_{\mathbb{R}^d} y \exp\left(\frac{f^\star(x) + g^\star(y) - 1/2\|x-y\|^2}{\varepsilon}\right)\mathrm{d}\nu(y). \end{aligned}$$

Vamos a introducir la relación entre los *potenciales entrópicos óptimos* y la *proyección baricéntrica*. De las ecuaciones (2.4) se tiene

$$\exp\left(\frac{f^\star(x) - 1/2\|x\|^2}{\varepsilon}\right)\int_{\mathbb{R}^d} \exp\left(\frac{g^\star(y) - 1/2\|y\|^2 + x\cdot y}{\varepsilon}\right)\mathrm{d}\nu(y) = 1.$$

Definimos las funciones, que se conocen como *potenciales entrópicos de Brenier*

$$\varphi^\star(x) = \left[\frac{\|x\|^2}{2} - f^\star(x)\right], \quad \psi^\star(y) = \left[\frac{\|y\|^2}{2} - g^\star(y)\right].$$

Por tanto la igualdad anterior se convierte en

$$\exp\left(\frac{\varphi^\star(x)}{\varepsilon}\right) = \int_{\mathbb{R}^d} \exp\left(\frac{x\cdot y - \psi^\star(y)}{\varepsilon}\right) \mathrm{d}\nu(y),$$

$$\Longrightarrow \varphi^\star(x) = \varepsilon\ \log\left(\int_{\mathbb{R}^d} \exp\left(\frac{x\cdot y - \psi^\star(y)}{\varepsilon}\right) \mathrm{d}\nu(y)\right).$$

Esa función es $\mathcal{C}^\infty$ (ver apéndice C) y se tiene

$$\nabla\varphi^\star(x) = \frac{\int_{\mathbb{R}^d} \exp\left(\frac{x\cdot y - \psi^\star(y)}{\varepsilon}\right)\ y\ \mathrm{d}\nu(y)}{\int_{\mathbb{R}^d} \exp\left(\frac{x\cdot y - \psi^\star(y)}{\varepsilon}\right)\ \mathrm{d}\nu(y)} \tag{2.12}$$

En virtud de la expresión de la *proyección baricéntrica* se deduce que

$$\begin{aligned} T_\varepsilon(x) &:= \mathbb{E}(Y|X=x) = \int_{\mathbb{R}^d} y \exp\left(\frac{f^\star(x) + g^\star(y) - 1/2\|x-y\|^2}{\varepsilon}\right) \mathrm{d}\nu(y) \\ &= \frac{\int_{\mathbb{R}^d} y \exp\left(\frac{f^\star(x)+g^\star(y)-1/2\|x-y\|^2}{\varepsilon}\right) \mathrm{d}\nu(y)}{\int_{\mathbb{R}^d} \exp\left(\frac{f^\star(x)+g^\star(y)-1/2\|x-y\|^2}{\varepsilon}\right) \mathrm{d}\nu(y)} = \frac{\int_{\mathbb{R}^d} \exp\left(\frac{x\cdot y-\psi^\star(y)}{\varepsilon}\right)\ y\ \mathrm{d}\nu(y)}{\int_{\mathbb{R}^d} \exp\left(\frac{x\cdot y-\psi^\star(y)}{\varepsilon}\right)\ \mathrm{d}\nu(y)} = \nabla\varphi^\star(x). \end{aligned}$$

Donde se ha divido por 1 (recordemos la condición sobre $r(x,y)$) y se ha utilizado la definición de las funciones φ y ψ.

Este gradiente es conocido como *aplicaciones entrópicas de Brenier*. Además, como propiedad adicional ([8]) se tiene

$$\nabla^2\varphi^\star(x) = \frac{1}{\varepsilon}\mathrm{Cov}_{\pi^\star}[Y|X=x]. \tag{2.13}$$

Esto se debe a

$$\nabla^2\varphi^\star(x) = \frac{1}{\varepsilon}\left[\frac{\int_{\mathbb{R}^d} yy^T \exp\left(\frac{x\cdot y-\psi^\star(y)}{\varepsilon}\right)\ \mathrm{d}\nu(y)}{\int_{\mathbb{R}^d} \exp\left(\frac{x\cdot y-\psi^\star(y)}{\varepsilon}\right)\ \mathrm{d}\nu(y)} - \nabla\varphi^\star(x)\nabla\varphi^\star(x)^T\right].$$

que justamente es $\frac{1}{\varepsilon}\mathrm{Cov}_{\pi^\star}[Y|X=x]$. Notemos que, por ende, $\nabla^2\varphi^\star \geq 0$ lo que implica que $\varphi^\star$ es convexa. En conclusión, la *proyección baricéntrica* es una aplicación de transporte óptimo, pero no lleva μ a ν sino a algo aproximado a ν (dado por el ε). Notemos que estos razonamientos son análogos para la otra función del par considerado.

2.2. Algoritmo de Sinkhorn

En esta sección veremos el *algoritmo de Sinkhorn*, que se utilizará para obtener las funciones óptimas $f^\star$ y $g^\star$. En primer lugar haremos una adaptación del teorema clásico que se encuentra en [24]. En este teorema se trabaja con espacios compactos, que no coincide con nuestro caso de estudio, es por ello que lo extenderemos a $\mathbb{R}^d$.

Teorema 2.3 (Algoritmo de Sinkhorn caso compacto)**.** *Sean M y N dos espacios compactos y μ y ν medidas Borel regulares no negativas en M y N respectivamente, tal que $\mu(M) > 0$ y $\nu(N) > 0$. Sea $H(x,y) = \exp\left(\frac{-1/2\|x-y\|^2}{\varepsilon}\right)$. Entonces existen funciones $f(x)$ y $g(x)$ positivas y continuas en M y N, respectivamente, tal que*

$$\begin{cases} \int_M f(x) \exp\left(\frac{-1/2\|x-y\|^2}{\varepsilon}\right) g(y)\, \mathrm{d}\mu(x) = 1 \quad y \in N, \\ \\ \int_N f(x) \exp\left(\frac{-1/2\|x-y\|^2}{\varepsilon}\right) g(y)\, \mathrm{d}\nu(y) = 1 \quad x \in M. \end{cases}$$

Las forma $f(x) \exp\left(\frac{-1/2\|x-y\|^2}{\varepsilon}\right) g(y)$ se puede obtener como límite de la iteración en la que se escala alternativamente las función H para que tengan las integrales adecuadas sobre M y N. Además, si cada conjunto abierto no vacío en M y N tiene medida positiva, entonces el producto $f(x) \exp\left(\frac{-1/2\|x-y\|^2}{\varepsilon}\right) g(y)$ es único. Las funciones f y g son únicas salvo un múltiplo escalar positivo.

La iteración comentada es la siguiente. Partimos de

$$f_0(x) = 1, \quad g_0(y) = \frac{1}{\int_M H(x,y)\, \mathrm{d}\mu(x)}.$$

Ahora, la iteración se construye empezando con (f_0, g_0) y siguiendo con

$$f_{n+1}(x) = a_n(x) f_n(x), \quad g_{n+1}(y) = b_n(y) g_n(y), \quad n = 0, 1, 2, \ldots,$$

donde

$$\begin{cases} a_n(x) = \frac{1}{\int_N f_n(x) H(x,y) g_n(y)\, \mathrm{d}\nu(y)}, \\ \\ b_n(x) = \frac{1}{\int_M a_n(x) f_n(x) H(x,y) g_n(y)\, \mathrm{d}\mu(x)}. \end{cases}$$

Teorema 2.4 (Algoritmo de Sinkhorn caso $\mathbb{R}^d$)**.** *Si suponemos $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ el teorema anterior se extiende a $\mathbb{R}^d$.*

Demostración. Sea $R_n \nearrow \infty$ una sucesión creciente y definamos $M_n = \overline{B}(0, R_n)$, $N_n = \overline{B}(0, R_n)$. Definimos las truncaciones normalizadas

$$\tilde{\mu}_n = \frac{\mu|_{M_n}}{\mu(M_n)}, \qquad \tilde{\nu}_n = \frac{\nu|_{N_n}}{\nu(N_n)}.$$

Como $\mu(M_n) \to 1$ y $\nu(N_n) \to 1$, se tiene que $\tilde{\mu}_n \to \mu$, $\tilde{\nu}_n \to \nu$ débilmente. Para cada n, aplicamos el teorema de Sinkhorn en $M_n \times N_n$, obteniendo funciones continuas positivas $f_n : M_n \to (0, \infty)$, $g_n : N_n \to (0, \infty)$ tales que

$$\int_{M_n} f_n(x)\, H(x,y)\, g_n(y)\, \mathrm{d}\tilde{\mu}_n(x) = 1, \qquad \int_{N_n} f_n(x)\, H(x,y)\, g_n(y)\, \mathrm{d}\tilde{\nu}_n(y) = 1,$$

para todo $x \in M_n, y \in N_n$, donde $H(x,y) = \exp(-\frac{1}{2\varepsilon}\|x-y\|^2)$. Si cada abierto no vacío tiene medida positiva con respecto a μ, ν, entonces f_n, g_n están determinadas salvo un factor escalar.

Sea $K \subset \mathbb{R}^d$ compacto. Tomando $K' \supset K$ también compacto, con $K' \subset M_n \cap N_n$ para n grande, como H es continua, existen constantes $m_{K,K'}, M_{K,K'} > 0$ tales que

$$m_{K,K'} \leq H(x,y) \leq M_{K,K'} \qquad \text{para todo } x \in K, y \in K'.$$

Además, si $\int_{K'} g_n \, \mathrm{d}\tilde{\nu}_n \geq c > 0$, entonces

$$f_n(x) \leq \frac{1}{m_{K,K'}\, c} \qquad \text{para todo } x \in K.$$

De modo similar, $g_n(y)$ está uniformemente acotada superiormente en K, y se obtienen cotas inferiores usando los mismos argumentos: si $f_n(x)\, H(x,y)\, g_n(y) \leq 1$ y $H \geq m > 0$, entonces $f_n(x) \geq \frac{1}{M_{K,K'} \int_{K'} g_n \, \mathrm{d}\tilde{\nu}_n}$. En conjunto, existen constantes $0 < c_K \leq C_K < \infty$ tales que

$$c_K \leq f_n(x) \leq C_K, \qquad c_K \leq g_n(y) \leq C_K \qquad \text{para todo } x, y \in K, n \text{ grande.}$$

Esto implica que $\{f_n\}, \{g_n\}$ son equicontinuas y uniformemente acotadas en compactos, y por Arzelà − Ascoli (ver [27]) existe una subsucesión que converge uniformemente en compactos a funciones continuas $f, g : \mathbb{R}^d \to (0, \infty)$.

Ahora pasamos al límite en las ecuaciones:

$$\int_{M_n} f_n(x)\, H(x,y)\, g_n(y) \, \mathrm{d}\tilde{\mu}_n(x) = 1 \quad \text{para cada } y.$$

Como $f_n \to f$, $g_n \to g$ uniformemente en compactos, y $H(x,y)$ es continua, el integrando converge puntualmente. Además, dado que f_n, g_n están acotadas y $H(x,y) \leq e^{-\|x-y\|^2/(2\varepsilon)}$, existe una cota integrable común para aplicar convergencia dominada:

$$f_n(x)\, H(x,y)\, g_n(y) \leq C \, \exp\left(-\frac{\|x-y\|^2}{2\varepsilon}\right),$$

la cual pertenece a $L^1(\mu)$ porque $\mu \in \mathcal{P}_2$. Entonces

$$\int_{\mathbb{R}^d} f(x)\, H(x,y)\, g(y) \, \mathrm{d}\mu(x) = 1, \qquad \forall y \in \mathbb{R}^d.$$

Un razonamiento análogo da la segunda ecuación:

$$\int_{\mathbb{R}^d} f(x)\, H(x,y)\, g(y) \, \mathrm{d}\nu(y) = 1, \qquad \forall x \in \mathbb{R}^d.$$

Para la unicidad, si $(\tilde{f}, \tilde{g})$ es otra solución continua positiva, entonces el producto $\tilde{f}(x)\, H(x,y)\, \tilde{g}(y) = f(x)\, H(x,y)\, g(y)$ para $\mu \otimes \nu$-casi todo (x,y), y por continuidad se extiende a todo $\mathbb{R}^d \times \mathbb{R}^d$. Esto implica que existe $\lambda > 0$ tal que $\tilde{f} = \lambda f$, $\tilde{g} = \lambda^{-1} g$, es decir, la solución es única salvo un factor escalar.

□

2.3. Formulación dinámica

El problema de transporte óptimo entrópico también se puede entender desde un punto de vista dinámico. Utilizaremos la siguiente notación:

Sea $(\mathsf{p}_t)_{t\in[0,1]}$ una familia de medidas en $\mathcal{P}_2(\mathbb{R}^d)$, $(v_t)_{t\in[0,1]}$ una familia de campos de vectores. Al conjunto de pares (p_t, v_t) que cumplen la *ecuación de continuidad* (1.25) lo denotaremos nuevamente por C.

El siguiente teorema relaciona la formulación estática y la formulación dinámica. Aparece en [15, Teorema 4.3].

Teorema 2.5 (Fórmula de Benamou-Brenier para el problema entrópico)**.** *Si ν y ν son medidas absolutamente continuas y tienen entropía finita entonces*

$$\mathcal{T}_\varepsilon(\mu,\nu)+\frac{\varepsilon}{2}(H(\mu)+H(\nu))=\inf_{(\mathsf{p}_t,v_t)\in\mathsf{C}}\int_0^1\int_{\mathbb{R}^d}\left(\|v_t(x)\|^2+\frac{\varepsilon^2}{8}\|\nabla\log\mathsf{p}_t(x)\|^2\right)\,\mathrm{d}\mathsf{p}_t(x)\,\mathrm{d}t, \tag{2.14}$$

donde $H(\mu):=\int\log(\,\mathrm{d}\mu)\,\mathrm{d}\mu$ y $H(\nu):=\int\log(\,\mathrm{d}\nu)\,\mathrm{d}\nu$.

Los cálculos para este resultado se encuentran en [9]. Notemos que cuando $\varepsilon=0$ se tiene la fórmula de Benamou-Brenier para el problema clásico (1.26).

2.3.1. Formulación estocástica

Existe otra formulación del problema dinámico, esta viene dada por la ecuación de *Fokker-Planck*. Sea (p_t, b_t) en las condiciones de la formulación dinámica, este par satisface la ecuación si

$$\partial\mathsf{p}_t+\nabla\cdot(b_t\mathsf{p}_t)=\frac{\varepsilon}{2}\Delta\mathsf{p}_t,\quad\forall\, t\ \in[0,1]. \tag{2.15}$$

Denotaremos por $(\mathsf{p}_t, b_t)\in\mathsf{F}$ si un par (p_t, b_t) satisface la ecuación con condiciones iniciales $\mathsf{p}_0=\mu$ y $\mathsf{p}_1=\nu$. Además, se tiene que

$$\mathcal{T}_\varepsilon(\mu,\nu)+\varepsilon H(\mu)=\inf_{(\mathsf{p}_t,b_t)\in\mathsf{F}}\int_0^1\int_{\mathbb{R}^d}\left(\frac{1}{2}\|b_t(x)\|^2\right)\,\mathrm{d}\mathsf{p}_t(x)\,\mathrm{d}t. \tag{2.16}$$

Notemos que la *ecuación de continuidad* y la *ecuación de Fokker-Planck* son equivalentes si elegimos $b_t=v_t+\frac{\varepsilon}{2}\nabla\log\mathsf{p}_t$ y reescribimos $\Delta\mathsf{p}_t=\nabla\cdot(\mathsf{p}_t\nabla\log\mathsf{p}_t)$. Entramos en detalles en la siguiente proposición:

Proposición 2.6. *Bajo las condiciones impuestas hasta ahora, se tiene que* (2.16) *es equivalente a* (2.14).

Demostración. De nuevo, escribimos $\Delta\mathsf{p}_t=\nabla\cdot(\mathsf{p}_t\nabla\log\mathsf{p}_t)$ la *ecuación de Fokker-Planck* se escribe

$$\frac{\partial\mathsf{p}_t}{\partial t}+\nabla\cdot((v_t-\frac{\varepsilon}{2}\nabla\log\mathsf{p}_t)\mathsf{p}_t)=0.$$

Insertamos $b_t:=v_t-\frac{\varepsilon}{2}\nabla\log\mathsf{p}_t$ en (2.16) y simplificamos

$$\inf_{(\mathsf{p}_t,b_t)\in\mathsf{F}}\int_0^1\int_{\mathbb{R}^d}\left(\frac{1}{2}\|b_t(x)\|^2\ +\ \frac{\varepsilon^2}{8}\|\nabla\log\mathsf{p}_t(x)\|^2\ +\ \frac{\varepsilon}{2}b_t^T\nabla\log\mathsf{p}_t(x)\right)\mathsf{p}_t(x)\ \,\mathrm{d}x\,\mathrm{d}t,$$

donde se utiliza (T) para denotar la traspuesta. Utilizamos integración por partes y la *ecuación de Fokker-Planck* anterior

$$\begin{aligned}
&\int_0^1 \int_{\mathbb{R}^d} \left(b_t^T \mathsf{p}_t \nabla \log \mathsf{p}_t\right) \ \mathrm{d}x\, \mathrm{d}t \\
&= -\int_0^1 \int_{\mathbb{R}^d} \left(\nabla \cdot (b_t^T \mathsf{p}_t \log \mathsf{p}_t)\right) \ \mathrm{d}x\, \mathrm{d}t \\
&= \int_0^1 \int_{\mathbb{R}^d} (\partial_t \mathsf{p}_t) \log \mathsf{p}_t \ \mathrm{d}x\, \mathrm{d}t \\
&= \int_0^1 \int_{\mathbb{R}^d} \partial_t (\mathsf{p}_t \log \mathsf{p}_t) \ \mathrm{d}x\, \mathrm{d}t - \int_0^1 \int_{\mathbb{R}^d} \partial_t \mathsf{p}_t \ \mathrm{d}x\, \mathrm{d}t \\
&= \int_0^1 \partial_t \int_{\mathbb{R}^d} (\mathsf{p}_t \log \mathsf{p}_t) \ \mathrm{d}x\, \mathrm{d}t - \int_0^1 \partial_t \int_{\mathbb{R}^d} \mathsf{p}_t \ \mathrm{d}x\, \mathrm{d}t.
\end{aligned}$$

donde el primer término es la definición de $H(\mathsf{p}_t)$ utilizando la función de densidad (notación introducida en (2.14)) y el segundo término es 0 porque se integra en todo el espacio. Por tanto,

$$\int_0^1 \partial_t H(\mathsf{p}_t)\, \mathrm{d}t + 0 = H(\mathsf{p}_1) - H(\mathsf{p}_0) = H(\nu) - H(\mu).$$

□

Una propiedad clave es la relación de los optimizadores $b_t = v_t + \frac{\varepsilon}{2} \nabla \log \mathsf{p}_t$, es decir, si tenemos un par que optimice un problema inmediatamente va a optimizar el otro vía esa igualdad. Notemos que p_t es el mismo en ambos casos.

Capítulo 3

Puentes de Schrödinger

En este punto, se establece una conexión natural con otro problema fundamental: el problema del Puente de Schrödinger. En lugar de abordar el transporte de masas entre dos distribuciones de forma determinista o mediante una regularización entrópica artificial, este problema plantea una pregunta diferente: ¿cuál es la evolución más probable de un sistema de partículas sometidas a un proceso de difusión, si conocemos su distribución en dos momentos diferentes?

Este problema, formulado originalmente por Schrödinger en 1931, busca reconstruir la dinámica más factible de un sistema de partículas en un medio aleatorio, basándose en el principio de mínima desviación respecto a un proceso de referencia, en nuestro caso un *Movimiento Browniano* reversible. Además, es equivalente al problema de transporte óptimo entrópico como veremos en el desarrollo del capítulo.

Buscaremos minimizar la divergencia de Kullback-Leibler entre una medida de probabilidad de transporte y una medida de referencia, bajo ciertas restricciones de marginales. Antes de entrar en detalles es necesario introducir terminología y notación. Para este capítulo seguiremos Pooladian y Niles-Weed [32] y lo complementaremos con Leonard [26]. Este último artículo presenta resultados en condiciones más generales con las que tratamos aquí, por ello será necesario hacer una adaptación de ellos.

Además, se utilizará el apéndice B, el cual incluye algunos conceptos y resultados acerca de procesos estocásticos y ecuaciones diferenciales estocásticas.

Seguimos trabajando en $\mathbb{R}^d$, a partir de aquí denotaremos $\Lambda_s := (2\pi s)^{-d/2}$. Llamaremos $\mathcal{C} = \mathcal{C}([0,1], \mathbb{R}^d)$ al espacio de funciones continuas en el intervalo de tiempo $[0,1]$, dotado con la métrica uniforme. En estas condiciones es un espacio métrico separable y completo. Sea $Y = \mathcal{C}$ o $\mathbb{R}^d$, denotaremos $M_+(Y)$ es el espacio de las medidas positivas en un espacio Y y $\mathcal{P}(Y)$ al espacio de las medidas de probabilidad en Y. Equipamos a $\mathbb{R}^d$ con lo usual, la σ-álgebra de Borel. Detallamos a continuación conceptos al situarnos en $\mathcal{C}$.

Dada $P \in \mathcal{P}(\mathcal{C})$ (o $P \in M_+(\mathcal{C})$) denotaremos P_t a la ley marginal en t, es decir, si

$$\pi_t(x) = x(t), \quad x \in \mathcal{C} \Rightarrow P_t = \pi_t \# P.$$

De manera similar se define $P_{s,t}$ para $s, t \in [0,1]$ (e incluso para más índices temporales). Puesto que $\mathcal{C}$ es un espacio Polaco las probabilidades de Borel en él están completamente determinadas por las distribuciones finito dimensionales (ver apéndice B), es decir,

$$\pi_{s_1,\dots,s_k} \# P, \quad s_1, \dots, s_k \in [0,1].$$

En $\mathcal{C}$ juega un papel destacado la medida de Wiener, es decir, la ley del Movimiento Browniano estándar (ver apéndice B). En algunos momentos de esta sección será útil pensar en el proceso canónico en $\mathcal{C}$, dado por $X(t) = \pi_t(x),\ t \in [0,1]$. Este proceso canónico es simplemente la identidad, la escribimos en mayúsculas para enfatizar el hecho de que es un proceso estocástico. Del hecho de que $\mathcal{C}$ sea un espacio Polaco, en virtud de [1, Teorema 6.6.6], se deduce la existencia de probabilidades condicionadas regulares. Esto quiere decir que para cualquier probabilidad P (o medida positiva) en $\mathcal{C}$, y cualquier sub-σ álgebra,$\mathcal{G}$, de la σ-álgebra de Borel existe una función $Q(x, B)$ con $x \in \mathcal{C}$ y $B \subset \mathcal{C}$, tal que

1. Para x fijo $Q(x, \cdot)$ es una probabilidad en $\mathcal{C}$.
2. Para B fijo $Q(\cdot, B)$ es $\mathcal{G}$-medible y

$$Q(x, B) = P(B|\mathcal{G})(x) \text{ c.s.}$$

Notemos que si $\mathcal{G}$ es la σ-álgebra generada por X_t esto significa que

$$P(B) = \int_{\mathbb{R}^d} Q(x_t, B)\, \mathrm{d}P_{X(t)}(x_t).$$

Escribiremos $P(\cdot|x_t) := P_{X(t)}(x_t)$ para denotar esta probabilidad condicionada regular por el valor de X_t, con notación similar para las probabilidades condicionadas dados X_s y X_t, $s < t \in [0,1]$. Para el *problema de Schrödinger* es necesario considerar una medida de referencia que denotaremos por R. Esta medida de referencia cumple que la primera marginal ha de ser igual a la medida de Lebesgue y la medida condicionada $R(\cdot|x_0)$ ha de coincidir con la ley de un Movimiento Browniano que parte de x_0, de forma que

$$R = \int_{\mathbb{R}^d} R(\cdot|x_0)\, \mathrm{d}x_0.$$

Calculamos la primera marginal en el instante t. Sea $A \subset \mathbb{R}^d$ Borel, de acuerdo con la fórmula de arriba

$$R_t(A) = R(\{x \in \mathcal{C} : x(t) \in A\}) = \int_{\mathbb{R}^d} R(\{x \in \mathcal{C} : x(t) \in A\}|x_0)\, \mathrm{d}x_0,$$

La distribución del Movimiento Browniano en el instante t es $\mathcal{N}(x_0, t\mathrm{Id})$ (ver apéndice B) luego, si seguimos con lo anterior y aplicamos Fubini

$$= \int_{\mathbb{R}^d} \left[\int_A \Lambda_t \exp(-\frac{1}{2t}\|x_t - x_0\|^2)\, \mathrm{d}x_t \right] \mathrm{d}x_0$$
$$= \int_A \left[\int_{\mathbb{R}^d} \Lambda_t \exp(-\frac{1}{2t}\|x_t - x_0\|^2)\, \mathrm{d}x_0 \right] \mathrm{d}x_t.$$

La integral de dentro es 1 por ser la densidad normal integrada en todo el espacio, luego la expresión es igual a $\ell_d(A) = R_0(A)$. En otras palabras, la medida de Lebesgue es la medida invariante al semigrupo de transición del Movimiento Browniano (o semigrupo del calor):

$$\mathcal{H}_s[Q](z) := \Lambda_s \int \exp(-\frac{1}{2s}\|x - z\|^2)\, \mathrm{d}Q(x).$$

Con esta notación se reescribe $\mathcal{H}_t[R_0] = R_0$.

3.1. Problema del Puente de Schrödinger

El problema del Puente de Schrödinger (o problema de Schrödinger) al que se dedica esta sección es el de encontrar, de entre todas las probabilidades en $\mathcal{C}$ con marginal inicial μ y marginal final ν, la que minimiza la divergencia de Kullback-Leibler con parámetro ε. Es decir, el problema consiste en

$$\min_{P:P_0=\mu,\ P_1=\nu} \{\varepsilon\, \mathrm{KL}(P|R)\} \equiv \min_{P:\pi_0\#P=\mu,\pi_1\#P=\nu} \{\varepsilon\, \mathrm{KL}(P|R)\}. \tag{3.1}$$

Los comentarios anteriores sobre probabilidades y medidas condicionadas regulares nos permite escribir

$$P(A \cap ((X(0), X(1)) \in B) = \int_B P(A|x_0, x_1)\, \mathrm{d}P_{0,1}(x_0, x_1),$$

donde $P_{0,1}$ es el proceso en los instantes inicial y final. Las condiciones $P_0 = \mu,\ P_1 = \nu$ significan que $P_{0,1}$ es un plan de transporte de μ a ν ($P_{0,1} \in \Pi(\mu, \nu)$). Análogamente,

$$R(A \cap ((X(0), X(1)) \in B) = \int_B R(A|x_0, x_1)\, \mathrm{d}R_{0,1}(x_0, x_1),$$

Aquí $R(\cdot|x_0, x_1)$ es la distribución del Movimiento Browniano condicionado a tomar x_0 en el instante inicial y x_1 en el instante final, mientras que $R_{0,1}$ es la medida dada por

$$\mathrm{d}R_{0,1}(x_0, x_1) = \Lambda_1 \exp(-\frac{1}{2}\|x_0 - x_1\|^2)\, \mathrm{d}x_0\, \mathrm{d}x_1. \tag{3.2}$$

Notemos que $R(\cdot|x_0, x_1)$ es la ley de un Puente Browniano de x_0 a x_1, (ver apéndice B).

Lema 3.1. *En las condiciones anteriores, si $P \ll R$ entonces*

a) $P_{0,1} \ll R_{0,1}$.

b) $P(\cdot|x_0, x_1) \ll R(\cdot|x_0, x_1)$.

Demostración. Sea $A \subset \mathbb{R}^d$-

a) Por definición

$$R_{0,1}(A) = R(\{x : (x(0), x(1)) \in A\}) = 0,$$

como $P_{0,1} \ll R_{0,1}$ se tiene que

$$P_{0,1}(A) = P(\{x : (x(0), x(1)) \in A\}) = P_{0,1}(A) = 0.$$

b) Se verifica

$$P(A \cap ((X(0), X(1)) \in B) = \int_B \left[\int I_A \frac{\mathrm{d}P}{\mathrm{d}R} \, \mathrm{d}R(\cdot|x_0, x_1) \right] \mathrm{d}R_{0,1}(x_0, x_1),$$

por otro lado, se cumple

$$\begin{aligned} P(A \cap ((X(0), X(1)) \in B) &= \int_B P(A|x_0, x_1) \, \mathrm{d}P_{0,1}(x_0, x_1) \\ &= \int_B P(A|x_0, x_1) \frac{\mathrm{d}P_{0,1}}{\mathrm{d}R_{0,1}} \, \mathrm{d}R_{0,1}(x_0, x_1). \end{aligned}$$

Juntando ambas igualdades y puesto que se verifican para todo B medible se deduce

$$\int_A \frac{\mathrm{d}P}{\mathrm{d}R} \, \mathrm{d}R(\cdot|x_0, x_1) = P(A|x_0, x_1) \frac{\mathrm{d}P_{0,1}}{\mathrm{d}R_{0,1}}, \quad R_{0,1}\text{-c.s.}$$

Además, en el conjunto de valores (x_0, x_1) tal que $\frac{\mathrm{d}P_{0,1}}{\mathrm{d}R_{0,1}}(x_0, x_1) > 0$ (que se da $P_{0,1}$-c.s.) se tiene

$$P(A|x_0, x_1) = \int_A \frac{\frac{\mathrm{d}P}{\mathrm{d}R}}{\frac{\mathrm{d}P_{0,1}}{\mathrm{d}R_{0,1}}(x_0, x_1)} \, \mathrm{d}R(\cdot|x_0, x_1).$$

Esto demuestra que $P(\cdot|x_0, x_1) \ll R(\cdot|x_0, x_1)$ $P_{0,1}$-c.s. y

$$\frac{\mathrm{d}P(\cdot|x_0, x_1)}{\mathrm{d}R(\cdot|x_0, x_1)} = \frac{\frac{\mathrm{d}P}{\mathrm{d}R}}{\frac{\mathrm{d}P_{0,1}}{\mathrm{d}R_{0,1}}(x_0, x_1)}.$$

□

Este lema nos permite expresar la entropía relativa de la siguiente forma

$$\begin{aligned} \varepsilon \, \mathbf{KL}(P|R) &= \varepsilon \int \log \left(\frac{\mathrm{d}P}{\mathrm{d}R} \right) \mathrm{d}P = \varepsilon \int \log \left(\frac{\mathrm{d}P_{0,1}}{\mathrm{d}R_{0,1}} \right) \mathrm{d}P_{0,1} \\ &+ \varepsilon \int \left[\int \frac{\mathrm{d}P(\cdot|x_0, x_1)}{\mathrm{d}R(\cdot|x_0, x_1)} \, \mathrm{d}P(\cdot|x_0, x_1) \right] \mathrm{d}P_{0,1}(x_0, x_1) \\ &= \varepsilon \, \mathbf{KL}(\, \mathrm{d}P_{0,1} | \, \mathrm{d}R_{0,1}) + \varepsilon \int \mathbf{KL}(P(\cdot|x_0, x_1)|R(\cdot|x_0, x_1)) \, \mathrm{d}P_{0,1}(x_0, x_1). \end{aligned}$$

Esta descomposición permite escribir (3.1) del siguiente modo

$$\min_{P:P_0=\mu,\ P_1=\nu}\left\{\varepsilon\,\mathrm{KL}(P_{0,1}|R_{0,1})+\min_{P(\cdot|x_0,x_1)}\varepsilon\int\mathrm{KL}(P(\cdot|x_0,x_1)|R(\cdot|x_0,x_1))\,\mathrm{d}P_{0,1}(x_0,x_1)\right\}. \tag{3.3}$$

Notemos que el mínimo dentro de los corchetes no impone ninguna restricción sobre $P(\cdot|x_0,x_1)$. Por lo tanto ese mínimo se alcanza tomando $P(\cdot|x_0,x_1)=R(\cdot|x_0,x_1)$, es decir, igual a la distribución del *Puente Browniano* de x_0 a x_1. Con esta elección el problema del Puente de Schrödinger se reduce al cálculo de

$$\min_{P_{0,1}\in\Pi(\mu,\nu)}\{\varepsilon\,\mathrm{KL}(P_{0,1}|R_{0,1})\}. \tag{3.4}$$

Teorema 3.2 (Equivalencia entre el problema del Puente de Schrödinger y el problema entrópico)**.** *El problema de transporte óptimo entrópico y el problema del Puente de Schrödinger son equivalentes.*

Demostración. En primer lugar notemos que $\mathrm{d}R_{0,1}$ siempre es positivo, gracias a la expresión (3.2). Además, $\Lambda_1\exp(-\frac{1}{2}\|x_0-x_1\|^2)$ es la densidad de la medida $R_{0,1}$ respecto a la medida de Lebesgue.

Ahora bien, como $P_{0,1}\ll R_{0,1}$ entonces $P_{0,1}$ también tiene que tener una densidad con respecto a la medida de Lebesgue que llamaremos $p(x_0,x_1)$. Teniendo en cuenta esto

$$\begin{aligned}
&\min_{P_{0,1}\in\Pi(\mu,\nu)}\{\varepsilon\,\mathrm{KL}(P_{0,1}|R_{0,1})\}=\min_{P_{0,1}\in\Pi(\mu,\nu)}\left\{\varepsilon\int\log\left(\frac{\mathrm{d}P_{0,1}}{\mathrm{d}R_{0,1}}\right)\mathrm{d}P_{0,1}\right\}\\
&=\min_{P_{0,1}\in\Pi(\mu,\nu)}\left\{\varepsilon\int_{\mathbb{R}^d\times\mathbb{R}^d}\log\left(\frac{p(x_0,x_1)}{\Lambda_1\exp(-\frac{1}{2}\|x_0-x_1\|^2)}\right)p(x_0,x_1)\,\mathrm{d}x_0\,\mathrm{d}x_1\right\}\\
&=\min_{P_{0,1}\in\Pi(\mu,\nu)}\left\{\varepsilon\int_{\mathbb{R}^d\times\mathbb{R}^d}\left[\log\left(p(x_0,x_1)\right)-\log\left(\Lambda_1\right)-\frac{1}{2}\|x_0-x_1\|^2\right]p(x_0,x_1)\,\mathrm{d}x_0\,\mathrm{d}x_1\right\}\\
&=\min_{P_{0,1}\in\Pi(\mu,\nu)}\left\{\varepsilon\int_{\mathbb{R}^d\times\mathbb{R}^d}\log\left(p(x_0,x_1)\right)p(x_0,x_1)\,\mathrm{d}x_0\,\mathrm{d}x_1\right.\\
&\left.+\varepsilon\int_{\mathbb{R}^d\times\mathbb{R}^d}\frac{1}{2}\|x_0-x_1\|^2p(x_0,x_1)\,\mathrm{d}x_0\,\mathrm{d}x_1-\varepsilon\log\left(\Lambda_1\right)\right\}.
\end{aligned}$$

Por un lado, notar que el factor $-\varepsilon\log(\Lambda_1)$ no afecta a un problema de minimización luego lo descartamos. Por otro lado, μ y ν necesariamente han de ser absolutamente continuas con respecto a la medida de Lebesgue para que la formulación tenga un valor finito. Llamaremos h_0 y h_1 a las respectivas densidades, luego la expresión

$$\begin{aligned}
&\varepsilon\int_{\mathbb{R}^d\times\mathbb{R}^d}\log\left(p(x_0,x_1)\right)p(x_0,x_1)\,\mathrm{d}x_0\,\mathrm{d}x_1\\
&=\int_{\mathbb{R}^d}\log(h_0(x_0))\left[\int_{\mathbb{R}^d}p(x_0,x_1)\,\mathrm{d}x_1\right]\mathrm{d}x_0+\int_{\mathbb{R}^d}\log(h_1(x_1))\left[\int_{\mathbb{R}^d}p(x_0,x_1)\,\mathrm{d}x_0\right]\mathrm{d}x_1\\
&=\varepsilon\int_{\mathbb{R}^d}\log\left(h_0(x_0)\right)h_0(x_0)\,\mathrm{d}x_0+\varepsilon\int_{\mathbb{R}^d}\log\left(h_1(x_1)\right)h_1(x_1)\,\mathrm{d}x_1,
\end{aligned}$$

esta expresión no depende, por tanto, de $p\in\Pi(\mu,\nu)$ luego el problema (3.1) es equivalente a

$$\min_{\pi\in\Pi(\mu,\nu)}\left\{\varepsilon\int_{\mathbb{R}^d\times\mathbb{R}^d}\frac{1}{2}\|x_0-x_1\|^2\,\mathrm{d}\pi(x_0,x_1)+\ \mathrm{KL}(\pi|\mu\otimes\nu)\right\},$$

que es el problema de transporte óptimo entrópico. □

Por tanto, (3.1) se minimiza cuando $P_{01}^\star = \pi^\star$ y $P^\star$ se escribe como una mezcla de *Puentes Brownianos* con las distribuciones iniciales y finales dadas por $\pi^\star$, es decir,

$$P^\star = \int_{\mathbb{R}^d \times \mathbb{R}^d} R(\cdot|x_0, x_1)\ \mathrm{d}\pi^\star(x_0, x_1). \tag{3.5}$$

Definición 3.3. Tomaremos (3.5) como definición de *Puente de Schrödinger*; para cualquier par de medidas de probabilidad en $\mathcal{P}_2(\mathbb{R}^d)$, su *Puente de Schrödinger* es la mezcla de *Puentes Brownianos* dado por (3.5), donde $\pi^\star$ es la solución del problema de transporte óptimo entrópico.

El objetivo es estimar de manera eficiente el *Puente de Schrödinger*. Denotaremos $P^\star$ al *Puente de Schrödinger* entre μ y ν, definimos el flujo tiempo-marginal del puente como

$$\mathsf{p}_t^\star := P_t^\star, \quad t \in [0, 1]. \tag{3.6}$$

Necesitamos probar un lema técnico antes de comenzar con el siguiente teorema:

Lema 3.4. *Sea M una medida dada por una mezcla de Puentes Brownianos con respecto a*

$$\mathrm{d}\pi(x_0, x_1) = \frac{1}{(2\pi\varepsilon)^{d/2}} \exp\left(\frac{(f(x_0) + g(x_1) - 1/2\|x_0 - x_1\|^2)}{\varepsilon}\right) \mathrm{d}\mu_0(x_0)\, \mathrm{d}\mu_1(x_1),$$

y definido como

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} M(\cdot|xy)\ \mathrm{d}\pi(x, y).$$

Si X_0 y X_1 son v.a bajo M entonces son condicionalmente independientes dado X_t.

Demostración. Calculamos la densidad conjunta de X_0, X_1, y X_t con respecto a $\mu_0 \otimes \mu_1 \otimes \mathcal{L}$. La denotaremos $p_{X_0,X_1,X_t}(x_0, x_1, x_t)$. Es sabido que ha de verificarse

$$p_{X_0,X_1,X_t}(x_0, x_1, x_t) = p_{X_0}(x_0) p_{X_1}(x_1) p_{X_t|X_0,X_1}(x_t|x_0, x_1).$$

Las densidades marginales de X_0 y X_1 con respecto a μ_0 y μ_1 son

$$p_{X_0}(x_0) = \Lambda_\varepsilon \exp\left(\frac{f(x_0)}{\varepsilon}\right),$$
$$p_{X_1}(x_1) = \Lambda_\varepsilon \exp\left(\frac{g(x_1)}{\varepsilon}\right).$$

La distribución de X_t dada $X_0 = x_0$ y $X_1 = x_1$ es:

$$p_{X_t|X_0,X_1}(x_t|x_0, x_1) = \frac{1}{(2\pi t(1-t))^{d/2}} \exp\left(-\frac{\|x_t - ((1-t)x_0 + tx_1)\|^2}{2t(1-t)}\right).$$

Finalmente, la densidad conjunta es

$$\Lambda_\varepsilon \Lambda_\varepsilon \exp\left(\frac{f(x_0) + g(x_1)}{\varepsilon}\right) \frac{1}{(2\pi t(1-t))^{d/2}} \exp\left(-\frac{\|x_t - ((1-t)x_0 + tx_1)\|^2}{2t(1-t)}\right).$$

Si juntamos exponenciales, hacemos denominador común y reorganizamos llegamos a la siguiente expresión

$$\Lambda_\varepsilon \Lambda_{t(1-t)\varepsilon} \exp\left(-\frac{\|x_t - ((1-t)x_0 + tx_1)\|^2}{2\varepsilon t(1-t)}\right) \exp\left(\frac{f(x_0) + g(x_1) - 1/2\|x_0 - x_1\|^2}{\varepsilon}\right).$$

Por último, según las fórmulas de las densidades condicionales se tiene que al dividir resulta:

$$p_{X_0,X_1|X_t}(x_0, x_1|x_t) = \Lambda_\varepsilon \exp\left(f(x_0) - \frac{1}{2\varepsilon}\|x_0 - x_1\|^2\right) \Lambda_\varepsilon \exp\left(g(x_1) - \frac{1}{2\varepsilon}\|x_0 - x_1\|^2\right).$$

Como la densidad condicional de X_0 y X_1 dado X_t se descompone en un producto de dos términos independientes podemos concluir. □

Estamos en condiciones de enunciar y demostrar el resultado principal de esta sección.

Teorema 3.5. *Sea π una probabilidad de la siguiente forma:*

$$\mathrm{d}\pi(x_0, x_1) = \frac{1}{(2\pi\varepsilon)^{d/2}} \exp\left(\frac{(f(x_0) + g(x_1) - 1/2\|x_0 - x_1\|^2)}{\varepsilon}\right) \mathrm{d}\mu_0(x_0)\, \mathrm{d}\mu_1(x_1), \tag{3.7}$$

para cualesquiera funciones f y g medibles y $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$. Sea M la medida dada por una mezcla de Puentes Brownianos con respecto a (3.7) como en (3.5), con $t-$marginales m_t para $t \in [0,1]$. Se tiene lo siguiente:

1. *La medida M es de Markov.*
2. *La marginal m_t tiene la siguiente expresión:*

$$\mathrm{d}\mathsf{m}_t(z) = \mathcal{H}_{(1-t)\varepsilon}[\exp(g/\varepsilon)\mu_1](z)\, \mathcal{H}_{t\varepsilon}[\exp(f/\varepsilon)\mu_0](z)\, \mathrm{d}z. \tag{3.8}$$

3. *M es la ley de la solución de la siguiente ecuación diferencial estocástica:*

$$\mathrm{d}X_t = \varepsilon \nabla \log \mathcal{H}_{(1-t)\varepsilon}[\exp(g/\varepsilon)\mu_1](X_t)\, \mathrm{d}t + \sqrt{\varepsilon}\, \mathrm{d}W_t, \quad X_0 \sim \mu_0. \tag{3.9}$$

4. *Si*

$$\mathrm{d}\rho(x_1) = \mathrm{d}\mu_1(x) \exp\left(\frac{g(x_1)}{\varepsilon} + \log(\mathcal{H}_\varepsilon[\exp(f/\varepsilon)\mu_0](x_1))\right),$$

y $\nabla\varphi_{1-t}$ es la aplicación de Brenier entrópica entre m_t y ρ con regularización $(1-t)\varepsilon$ el coeficiente de deriva de la expresión de arriba se expresa como

$$b_t(z) = \frac{1}{1-t}(z - \nabla\varphi_{1-t}(z)).$$

Si (3.7) es el plan de transporte óptimo entrópico entre μ_0 y μ_1 entonces $\rho \equiv \mu_1$.

Demostración. 1. En primer lugar veremos que M es de Markov:

Sea $(X_t)_{t\in[0,1]}$ un proceso estocástico distribuido según M. Es condición suficiente probar que para cualquier $a \in \sigma(X_{[0,t]})$, $b \in \sigma(X_{[t,1]})$ se tiene

$$\mathbb{E}[ab|X_t] = \mathbb{E}[a|X_t]\mathbb{E}[b|X_t]$$

casi siempre. Es suficiente ya que garantiza que la información futura (en el intervalo $[t,1]$) es independiente del pasado (en $[0,t]$) dado el presente X_t, lo cual es precisamente la propiedad de Markov.

En virtud de la propiedad de la torre $\mathbb{E}[ab|X_t] = \mathbb{E}[\mathbb{E}[ab|X_0, X_t, X_1]|X_t]$ y, puesto que la ley de un camino entre X_0 y X_1 condicionado a X_t es un *Puente Browniano* -por tanto de Markov, ver apéndice B- tenemos

$$\mathbb{E}[\mathbb{E}[ab|X_0, X_t, X_1]|X_t] = \mathbb{E}[\mathbb{E}[a|X_0, X_t]\mathbb{E}[b|X_t, X_1]|X_t].$$

Ahora bien, por el lema anterior se deduce que

$$\mathbb{E}[\mathbb{E}[a|X_0, X_t]\mathbb{E}[b|X_t, X_1]|X_t] = \mathbb{E}[\mathbb{E}[a|X_0, X_t]|X_t]\mathbb{E}[\mathbb{E}[b|X_0, X_t]|X_t].$$

Según la ley de la esperanza total la última expresión es $\mathbb{E}[a|X_t]\mathbb{E}[b|X_t]$. Como queríamos.

2. Sabemos que la expresión de la marginal de la distribución m_t viene determinada directamente por (3.5):

$$\mathrm{d}\mathsf{m}_t(z) = \iint M_t(z|X_0 = x_0,\ X_1 = x_1)\,\mathrm{d}\pi(x,y).$$

Un proceso del tipo *Puente Browniano* cumple que la densidad condicionada en t dado $X_0 = x_0$ y $X_1 = x_1$ es una distribución normal multivariante con media $(1-t)x_0 + tx_1$ y varianza $\varepsilon t(1-t)\mathrm{Id}$. (Ver apéndice B). Luego

$$= \iint \mathcal{N}(z|(1-t)x_0 + tx_1, \varepsilon t(1-t)\mathrm{Id})\,\mathrm{d}\pi(x_0,x_1).$$

A continuación utilizamos la expresión de π:

$$\Lambda_\varepsilon \iint \exp\Big(\frac{f(x_0) + g(x_1) - \frac{1}{2}\|x_0 - x_1\|^2}{\varepsilon}\Big)\mathcal{N}(z|(1-t)x_0 + tx_1, \varepsilon t(1-t)\mathrm{Id})\,\mathrm{d}\mu_0(x_0)\,\mathrm{d}\mu_1(x_1).$$

Según la fórmula explícita de la densidad Gaussiana y reorganizando la expresión,

$$\int \exp((g(x_1)/\varepsilon)\mathcal{N}(z|x_1, \varepsilon t(1-t)\mathrm{Id})\,\mathrm{d}\mu_1(x_1) \int \exp((f(x_0)/\varepsilon)\mathcal{N}(z|x_0, \varepsilon t\mathrm{Id})\,\mathrm{d}\mu_0(x_0)$$

$$= \mathcal{H}_{(1-t)\varepsilon}[\exp(g/\varepsilon)\mu_1](z)\ \mathcal{H}_{t\varepsilon}[\exp(f/\varepsilon)\mu_0](z)\,\mathrm{d}z.$$

3. Demostremos ahora la tercera parte de este resultado. Este apartado esta dividido en varios pasos. Para facilitar la notación nos restringimos al caso que $\varepsilon = 1$, el resto de casos se deducen de este. Fijamos $t < 1$.

En primer lugar enunciaremos una versión adaptada de [14, Teorema 2.4] y lo combinaremos con [13, Proposición 2.5].

Teorema 3.6. *Si X es un proceso con trayectorias continuas en $[0,1]$, tal que $X(0)=0$ c.s. y, además, P_X es la ley inducida de dicho proceso que satisface que la entropía relativa con respecto a la medida de Wiener es finita, i.e., $H(P_X|\mu_W)<\infty$ entonces existe un proceso adaptado $\{b_t\}_{t\in[0,1]}$ con respecto a la filtración natural tal que*

$$X_t=\int_0^t b_s\,\mathrm{d}s+W_t,$$

donde W_t es un movimiento browniano. Además,

$$b_t=\lim_{h\searrow 0}\frac{1}{h}\mathbb{E}[X_{t+h}-X_t|\mathcal{F}_t].$$

Recordemos que M viene determinada por la mezcla de *Puentes Brownianos*, es decir, si $X_{[0,t]}$ es su proceso no se cumple que $X(0)=0$, luego necesitamos centrarlo al origen. Denotaremos $\tilde{X}$ al proceso desplazado

$$\tilde{X}_t=X_t-X_0, \tag{3.10}$$

y $\tilde{M}$ a la ley inducida. Además, condicionalmente dado $(X_0=x_0,X_1=x_1)$, $\tilde{X}_t$ es un Puente Browniano de 0 a x_1-x_0. Esto implica que $\tilde{X}_t$ tiene trayectorias continuas c.s. Luego $\tilde{M}$ verifica la hipótesis de trayectorias continuas.

Supongamos que se verifica que $H(\tilde{M}|\mu_W)<\infty$ (luego lo demostraremos), entonces en virtud del teorema 3.6 se tiene:

$$\tilde{X}_t=\int_0^t b_s\,\mathrm{d}s+W_t,$$

donde W_t es Movimiento Browniano bajo $\tilde{M}$ y,

$$b_t=\lim_{h\searrow 0}\frac{1}{h}\mathbb{E}[\tilde{X}_{t+h}-\tilde{X}_t|\tilde{\mathcal{F}}_t].$$

Aquí $\tilde{\mathcal{F}}_t=\sigma(X_0,X_s-X_0;s\le t)$. Si volvemos a X_t se tiene

$$X_t=X_0+\int_0^t b_s\,\mathrm{d}s+W_t,$$

donde $X_0\sim\mu_0$, b_s adaptado a la filtración de $\mathcal{F}_t=\sigma(X_s;0\le s\le t)$ (coincide con la filtración $\tilde{\mathcal{F}}_t$). Y

$$b_t=\lim_{h\searrow 0}\frac{1}{h}\mathbb{E}[X_{t+h}-X_t|\mathcal{F}_t].$$

Por tanto para verificar la representación en términos de la ecuación diferencial estocástica es suficiente calcular

$$b_t=\lim_{h\searrow 0}\frac{1}{h}\mathbb{E}[X_{t+h}-X_t|X_{[0,t]}],\qquad (\text{en } L^2).$$

En primer lugar, notemos que, de manera general, si $\{Y_t\}_{0\le t\le 1}$ es un Puente Browniano entonces dado $Y_t=y_t, Y_r=y_r$ el proceso $\{Y_t\}_{s\le t\le r}$ sigue siendo un Puente Browniano de y_t a y_r (en $[s,t]$).

Para continuar, vamos a usar que el proceso es de Markov (demostración análoga a la del apartado 1) y que condicionado a X_0 y X_1 el camino es un *Puente Browniano*.

$$\mathbb{E}[X_{t+h} - X_t \mid X_{[0,t]}] = \mathbb{E}[\mathbb{E}[X_{t+h} - X_t \mid X_0, X_t, X_1] \mid X_{[0,t]}] \tag{3.11}$$

Nos centramos en la esperanza de dentro, el X_0 no es relevante (estamos trabajando con un proceso de Markov), además, sabemos que por ser un *Puente Browniano* se verifica

$$\mathbb{E}\left(X_{t+h} \mid X_t, X_1\right) = \tfrac{1-t-h}{1-t} X_t + \tfrac{h}{1-t} X_1,$$

$$\mathbb{E}\left(X_t \mid X_t, X_1\right) = X_t.$$

Restando ambas expresiones y simplificando:

$$\mathbb{E}\left(X_{t+h} \mid X_t, X_1\right) - X_t = \frac{h}{1-t}(X_1 - X_t).$$

Si lo llevamos a la expresión (3.11) y teniendo en cuenta la linealidad de la esperanza nos queda:

$$\mathbb{E}[\mathbb{E}[X_{t+h} - X_t | X_0, X_t, X_1] | X_{[0,t]}] = \frac{h}{1-t}\mathbb{E}[X_1 - X_t | X_{[0,t]}].$$

Ahora bien, si utilizamos de nuevo que el proceso es de Markov se cumple la siguiente igualdad

$$\frac{h}{1-t}\mathbb{E}[X_1 - X_t | X_{[0,t]}] = \frac{h}{1-t}\mathbb{E}[X_1 - X_t | X_t]$$

Y recuperando el límite, se tiene

$$\lim_{h \searrow 0} \frac{1}{h}\mathbb{E}[\mathbb{E}[X_{t+h} - X_t | X_0, X_t, X_1] | X_t] = \frac{1}{1-t}\mathbb{E}[X_1 - X_t | X_t].$$

A continuación calculamos última esperanza. En virtud del lema anterior, la densidad conjunta de X_1, X_t es

$$\begin{aligned}
&\int_{\mathbb{R}^d} \Lambda_\varepsilon \Lambda_{t(1-t)\varepsilon} e^{\left(-\frac{\|x_t - ((1-t)x_0 + tx_1)\|^2}{2\varepsilon t(1-t)}\right)} e^{\left(\frac{f(x_0)+g(x_1)-1/2\|x_0-x_1\|^2}{\varepsilon}\right)} \, \mathrm{d}\mu_0(x_0) \\
=&\Lambda_{t(1-t)\varepsilon} e^{\left(\frac{g(x_1)}{\varepsilon}\right)} e^{\left(-\frac{1}{2\varepsilon(1-t)}\|x_t - x_1\|^2\right)} \int_{\mathbb{R}^d} \Lambda_{t\varepsilon} e^{\left(\frac{f(x_0)}{\varepsilon}\right)} e^{\left(-\frac{1}{2\varepsilon t}\|x_t - x_0\|^2\right)} \, \mathrm{d}\mu_0(x_0).
\end{aligned}$$

La densidad marginal de X_t es

$$\begin{aligned}
&\iint_{\mathbb{R}^d \times \mathbb{R}^d} \Lambda_\varepsilon \Lambda_{t(1-t)\varepsilon} e^{\left(-\frac{\|x_t - ((1-t)x_0 + tx_1)\|^2}{2\varepsilon t(1-t)}\right)} e^{\left(\frac{f(x_0)+g(x_1)-1/2\|x_0-x_1\|^2}{\varepsilon}\right)} \, \mathrm{d}\mu_0(x_0) \, \mathrm{d}\mu_1(x_1) \\
&\int_{\mathbb{R}^d} \Lambda_{t(1-t)\varepsilon} e^{\left(\frac{g(x_1)}{\varepsilon}\right)} e^{\left(-\frac{1}{2\varepsilon(1-t)}\|x_t - x_1\|^2\right)} \, \mathrm{d}\mu_1(x_1) \int_{\mathbb{R}^d} \Lambda_{t\varepsilon} e^{\left(\frac{f(x_0)}{\varepsilon}\right)} e^{\left(-\frac{1}{2\varepsilon t}\|x_t - x_0\|^2\right)} \, \mathrm{d}\mu_0(x_0)
\end{aligned}$$

Luego la densidad de X_1 condicionada por $X_t = x_t$ es

$$\begin{aligned}
&\frac{\Lambda_{t(1-t)\varepsilon} e^{\left(\frac{g(x_1)}{\varepsilon}\right)} e^{\left(-\frac{1}{2\varepsilon(1-t)}\|x_t - x_1\|^2\right)} \int_{\mathbb{R}^d} \Lambda_{t\varepsilon} e^{\left(\frac{f(x_0)}{\varepsilon}\right)} e^{\left(-\frac{1}{2\varepsilon t}\|x_t - x_0\|^2\right)} \, \mathrm{d}\mu_0(x_0)}{\int_{\mathbb{R}^d} \Lambda_{t(1-t)\varepsilon} e^{\left(\frac{g(x_1)}{\varepsilon}\right)} e^{\left(-\frac{1}{2\varepsilon(1-t)}\|x_t - x_1\|^2\right)} \, \mathrm{d}\mu_1(x_1) \int_{\mathbb{R}^d} \Lambda_{t\varepsilon} e^{\left(\frac{f(x_0)}{\varepsilon}\right)} e^{\left(-\frac{1}{2\varepsilon t}\|x_t - x_0\|^2\right)} \, \mathrm{d}\mu_0(x_0)} \\
&= \frac{e^{\left(\frac{g(x_1)}{\varepsilon}\right)} e^{\left(-\frac{1}{2\varepsilon(1-t)}\|x_t - x_1\|^2\right)}}{\int_{\mathbb{R}^d} e^{\left(\frac{g(x_1)}{\varepsilon}\right)} e^{\left(-\frac{1}{2\varepsilon(1-t)}\|x_t - x_1\|^2\right)} \, \mathrm{d}\mu_1(x_1)}.
\end{aligned}$$

En conclusión,

$$\frac{1}{1-t}\mathbb{E}[X_1 - X_t | X_t = x_t] = \frac{\int_{\mathbb{R}^d} \frac{x_1 - x_t}{1-t} e^{\left(\frac{g(x_1)}{\varepsilon}\right)} e^{\left(-\frac{1}{2\varepsilon(1-t)}\|x_t - x_1\|^2\right)}}{\int_{\mathbb{R}^d} e^{\left(\frac{g(x_1)}{\varepsilon}\right)} e^{\left(-\frac{1}{2\varepsilon(1-t)}\|x_t - x_1\|^2\right)} \, \mathrm{d}\mu_1(x_1)} \tag{3.12}$$

Esta expresión se puede escribir como:

$$\varepsilon \nabla \log \mathcal{H}_{(1-t)\varepsilon}[\exp(g/\varepsilon)](x_t).$$

Con esto finalizaría la prueba, únicamente queda demostrar que

$$H(\tilde{M}|\mu_W) < \infty.$$

Para cada par (x_0, x_1), sea $\tilde{M}(\cdot|x_0, x_1)$ la ley del *Puente Browniano* centrado de 0 a $x := x_1 - x_0$, es decir, del proceso

$$\tilde{B}_t := B_t^{x_0, x_1} - x_0, \quad \text{con} \quad \tilde{B}_0 = 0, \ \tilde{B}_1 = x.$$

Entonces:

$$\tilde{M} = \int \tilde{M}(\cdot|x_0, x_1) \, \mathrm{d}\pi(x_0, x_1).$$

El *Puente Browniano* $\tilde{B}_t$ satisface

$$\mathrm{d}\tilde{B}_t = \frac{x - \tilde{B}_t}{1-t} \, \mathrm{d}t + \, \mathrm{d}\mathbf{W}_t.$$

El [28, Teorema 7.7] nos da una expresión de la derivada de Radon-Nikodym $\frac{\mathrm{d}\tilde{M}(\cdot|x_0,x_1)}{\mathrm{d}\mu_W}$. Utilizando [13, Proposición 2.11] se tiene que

$$H(\tilde{M}(\cdot|x_0, x_1)|\mu_W) = \frac{1}{2}\mathbb{E}_{\tilde{M}(\cdot|x_0,x_1)}\left[\int_0^t \left\| \frac{x - \tilde{B}_s}{1-s} \right\|^2 \, \mathrm{d}s\right],$$

siempre que

$$\mathbb{E}_{\tilde{M}(\cdot|x_0,x_1)}\left[\int_0^t \left\| \frac{x - \tilde{B}_s}{1-s} \right\|^2 \, \mathrm{d}s\right] < \infty.$$

En primer lugar veamos que este término es finito. Por ser *Puente Browniano* centrado, sabemos que:

$$\mathbb{E}[\tilde{B}_s] = sx, \quad \mathrm{Var}(\tilde{B}_s) = s(1-s)\mathrm{Id}.$$

Entonces:

$$\begin{aligned}\mathbb{E}\left[\left\| \frac{x - \tilde{B}_s}{1-s} \right\|^2\right] &= \frac{1}{(1-s)^2}\mathbb{E}\left[\|x - \tilde{B}_s\|^2\right] \\ &= \|x\|^2 + \frac{ds}{1-s}.\end{aligned}$$

Aquí se ha utilizado que $\mathbb{E}[\|z - a\|]^2 = \|a - \mu\|^2 + \mathrm{Tr}(\Sigma_Z)$.

Integrando:

$$\mathbb{E}_{\tilde{M}(\cdot|x_0,x_1)}\left[\int_0^t \left\|\frac{x-\tilde{B}_s}{1-s}\right\|^2 \mathrm{d}s\right] = \int_0^t \left(\|x\|^2 + \frac{ds}{1-s}\right) \mathrm{d}s = \|x\|^2 + \int_0^t \frac{ds}{1-s}\,\mathrm{d}s.$$

La última integral es finita puesto que $t < 1$. En conclusión,

$$H(\tilde{M}(\cdot|x_0,x_1)|\mu_W) \leq \frac{1}{2}\left(\|x_1 - x_0\|^2 + C\right), \tag{3.13}$$

para una constante $C > 0$. Esto también implica que $\tilde{M}(\cdot|x_0,x_1) \ll \mu_W$ en $[0,t]$.

Lo siguiente será relacionar esto con $\tilde{M}$. Como bien se ha comentado antes

$$\tilde{M}(A) = \int \tilde{M}(\cdot|x_0,x_1)\,\mathrm{d}\pi(x_0,x_1), \quad \forall A \subset \mathcal{C}([0,1],\mathbb{R}^d).$$

En virtud del teorema de Radon-Nikodym y Tonelli-Fubini se tiene:

$$\frac{\mathrm{d}\tilde{M}}{\mathrm{d}\mu_W}(\omega) = \int \frac{\mathrm{d}\tilde{M}(\cdot|x_0,x_1)}{\mathrm{d}\mu_W}(\omega)\,\mathrm{d}\pi(x_0,x_1). \tag{3.14}$$

Por otro lado,

$$\begin{aligned} H(\tilde{M}|\mu_W) &= \int \log\frac{\mathrm{d}\tilde{M}}{\mathrm{d}\mu_W}\,\mathrm{d}\tilde{M} \\ &= \int \left[\log\left(\int \frac{\mathrm{d}\tilde{M}(\cdot|x_0,x_1)}{\mathrm{d}\mu_W}\,\mathrm{d}\pi(x_0,x_1)\right)\right]\mathrm{d}\tilde{M}. \end{aligned}$$

Utilizando (3.14) se tiene que la última expresión es

$$\int \left[\left(\int \frac{\mathrm{d}\tilde{M}(\cdot|x_0,x_1)}{\mathrm{d}\mu_W}(\omega)\,\mathrm{d}\pi(x_0,x_1)\right)\log\left(\int \frac{\mathrm{d}\tilde{M}(\cdot|x_0,x_1)}{\mathrm{d}\mu_W}(\omega)\,\mathrm{d}\pi(x_0,x_1)\right)\right]\mathrm{d}\mu_W(\omega). \tag{3.15}$$

Notemos que la función $\varphi(s) = s\log(s)$ es convexa luego en virtud de la desigualdad de Jensen obtenemos

$$\begin{aligned} &\int \frac{\mathrm{d}\tilde{M}(\cdot|x_0,x_1)}{\mathrm{d}\mu_W}\,\mathrm{d}\pi(x_0,x_1)\log\left(\int \frac{\mathrm{d}\tilde{M}(\cdot|x_0,x_1)}{\mathrm{d}\mu_W}\,\mathrm{d}\pi(x_0,x_1)\right) \\ &\leq \int \frac{\mathrm{d}\tilde{M}(\cdot|x_0,x_1)}{\mathrm{d}\mu_W}\log\left(\frac{\mathrm{d}\tilde{M}(\cdot|x_0,x_1)}{\mathrm{d}\mu_W}\right)\mathrm{d}\pi(x_0,x_1). \end{aligned} \tag{3.16}$$

Por la monotonía de la integral (integramos respecto a la medida de Wiener)

$$\begin{aligned} &\int \left[\int \frac{\mathrm{d}\tilde{M}(\cdot|x_0,x_1)}{\mathrm{d}\mu_W}\,\mathrm{d}\pi(x_0,x_1)\log\left(\int \frac{\mathrm{d}\tilde{M}(\cdot|x_0,x_1)}{\mathrm{d}\mu_W}\,\mathrm{d}\pi(x_0,x_1)\right)\right]\mathrm{d}\mu_W \\ &\leq \int \left[\int \frac{\mathrm{d}\tilde{M}(\cdot|x_0,x_1)}{\mathrm{d}\mu_W}\log\left(\frac{\mathrm{d}\tilde{M}(\cdot|x_0,x_1)}{\mathrm{d}\mu_W}\right)\mathrm{d}\pi(x_0,x_1)\right]\mathrm{d}\mu_W. \end{aligned}$$

Si recuperamos (3.15) e intercambiamos el orden de integración en virtud de Fubini-Tonelli resulta

$$\begin{aligned} H(\tilde{M}|\mu_W) &\leq \int \left[\int \frac{\mathrm{d}\tilde{M}(\cdot|x_0,x_1)}{\mathrm{d}\mu_W} \log \left(\frac{\mathrm{d}\tilde{M}(\cdot|x_0,x_1)}{\mathrm{d}\mu_W} \right) \mathrm{d}\pi(x_0,x_1) \right] \mathrm{d}\mu_W \\ &= \int \int \frac{\mathrm{d}\tilde{M}(\cdot|x_0,x_1)}{\mathrm{d}\mu_W} \log \left(\frac{\mathrm{d}\tilde{M}(\cdot|x_0,x_1)}{\mathrm{d}\mu_W} \right) \mathrm{d}\mu_W \, \mathrm{d}\pi(x_0,x_1) \\ &= \int H(\tilde{M}(\cdot|x_0,x_1)|\mu_W) \, \mathrm{d}\pi(x_0,x_1). \end{aligned}$$

Sustituyendo (3.13) se tiene

$$H(M|\mu_W) \leq \frac{1}{2} \int \left(\|x_1 - x_0\|^2 + C \right) \mathrm{d}\pi(x_0,x_1).$$

Como μ_0 y μ_1 tienen momento de orden 2 finito. Se verifica que es finito.

4. Denotamos

$$f_{1-t}(z) := \varepsilon \log \mathcal{H}_{(1-t)\varepsilon}[e^{g/\varepsilon}\mu_1](z).$$

y $\pi_t(z,x_1)$ definida por la siguiente expresión

$$\mathrm{d}\pi_t(z,x_1) = \Lambda_{(1-t)\varepsilon} \exp \left(\frac{-(1-t)f_{1-t}(z) + (1-t)g(x_1) - \frac{1}{2}\|z-x_1\|^2}{(1-t)\varepsilon} \right) \mathrm{d}\mathsf{m}_t(z) \, \mathrm{d}\mu_1(x_1),$$

Observamos que la primera marginal de π_t es m_t ya que

$$\begin{aligned} &\int \pi_t(z,x_1) \, \mathrm{d}\mu_1(x_1) = \int \Lambda_{(1-t)\varepsilon} \exp \left(\frac{-(1-t)f_{1-t}(z) + (1-t)g(x_1) - \frac{1}{2}\|z-x_1\|^2}{(1-t)\varepsilon} \right) \mathrm{d}\mu_1(x_1) \\ &= \int \Lambda_{(1-t)\varepsilon} \exp \left(\frac{-f_{1-t}(z)}{\varepsilon} \right) \exp \left(\frac{(1-t)g(x_1) - \frac{1}{2}\|z-x_1\|^2}{(1-t)\varepsilon} \right) \mathrm{d}\mu_1(x_1) \\ &= \frac{1}{\mathcal{H}_{(1-t)\varepsilon}[e^{g/\varepsilon}\mu_1](z)} \mathcal{H}_{(1-t)\varepsilon}[e^{g/\varepsilon}\mu_1](z) = 1. \end{aligned}$$

Luego, en efecto, la primera marginal de π_t es m_t y, además, π_t es una probabilidad. La segunda marginal tiene densidad respecto μ_1 dada por

$$\int \pi_t(z,x_1) \, \mathrm{d}\mathsf{m}_t(z)$$

Por el apartado segundo tenemos

$$\mathrm{d}\mathsf{m}_t(z) = \mathcal{H}_{(1-t)\varepsilon}[e^{g/\varepsilon}\mu_1](z) \mathcal{H}_{t\varepsilon}[e^{f/\varepsilon}\mu_0](z) \, \mathrm{d}z.$$

Si integramos para obtener la otra marginal (simplificamos directamente el factor $\mathcal{H}_{(1-t)\varepsilon}[e^{g/\varepsilon}\mu_1](z)$)

$$\int \pi_t(z,x_1) \, \mathrm{d}\mathsf{m}_t(z) = e^{g(x_1)/\varepsilon} \int e^{-\frac{\|z-x_1\|^2}{2(1-t)\varepsilon}} \mathcal{H}_{t\varepsilon}[e^{f/\varepsilon}\mu_0](z) \, \mathrm{d}z.$$

Aplicamos la definición de semigrupo y se tiene

$$= e^{g(x_1)/\varepsilon}\mathcal{H}_{(1-t)\varepsilon}[\mathcal{H}_{t\varepsilon}[e^{f/\varepsilon}\mu_0]](x_1),$$

y, utilizamos la propiedad de semigrupo $T_{t+s} = T_t \circ T_s$

$$= e^{g(x_1)/\varepsilon}\mathcal{H}_{\varepsilon}[e^{f/\varepsilon}\mu_0](x_1).$$

Denotamos esta segunda marginal por ρ, es decir, ρ es la probabilidad

$$\mathrm{d}\rho(x_1) = \exp\left(\frac{g(x_1)}{\varepsilon} + \log \mathcal{H}_{\varepsilon}[e^{f/\varepsilon}\mu_0](x_1)\right)\,\mathrm{d}\mu_1(x_1)$$

Notemos que $\rho \ll \mu_1$, pero el hecho de que $e^{g(x_1)/\varepsilon}\mathcal{H}_{\varepsilon}[e^{f/\varepsilon}\mu_0]$ sea siempre mayor que 0 también implica que $\mu_1 \ll \rho$. La derivada de μ_1 con respecto a ρ es

$$\frac{1}{e^{g(x_1)/\varepsilon}\mathcal{H}_{\varepsilon}[e^{f/\varepsilon}\mu_0]}.$$

Entonces podemos escribir $\mathrm{d}\pi_t$ de la siguiente forma

$$\Lambda_{(1-t)\varepsilon}\exp\left(\frac{-(1-t)f_{1-t}(z) - (1-t)\varepsilon\log\mathcal{H}_{\varepsilon}[e^{f/\varepsilon}\mu_0](x_1) - \frac{1}{2}\|z-x_1\|^2}{(1-t)\varepsilon}\right)\,\mathrm{d}\mathsf{m}_t(z)\,\mathrm{d}\rho(x_1)$$

Esto quiere decir que esta probabilidad es el transporte óptimo entrópico entre m_t y ρ.

Teniendo en cuenta esto, recordemos la definición de $\nabla\varphi_{1-t}$:

$$\frac{\int x_1 \exp\left(\frac{g(x_1)}{\varepsilon} - \frac{1}{2\varepsilon(1-t)}\|z-x_1\|^2\right)\,\mathrm{d}\mu_1(x_1)}{\int \exp\left(\frac{g(x_1)}{\varepsilon} - \frac{1}{2\varepsilon(1-t)}\|z-x_1\|^2\right)\,\mathrm{d}\mu_1(x_1)}$$

se verifica que es la esperanza condicionada dado z.

En el caso que (3.7) sea el plan entrópico óptimo entre μ_0 y μ_1, en particular se tiene que sus marginales son μ_0 y μ_1 entonces si su segunda marginal es μ_1 los cálculos hechos más arriba demuestran que $\rho \equiv \mu_1$.

□

Cuando el par $(\mathsf{p}_t^\star, b_t^\star)$ satisface la *ecuación de Fokker-Planck* entonces

$$\begin{aligned} b_t^\star(z) &= \varepsilon\nabla\log\mathcal{H}_{(1-t)\varepsilon}[e^{g^\star/\varepsilon}\nu](z) \\ &= \frac{1}{1-t}\left(-z + \frac{\int y\exp\left(\frac{g^\star(y)}{\varepsilon} - \frac{1}{2\varepsilon(1-t)}\|z-y\|^2\right)\,\mathrm{d}\nu(y)}{\int\exp\left(\frac{g^\star(y)}{\varepsilon} - \frac{1}{2\varepsilon(1-t)}\|z-y\|^2\right)\,\mathrm{d}\nu(y)}\right) \\ &=: \frac{1}{1-t}(-z + \nabla\varphi_{1-t}^\star(z)). \end{aligned}$$

Esto implica que si X_t resuelve

$$\mathrm{d}X_t = b_t^\star(X_t)\,\mathrm{d}t + \sqrt{\varepsilon}\,\mathrm{d}\mathbf{W}_t, \quad X_0 \sim \mu, \tag{3.17}$$

entonces la ley de X_t es $\mathsf{p}_t^\star$. Otra propiedad muy importante es que la función $b_t^\star$ se escribe en términos de $g^\star$.

De hecho, la ecuación diferencial estocástica anterior nos lleva a una *medida* que coincide exactamente con el *Puente de Schrödinger*. Y, además, la expresión de la ecuación que produce el *Puente de Schrödinger* es

$$\mathrm{d}X_t = \left(\frac{-1}{1-t}X_t + \frac{1}{1-t}\nabla\varphi_{1-t}^\star(X_t)\right)\,\mathrm{d}t + \sqrt{\varepsilon}\,\mathrm{d}\mathbf{W}_t, \quad X_0 \sim \mu \tag{3.18}$$

donde $\nabla\varphi_{1-t}^\star$ es la *aplicación de Brenier entrópica* entre $\mathsf{p}_t^\star$ y ν con parámetro de regularización $(1-t)\varepsilon$.

A modo de conclusión veamos cómo se obtiene la función $b_t^\star$, con lo visto en la sección anterior (cambiando el parámetro de regularización de ε a $(1-t)\varepsilon$ y x a z) se tiene:

1. Mediante el *algoritmo de Sinkhorn* obtenemos $g^\star$ y $f^\star$ (aunque luego sólo se use $g^\star$).
2. Los *potenciales entrópicos de Brenier* se definían como

$$\varphi_{1-t}^\star(z) = \left[\frac{\|z\|^2}{2} - f^\star(z)\right],\ \psi_{1-t}^\star(y) = \left[\frac{\|y\|^2}{2} - g^\star(y)\right].$$

3. Se tenía

$$\varphi_{1-t}^\star(z) = (1-t)\varepsilon\ \log\left(\int_{\mathbb{R}^d}\exp\left(\frac{z\cdot y - \psi_{1-t}^\star(y)}{(1-t)\varepsilon}\right)\,\mathrm{d}\nu(y)\right).$$

4. La *aplicación entrópica de Brenier* es entonces

$$\nabla\varphi_{1-t}^\star(z) = \frac{\int_{\mathbb{R}^d}\exp\left(\frac{z\cdot y-\psi^\star(y)}{(1-t)\varepsilon}\right)\ y\ \mathrm{d}\nu(y)}{\int_{\mathbb{R}^d}\exp\left(\frac{z\cdot y-\psi^\star(y)}{(1-t)\varepsilon}\right)\ \mathrm{d}\nu(y)}. \tag{3.19}$$

5. Utilizando la relación $\frac{1}{2}\|z-y\|^2 = \frac{1}{2}\|z\|^2 + \frac{1}{2}\|y\|^2 - z\cdot y$ y la definición del potencial entrópico, se llega a la expresión

$$\nabla\varphi_{1-t}^\star(z) = \frac{\int y\exp\left(\frac{g^\star(y)}{\varepsilon} - \frac{1}{2\varepsilon(1-t)}\|z-y\|^2\right)\,\mathrm{d}\nu(y)}{\int\exp\left(\frac{g^\star(y)}{\varepsilon} - \frac{1}{2\varepsilon(1-t)}\|z-y\|^2\right)\,\mathrm{d}\nu(y)}.$$

6. Por último

$$b_t^\star(z) = \frac{1}{1-t}(-z + \nabla\varphi_{1-t}^\star(z)). \tag{3.20}$$

3.2. Puente de Sinkhorn

Siguiendo la notación del epígrafe anterior, nuestro objetivo es estimar el camino descrito por $P^\star$ a partir de la base de muestras. El método que presentamos, el *Puente de Sinkhorn* descrito por [32], supone una mejora con respecto a los enfoques existentes:

Finlay [12] propuso el uso de los *potenciales de Brenier* para estimar el coeficiente de deriva en la ecuación asociada al *Puente de Schrödinger*. No obstante, su estimador resulta intratable en la práctica. Por otro lado, Stromme [34] adopta una estrategia diferente basada en el *algoritmo de Sinkhorn* para calcular un plan óptimo entrópico entre las medidas de origen y destino. Sin embargo, su enfoque requiere la generación de n nuevas muestras de las distribuciones μ y ν para obtener una única muestra aproximada del *Puente de Schrödinger*, lo que conlleva un alto costo computacional debido a la necesidad de regenerar datos en cada iteración.

En contraste, el método presentado permite obtener muestras aproximadas del *Puente de Schrödinger* de manera más eficiente. También utilizamos el *algoritmo de Sinkhorn* para estimar el plan óptimo entrópico y, una vez calculados los coeficientes de deriva, podemos reutilizar las muestras de origen, reduciendo significativamente la complejidad computacional.

En esta sección supondremos variables aleatorias independientes e idénticamente distribuidas (i.i.d.) según una distribución μ: $X_1, \dots, X_m \sim \mu$ y , análogamente, $Y_1, \dots, Y_n \sim \nu$. Definimos las siguientes medidas empíricas:

$$\begin{cases} \mu_m := \frac{1}{m} \sum_{i=1}^m \delta_{X_i}, \\ \nu_n := \frac{1}{n} \sum_{j=1}^n \delta_{Y_j}. \end{cases}$$

En virtud de (3.20) es fácil definir un estimador en la base de muestras. Sean $(\hat{f}, \hat{g}) \in \mathbb{R}^m \times \mathbb{R}^n$ los *potenciales óptimos entrópicos* del problema

$$\mathcal{T}_\varepsilon(\mu_m, \nu_n) := \min_{\pi \in \Pi(\mu_m, \nu_n)} \left\{ \int_{\mathbb{R}^d \times \mathbb{R}^d} \frac{1}{2} \|x - y\|^2 \, \mathrm{d}\pi(x, y) \; + \; \varepsilon \, \mathrm{KL}(\pi | \mu_m \otimes \nu_n) \right\}, \tag{3.21}$$

que se pueden obtener de manera eficiente si se utiliza el *algoritmo de Sinkhorn*. En virtud de la expresión de $b_t^\star$ de la sección anterior ((3.20)) se tiene que en este caso

$$\nabla \hat{\varphi}_{1-t}(z) = \frac{\int y \exp\left(\frac{\hat{g}(y)}{\varepsilon} - \frac{1}{2\varepsilon(1-t)} \|z - y\|^2 \right) \mathrm{d}\nu_n(y)}{\int \exp\left(\frac{\hat{g}(y)}{\varepsilon} - \frac{1}{2\varepsilon(1-t)} \|z - y\|^2 \right) \mathrm{d}\nu_n(y)}$$

$$= \frac{\int y \exp\left(\frac{\hat{g}(y)}{\varepsilon} - \frac{1}{2\varepsilon(1-t)} \|z - y\|^2 \right) \frac{1}{n} \sum_{j=1}^n \mathrm{d}\delta_{Y_j}}{\int \exp\left(\frac{\hat{g}(y)}{\varepsilon} - \frac{1}{2\varepsilon(1-t)} \|z - y\|^2 \right) \frac{1}{n} \sum_{j=1}^n \mathrm{d}\delta_{Y_j}}$$

$$= \frac{\sum_{j=1}^n Y_j \exp\left(\frac{\hat{g}_j}{\varepsilon} - \frac{1}{2\varepsilon(1-t)} \|z - Y_j\|^2 \right)}{\sum_{j=1}^n \exp\left(\frac{\hat{g}_j}{\varepsilon} - \frac{1}{2\varepsilon(1-t)} \|z - Y_j\|^2 \right)}.$$

Luego la expresión del estimador del *drift* resulta

$$\hat{b}_t(z) = \frac{1}{1-t}(-z + \nabla\hat{\varphi}_{1-t}(z)).$$

En conclusión, nuestro estimador viene dado por la solución de la siguiente ecuación diferencial estocástica (3.18) discretizada vía Euler-Maruyama

$$\begin{aligned} &\hat{X}_{(j+1)\eta} = \hat{b}_{j\eta}(\hat{X}_{j\eta})\eta + \sqrt{\varepsilon\eta}\,\xi_j, \quad j = 0, \dots, k-1, \\ &\hat{X}_0 = X_0 \sim \mu, \\ &\xi_0, \dots, \xi_{k-1} \sim \mathcal{N}(0, \mathrm{Id}) \text{ i.i.d.} \end{aligned} \tag{3.22}$$

donde $\eta \in (0, 1)$ es el tamaño del paso.

Nota. Notemos que nuestro *drift* es función de un potencial $\hat{g}$ que se obtiene a partir de una ejecución del *algoritmo de Sinkhorn*.

Capítulo 4

Análisis estadístico del Puente de Sinkhorn

Este capítulo trata del error cometido al utilizar el *Puente de Sinkhorn*, lo compararemos en variación total con el *Puente de Schrödinger*. En primer lugar, nos enfocaremos en la tarea de estimación con una sola muestra, ya que refleja mejor las situaciones del mundo real. En segundo lugar introduciremos el error de discretización.

En estos casos de estimación con una sola muestra, se conoce la distribución de origen, como una distribución normal estándar, pero solo se dispone de un conjunto limitado de datos de la distribución de interés, como imágenes en un conjunto de datos. Por tanto supondremos completo acceso a μ y acceso a ν a través de datos independientes e igualmente distribuidos. Por otro lado, asumiremos que ya hemos computado el *algoritmo de Sinkhorn* del siguiente problema entrópico

$$\mathcal{T}_\varepsilon(\mu,\nu_n) := \min_{\pi\in\Pi(\mu,\nu_n)} \left\{ \int_{\mathbb{R}^d\times\mathbb{R}^d} \frac{1}{2}\|x-y\|^2 \,\mathrm{d}\pi(x,y) \;+\; \varepsilon\,\mathrm{KL}(\pi|\mu\otimes\nu_n) \right\},$$

y de este modo hemos obtenido $(\hat{f},\hat{g})$. Estos potenciales permiten llegar a un plan de transporte óptimo entrópico que denotaremos π_n (esto se debe a la ecuación (2.6)). Desde un punto de vista formal esta situación corresponde al caso $m\to\infty$ en la sección anterior; el estimador no cambia (notemos la independencia del parámetro m).

Sea $\tilde{\mathsf{P}}$ la medida de Markov asociada a la mezcla de *Puentes Brownianos* definidos con respecto a π_n. Por el teorema (3.5), las $t-$marginales son

$$\tilde{\mathsf{p}}_t(z) = \mathcal{H}_{(1-t)\varepsilon}[e^{\hat{g}/\varepsilon}\nu_n](z)\mathcal{H}_{t\varepsilon}[e^{\hat{f}/\varepsilon}\mu](z),$$

y el coeficiente deriva empírico de una muestra es

$$\hat{b}_t(z) = \varepsilon\nabla\log\mathcal{H}_{(1-t)\varepsilon}[e^{\hat{g}/\varepsilon}\nu_n](z).$$

Por tanto $\tilde{\mathsf{P}}$ es la ley del proceso

$$\mathrm{d}\tilde{X}_t = \hat{b}_t(\tilde{X}_t)\,\mathrm{d}t + \sqrt{\varepsilon}\;\mathrm{dW}_t, \tag{4.1}$$

con $\tilde{X}_0\sim\mu$.

Nota. Démonos cuenta que esto coincide con nuestra ecuación a estimar (3.22) salvo que en este caso no hay discretización, este proceso no es implementable pero es una herramienta útil para el estudio teórico.

Denotamos $\mathsf{P}^\star$ a la ley del proceso óptimo, $\tilde{\mathsf{P}}$ al igual que antes y $\hat{\mathsf{P}}$ a la ley de nuestra estimación vía *Puente de Sinkhorn*. Estimaremos la diferencia entre el *Puente de Schrödinger* óptimo del siguiente modo.

1. Compararemos $\tilde{\mathsf{P}}$ con $\mathsf{P}^\star$. (Sección 4.1).
2. Luego, la diferencia entre $\tilde{\mathsf{P}}$ y $\hat{\mathsf{P}}$. (Sección 4.2).
3. Se usará la desigualdad triangular para relacionar $\hat{\mathsf{P}}$ y $\mathsf{P}^\star$:

$$\mathbb{E}[\,\mathrm{TV}^2(\hat{\mathsf{P}}_{[0,\tau]}|\mathsf{P}^\star_{[0,\tau]})] \lesssim \mathbb{E}[\,\mathrm{TV}^2(\hat{\mathsf{P}}_{[0,\tau]}|\tilde{\mathsf{P}}_{[0,\tau]})] + \mathbb{E}[\,\mathrm{TV}^2(\tilde{\mathsf{P}}_{[0,\tau]}|\mathsf{P}^\star_{[0,\tau]})].$$

Notemos que las medidas están restringidas al intervalo $[0,\tau] \subset [0,1]$, aquí τ es un hiperparámetro que controla el tamaño total del trayecto de los puentes.

Como ya hemos comentado existen otros métodos para estimar el *Puente de Schrödinger*, sin embargo, la mayoría se han probado en casos simples, por ejemplo en espacios con baja dimensión o estimación de distribuciones sencillas. Se pueden citar [3], [7], [10],[19]. En este punto esos algoritmos sufren de la maldición de la dimensionalidad puesto que se suelen utilizar redes neuronales para estimar el coeficiente deriva y las redes escalan mal al aumentar el tamaño de los datos. En el capítulo 5 se comentará cómo funciona de bien este nuevo método.

Según la formulación del problema del Puente de Schrödinger, cualquier algoritmo para resolverlo puede interpretarse como un aprendizaje de interpolación de acuerdo con el transporte óptimo entrópico de marginales μ y ν. Sin embargo, en la práctica, solo disponemos de un número finito de muestras $\{X_i\}_{n=1}^m \sim \mu$ e $\{Y_j\}_{j=1}^n \sim \nu$. Esto significa que el *Puente de Schrödinger* se entrena para transportar muestras entre las distribuciones empíricas

$$\frac{1}{m}\sum_{i=1}^m \delta_{X_i} \quad \text{y} \quad \frac{1}{n}\sum_{j=1}^n \delta_{Y_j}.$$

Debido a la maldición de la dimensionalidad, las muestras no logran describir correctamente las variedades de imágenes en alta dimensión.

En contraste el método aquí presentado estima el *Puente de Schrödinger* con una tasa de convergencia que depende de la dimensión intrínseca de la medida objetivo (sin embargo esta puede ser alta). Esto lleva al resultado más importante del capítulo, que separaremos en otros dos:

1. El teorema 4.1 en el que, como su propio nombre indica compararemos el *Puente de Schrödinger* con un estimación de una muestra sin discretización, es decir, este es el error que se comete al aproximar.
2. Y luego incorporaremos el error que se produce por introducir la discretización vía Euler-Maruyama.

Teorema 4.1 (Estimación de una muestra sin discretización)**.** *Sean $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ con $supp(\nu)$ contenido en una subvariedad de dimensión k en $\mathbb{R}^d$, dicha subvariedad está contenida en una bola de radio $R > 0$. Sean $\tilde{\mathsf{P}}$ la ley inducida por*

$$\mathrm{d}\tilde{X}_t = \hat{b}_t(\tilde{X}_t)\,\mathrm{d}t + \sqrt{\varepsilon}\ \mathrm{d}W_t,$$

y $\mathsf{P}^\star$ la correspondiente a

$$\mathrm{d}X_t = b^\star_t(X_t)\,\mathrm{d}t + \sqrt{\varepsilon}\,\mathrm{d}W_t.$$

Entonces se tiene para todo $\tau \in [0,1)$,

$$\mathbb{E}[\,TV^2(\tilde{\mathsf{P}}_{[0,\tau]}, \mathsf{P}^\star_{[0,\tau]})] \lesssim \left(\frac{\varepsilon^{-\mathsf{k}/2-1}}{\sqrt{n}} + \frac{R^2\varepsilon^{-\mathsf{k}}}{n(1-\tau)^{\mathsf{k}+2}}\right). \tag{4.2}$$

Nota. El sumando de la derecha diverge exponencialmente en k a medida que $\tau \to 1$; esto es una consecuencia del hecho de que el estimador del coeficiente deriva $\hat{b}_t$ obliga a que las muestras colapsen exactamente sobre los datos de entrenamiento en el tiempo terminal, lo cual está lejos de la verdadera medida objetivo.

Demostración. Introducimos el siguiente el plan entrópico

$$\bar{\pi}_n(x,y) := \Lambda_\varepsilon \exp\left(\frac{\bar{f}(x) + g^\star(y) - \frac{1}{2}\|x-y\|^2}{\varepsilon}\right) \mathrm{d}\mu(x)\,\mathrm{d}\nu_n(y), \tag{4.3}$$

donde $g^\star$ es el *potencial entrópico óptimo* para las medidas (ν,μ) y $\bar{f}:\mathbb{R}^d \longrightarrow \mathbb{R}$ es la siguiente iteración del *algoritmo de Sinkhorn* cuando se considera el potencial $g^\star$ y ν_n, es decir,

$$\bar{f}(x) := -\varepsilon \log\left(\Lambda_\varepsilon \frac{1}{n}\sum_{j=1}^N \exp((g^\star(Y_j) - \frac{1}{2}\|x-Y_j\|^2)/\varepsilon\right). \tag{4.4}$$

Démonos cuenta, también, que $\bar{\pi}_n \in \Pi(\mu,\bar{\nu}_n)$ donde $\bar{\nu}_n$ es una versión escalada de ν_n.

Consideramos ahora la medida asociada a la mezcla de *Puentes Brownianos* con respecto a $\bar{\pi}_n$, la denotaremos como $\bar{\mathsf{P}}$ (con $t-$marginales $\bar{\mathsf{p}}_t$); en virtud del teorema (3.5) corresponde a la solución de una ecuación diferencial estocástica con coeficiente deriva

$$\begin{aligned}\bar{b}_t(z) &= \varepsilon\nabla \log \mathcal{H}_{(1-t)\varepsilon}[e^{g^\star/\varepsilon}\nu_n](z)\\ &= \frac{1}{1-t}\left(-z + \frac{\sum_{j=1}^N Y_j \exp\left(\frac{g^\star(Y_j)}{\varepsilon} - \frac{1}{2\varepsilon(1-t)}\|z-Y_j\|^2\right)}{\sum_{j=1}^N \exp\left(\frac{g^\star(Y_j)}{\varepsilon} - \frac{1}{2\varepsilon(1-t)}\|z-Y_j\|^2\right)}\right).\end{aligned}$$

Este coeficiente sale directamente de aplicar el teorema a la medida $\bar{\mathsf{P}}$.

Insertamos $\bar{P}_{[0,\tau]}$ vía la desigualdad triangular

$$\mathbb{E}[\,\mathrm{TV}^2(\tilde{\mathsf{P}}_{[0,\tau]}, \mathsf{P}^\star_{[0,\tau]})] \lesssim \mathbb{E}[\,\mathrm{TV}^2(\tilde{\mathsf{P}}_{[0,\tau]}, \bar{\mathsf{P}}_{[0,\tau]})] + \mathbb{E}[\,\mathrm{TV}^2(\bar{\mathsf{P}}_{[0,\tau]}, \mathsf{P}^\star_{[0,\tau]})].$$

La constante que sale en la cota viene determinada por trabajar con la variación total cuadrática (ver apéndice B). Utilizamos ahora la *desigualdad de Pinsker* (ver apéndice B).

$$\lesssim \mathbb{E}[\,\mathrm{KL}(\tilde{\mathsf{P}}_{[0,\tau]}|\bar{\mathsf{P}}_{[0,\tau]})] + \mathbb{E}[\,\mathrm{KL}(\bar{\mathsf{P}}_{[0,\tau]}|\mathsf{P}^\star_{[0,\tau]})].$$

Analizamos ambos sumandos por separado:

1. $\mathbb{E}[\,\mathrm{KL}(\tilde{\mathsf{P}}_{[0,\tau]}|\bar{\mathsf{P}}_{[0,\tau]})]$:

   En virtud del teorema (3.5) un proceso con ley $\tilde{\mathsf{P}}$ se puede obtener muestreando un *Puente Browniano* entre $(X_0,X_1) \sim \pi_n$ (análogamente con $\bar{\mathsf{P}}$ y $\bar{\pi}_n$). Hacemos uso de la desigualdad de procesamiento de datos (procesar datos no puede aumentar la información sobre la fuente original.

Todo procesamiento de datos reduce o conserva la información; nunca la aumenta). Esto se aplica al aumentar el intervalo de $[0, \tau]$ a $[0, 1]$.

$$\mathbb{E}[\,\mathbf{KL}(\tilde{\mathsf{P}}_{[0,\tau]}|\bar{\mathsf{P}}_{[0,\tau]})] \leq \mathbb{E}[\,\mathbf{KL}(\tilde{\mathsf{P}}|\bar{\mathsf{P}})],$$

Como π_n y $\bar{\pi}_n$ tienen densidades con respecto $\mu \otimes \nu_n$ se verifica:

$$\leq \mathbb{E}[\,\mathbf{KL}(\pi_n|\bar{\pi}_n)] = \mathbb{E}\left[\int \log\left(\frac{\pi_n}{\bar{\pi}_n}\right) \mathrm{d}\pi_n\right].$$

Recordando las expresiones que tienen π_n y $\bar{\pi}_n$:

$$\begin{cases} \Lambda_\varepsilon \exp\left(\frac{\hat{f}(x)+\hat{g}(y)-\frac{1}{2}\|x-y\|^2}{\varepsilon}\right) \mathrm{d}\mu(x)\,\mathrm{d}\nu_n(y), \\ \\ \Lambda_\varepsilon \exp\left(\frac{\bar{f}(x)+g^\star(y)-\frac{1}{2}\|x-y\|^2}{\varepsilon}\right) \mathrm{d}\mu(x)\,\mathrm{d}\nu_n(y). \end{cases}$$

y al simplificar las exponenciales y el logaritmo queda:

$$\frac{1}{\varepsilon}(\hat{f}(x) + \hat{g}(y) + \bar{f}(x) + g^\star(y)).$$

Retomamos la última expresión

$$\frac{1}{\varepsilon}\mathbb{E}\left[\int \hat{f}(x) + \hat{g}(y)\,\mathrm{d}\pi_n(x,y) - \int \bar{f}(x)\,\mathrm{d}\mu(x) - \int g^\star\,\mathrm{d}\nu_n(y)\right],$$

y, por definición,

$$\frac{1}{\varepsilon}\mathbb{E}\left[\mathcal{T}_\varepsilon(\mu,\nu_n) - \int \bar{f}(x)\,\mathrm{d}\mu(x) - \int g^\star\,\mathrm{d}\nu_n(y)\right].$$

Por otro lado, $\bar{f}$ satisface

$$\bar{f} = \max_{f \in L^1(\mu)} \left\{ \int_{\mathbb{R}^d} f(x)\ \mathrm{d}\mu(x) + \int_{\mathbb{R}^d} g^\star(y)\ \mathrm{d}\nu_n(y) \right.$$

$$\left. +\varepsilon \int_{\mathbb{R}^d\times\mathbb{R}^d} \left[\exp\left(\frac{f(x)+g^\star(y)-\frac{1}{2}\|x-y\|^2}{\varepsilon}\right) - 1\right]\ \mathrm{d}\mu(x)\ \mathrm{d}\nu_n(y)\right\}.$$

En particular, se tiene:

$$\int_{\mathbb{R}^d} \bar{f}(x)\ \mathrm{d}\mu(x) + \int_{\mathbb{R}^d} g^\star(y)\ \mathrm{d}\nu_n(y)$$

$$+\varepsilon \int_{\mathbb{R}^d\times\mathbb{R}^d} \left[\exp\left(\frac{\bar{f}(x)+g^\star(y)-\frac{1}{2}\|x-y\|^2}{\varepsilon}\right) - 1\right]\ \mathrm{d}\mu(x)\ \mathrm{d}\nu_n(y)$$

$$\leq \int_{\mathbb{R}^d} f^\star(x)\ \mathrm{d}\mu(x) + \int_{\mathbb{R}^d} g^\star(y)\ \mathrm{d}\nu_n(y)$$

$$+\varepsilon \int_{\mathbb{R}^d\times\mathbb{R}^d} \left[\exp\left(\frac{f^\star(x)+g^\star(y)-\frac{1}{2}\|x-y\|^2}{\varepsilon}\right) - 1\right]\ \mathrm{d}\mu(x)\ \mathrm{d}\nu_n(y)$$

Siguiendo con lo de arriba

$$\begin{aligned}\mathbb{E}[\,\mathrm{KL}(\tilde{\mathsf{P}}_{[0,\tau]}|\bar{\mathsf{P}}_{[0,\tau]})] &\leq \frac{1}{\varepsilon}\mathbb{E}\left[\mathcal{T}_\varepsilon(\mu,\nu_n) - \int f^\star \,\mathrm{d}\mu - \int g^\star \,\mathrm{d}\nu_n\right] \\ &= \frac{1}{\varepsilon}\mathbb{E}\left[\mathcal{T}_\varepsilon(\mu,\nu_n) - \int f^\star \,\mathrm{d}\mu - \int g^\star \,\mathrm{d}\nu\right],\end{aligned}$$

donde se utiliza que g es independiente de $Y_1, \dots Y_n$. Recordemos la expresión

$$\begin{aligned}I_\varepsilon[\pi] = &\int_{\mathbb{R}^d} f(x)\ \mathrm{d}\mu(x) + \int_{\mathbb{R}^d} g(y)\ \mathrm{d}\nu(y) \\ &+ \varepsilon \int_{\mathbb{R}^d\times\mathbb{R}^d} r(x,y)\log\left(\frac{r(x,y)}{\exp\left(\frac{f(x)+g(y)-c_2(x,y)}{\varepsilon}\right)}\right)\ \mathrm{d}\mu(x)\ \mathrm{d}\nu(y).\end{aligned}$$

En el caso de que estemos con los potenciales óptimos entonces el logaritmo es 0 y por tanto

$$\int f^\star \,\mathrm{d}\mu + \int g^\star \,\mathrm{d}\nu = \mathcal{T}_\varepsilon(\mu,\nu)$$

En definitiva,

$$\mathbb{E}[\,\mathrm{KL}(\tilde{\mathsf{P}}_{[0,\tau]}|\bar{\mathsf{P}}_{[0,\tau]})] \leq \frac{1}{\varepsilon}\mathbb{E}\left[\mathcal{T}_\varepsilon(\mu,\nu_n) - \mathcal{T}_\varepsilon(\mu,\nu)\right],$$

Utilizando el resultado [17, Teorema 2.6] se finaliza esta parte.

2. Continuamos con el siguiente término. Para esta parte nos remitimos al C.3 del apéndice C lema . En virtud de dicho resultado se tiene

$$\mathbb{E}[\,\mathrm{KL}(\bar{\mathsf{P}}_{[0,\tau]}|\mathsf{P}^\star_{[0,\tau]})] \leq \frac{R^2\varepsilon^{-\mathsf{k}}}{n}(1-\tau)^{-\mathsf{k}-2}.$$

En conclusión, al unir ambos se consigue la tesis

$$\mathbb{E}[\,\mathrm{TV}^2(\hat{\mathsf{P}}_{[0,\tau]}, \mathsf{P}^\star_{[0,\tau]})] \lesssim \left(\frac{\varepsilon^{-\mathsf{k}/2-1}}{\sqrt{n}} + \frac{R^2\varepsilon^{-\mathsf{k}}}{n(1-\tau)^{\mathsf{k}+2}}\right).$$

□

En definitiva, hemos conseguido una cota del error de aproximación que depende de la dimensión intrínseca.

Ahora bien, también tenemos que tener en cuenta el error que se comete al discretizar la ecuación diferencial estocástica, remitimos a la expresión que teníamos en (3.22). Nos limitaremos a presentar directamente la cota, ya que los detalles técnicos quedan fuera del alcance de este trabajo y pueden consultarse en las referencias [32], que a la vez se apoya en [6].

Si $\mathrm{supp}(\nu) \subset B(0,R)$, se verifica

$$\mathbb{E}[TV^2(\hat{\mathsf{P}}_{[0,\tau]}|\tilde{\mathsf{P}}_{[0,\tau]})] \lesssim (\varepsilon+1)(1-\tau)^{-2}d\eta\ (1\ \vee\ R^4(1-\tau)^{-2}\varepsilon^{-2}).$$

Teorema 4.2 (Error total del Puente de Sinkhorn)**.** *Sean $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ con $supp(\nu) \subset B(0,R) \subset \mathcal{M}$, donde $\mathcal{M}$ es una subvariedad de dimensión k en $\mathbb{R}^d$. Dadas n muestras i.i.d de ν, el Puente de Sinkhorn de una muestra $\hat{\mathsf{P}}$ estima el Puente de Schrödinger $\mathsf{P}^\star$ con el siguiente error:*

$$\mathbb{E}[\, TV^2(\hat{\mathsf{P}}_{[0,\tau]}, \mathsf{P}^\star_{[0,\tau]})] \lesssim \left(\frac{\varepsilon^{-\mathsf{k}/2-1}}{\sqrt{n}} + \frac{R^2 \varepsilon^{-\mathsf{k}}}{n(1-\tau)^{\mathsf{k}+2}} \right) + (\varepsilon+1)(1-\tau)^{-2} \eta d \, (1 \ \vee \ R^4 (1-\tau)^{-2} \varepsilon^{-2}). \tag{4.5}$$

Demostración. Esta prueba resulta de utilizar las cotas obtenidas en los resultados anteriores. □

El teorema 4.5 no implica generar de la distribución objetivo ν. Obtener tales garantías requiere argumentar que simular el *Puente de Sinkhorn* en un intervalo adecuado $[0, \tau]$, para un τ cercano a 1, produce muestras cercanas a la densidad real (sin colapsar completamente sobre los datos de entrenamiento).

Capítulo 5

Aplicaciones prácticas

Hasta aquí nos hemos centrado en los aspectos teóricos y el análisis estadístico necesarios para comprender el *Puente de Sinkhorn* y su relevancia en el contexto del transporte óptimo regularizado y el problema del Puente de Schrödinger. En este capítulo, dejaremos de lado las cuestiones teóricas y exploraremos la implementación práctica del método, con ejemplos concretos que nos ayuden a comprender su funcionamiento.

Comenzaremos viendo cómo los distintos parámetros del algoritmo «como el coeficiente de regularización, ε, y el tamaño total del puente, τ» pueden afectar a la calidad de los resultados obtenidos. Además, comentaremos los diferentes métodos para computar el *algoritmo de Sinkhorn* y, de este modo, calcular los potenciales entrópicos óptimos.

A continuación, presentaremos un ejemplo básico de aplicación en distribuciones tridimensionales, lo cual servirá como punto de partida accesible para visualizar cómo se comporta el algoritmo en un entorno controlado. Este ejemplo tendrá como objetivo facilitar la comprensión intuitiva del método.

Finalmente, trasladaremos lo aprendido a un caso de uso más ambicioso: la generación de imágenes. Aquí mostraremos cómo el *Puente de Sinkhorn* puede aplicarse para interpolar distribuciones complejas en espacios de alta dimensión. En nuestro experimento trabajaremos con imágenes RGB de tamaño 128×128, este experimento se ha realizado en una tarjeta gráfica NVIDIA GeForce RTX 4090 con 24GB de VRAM.

En este punto detallaremos la influencia de los hiperparámetros y corroboraremos las hipótesis antes descritas. Además, anticipamos que en espacios de dimensión alta el algoritmo no funciona bien, precisamente por la alta dimensión. Además, expondremos heurísticamente otro motivo por el que no funciona correctamente. Es nuestro objetivo como trabajo futuro solventar estos problemas ya sea por métodos de reducción de dimensión, como los *autoencoders*, o por una nueva propuesta de estimador.

Siguiendo las hipótesis asumidas hasta ahora y basándonos en los teoremas expuestos en el capítulo anterior se deduce la repercusión que tienen los parámetros a la hora de la implementación. Para una exposición más clara las resumimos en la siguiente tabla.

Parámetro	**Descripción**	**Recomendaciones**
ε	Parámetro de regularización. A medida que crece, el *Puente de Schrödinger* se vuelve menos informativo, pero el *algoritmo de Sinkhorn* converge más rápido. Para valores pequeños, puede no converger.	Usar valores moderados que equilibren entre calidad informativa y estabilidad numérica. Evitar valores muy pequeños.
τ	Cuando ε es alto, la marginal $\mathsf{p}_\tau^\star$ solo reconstruye ν si $\tau \to 1$. En ese caso, k diverge por el colapso de las muestras en los datos de entrenamiento al final del tiempo, lo que se aleja de la medida objetivo.	Si se usa ε alto, no llevar τ demasiado cerca de 1. Buscar un valor de compromiso que evite la divergencia de k.

Cuadro 5.1: Descripción y recomendaciones para los parámetros del modelo.

En cuanto a los parámetros R y k no son parámetros a elegir, pero cabe comentar que su existencia es condición necesaria para los resultados teóricos. Resulta razonable suponer que las imágenes se encuentran contenidas en una variedad de dimensión inferior al espacio ambiente; sin embargo, conocer su valor exacto no resulta crucial para la implementación. Lo mismo puede decirse respecto a R.

Por otro lado, el algoritmo genérico del *Puente de Sinkhorn* lo estableceremos a continuación no sin antes hacer un breve comentario. Se ha decidido implementarlo en *PyTorch* para mejorar la eficiencia de las operaciones.

Cabe comentar que la parte de discretización de la ecuación diferencial estocástica (ecuación de Langevin discretizada vía Euler-Maruyama) en el algoritmo es estándar en la comunidad de la generación de nuevas muestras a partir de muestras de entrenamiento. La novedad que presenta este método es la estimación del *drift* de la ecuación

$$\mathrm{d}X_t = \left(\frac{-1}{1-t} X_t + \frac{1}{1-t} \nabla \varphi^\star_{1-t}(X_t) \right) \mathrm{d}t + \sqrt{\varepsilon}\, \mathrm{d}\mathbf{W}_t, \tag{5.1}$$

mediante una única ejecución del *algoritmo de Sinkhorn*. Ya lo hemos comentado en el comienzo de la sección relativa al *Puente de Sinkhorn* 3.2.

El enfoque más común es estimar el *drift* mediante redes neuronales, aunque esto es ineficiente y necesita de tomar la ecuación hacia delante como determinista en algunos casos.

5.1. Ejemplo básico 3 dimensiones

Se presenta a continuación un ejemplo en un entorno controlado para vislumbrar el potencial de este estimador y hacernos un esquema mental de cómo se aplicará para la generación de imágenes. Consideramos tres distribuciones tridimensionales con las siguientes formas: un cubo, un toro y una hélice.

Algorithm 1 Puente de Sinkhorn

Input: Datos $\{X_i\}_{i=1}^m \sim \mu$, $\{Y_j\}_{j=1}^n \sim \nu$, parámetros $\varepsilon > 0$, $\tau \in (0,1)$, y $N \geq 1$

Computar: Potenciales de Sinkhorn mediante *algoritmo de Sinkhorn* $(\hat{f}, \hat{g}) \in \mathbb{R}^m \times \mathbb{R}^n$ ▷ Usando POT

Inicializar: $x^{(0)} \sim \mu$, $k = 0$, tamaño de paso $\eta = \tau/N$

while $k \leq N - 1$ **do**

$x^{(k+1)} = x^{(k)} + \eta\, \hat{b}_\eta(x^{(k)}) + \sqrt{\eta\, \varepsilon}\, \xi_k$ ▷ $\xi_k \sim \mathcal{N}(0, I)$

$k \leftarrow k + 1$

Return: $x^{(N)}$

Generamos 2000 puntos de cada distribución. Queremos transportar nuevas muestras de la distribución inicial a la distribución final.

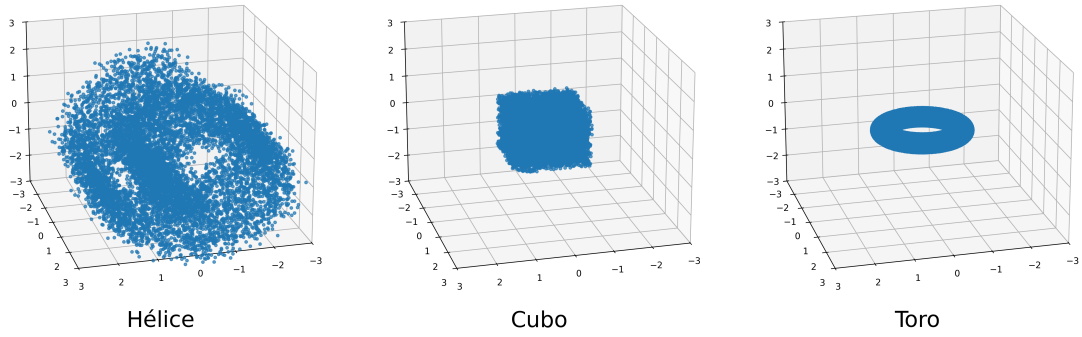


Figura 5.1: Conjunto de datos en un entorno simulado

Hemos ejecutado el transporte de la distribución con forma de hélice a cubo, de cubo a toro y de toro a cubo. Con $\tau = 0{,}9$, 50 pasos y $\varepsilon = 0{,}01$.

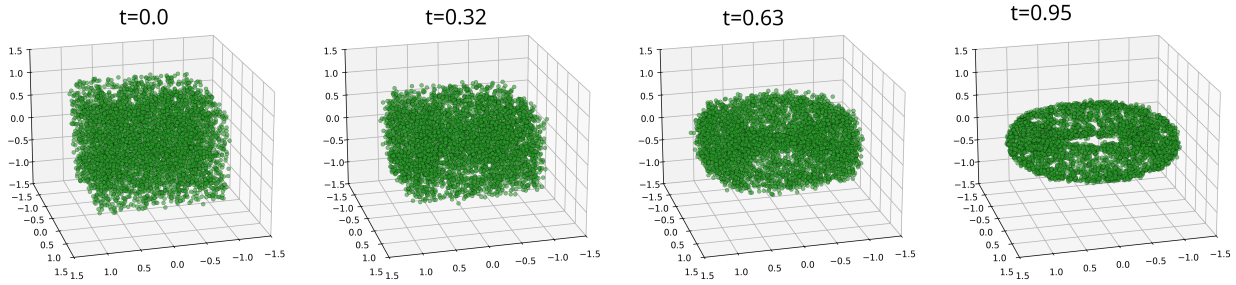


Figura 5.2: Evolución de cubo a toro

En las transformaciones entre las distribuciones con forma de toro y de cubo, se observa que las representaciones generadas no reproducen exactamente la distribución original. Esta discrepancia, en el caso de datos visuales como imágenes, podría tener implicaciones relevantes en la calidad de la reconstrucción, las cuales se abordarán más adelante en el trabajo cuando corresponda.

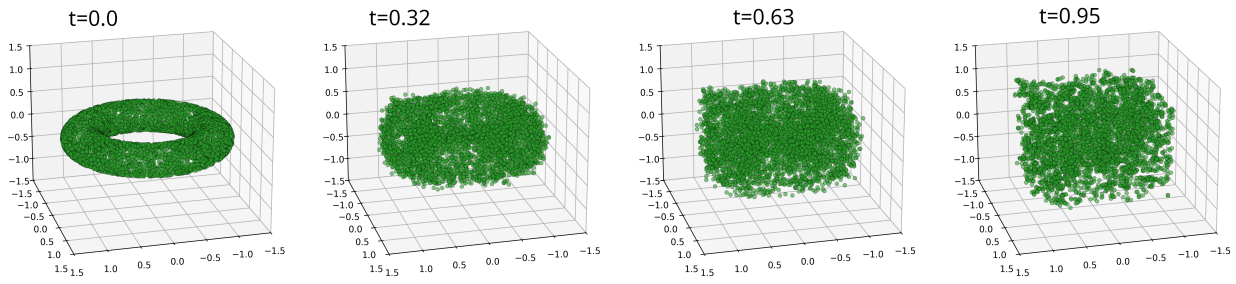


Figura 5.3: Evolución de toro a cubo

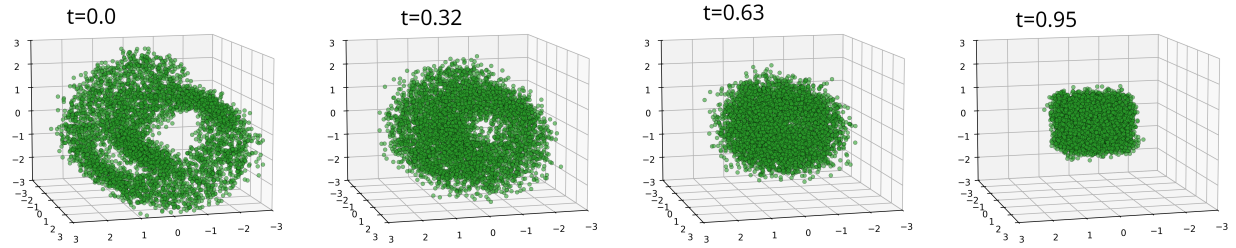


Figura 5.4: Evolución de hélice a cubo

Como se ha advertido este es un ejemplo sencillo e ilustra el funcionamiento del algoritmo, en la siguiente sección trataremos de extrapolar esta idea a las imágenes.

5.2. Generación de imágenes

Como ya se ha introducido a lo largo del trabajo nuestra intención es generar imágenes a través del algoritmo del *Puente de Sinkhorn*. Este es un objetivo ambicioso pues las imágenes son datos de gran dimensión (alto $\times$ ancho $\times$ número de canales) y esto hay que multiplicarlo por el número de imágenes necesario para obtener la estimación del *drift* mediante el *algoritmo de Sinkhorn*.

En cuanto a este, se utilizará la librería POT, sin embargo dentro de esta hay diferentes algoritmos los cuales se resumen en la siguiente lista:

1. `sinkhorn`: el algoritmo clásico de Sinkhorn.
2. `sinkhorn_log`: una versión en espacio logarítmico. Es más estable pero puede ser más lenta.
3. `sinkhorn_stabilized`: una versión estabilizada con logaritmos. Útil para evitar errores numéricos cuando la regularización es muy pequeña.
4. `sinkhorn_epsilon_scaling`: una variante que escala el valor de ε (regularización). Suele obtener mejores resultados en casos donde otros métodos fallan en converger.
5. `greenkhorn`: versión avaraciosa (greedy) del *algoritmo de Sinkhorn*. Puede acelerar el proceso en algunos contextos.

6. `screenkhorn`: versión con preselección (screening) que reduce el tamaño del problema. Aproximación rápida ideal para problemas grandes.

En nuestro caso hemos decidido utilizar el algoritmo `sinkhorn_log` pues el trabajar con píxeles de ruido gaussiano los valores están cerca de 0 lo que puede hacer que el ordenador colapse al tratar con números muy pequeños. Y además, a priori al no conocer las distribuciones las imágenes pueden diferir mucho entre ellas luego necesitamos estabilidad.

Por lo general aumentar la cantidad de muestras hará que los resultados sean mejores, sin embargo la capacidad de la GPU es limitada, por ello hay que buscar el equilibrio. En cuanto a las muestras de partida cabe mencionar que serán ruido gaussiano, es decir, los píxeles de cada imagen se distribuyen siguiendo una distribución normal.

Basándonos en el ejemplo de la sección anterior asumiremos que el ruido y las imágenes siguen ciertas distribuciones respectivamente en un espacio de imágenes abstracto. La clave es buscar el camino óptimo que unan dichas distribuciones. Con un número de imágenes, que luego especificaremos, estimaremos el *drift* y con ello el camino que relacione el ruido con las imágenes.

Para generar una nueva imagen generaremos un nuevo ruido y con el *drift* obtenido en el paso anterior y mediante el algoritmo descrito arriba (que incluye la discretización de la ecuación diferencial estocástica de Langevin según Euler-Maruyama) desplazar las imágenes de ruido a imágenes que deseamos generar.

Utilizaremos el conjunto de datos CelebAI, conjunto de datos típico en la literatura de *Machine Learning*, que incluyen 200.000 fotos de tamaño 178 x 218 de caras de famosos. Tiene sentido considerar que estas imágenes siguen una distribución (aunque sea desconocida) dado el parecido entre ellas. Además, hemos de suponer que subyacen en una subvariedad de dimensión k para aplicar los resultados del capítulo anterior. Para hacer los experimentos hemos reescalado las imágenes a tamaño 128 x 128. Un esquema intuitivo de lo que tratamos de hacer es la figura 5.5

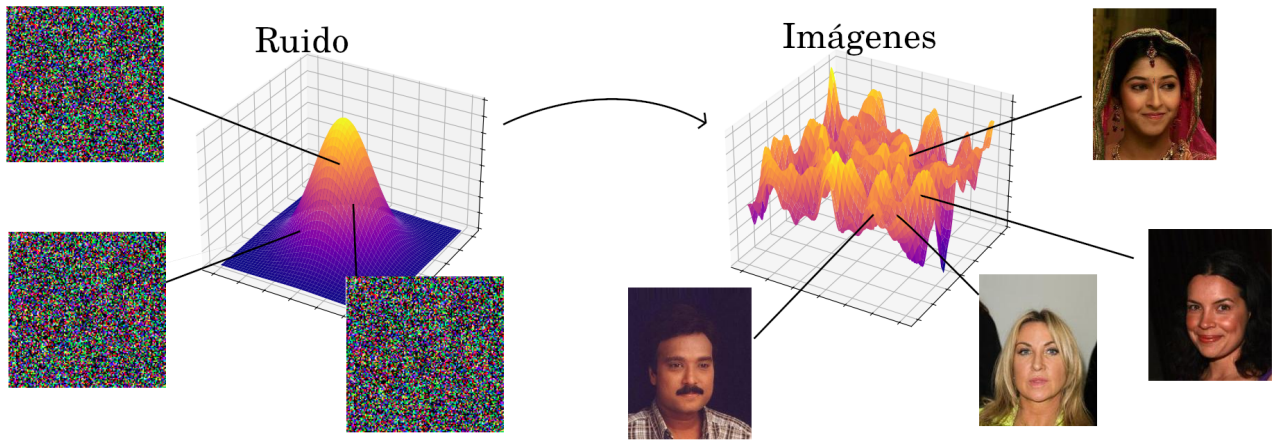


Figura 5.5: Esquema intuitivo del transporte de imágenes

Como ya se ha advertido supondremos que existe un espacio abstracto de imágenes, en una parte de este están las imágenes de ruido gaussiano con cierta distribución μ y en otra parte están las imágenes de caras con otra distribución ν. Este espacio está formado por toda la posible combinación de píxeles de cada imagen, es decir, la mayoría de elementos serán ruido.

Para comprobar que las imágenes de ruido y las imágenes de caras se diferencian bien en el espacio de imágenes hemos hecho un t-SNE con 3000 muestras de cada conjunto.

El algoritmo *t-SNE* parte de un conjunto de puntos en un espacio de alta dimensión y tiene como objetivo representar esos puntos en un espacio de baja dimensión, preservando sus relaciones de proximidad.

1. Para cada punto x_i, se calcula una distribución de probabilidad condicional $p_{j|i}$ sobre todos los demás puntos, utilizando una distribución gaussiana centrada en x_i. Esta distribución refleja qué tan similares son los puntos cercanos a x_i en el espacio original.

2. Se simetrizan estas probabilidades para obtener p_{ij}, una medida de similitud conjunta entre puntos, mediante:
$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$$
donde n es el número total de puntos.

3. En el espacio de baja dimensión, se define una distribución análoga q_{ij} utilizando una distribución t de Student con un grado de libertad (equivalente a una Cauchy).

4. El objetivo del algoritmo es que q_{ij} se aproxime lo más posible a p_{ij}. Para ello, se minimiza la divergencia de Kullback-Leibler entre ambas distribuciones:
$$KL(P \,||\, Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

5. Esta función de coste se minimiza mediante descenso por gradiente.

En resumen, t-SNE busca conservar las similitudes locales entre puntos, proyectando datos de alta dimensión en un espacio de menor dimensión de forma que se respeten las relaciones de proximidad. En la figura 5.6 se puede observar

En cuanto a los parámetros ε y τ los dejaremos en un inicio sin determinar, más adelante probaremos con varios y comentaremos las diferencias obtenidas y cuál puede ser lo óptimo en este ejemplo de los rostros. El número de pasos lo establecemos en un inicio en 500, i.e., $N = 500$.

En el experimento se han aplanado las imágenes para convertirlas en vectores y, de este modo, hacerlo compatible con las funciones de la librería de Python dedicada a transporte óptimo. Esto no es lo más recomendable pues al aplanar las imágenes se pierde información espacial entre los píxeles. Notemos que al hacer esta operación estaríamos trabajando con elementos de la forma $[\text{nro. de muestras}, 1, 128 \times 128 \times 3]$.

Puesto que la memoria de la GPU es limitada y estamos tratando con vectores de gran tamaño el número de muestras no puede exceder las 1400 imágenes en cada conjunto, en caso contrario el sistema colapsa. Es por ello que en este experimento hemos utilizado 500 imágenes de ruido y otras 500 imágenes de caras.

Situémonos en el algoritmo 0. Comencemos identificando los datos:

en primer lugar, $\{X_i\}_m$ serán los datos de partida es decir las imágenes de ruido gaussiano cuyos píxeles se distribuyen normalmente. Como ya hemos comentado arriba se supondrá que en el espacio de imágenes estas imágenes ruidosas «están cerca». Por otro lado, $\{Y_j\}_n$ serán las imágenes de las caras que

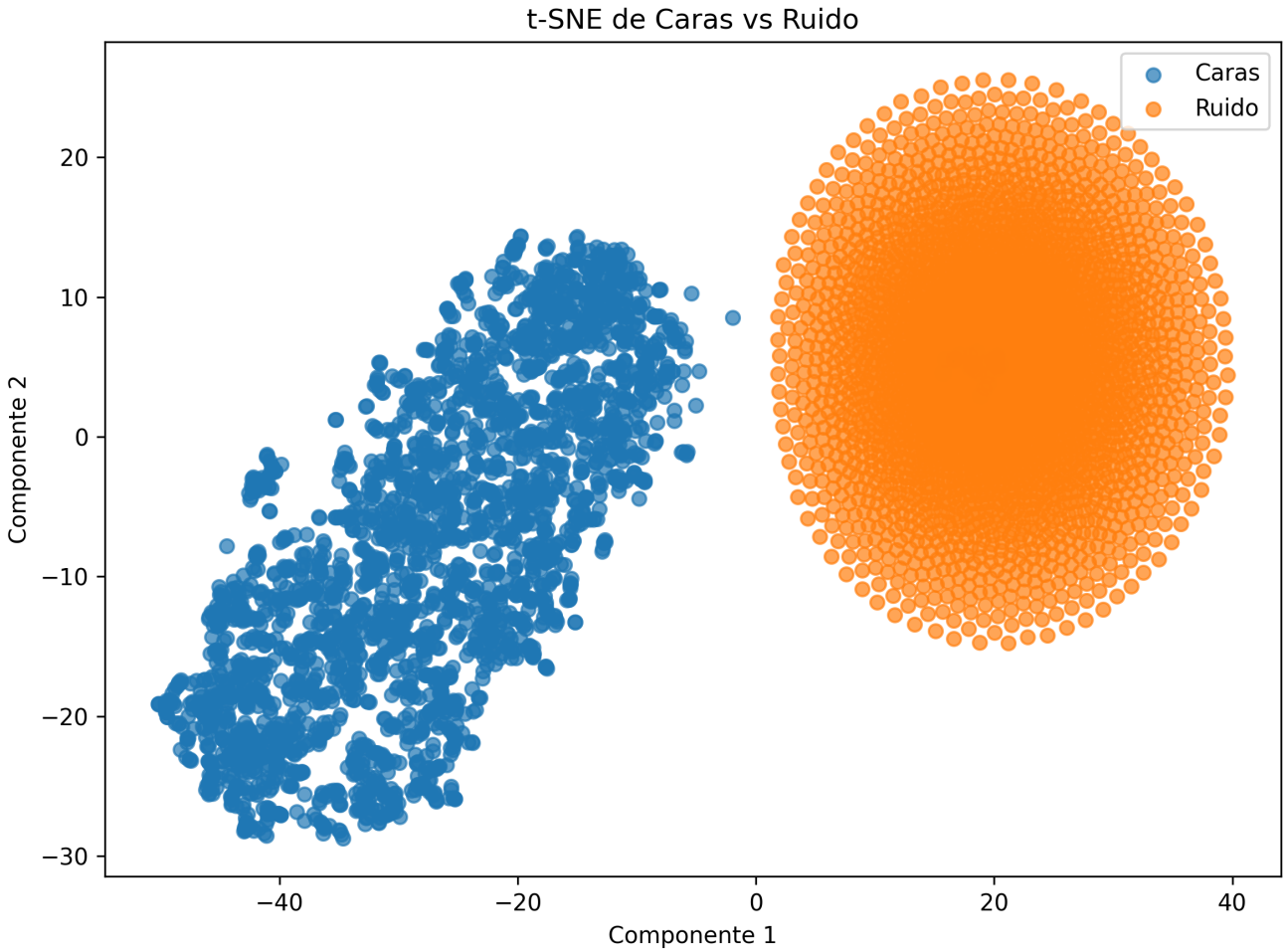


Figura 5.6: t-SNE: Ruido - Caras

también supondremos que están cerca entre ellas. Se puede observar de nuevo un esquema intuitivo en la figura 5.5.

El tratar con un número de muestras limitado trae problemas pues recordando la expresión del teorema 4.5

$$\mathbb{E}[\,\mathrm{TV}^2(\hat{\mathsf{P}}_{[0,\tau]}, \mathsf{P}^{\star}_{[0,\tau]})] \lesssim \left(\frac{\varepsilon^{-\mathsf{k}/2-1}}{\sqrt{n}} + \frac{R^2\varepsilon^{-\mathsf{k}}}{n(1-\tau)^{\mathsf{k}+2}}\right) + (\varepsilon+1)(1-\tau)^{-2}\eta d\,(1 \vee R^4(1-\tau)^{-2}\varepsilon^{-2}), \tag{5.2}$$

la diferencia entre el *Puente de Schrödinger* exacto y el estimado disminuye a medida que el número de muestras crece, de hecho disminuye bastante despacio. Debido a esto se espera que:

1. el número de muestras dentro del margen de la GPU no va a afectar excesivamente.
2. no haya buenos resultados en la generación de imágenes si no conseguimos reducir la dimensión.

Por último, en el paso de generar muestras partimos de una imagen de ruido aleatoria, al generarla aleatoriamente no va a estar en $\{X_i\}_{i=1}^m$ pero está muy cerca de una que sí que lo está. Como hemos

«aprendido» el camino más probable que lleve imágenes de ruido a imágenes de caras, al darle una nueva muestra de ruido y transportarla por dicho camino nos debería llevar a una imagen cercana de una cara pero no exactamente una de $\{Y_j\}_{j=1}n$ sino una muy parecida (ya que hemos supuesto que se distribuyen en un espacio de imágenes y todas las caras estarán cerca). Y con ello se generaría una nueva imagen.

No se puede esperar que esta imagen sea completamente nueva pues, lo primero hemos usado un número de muestras limitadas (lo normal es tratar con orden 10^3) y lo segundo es un problema del estimador. Lo desarrollamos de manera heurística a continuación: recordemos la expresión del *drift* definido como

$$\hat{b}_t(z) = \frac{1}{1-t}\left(-z + \frac{\sum_{j=1}^{N} Y_j\, w_j(z)}{\sum_{j=1}^{N} w_j(z)}\right),$$

donde

$$w_j(z) = \exp\left(\frac{\hat{g}(Y_j)}{\varepsilon} - \frac{1}{2\varepsilon(1-t)}\|z - Y_j\|^2\right).$$

Este *drift* constituye un campo vectorial que depende de una combinación ponderada de las muestras $\{Y_j\}_{j=1}^{N}$, con pesos $w_j(z)$ que decrecen exponencialmente con la distancia entre z y cada Y_j, modulados adicionalmente por el término $\hat{g}(Y_j)$.

En particular, el término $-\frac{1}{2\varepsilon(1-t)}\|z - Y_j\|^2$ en el exponente induce que para muestras Y_j alejadas de z, el valor de $\|z - Y_j\|^2$ sea grande, haciendo que $w_j(z)$ se reduzca exponencialmente hacia cero. Esta rápida decaída hace que las contribuciones de las muestras distantes sean despreciables frente a las de aquellas Y_j que se encuentran próximas a z. Dicho formalmente,

Consideremos la expresión

$$\frac{\sum_{j=1}^{N} Y_j \exp\left(\frac{\hat{g}(Y_j)}{\varepsilon} - \frac{1}{2\varepsilon(1-t)}\|z - Y_j\|^2\right)}{\sum_{j=1}^{N} \exp\left(\frac{\hat{g}(Y_j)}{\varepsilon} - \frac{1}{2\varepsilon(1-t)}\|z - Y_j\|^2\right)}.$$

Separando los sumandos, para cada j el término en numerador y denominador es

$$Y_j \exp\left(\frac{\hat{g}(Y_j)}{\varepsilon} - \frac{1}{2\varepsilon(1-t)}\|z - Y_j\|^2\right) \quad \text{y} \quad \exp\left(\frac{\hat{g}(Y_j)}{\varepsilon} - \frac{1}{2\varepsilon(1-t)}\|z - Y_j\|^2\right).$$

Al hacer $t \to 1$, observamos el comportamiento del factor

$$\exp\left(-\frac{1}{2\varepsilon(1-t)}\|z - Y_j\|^2\right).$$

- Si $z = Y_j$, entonces $\|z - Y_j\|^2 = 0$ y el factor vale

  $$\exp(0) = 1,$$

  por lo que el peso es

  $$\exp\left(\frac{\hat{g}(Y_j)}{\varepsilon}\right),$$

  un valor finito y no nulo.

- Si $z \neq Y_j$, entonces $\|z - Y_j\|^2 > 0$ y como $t \to 1$ (entonces $1 - t \downarrow 0$), se tiene

$$\frac{1}{2\varepsilon(1-t)}\|z - Y_j\|^2 \to +\infty,$$

  por lo que

$$\exp\left(-\frac{1}{2\varepsilon(1-t)}\|z - Y_j\|^2\right) \to 0.$$

Esto implica que, en el límite

$$\frac{\sum_{j=1}^{N} Y_j \exp\left(\frac{\hat{g}(Y_j)}{\varepsilon} - \frac{1}{2\varepsilon(1-t)}\|z - Y_j\|^2\right)}{\sum_{j=1}^{N} \exp\left(\frac{\hat{g}(Y_j)}{\varepsilon} - \frac{1}{2\varepsilon(1-t)}\|z - Y_j\|^2\right)} \xrightarrow[t\to 1]{} Y_{j^*},$$

donde

$$j^* = \min_j \|z - Y_j\|^2,$$

es decir, el valor Y_j más cercano a z.

Por lo tanto, al acercarse t a 1, los pesos asignados a muestras Y_j distantes de z tienden a cero y la expresión se concentra en la muestra más cercana. Así, la expresión

$$\frac{\sum_{j=1}^{N} Y_j\, w_j(z)}{\sum_{j=1}^{N} w_j(z)}$$

puede interpretarse como una combinación convexa ponderada principalmente por las muestras localmente cercanas a z.

En consecuencia, la dirección del *drift* orienta z hacia un punto dentro de la envolvente convexa del conjunto de muestras $\{Y_j\}$. Mostramos en 5.7 una ilustración.

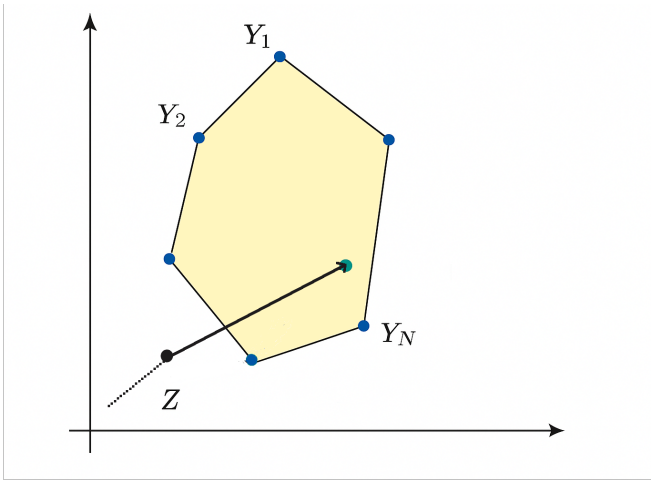


Figura 5.7: Esquema intuitivo de la envolvente convexa

Este hecho implica que, aunque se inicialice el proceso con muestras z provenientes de una distribución de ruido distante (por ejemplo, X_j), el *drift* inducirá una trayectoria que converge hacia puntos que

son combinaciones convexas de las muestras de la distribución objetivo, es decir, imágenes muy similares a las Y_j observadas.

Esta propiedad subraya una propiedad inherente a la construcción del *drift*: la incapacidad para generar nuevas representaciones fuera de la envolvente convexa formada por las muestras finitas de Y_j. Generar algo fuera de la envolvente convexa podría ser poco recomendable, lo que hay que tener en cuenta es que con más muestras la envolvente convexa se los datos se aproximaría a cubrir todo el soporte de la distribución.

Finalmente, cabe destacar que la ponderación basada en $w_j(z)$ actúa como un mecanismo de selección local, enfatizando la influencia de las muestras más cercanas y su correspondiente $\hat{g}(Y_j)$, mientras que minimiza la contribución de aquellas más lejanas.

De hecho, lo que ocurre es no solo que caiga dentro de la envolvente convexa sino que cae muy próximo a las muestras, forzando a que las nuevas generaciones se parezcan mucho a una imagen del conjunto de muestras. Representamos con círculos rojos las zonas donde caerían las nuevas muestras. Se puede observar en la figura 5.8.

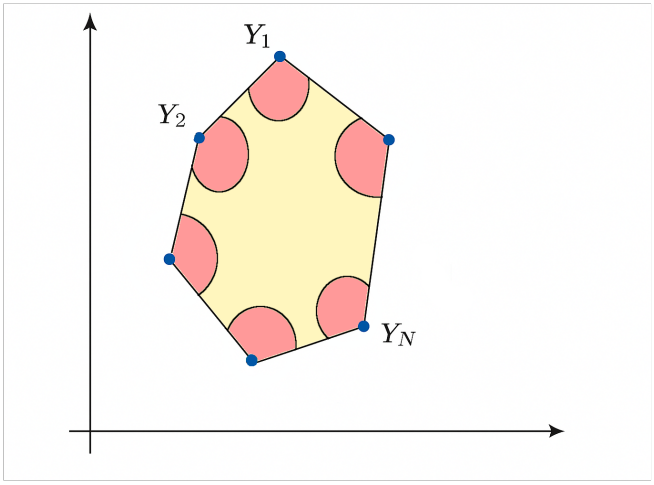
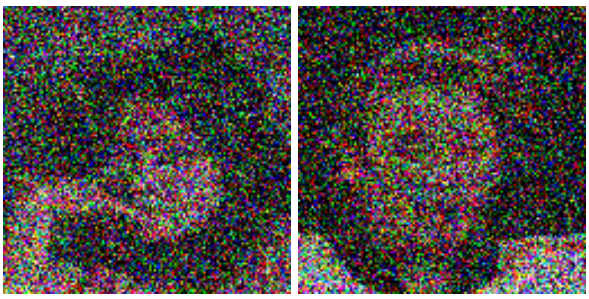


Figura 5.8: Esquema intuitivo de cómo se distribuyen las generaciones

Por último pasamos a mostrar las imágenes que hemos generado según el experimento en la figura 5.9. Los valores de ε y τ se especifican en cada imagen. Comentaremos la diferencia al variar estos parámetros.

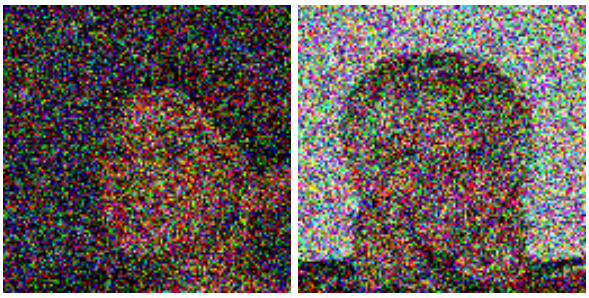
En primer lugar, comentar que los resultados son muy pobres y se podría decir que este método no funciona como se esperaría. Sin embargo, ya comentamos antes cuáles son los problemas que subyacen. Según la figura 5.9 se puede observar que son imágenes del *dataset* de caras con ruido. Ese ruido se interpreta como una desviación de la muestra original acorde con 5.8. Nada asegura que en la distribución de imágenes de caras una cara con ruido no pertenezca a esa misma distribución.

En segundo lugar, en cuanto al parámetro τ se aprecia que al disminuir este valor «nos quedamos a mitad de camino» y por eso la imagen es casi solo ruido. Si nos fijamos en ε, que recordemos que era el parámetro de regularización, al aumentar dicho valor el ruido es un poco mayor.



Imágenes con $\varepsilon = 0{,}1$ y $\tau = 0{,}6$.

Imágenes con $\varepsilon = 0{,}1$ y $\tau = 0{,}9$.



Imágenes con $\varepsilon = 1$ y $\tau = 0{,}6$.

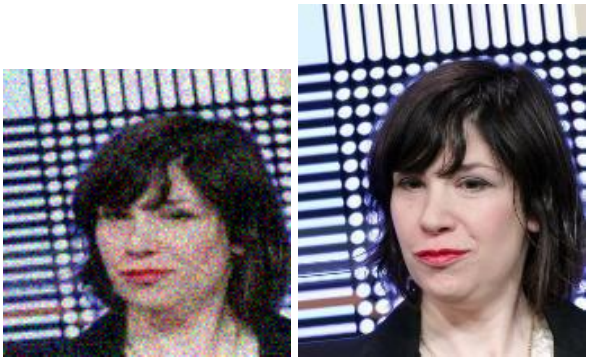
Imágenes con $\varepsilon = 1$ y $\tau = 0{,}9$.

Imágenes con $\varepsilon = 4$ y $\tau = 0{,}6$.

Imágenes con $\varepsilon = 4$ y $\tau = 0{,}9$.

Figura 5.9: Imágenes generadas del experimento con diferentes parámetros de ε y τ.

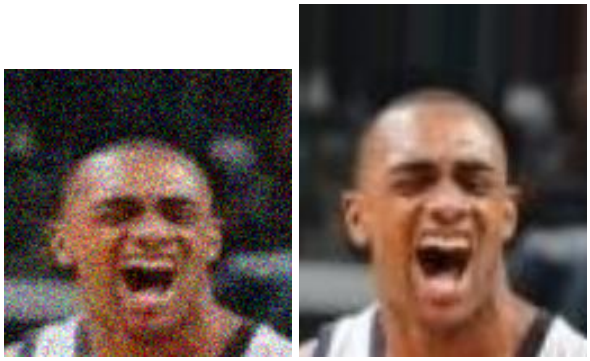
Ejemplos de las imágenes originales correspondientes a las imágenes de parámetro $\tau = 0{,}9$ se pueden observar en la figura 5.10. El único cambio perceptible es el ojo de la mujer generada con $\varepsilon = 4$.



Par $\varepsilon = 0{,}1$, $\tau = 0{,}9$, iteración 1

Par $\varepsilon = 0{,}1$, $\tau = 0{,}9$, iteración 2

Par $\varepsilon = 1$, $\tau = 0{,}9$, iteración 1



Par $\varepsilon = 1$, $\tau = 0{,}9$, iteración 2

Par $\varepsilon = 4$, $\tau = 0{,}9$, iteración 1

Par $\varepsilon = 4$, $\tau = 0{,}9$, iteración 2

Figura 5.10: Comparación directa entre imágenes generadas y sus correspondientes originales para $\tau = 0{,}9$ y distintos valores de ε.

Capítulo 6

Conclusiones y trabajo futuro

En cuanto a la parte teórica de esta memoria: hemos desarrollado el problema del *Puente de Schrödinger*, estableciendo sus vínculos con el transporte óptimo clásico y su versión regularizada: el transporte óptimo entrópico (capítulos 1,2 y 3).

Con respecto a la parte de aplicaciones: hemos estudiado el *Puente de Sinkhorn*, que surge como estimador al *Puente de Schrödinger*, y su comportamiento estadístico (capítulo 4). En el último capítulo hemos tratado de aplicar dicho estimador a la generación de imágenes. Aunque los resultados no hayan sido satisfactorios, el experimento nos ha permitido comprender mejor los problemas que surgen al tratar con el *Puente de Sinkhorn*: dimensionalidad y definición del *drift* estimado. Además, hemos expuesto los cambios en los resultados al variar los hiperparámetros de regularización y de longitud del camino.

En la actualidad existen métodos como *Generative Adversarial Networks (GANs)* [16] y *Diffusion Models* [18] que generan imágenes realmente sorprendentes, sin embargo, el coste computacional es muy elevado. Esta vía propuesta podría ser útil pero aun no está completamente desarrollada. Comentaremos en la siguiente sección una idea que puede solventar el problema de dimensionalidad.

6.1. Líneas futuras

En cuanto a las siguientes vías de investigación nos gustaría probar con un método para reducir la dimensión, para ello nos apoyaremos en un *VAE*. Sin entrar en detalles, explicaremos qué es esto. En el contexto de generación de imágenes, los *Variational Autoencoders* (*VAE*s) constituyen una familia de modelos generativos basados en arquitecturas de *autoencoder* probabilísticos. A diferencia de los autoencoders clásicos, los *VAE*s no aprenden una codificación determinista del dato, sino una distribución latente probabilística, lo que los convierte en herramientas clave para tareas de muestreo y generación.

El objetivo principal de un *VAE* es modelar una distribución de datos compleja (como la de imágenes naturales) a través de una distribución latente de dimensión reducida. Para ello, introduce dos componentes principales:

- *Encoder*: una red neuronal que aproxima la distribución posterior $q_\phi(z|x)$, mapeando la imagen de entrada x a una distribución (normalmente Gaussiana) sobre el espacio latente z.
- *Decoder*: otra red que toma una muestra $z \sim q_\phi(z|x)$ y reconstruye una imagen $\hat{x} \sim p_\theta(x|z)$.

La función de pérdida es:

$$\mathcal{L}(x;\theta,\phi) = \|x - \hat{x}\|_2^2 - \text{KL}(q_\phi(z|x) \,||\, p(z)) \tag{6.1}$$

1. El primer término es la pérdida de reconstrucción, que mide como de bien el *decoder* reproduce los datos originales. En este trabajo, se ha utilizado el error cuadrático medio (MSE).
2. La divergencia de Kullback-Leibler entre la distribución latente aprendida y un prior (normalmente una Gaussiana estándar), que regulariza el espacio latente.

En este trabajo, el *VAE* se emplea para:

- Aprender una representación latente suave y estructurada de las imágenes, ideal para reducir la dimensión del problema.
- Facilitar la interpolación entre imágenes de manera coherente, aprovechando que el espacio latente gaussiano permite trayectorias continuas.

Al hacer uso de este modelo pasamos de imágenes de tamaño $[\text{nro. de muestras}, 1, 128 \times 128 \times 3]$ a vectores de menor dimensión.

Una vez que pasemos las muestras de ruido y muestras de caras por el *encoder* hacemos el algoritmo 0 para transportar las distribuciones del ruido latentes a la distribución de las caras latentes. Para generar una nueva imagen: damos una nueva imagen de ruido gaussiano, la pasamos por el *encoder* y repetimos el razonamiento que hemos descrito en el experimento 1. Después de conseguir el ruido transportado, es decir, la nueva muestra, lo pasamos por el *decoder* y, con ello, conseguimos una nueva imagen.

Nos proponemos estudiar el algoritmo aquí descrito, tanto en el aspecto práctico de comprobar si genera imágenes realistas, diferentes de las usadas en el conjunto de entrenamiento, como a nivel teórico, para poder orientar posibles mejoras posteriores.

Apéndice A

Algunos resultados de análisis convexo

En este apéndice introduciremos algún resultado que será necesario en el capítulo 1 sobre funciones convexas.

Definición A.1. Una función convexa φ en $\mathbb{R}^d$ es una función $\varphi : \mathbb{R}^d \longrightarrow \mathbb{R} \cup \{+\infty\}$, no idénticamente ∞, tal que para todo $x, y \in \mathbb{R}^d$

$$\forall t \in [0,1], \quad \varphi(tx + (1-t)y) \leq t\varphi(x) + (1-t)\varphi(y).$$

Se dice que es estrictamente convexa si se da la desigualdad en sentido estricto.

1. Se puede probar que una función convexa φ es automáticamente continua y localmente Lipschitz en el mayor abierto incluido en el dominio de φ. Por tanto, en virtud del teorema de Rademacher [11, subsección 3.1.2], una función convexa es diferenciable en casi todo punto de dicho abierto.

2. El grafo de φ está por encima de su tangente, es decir, para todo x donde φ es diferenciable, se tiene
$$\forall z \in \mathbb{R}^d, \quad \varphi(z) \geq \varphi(x) + \nabla\varphi(x) \cdot (z - x),$$
que expresa el hecho geométrico de que todo el grafo de φ está por encima de el hiperplano tangente en el punto x. En particular $\nabla\varphi$ es monótona.

3. **Definición A.2.** Se define la subdiferencial de una función convexa como el siguiente conjunto
$$y \in \partial\varphi(x) \iff \left\{\forall x \in \mathbb{R}^d, \varphi(z) \geq \varphi(x) + \langle y, z - x\rangle\right\}.$$
Identificaremos la aplicación subdiferencial con su grafo.

4. Es consecuencia inmediata de la definición que la subdiferencial de una función convexa φ es una aplicación monótona: para todo $y_1 \in \partial\varphi(x_1), y_2 \in \partial\varphi(x_2)$,
$$\langle y_2 - y_1, x_2 - x_1\rangle \geq 0.$$

5. Para cualquier función $\varphi : \mathbb{R}^d \longrightarrow \mathbb{R}$, se puede definir la función convexa conjugada, o la transformada de Legendre, por
$$\varphi^*(y) = \sup_x (xy - \varphi(x)).$$

$\varphi^*(y)$ es una función convexa inferiormente semicontinua.

Proposición A.3. *Sea φ una función convexa inferiormente semicontinua en $\mathbb{R}^d$. Entonces, para todo $x, y \in \mathbb{R}^d$,*

$$x \cdot y = \varphi(x) + \varphi^*(y) \iff y \in \partial\varphi(x) \iff x \in \partial\varphi^*(y).$$

Demostración.

$$\begin{aligned} xy = \varphi(x) + \varphi^*(y) &\iff xy \geq \varphi(x) + \varphi^*(y) \\ &\iff \forall z \in \mathbb{R}^d, \quad xy \geq \varphi(x) + yz - \varphi(z) \\ &\iff \forall z \in \mathbb{R}^d, \quad \varphi(z) \geq \varphi(x) + \langle y, z - x \rangle \\ &\iff y \in \partial\varphi(x). \end{aligned}$$

Por simetría, y la siguiente propiedad se deduce la equivalencia con $x \in \partial\varphi^*(y)$. □

6. La transformada de Legendre induce una dualidad entre funciones convexas. Para ello introducimos el siguiente resultado.

Proposición A.4. *La transformada de Legendre se define como $\varphi^*(y) = \sup_x(xy - \varphi(x))$. Además se tienen las siguientes equivalencias:*

a) *φ^* es convexa e inferiormente semicontinua (de hecho continua y localmente Lipchitz en $Int(Dom(f))$).*

b) *$\varphi = \psi^*$ para alguna función ψ.*

c) *$\varphi^{**} = \varphi$.*

Apéndice B

Procesos estocásticos

En este apéndice vamos a ver los resultados con respecto a procesos estocásticos que sean necesarios en este trabajo. Para ello seguiremos principalmente [30].

Definición B.1. Un proceso estocástico es una familia parametrizada de variables aleatorias $\{X_t\}_{t\in T}$ definido en un espacio de probabilidad $(\Omega, \mathcal{F}, P)$.

El espacio T será normalmente $[0, \infty)$ o un intervalo $[a, b]$. Notemos que si fijamos $t \in T$ tenemos una variable aleatoria:

$$\omega \longrightarrow X_t(\omega), \quad \omega \in \Omega.$$

Y, si fijamos $\omega \in \Omega$ podemos considerar la función

$$t \longrightarrow X_t(\omega), \quad t \in T,$$

que se llamará *camino* de X_t. De manera intuitiva es conveniente interpretar t como el *tiempo* y ω como una *partícula* o *experimento* individual. Bajo estas premisas $X_t(\omega)$ representa la posición a tiempo t de la partícula ω.

A veces resulta conveniente escribir $X(t, \omega)$ en lugar de $X_t(\omega)$. De este modo, también se puede considerar el proceso como una función de dos variables:

$$(t, \omega) \to X(t, \omega)$$

donde $t \in T$, $\omega \in \Omega$, y $X(t, \omega) \in \mathbb{R}^n$. Esta es una perspectiva natural en el análisis estocástico, ya que (como veremos) es crucial que $X(t, \omega)$ sea medible conjuntamente en (t, ω).

Finalmente, observamos que se puede identificar cada ω con la función $t \mapsto X_t(\omega)$ de T en $\mathbb{R}^n$. Así, podemos considerar Ω como un subconjunto del espacio $\widetilde{\Omega} = (\mathbb{R}^n)^T$, es decir, el conjunto de todas las funciones de T en $\mathbb{R}^n$. Entonces, la σ-álgebra $\mathcal{F}$ contendrá a la σ-álgebra $\mathcal{B}$ generada por los conjuntos de la forma:

$$\{\omega \colon \omega(t_1) \in F_1, \ldots, \omega(t_k) \in F_k\}, \quad \text{con } F_i \subset \mathbb{R}^n \text{ Borel.}$$

Por lo tanto, también se puede adoptar la perspectiva de que un proceso estocástico es una medida de probabilidad P definida sobre el espacio medible $((\mathbb{R}^n)^T, \mathcal{B})$.

Las distribuciones finito dimensionales del proceso $X = \{X_t\}_{t\in T}$ son las medidas $\mu_{t_1,\dots,t_k}$, definidas en $\mathbb{R}^{nk}$, para $k = 1, 2, \dots$, mediante:

$$\mu_{t_1,\dots,t_k}(F_1 \times F_2 \times \cdots \times F_k) = P\left[X_{t_1} \in F_1, \dots, X_{t_k} \in F_k\right], \quad t_i \in T.$$

Aquí, $F_1, \dots, F_k$ son conjuntos Borel de $\mathbb{R}^n$.

Dos procesos tienen la misma distribución si y sólo si tienen las mismas distribuciones finito dimensionales. De hecho, dada una familia $\{\nu_{t_1,\dots,t_k} : k \in \mathbb{N}, t_i \in T\}$ de medidas de probabilidad sobre $\mathbb{R}^{nk}$, es importante poder construir un proceso estocástico $Y = \{Y_t\}_{t\in T}$ que tenga $\nu_{t_1,\dots,t_k}$ como sus distribuciones finito dimensionales. Uno de los teoremas famosos de Kolmogórov afirma que esto puede hacerse, siempre que la familia $\{\nu_{t_1,\dots,t_k}\}$ satisfaga dos condiciones de consistencia natural.

Junto con los procesos de interés tendremos una filtración: una colección creciente de σ-álgebras. $\{\mathcal{F}_t\}_{t\in T}$. El sentido de la filtración es la de modelar la información disponible hasta un instante dado. Se dice que un proceso $\{X_t\}_{t\in T}$ es adaptado a la filtración $\{\mathcal{F}_t\}_{t\in T}$ si X_t es $\mathcal{F}_t$-medible para todo $t \in T$. Un proceso siempre es adaptado respecto a la filtración natural: $\mathcal{F}_t^X = \sigma(X_s : s \leq t)$, pero puede ser más interesante considerar otras filtraciones.

Por otro lado, una martingala respecto de la filtración $\{\mathcal{F}_t\}_{t\in T}$ es un proceso, $\{M_t\}_{t\in T}$ adaptado a $\{\mathcal{F}_t\}_{t\in T}$, tal que

$$\mathbb{E}[|M_t|] < \infty \quad \text{y} \quad \mathbb{E}[M_t|\mathcal{F}_s] = M_s \text{ c.s.}$$

para $s \leq t \in T$.

B.1. El Movimiento Browniano

El Movimiento Browniano (o proceso de Wiener) es el ejemplo clásico de proceso estocástico, con un papel entre los procesos estocásticos similar al de la distribución normal entre las leyes de variables aleatorias. Se dice que un proceso $\{\mathrm{W}_t\}_{t\geq 0}$ con trayectorias continuas es un Movimiento Browniano estándar si

1. $\mathrm{W}_0 = 0$, c.s. ,
2. W_t tiene incrementos independientes: si $s < t$, $\mathrm{W}_t - \mathrm{W}_s$ es independiente de $\mathcal{F}_s$,
3. si $s < t$, $\mathrm{W}_t - \mathrm{W}_s$ tiene distribución normal, centrada, con varianza $t - s$.

De la segunda y tercera propiedad se deduce que si $t_1 < t_2 \dots < t_n$ entonces $\mathrm{W}_{t_1}, \dots \mathrm{W}_{t_n} - \mathrm{W}_{t_{n-1}}$ son variables aleatorias normales independientes, por lo que $(\mathrm{W}_{t_1}, \dots \mathrm{W}_{t_{n-1}}, \mathrm{W}_{t_n})$ tienen distribución normal. En otras palabras, $\{\mathrm{W}_t\}_{t\geq 0}$ es un proceso con trayectorias continuas, Gaussiano, centrado, con función de covarianza $K(s,t) = s \wedge t$.

Análogamente se define para el caso del Movimiento Browniano d-dimensional.

Por otro lado definimos la medida de Wiener. Sea $C_0([0,T];\mathbb{R})$ el espacio de funciones continuas $\omega : [0,T] \to \mathbb{R}$ tales que $\omega(0) = 0$. Este espacio está equipado con la topología de la convergencia uniforme, y su σ-álgebra de Borel correspondiente se denota por $\mathcal{B}$.

La **medida de Wiener** es una medida de probabilidad en $(C_0([0,T];\mathbb{R}),\mathcal{B})$ tal que el proceso canónico

$$W_t(\omega) := \omega(t), \quad \omega \in C_0([0,T];\mathbb{R}),$$

es un *proceso de Wiener estándar*, es decir, cumple:

1. $W_0 = 0$ casi seguro.
2. Los incrementos $W_t - W_s$ son independientes de $\{W_r : r \leq s\}$ para $0 \leq s < t \leq T$.
3. Para $0 \leq s < t \leq T$, el incremento $W_t - W_s \sim \mathcal{N}(0, t-s)$, distribución normal con media cero y varianza $t-s$.
4. Las trayectorias de W_t son continuas casi seguramente.

La medida de Wiener describe por tanto la distribución de probabilidad sobre el espacio de trayectorias continuas que corresponden a un Movimiento Browniano estándar. Una prueba de la existencia de la medida de Wiener, así como un desarrollo detallado de conceptos relacionados puede encontrarse en [4] o [22].

B.2. Fórmula de Itô

Sea W_t un Movimiento Browniano en $(\Omega, \mathcal{F}, P)$. Bajo cierta regularidad, un proceso de Itô es un proceso estocástico X_t de la forma

$$X_t = X_0 + \int_0^t u(s,\omega)\,\mathrm{d}s + \int_0^t v(s,\omega)\,\mathrm{dW}_s.$$

Teorema B.2. *Sea X_t un proceso de Itô dado por*

$$\mathrm{d}X_t = u\,\mathrm{d}t + v\,\mathrm{d}W_t.$$

Sea $g(t,x) \in \mathcal{C}([0,\infty)\times\mathbb{R})$. Entonces

$$Y_t = g(t, X_t),$$

es de nuevo un proceso de Itô, y

$$\mathrm{d}Y_t = \frac{\partial g}{\partial t}(t,X_t)\,\mathrm{d}t + \frac{\partial g}{\partial x}(t,X_t)\,\mathrm{d}X_t + \frac{1}{2}\frac{\partial^2 g}{\partial x^2}(t,X_t)\cdot(\mathrm{d}X_t)^2,$$

donde $(\mathrm{d}X_t)^2 = (\mathrm{d}X_t)\cdot(\mathrm{d}X_t)$ se obtiene a partir de las reglas

$$\mathrm{d}t\cdot\mathrm{d}t = \mathrm{d}t\cdot\mathrm{d}W_t = \mathrm{d}W_t\cdot\mathrm{d}t = 0, \quad \mathrm{d}W_t\cdot\mathrm{d}W_t = \mathrm{d}t.$$

B.3. Ecuaciones diferenciales estocásticas

Nos planteamos ahora buscar las posibles soluciones $X_t(\omega)$ de la ecuación diferencial estocástica

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = b(t,X_t) + \sigma(t,X_t)\mathrm{W}_t, \quad b(t,x)\in\mathbb{R}, \sigma(t,x)\in\mathbb{R}.$$

En virtud de la interpretación de Itô X_t satisface la ecuación integral estocástica

$$X_t = X_0 + \int_0^t b(s, X_s)\,\mathrm{d}s + \int_0^t \sigma(s, X_s)\,\mathrm{d}\mathbf{W}_s,$$

o en forma diferencial

$$\mathrm{d}X_t = b(t, X_t)\,\mathrm{d}t + \sigma(t, X_t)\,\mathrm{d}\mathbf{W}_t.$$

Se llamará *drift* (o coeficiente deriva) a b y *coeficiente de difusión* a σ.

Se puede asegurar existencia y unicidad de soluciones

Teorema B.3. *Si existe una constante $K > 0$ tal que*

$$\|b(t,x) + b(t,y)\| + \|\sigma(t,x) + \sigma(t,y)\| \leq K\|x - y\|,$$
$$\|b(t,x)\|^2 + \|\sigma(t,x)\|^2 \leq K(1 + \|x\|^2),$$

y ξ es un vector aleatorio d-dimensional, independiente de B, con $\mathbb{E}\|\xi\|^2 < \infty$ entonces existe una única solución con condición inicial $X(0) = \xi$.

B.3.1. Ecuaciones lineales

Consideramos la ecuación

$$\mathrm{d}X(t) = [A(t)X(t) + a(t)]\,\mathrm{d}t + \sigma(t)\,\mathrm{d}\mathbf{W}(t), \quad t \geq 0, \tag{B.1}$$

con condición inicial $X(0) = \xi$. Aquí B es un Movimiento Browniano r-dimensional, $A(t)$ es una matriz $d \times d$, $a(t)$ un vector d-dimensional y $\sigma(t)$ una matriz $d \times r$. Asumimos que $A(t)$, $a(t)$, y $\sigma(t)$ son no aleatorias, medibles y localmente acotadas.

La ecuación determinista asociada es

$$\dot{\xi}(t) = A(t)\xi(t) + a(t), \quad t \geq 0. \tag{B.2}$$

Bajo ciertas condiciones estándar (que no detallaremos aquí) sobre A, la ecuación (matricial) homogénea asociada

$$\dot{\Phi}(t) = A(t)\Phi(t), \quad t \geq 0, \tag{B.3}$$

con condición inicial $\Phi(0) = I_d$ (la matriz identidad de dimensión d) tiene una única solución, que denotaremos $\Phi(t)$. Esta es la *solución fundamental* de la ecuación homogénea. Generalmente, bajo las condiciones estándar, se tiene que la matriz $\Phi(t)$ es no singular. Entonces, la solución a con condición inicial $\xi(0) = \xi$ es:

$$\xi(t) = \Phi(t)\left[\xi + \int_0^t \Phi^{-1}(s)a(s)\,\mathrm{d}s\right].$$

Usando la fórmula de integración por partes, se puede comprobar que las ecuaciones del tipo tienen una solución que se expresa de forma explícita en términos de Φ.

Teorema B.4. *Con la notación y condiciones anteriores, la solución de (4.12) con condición inicial $X(0) = \xi$ es*

$$X(t) = \Phi(t)\left[\xi + \int_0^t \Phi^{-1}(s)a(s)\,\mathrm{d}s + \int_0^t \Phi^{-1}(s)\sigma(s)\,\mathrm{d}W(s)\right], \quad t \geq 0.$$

Si $A(t) \equiv A$ con autovalores con parte real negativa, $\sigma(t) \equiv \sigma$ y $a(t) \equiv 0$, entonces la solución con condición inicial $X(0) = \xi \sim \mathcal{N}(0, V)$, independiente de W y $V = \int_0^\infty e^{sA}\sigma\sigma^T e^{sA^T} ds$, es un proceso Gaussiano, centrado con covarianza

$$\rho(s,t) = \begin{cases} e^{(s-t)A}V & \text{si } 0 \le t \le s \\ Ve^{(s-t)A^T} & \text{si } 0 \le s \le t \end{cases}$$

Ecuación de Langevin. Proceso de Ornstein-Uhlenbeck

Si en lo anterior tomamos $d = r = 1$, $a(t) \equiv 0$, $A(t) = -\alpha < 0$ y $\sigma(t) = \sigma > 0$ obtenemos la ecuación de **Langevin** para el Movimiento Browniano de una partícula con fricción:

$$\mathrm{d}X(t) = -\alpha X(t)\,\mathrm{d}t + \sigma\,\mathrm{d}\mathbf{W}(t).$$

Por el Teorema B.4 la solución de esta ecuación es

$$X(t) = X(0)e^{-\alpha t} + \sigma \int_0^t e^{-\alpha(t-s)}\,\mathrm{d}\mathbf{W}(s), \quad t \ge 0.$$

Si $\mathbb{E}(X^2(0)) < \infty$ entonces

$$m(t) = m(0)e^{-\alpha t} \quad \text{y} \quad \rho(s,t) = \left[V(0) + \frac{\sigma^2}{2\alpha}(e^{2\alpha(s\wedge t)} - 1)\right] e^{-\alpha(t+s)}.$$

Si $X(0) \sim \mathcal{N}(0, \frac{\sigma^2}{2\alpha})$ entonces $X(t)$ es un proceso Gaussiano centrado estacionario con función de covarianza

$$\rho(s,t) = \frac{\sigma^2}{2\alpha} e^{-\alpha|s-t|}.$$

A este proceso se le conoce como *proceso de Ornstein-Uhlenbeck*.

Puente Browniano

Como ejemplo particular, y que se utilizará en la memoria frecuentemente, consideremos la ecuación diferencial estocástica:

$$\mathrm{d}X(t) = \frac{b - X(t)}{T - t}\,\mathrm{d}t + \,\mathrm{d}\mathbf{W}(t), \quad 0 \le t \le T$$

con condición inicial $X(0) = a$. En este caso se tiene:

$$A(t) = -\frac{1}{T-t}, \quad a(t) = \frac{b}{T-t}, \quad \sigma(t) = 1.$$

Con un cálculo simple se ve que $\Phi(t) = 1 - \frac{t}{T}$. El argumento del teorema anterior se puede aplicar en cualquier intervalo $[0, \tilde{T}]$ con $\tilde{T} \in (0, T)$. Obtenemos entonces que:

$$X(t) = a\left(1 - \frac{t}{T}\right) + b\left(\frac{t}{T}\right) + (T-t)\int_0^t \frac{\mathrm{d}\mathbf{W}(s)}{T-s}$$

es solución en $[0, \tilde{T}]$. Observemos que $(T-t)\int_0^t \frac{\mathrm{d}\mathbf{W}(s)}{T-s}$ es una variable normal centrada con varianza $\frac{t}{T}(T-t)$, de forma que $X(t) \to b$ en probabilidad cuando $t \to T$.

Se puede probar que la convergencia es en realidad c.s. (casi segura), y podemos considerar que X está definido y es continuo en $[0,T]$. El proceso X satisface $X(0) = a$, $X(T) = b$, y es Gaussiano, centrado, con trayectorias continuas y covarianza:

$$\rho(s,t) = s \wedge t - \frac{st}{T}.$$

A tal proceso se le llama *Puente Browniano de a a b en* $[0,T]$. En el caso $a = b = 0$ y $T = 1$, se dice simplemente *Puente Browniano*.

B.3.2. Ecuación de Fokker-Planck

Una vez que se obtiene la solución explícita de la ecuación estocástica, es natural preguntarse cómo evoluciona la densidad de probabilidad del proceso $X(t)$ a lo largo del tiempo. Esta evolución está gobernada por la *ecuación de Fokker–Planck*, también conocida como ecuación de Kolmogórov hacia adelante.

Consideremos la ecuación estocástica general:

$$\mathrm{d}X(t) = b(t, X(t))\ \mathrm{d}t + \sigma(t, X(t))\ \mathrm{dW}(t),$$

donde $X(t) \in \mathbb{R}^d$, $b : \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}^d$ es el coeficiente de deriva o *drift*, y $\sigma : \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}^{d\times r}$ es la matriz de difusión.

Entonces, la densidad de probabilidad $p(x,t)$ de $X(t)$ (cuando existe) satisface la ecuación de Fokker–Planck:

$$\frac{\partial p}{\partial t} = -\sum_{i=1}^{d} \frac{\partial}{\partial x_i}\left[b_i(x,t)p(x,t)\right] + \frac{1}{2}\sum_{i,j=1}^{d} \frac{\partial^2}{\partial x_i \partial x_j}\left[\left(D(x,t)\right)_{ij} p(x,t)\right],$$

donde $D(x,t) = \sigma(x,t)\sigma(x,t)^\top$ es la matriz de covarianza de la difusión.

Esta ecuación proporciona una descripción determinista de cómo se distribuye la masa de probabilidad en el espacio de estados a medida que avanza el tiempo.

B.3.3. Ecuación de Fokker–Planck para el Puente Browniano

Consideremos un *Puente Browniano* en el intervalo $[0,1]$, que parte de un valor inicial x_0 y termina en x_1. Este proceso puede describirse mediante la ecuación estocástica:

$$\mathrm{d}X(t) = (x_1 - X(t))\frac{1}{1-t}\ \mathrm{d}t + \ \mathrm{dW}(t), \qquad X(0) = x_0, \quad 0 \le t < 1.$$

El coeficiente de deriva es $b(x,t) = \frac{x_1 - x}{1-t}$, y el coeficiente de difusión es constante $\sigma(x,t) = 1$.

Entonces, la densidad de probabilidad $p(x,t)$ de $X(t)$ satisface la ecuación de Fokker–Planck:

$$\frac{\partial p}{\partial t} = -\frac{\partial}{\partial x}\left[\frac{x_1 - x}{1-t}\, p(x,t)\right] + \frac{1}{2}\frac{\partial^2 p}{\partial x^2}.$$

Esta ecuación describe la evolución temporal de la distribución de probabilidad del *Puente Browniano*, y refleja el hecho de que, conforme $t \to 1$, la densidad se concentra en torno al valor final x_1.

B.3.4. Marginal con respecto al tiempo de un Puente Browniano

Definimos el proceso desplazado:

$$R_t = x_0 + \mathrm{W}_t.$$

Entonces:

$$R_0 = x_0 + \mathrm{W}_0 = x_0, \quad R_1 = x_0 + \mathrm{W}_1.$$

Condicionamos ahora a que $R_1 = x_1$, es decir:

$$\mathrm{W}_1 = x_1 - x_0.$$

Sabemos que el vector aleatorio $(\mathrm{W}_t, \mathrm{W}_1)$ es normal. Por la fórmula de la normal condicional, se tiene:

$$\mathcal{L}(\mathrm{W}_t \mid (\mathrm{W}_1 = b)) \sim \mathcal{N}\left(\frac{\mathrm{Cov}(\mathrm{W}_t, \mathrm{W}_1)}{\mathrm{Var}(\mathrm{W}_1)} b,\ \mathrm{Var}(\mathrm{W}_t) - \frac{\mathrm{Cov}(\mathrm{W}_t, \mathrm{W}_1)^2}{\mathrm{Var}(\mathrm{W}_1)}\right).$$

Sustituyendo:

$$\mathbb{E}[\mathrm{W}_t \mid \mathrm{W}_1 = b] = \frac{\epsilon t}{\epsilon} b = tb,$$
$$\mathrm{Var}(\mathrm{W}_t \mid \mathrm{W}_1) = \epsilon t(1-t)\mathrm{Id}.$$

Aplicando esto a nuestro caso $b = x_1 - x_0$, obtenemos:

$$\mathcal{L}(\mathrm{W}_t \mid (\mathrm{W}_1 = x_1 - x_0)) \sim \mathcal{N}\left(t(x_1 - x_0),\ \epsilon t(1-t)\mathrm{Id}\right).$$

Recordamos que:

$$R_t = x_0 + \mathrm{W}_t \Rightarrow \mathcal{L}(R_t \mid (x_0, x_1)) \sim \mathcal{N}\left(x_0 + t(x_1 - x_0),\ \epsilon t(1-t)\mathrm{Id}\right).$$

La media puede reescribirse como:

$$x_0 + t(x_1 - x_0) = (1-t)x_0 + tx_1.$$

Por lo tanto:

$$\mathcal{L}(R_t \mid (x_0, x_1)) \sim \mathcal{N}\left((1-t)x_0 + tx_1,\ \epsilon t(1-t)\mathrm{Id}\right).$$

B.3.5. Propiedad de Markov para procesos de Itô

El sentido de esta subsección es presentar que la propiedad de Markov se cumple para los procesos de difusión, en particular, para el *Puente Browniano*, que es lo que se utilizará.

Definición B.5. Se dice que un proceso es de Markov si su distribución en tiempos futuros, conocido el valor actual, es independiente del tiempo anterior. Más formalmente, el proceso $\{X_t, \mathcal{F}_t\}_{t\in T}$ es de Markov si para cualquier función medible y acotada y para $s, t \geq 0$ se tiene que

$$\mathbb{E}(f(X_{t+s})|\mathcal{F}_s) = \mathbb{E}(f(X_{t+s})|X_s).$$

El siguiente teorema es [30, Teorema 7.1.2]

Teorema B.6. *Una difusión de Itô es un proceso de Markov.*

Nota. En particular, el *Puente Browniano* es un proceso de Markov.

B.3.6. Otros teoremas importantes

Teorema B.7 (Teorema de Girsanov)**.** *Sea $Y(t) \in \mathbb{R}^n$ un proceso de Itô de la forma*

$$\mathrm{d}Y(t) = a(t,\omega)\,\mathrm{d}t + \mathrm{d}W(t); \quad t \leq T, \quad Y_0 = 0,$$

donde $T \leq \infty$ es una constante dada y $W(t)$ es un Movimiento Browniano d-dimensional.

Definimos

$$M_t = \exp\left(-\int_0^t a(s,\omega)\,\mathrm{d}W_s - \frac{1}{2}\int_0^t a^2(s,\omega)\,\mathrm{d}s\right); \quad 0 \leq t \leq T.$$

Supongamos que M_t es una martingala con respecto a $\mathcal{F}_t^{(n)}$ y P. Definimos la medida Q sobre $\mathcal{F}_T^{(n)}$ por

$$dQ(\omega) = M_T(\omega)\,\mathrm{d}p(\omega).$$

Entonces Q es una medida de probabilidad sobre $\mathcal{F}_T^{(n)}$ y $Y(t)$ es un Movimiento Browniano n-dimensional con respecto a Q, para $0 \leq t \leq T$.

Nota.

1. La transformación $P \to Q$ se llama la *transformación de medidas de Girsanov*.
2. La siguiente *condición de Novikov*, la del enunciado, es suficiente para garantizar que $\{M_t\}_{t\leq T}$ sea una martingala (con respecto a $\mathcal{F}_t^{(n)}$ y P):

   $$\mathbb{E}\left[\exp\left(\frac{1}{2}\int_0^T a^2(s,\omega)\,\mathrm{d}s\right)\right] < \infty$$

   donde $\mathbb{E} = \mathbb{E}_P$ es la esperanza respecto a P.
3. Nótese que, como M_t es una martingala, de hecho tenemos que

   $$M_T\,\mathrm{d}p = M_t\,\mathrm{d}p \quad \text{sobre } \mathcal{F}_t^{(n)}, \quad t \leq T.$$

B.4. Variación total

Esta mide la variabilidad entre las distribuciones P y Q en términos de la variación total, con un enfoque particular en las diferencias grandes debido al cuadrado. La variación total entre dos distribuciones P y Q se define como:

$$\mathrm{TV}(P,Q) = \frac{1}{2}\sum_{\omega\in\Omega} |P(\omega) - Q(\omega)|$$

en el caso discreto, o bien:

$$\mathrm{TV}(P,Q) = \frac{1}{2}\int |p(x) - q(x)|dx$$

en el caso continuo, donde $p(x)$ y $q(x)$ son las funciones de densidad de P y Q, respectivamente. La esperanza se refiere al valor promedio de $\mathrm{TV}^2(P,Q)$, y depende de la distribución sobre P y Q. Este valor refleja la variabilidad promedio entre las distribuciones, amplificando las diferencias más grandes.

Tiene las siguientes propiedades:

1. **No negatividad**:
$$\text{TV}(P, Q) \geq 0$$

2. **Simetría**:
$$\text{TV}(P, Q) = \text{TV}(Q, P)$$

3. **Identidad de la variación total**:
$$\text{TV}(P, P) = 0$$
La variación total entre una distribución y ella misma es cero.

4. **Desigualdad triangular**:
$$\text{TV}(P, Q) \leq \text{TV}(P, R) + \text{TV}(R, Q)$$

5. **Invarianza bajo transformaciones de medida**: Si P y Q son distribuciones de probabilidad sobre Ω y $f : \Omega \to \mathbb{R}$ es una función medible, entonces:
$$\text{TV}(P \circ f^{-1}, Q \circ f^{-1}) = \text{TV}(P, Q)$$

6. **Desigualdad de Pinsker**: Establece la siguiente relación entre la variación total $\text{TV}(P, Q)$ y la divergencia de Kullback-Leibler $\text{KL}(P|Q)$:
$$\text{TV}(P, Q) \leq \sqrt{\frac{1}{2} \text{KL}(P|Q)}$$

7. **Propiedad subaditiva**: Si P y Q son distribuciones sobre espacios de probabilidad Ω_1 y Ω_2, entonces la variación total de la distribución conjunta de P y Q sobre $\Omega_1 \times \Omega_2$ es menor o igual que la suma de las variaciones totales sobre cada espacio individual:
$$\text{TV}(P \times Q, R \times S) \leq \text{TV}(P, R) + \text{TV}(Q, S)$$

Apéndice C

Lemas técnicos

Lema C.1. *La función*

$$\varphi^\star(x) = \epsilon \, \log\left(\int_{\mathbb{R}^d} \exp\left(\frac{x\cdot y - \psi^\star(y)}{\epsilon}\right) \, \mathrm{d}\nu(y)\right).$$

es $\mathcal{C}^\infty$.

Demostración. En primer lugar notemos que tenemos una composición de logaritmo con otra función positiva (una integral de una exponencial es mayor que 0). En virtud de la regla de la cadena $\varphi^\star(x)$ será $\mathcal{C}^\infty$ si lo es cada función por separado. En cuanto al logaritmo por lo comentado arriba no hay más que decir, entonces nos centraremos en la integral.

Para ello se utilizarán argumentos similares a prueba de que la función generadora de momentos es $\mathcal{C}^\infty$. En virtud del Teorema de dualidad de Kantorovich y que μ y ν tienen momento de orden 2 finito se tiene que $\varphi^\star(x) < \infty$ y, por tanto,

$$\int_{\mathbb{R}^d} \exp\left(\frac{x\cdot y - \psi^\star(y)}{\epsilon}\right) \, \mathrm{d}\nu(y) < \infty.$$

Podemos escribir

$$\int_{\mathbb{R}^d} \exp\left(\frac{x\cdot y - \psi^\star(y)}{\epsilon}\right) \, \mathrm{d}\nu(y) = \int_{\mathbb{R}^d} \exp\left(\frac{x\cdot y}{\epsilon}\right) \exp\left(\frac{-\psi^\star(y)}{\epsilon}\right) \, \mathrm{d}\nu(y),$$

y definiendo $\frac{\mathrm{d}\rho}{\mathrm{d}\nu}(y) := \exp\left(\frac{-\psi^\star(y)}{\epsilon}\right)$ se tiene que la integral es

$$\int_{\mathbb{R}^d} \exp\left(\frac{x\cdot y}{\epsilon}\right) \, \mathrm{d}\rho(y).$$

Definimos la función

$$M_\varepsilon(x) := \int_{\mathbb{R}} e^{\frac{xy}{\varepsilon}} \, d\rho(y),$$

la cual está bien definida para todo $x \in (-\delta\varepsilon, \delta\varepsilon)$. En efecto, para tales valores de x se cumple que

$$\left|\frac{xy}{\varepsilon}\right| \leq \delta|y| \quad \text{y por tanto} \quad \left|e^{\frac{xy}{\varepsilon}}\right| \leq e^{\delta|y|}.$$

Dado que $e^{\delta|y|} \in L^1(\rho)$, se concluye que $M_\varepsilon(x)$ está bien definida.

Además, para cada $n \in \mathbb{N}$, se verifica que la derivada de orden n está dada por

$$\frac{\mathrm{d}^n}{\mathrm{d}x^n} M_\varepsilon(x) = \int_{\mathbb{R}} \frac{y^n}{\varepsilon^n} e^{\frac{xy}{\varepsilon}} \, \mathrm{d}\rho(y),$$

y esta expresión es válida por el teorema de derivación bajo el signo integral, ya que

$$\left| \frac{y^n}{\varepsilon^n} e^{\frac{xy}{\varepsilon}} \right| \leq \frac{1}{\varepsilon^n} |y|^n e^{\delta|y|},$$

y $|y|^n e^{\delta|y|} \in L^1(\rho)$ por hipótesis. En consecuencia, la función M_ε pertenece es de clase C^∞, y sus derivadas están dadas por

$$M_\varepsilon^{(n)}(x) = \frac{1}{\varepsilon^n} \int_{\mathbb{R}} y^n e^{\frac{xy}{\varepsilon}} \, \mathrm{d}\rho(y).$$

□

Lema C.2 (Push-forward en términos de la subdiferencial)**.** *Sea φ una función convexa y μ una medida de probabilidad en $\mathbb{R}^d$, absolutamente continua con respecto a la medida de Lebesgue. Entonces para todo conjunto de Borel $A \subset \mathbb{R}^d$,*

$$\nabla\varphi \# \mu[A] = \mu[\partial\varphi^*(A)].$$

Demostración. 1. Es suficiente probar que para cualquier conjunto medible A, se tiene

$$\mu[(\nabla\varphi)^{-1}(A)] = \mu[\partial\varphi^*(A)].$$

A partir de propiedades generales de funciones convexas (véase el apéndice A), $\nabla\varphi(x) = y \Rightarrow x \in \partial\varphi^*(y)$; por lo tanto,

$$(\nabla\varphi)^{-1}(A) \subset \partial\varphi^*(A).$$

Dado que μ es absolutamente continua, basta con verificar que el conjunto

$$Z = \partial\varphi^*(A) \setminus (\nabla\varphi)^{-1}(A)$$

tiene medida de Lebesgue cero.

2. Sea $z \in \partial\varphi^*(A)$; entonces existe $x \in A$ tal que $z \in \partial\varphi^*(x)$, lo cual implica que $x \in \partial\varphi(z)$. Si z es un punto de diferenciabilidad de φ, entonces necesariamente $\nabla\varphi(z) = x \in A$, por lo que $z \in (\nabla\varphi)^{-1}(A)$. Por lo tanto, Z está contenido en el conjunto de puntos donde φ no es diferenciable, así que tiene medida de Lebesgue cero. □

C.1. Lemas capítulo 4

Lema C.3. *En las condiciones del teorema de 4.1 se tiene que*

$$\mathbb{E}[\, KL(\bar{\mathsf{P}}_{[0,\tau]} | \mathsf{P}^\star_{[0,\tau]})] \leq \frac{R^2 \varepsilon^{-\mathsf{k}}}{n} (1-\tau)^{-\mathsf{k}-2}.$$

Demostración. Para la prueba de este resultado nos basamos en [35], comentaremos donde lo utilizaremos.

Empezamos aplicando el teorema de Girsanov (ver apéndice B) para obtener una diferencia de los coeficientes derivas, los cuales pueden reescribirse en forma de las *aplicaciones entrópicas de Brenier*

$$\begin{aligned}\mathbb{E}[\,\mathrm{KL}(\bar{\mathsf{P}}_{[0,\tau]}|\mathsf{P}^{\star}_{[0,\tau]})] &\lesssim \int_0^{\tau} \mathbb{E}\|\bar{b}_t - b_t^{\star}\|^2_{L^2(\mathsf{p}_t)}\,\mathrm{d}t \\ &= \int_0^{\tau} \frac{1}{(1-t)^2}\mathbb{E}\|\nabla\bar{\varphi}_{1-t} - \nabla\varphi^{\star}_{1-t}\|^2_{L^2(\mathsf{p}_t)}\,\mathrm{d}t. \end{aligned} \tag{C.1}$$

La primera desigualdad viene justificada por lo siguiente. Recordemos que tenemos:

$$\begin{cases} \mathrm{d}X_t = b_t^{\star}\,\mathrm{d}t + \sqrt{\varepsilon}\,\mathrm{d}B_t^{\star}, \\ \\ \mathrm{d}X_t = \bar{b}_t\,\mathrm{d}t + \sqrt{\varepsilon}\,\mathrm{d}\bar{B}_t. \end{cases}$$

El cambio de medida se describe a través de un proceso θ_t que relaciona las dos medidas:

$$\theta_t = \frac{1}{\sqrt{\varepsilon}}(\bar{b}_t - b_t^{\star}).$$

Por el teorema de Girsanov (ver apéndice B), la densidad de Radon-Nikodym $\frac{d\bar{\mathsf{P}}}{d\mathsf{P}^{\star}}$ entre las medidas $\bar{\mathsf{P}}$ y $\mathsf{P}^{\star}$ está dada por:

$$\frac{d\bar{\mathsf{P}}}{d\mathsf{P}^{\star}} = \exp\left(\int_0^{\tau}\theta_t\,\mathrm{d}B_t^{\star} - \frac{1}{2}\int_0^{\tau}\|\theta_t\|^2\,\mathrm{d}t\right).$$

La divergencia Kullback-Leibler entre $\bar{\mathsf{P}}$ y $\mathsf{P}^{\star}$ es:

$$\mathrm{KL}(\bar{\mathsf{P}}_{[0,\tau]}|\mathsf{P}^{\star}_{[0,\tau]}) = \mathbb{E}_{\bar{\mathsf{P}}}\left[\int_0^{\tau}\theta_t\,\mathrm{d}B_t^{\star} - \frac{1}{2}\int_0^{\tau}\|\theta_t\|^2\,\mathrm{d}t\right].$$

Dado que la esperanza de $\int_0^{\tau}\theta_t\,\mathrm{d}B_t^{\star}$ bajo $\bar{\mathsf{P}}$ es cero:

$$\mathbb{E}_{\bar{\mathsf{P}}}\left[\int_0^{\tau}\theta_t\,\mathrm{d}B_t^{\star}\right] = 0,$$

la divergencia se simplifica a:

$$\mathrm{KL}(\bar{\mathsf{P}}_{[0,\tau]}|\mathsf{P}^{\star}_{[0,\tau]}) = \frac{1}{2}\mathbb{E}_{\bar{\mathsf{P}}}\left[\int_0^{\tau}\|\theta_t\|^2\,\mathrm{d}t\right].$$

Sustituyendo $\theta_t = \frac{1}{\sqrt{\varepsilon}}(\bar{b}_t - b_t^{\star})$, obtenemos:

$$\mathrm{KL}(\bar{\mathsf{P}}_{[0,\tau]}|\mathsf{P}^{\star}_{[0,\tau]}) = \frac{1}{2}\mathbb{E}_{\bar{\mathsf{P}}}\left[\int_0^{\tau}\frac{1}{\varepsilon}\|\bar{b}_t - b_t^{\star}\|^2\,\mathrm{d}t\right].$$

Finalmente, tomando la esperanza bajo $\mathsf{P}^{\star}$, obtenemos la cota que queríamos:

$$\begin{aligned}\mathbb{E}[\,\mathrm{KL}(\bar{\mathsf{P}}_{[0,\tau]}|\mathsf{P}^{\star}_{[0,\tau]})] &\le \frac{1}{2\varepsilon}\int_0^{\tau}\mathbb{E}\|\bar{b}_t - b_t^{\star}\|^2_{L^2(\mathsf{p}_t)}\,\mathrm{d}t \\ &= \int_0^{\tau}\frac{1}{(1-t)^2}\mathbb{E}\|\nabla\bar{\varphi}_{1-t} - \nabla\varphi^{\star}_{1-t}\|^2_{L^2(\mathsf{p}_t)}\,\mathrm{d}t.\end{aligned}$$

Aquí $\nabla\varphi^\star_{1-t}$ lo interpretamos como la esperanza condicional del plan entrópico $\pi^\star_t$ entre $\mathsf{p}^\star_t$ y ν, justificado en virtud del teorema (3.5), donde

$$\pi^\star_t(z,y) = \gamma^\star_t(z,y)\,\mathrm{d}\mathsf{p}^\star_t(z)\,\mathrm{d}\nu(y).$$

Nos centramos en la esperanza del integrando, aplicando la desigualdad triangular se tiene

$$\begin{aligned}\mathbb{E}\|\nabla\bar{\varphi}_{1-t} - \nabla\varphi^\star_{1-t}\|^2_{L^2(\mathsf{p}_t)} \lesssim \mathbb{E}\|\nabla\bar{\varphi}_{1-t} - \frac{1}{n}\sum_{j=1}^n Y_j\gamma^\star_t(\cdot,Y_j)\|^2_{L^2(\mathsf{p}^\star_t)} \\ + \mathbb{E}\|\frac{1}{n}\sum_{j=1}^n Y_j\gamma^\star_t(\cdot,Y_j) - \nabla\varphi^\star_{1-t}\|^2_{L^2(\mathsf{p}^\star_t)}.\end{aligned}$$

Para el segundo término, utilizando las mismas manipulaciones que Stromme [35, Lema 20] obtenemos una cota final:

$$\begin{aligned}&\mathbb{E}\|\frac{1}{n}\sum_{j=1}^n Y_j\gamma^\star_t(\cdot,Y_j) - \nabla\varphi^\star_{1-t}\|^2_{L^2(\mathsf{p}^\star_t)} \\ &= \mathbb{E}\left[\frac{1}{n^2}\sum_{j,k=1}^n \langle Y_j\gamma^\star_t(\cdot,Y_j) - \nabla\varphi^\star_{1-t}, Y_k\gamma^\star_t(\cdot,Y_k) - \nabla\varphi^\star_{1-t}\rangle_{L^2(\mathsf{p}^\star_t)}\right].\end{aligned}$$

Aquí utilizamos que para $j \neq k$, Y_j e Y_k son muestras i.i.d, se tiene que la última expresión es

$$\frac{1}{n}\|Y\gamma^\star_t(\cdot,Y) - \nabla\varphi^\star_{1-t}\|^2_{L^2(\mathsf{p}^\star_t\otimes\nu)} \leq \frac{R^2}{n}\|\gamma^\star_t\|^2_{L^2(\mathsf{p}^\star_t\otimes\nu)}.$$

Por último

$$\frac{R^2}{n}\|\gamma^\star_t\|^2_{L^2(\mathsf{p}^\star_t\otimes\nu)} \leq \frac{R^2}{n}((1-t)\varepsilon)^{-\mathsf{k}}.$$

Para el primer término

$$\begin{aligned}\nabla\bar{\varphi}_{1-t}(z) &= \frac{1}{n}\sum_{j=1}^n \frac{Y_j \exp\left(\frac{g^\star(Y_j)}{\varepsilon} - \frac{1}{2\varepsilon(1-t)}\|z-Y_j\|^2\right)}{\frac{1}{n}\sum_{j=1}^n \exp\left(\frac{g^\star(Y_j)}{\varepsilon} - \frac{1}{2\varepsilon(1-t)}\|z-Y_j\|^2\right)} \\ &= \frac{1}{n}\sum_{j=1}^n Y_j\bar{\gamma}_t(z,Y_j).\end{aligned}$$

Como se tiene la siguiente igualdad

$$\bar{\gamma}_t(z,Y_j) = \frac{\gamma^\star_t(z,Y_j)}{\frac{1}{n}\sum_{k=1}^n \gamma^\star_t(z,Y_k)}.$$

Repitiendo paso a paso las manipulaciones de Stromme [35, Lema 20] se tiene, de hecho, para $x \in \mathbb{R}^d$ fijo

$$\|\frac{1}{n}\sum_{j=1}^n Y_j(\gamma^\star_t(x,Y_j) - \bar{\gamma}_t(x,Y_j))\|^2 \leq R^2|\sum_{j=1}^n \gamma^\star_t(x,Y_j) - 1|^2.$$

Tomando la norma de $L^2(\mathsf{p}_t^\star)$ y la esperanza externa, vemos que el término que nos queda no es más que la primera componente del gradiente de la función objetivo del problema entrópico dual (ver definición C.4), que puede ser acotada vía lema C.5, resultando

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{j=1}^{n} Y_j(\gamma_t^\star(\cdot,Y_j)-\bar{\gamma}_t(\cdot,Y_j))\right\|^2_{L^2(\mathsf{p}_t^\star)}\right] \lesssim \frac{R^2}{n}\|\gamma_t^\star\|^2_{L^2(\mathsf{p}_t^\star\otimes\nu)} \leq \frac{R^2}{n}((1-t)\varepsilon)^{-\mathsf{k}},$$

donde la última desigualdad es en virtud del Stromme [35, Lema 16].

Seguimos con (C.1), aplicamos directamente el comentario anterior

$$\begin{aligned}\mathbb{E}[\,\mathrm{KL}(\bar{\mathsf{P}}_{[0,\tau]}|\mathsf{P}^\star_{[0,\tau]})] &\lesssim \frac{R^2\varepsilon^{-\mathsf{k}}}{n}\int_0^\tau (1-t)^{-\mathsf{k}-2}\,\mathrm{d}t \\ &\leq \frac{R^2\varepsilon^{-\mathsf{k}}}{n}(1-\tau)^{-\mathsf{k}-2}.\end{aligned}$$

□

Definición C.4. Sean P, Q medidas de probabilidad en $\mathbb{R}^d$. Para cada par $h_1 = (f_1, g_1) \in L^\infty(P) \times L^\infty(Q)$, existe un elemento de $L^\infty(P) \times L^\infty(Q)$ que denotaremos por $\nabla J_\epsilon^{PQ}(f_1, g_1)$ tal que para todo $h_0 = (f_0, g_0) \in L^\infty(P) \times L^\infty(Q)$ se tiene

$$\begin{aligned}\langle \nabla J_\epsilon^{PQ}(h_1), h_0\rangle_{L^2(P)\times L^2(Q)} &= \int f_0(x)\left(1-\int e^{\frac{1/2\|x-y\|^2-f_1(x)-g_1(y)}{\epsilon}}\,\mathrm{d}Q(y)\right)\mathrm{d}P(x) \\ &+ \int g_0(y)\left(1-\int e^{\frac{1/2\|x-y\|^2-f_1(x)-g_1(y)}{\epsilon}}\,\mathrm{d}P(x)\right)\mathrm{d}Q(y).\end{aligned}$$

Dicho de otro modo, el gradiente de J_ϵ^{PQ} en (f_1, g_1) es el error marginal correspondiente a (f_1, g_1).

Lema C.5. *Siguiendo con la definición anterior, suponemos $P = \mu$ y $Q = \nu_n$, donde ν_n es la medida empírica en la base de n muestras i.i.d de alguna medida ν. Sean (f, g) los potenciales óptimos entrópicos entre μ y ν, que inducen un plan de transporte óptimo entrópico π. Entonces*

$$\mathbb{E}\|\nabla J_\epsilon(f,g)\|^2_{L^2(\mu)\times L^2(\nu_n)} \lesssim \frac{\|\gamma\|^2_{L^2(\mu\otimes\nu)}}{n},$$

donde la esperanza es con respecto a los datos y $\gamma = \frac{\mathrm{d}\pi}{\mathrm{d}(\mu\otimes\nu)}$.

Demostración.

$$\begin{aligned}\mathbb{E}\|\nabla J_\epsilon(f,g)\|^2_{L^2(\mu)\times L^2(\nu_n)} &= \mathbb{E}\int\left(\frac{1}{n}\sum_{j=1}^n \gamma(x,Y_j)-1\right)^2\mathrm{d}\mu(x) \\ &+ \mathbb{E}\,\frac{1}{n}\sum_{j=1}^n\left(\int\gamma(x,Y_j)\,\mathrm{d}\mu(x)-1\right)^2.\end{aligned}$$

Notemos que por las condiciones de optimalidad

$$\int\gamma(x,Y_j)\,\mathrm{d}\mu(x)=1 \quad ,\forall\, Y_j \Longrightarrow \mathbb{E}[\gamma(x,Y_j)]=1$$

Por tanto, si escribimos $Z_j := \gamma(x, Y_j)$, que son i.i.d., vemos que

$$\mathbb{E}\int\left(\frac{1}{n}\sum_{j=1}^{n}\gamma(x,Y_j)-1\right)^2\,\mathrm{d}\mu(x)=\int\mathbb{E}\left(\frac{1}{n}\sum_{j=1}^{n}(Z_j-\mathbb{E}[Z_j])\right)^2$$

De esta expresión se deduce

$$\mathbb{E}\left((Z_j-\mathbb{E}[Z_j])(Z_k-\mathbb{E}[Z_k])\right)=0\quad\text{para}\quad j\neq k.$$

Por lo tanto, solo nos queda el primer término:

$$\mathbb{E}\left(\frac{1}{n}\sum_{j=1}^{n}(Z_j-\mathbb{E}[Z_j])\right)^2=\frac{1}{n^2}\sum_{j=1}^{n}\mathbb{E}\left((Z_j-\mathbb{E}[Z_j])^2\right).$$

Sabemos que $\mathbb{E}\left((Z_j-\mathbb{E}[Z_j])^2\right)=\mathrm{Var}(Z_j)$, por lo que la expresión inicial se convierte en:

$$\mathbb{E}\int\left(\frac{1}{n}\sum_{j=1}^{n}\gamma(x,Y_j)-1\right)^2\,\mathrm{d}\mu(x)=\int\frac{1}{n}\mathrm{Var}(Z_j)\,\mathrm{d}\mu(x)=\frac{1}{n}\mathrm{Var}_{\mu\otimes\nu}(Z_1).$$

En conclusión,

$$\mathbb{E}\|\nabla J_\epsilon(f,g)\|^2_{L^2(\mu)\times L^2(\nu_n)}=\frac{1}{n}\mathrm{Var}_{\mu\otimes\nu}(\gamma)\leq\frac{\|\gamma\|^2_{L^2(\mu\otimes\nu)}}{n},$$

ya que observemos que la varianza de γ se puede escribir de la siguiente forma:

$$\mathrm{Var}_{\mu\otimes\nu}(\gamma)=\|\gamma\|^2_{L^2(\mu\otimes\nu)}-\left(\mathbb{E}_{\mu\otimes\nu}[\gamma(x,Y_1)]\right)^2.$$

Es decir, la varianza de γ está siempre acotada superiormente por $\|\gamma\|^2_{L^2(\mu\otimes\nu)}$. □

Bibliografía

[1] Robert B Ash. *Real analysis and probability: probability and mathematical statistics: a series of monographs and textbooks*. Academic press, 2014.

[2] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[3] Espen Bernton, Jeremy Heng, Arnaud Doucet, and Pierre E Jacob. Schr\.ºdinger bridge samplers. *arXiv preprint arXiv:1912.13170*, 2019.

[4] Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

[5] Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 2017.

[6] Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*, 2022.

[7] Tianrong Chen, Guan-Horng Liu, and Evangelos A Theodorou. Likelihood training of schr\.ºdinger bridge using forward-backward sdes theory. *arXiv preprint arXiv:2110.11291*, 2021.

[8] Sinho Chewi and Aram-Alexandre Pooladian. An entropic generalization of caffarellis contraction theorem via covariance inequalities. *Comptes Rendus. Mathématique*, 361(G9):1471–1482, 2023.

[9] Lenaic Chizat, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, and Gabriel Peyré. Faster wasserstein distance estimation with the sinkhorn divergence, 2020.

[10] Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709, 2021.

[11] LawrenceCraig Evans. *Measure theory and fine properties of functions*. Routledge, 2018.

[12] Chris Finlay, Augusto Gerolin, Adam M Oberman, and Aram-Alexandre Pooladian. Learning normalizing flows from entropy-kantorovich potentials. *arXiv preprint arXiv:2006.06033*, 2020.

[13] Hans Föllmer. Time reversal on wiener space. In *Stochastic ProcessesMathematics and Physics: Proceedings of the 1st BiBoS-Symposium held in Bielefeld, West Germany, September 10–15, 1984*, pages 119–129. Springer, 2006.

[14] H. Föllmer and A. Wakolbinger. Time reversal of infinite-dimensional diffusions. *Stochastic Processes and their Applications*, 22(1):59–77, 1986.

[15] Nicola Gigli and Luca Tamanini. Benamou-brenier and duality formulas for the entropic cost on $rcd^*(k, n)$ spaces, 2018.

[16] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[17] Michel Groppe and Shayan Hundrieser. Lower complexity adaptation for empirical entropic optimal transport. *Journal of Machine Learning Research*, 25(344):1–55, 2024.

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[19] Hanwen Huang. One-step data-driven generative model via schr\.ºdinger bridge. *arXiv preprint arXiv:2405.12453*, 2024.

[20] Leonid V Kantorovich. On the translocation of masses. *Journal of mathematical sciences*, 133(4), 2006.

[21] LK Kantorovich. On a problem of monge. *Journal of Mathematical Sciences*, 133(4), 2006.

[22] Ioannis Karatzas and Steven Shreve. *Brownian motion and stochastic calculus*, volume 113. Springer Science & Business Media, 1991.

[23] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

[24] Paul Knopp and Richard Sinkhorn. A note concerning simultaneous integral equations. *Canadian Journal of Mathematics*, 20:855–861, 1968.

[25] Christian Léonard. From the schrödinger problem to the monge–kantorovich problem. *Journal of Functional Analysis*, 262(4):1879–1920, 2012.

[26] Christian Léonard. A survey of the schr\.ºdinger problem and some of its connections with optimal transport. *arXiv preprint arXiv:1308.0215*, 2013.

[27] Balmohan Vishnu Limaye. *Functional analysis*. New Age International, 1996.

[28] R Lipster and A Shiryaev. Statistics of random processes, i general theory, 2001, 2013.

[29] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pages 666–704, 1781.

[30] Bernt Øksendal and Bernt Øksendal. *Stochastic differential equations*. Springer, 2003.

[31] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[32] Aram-Alexandre Pooladian and Jonathan Niles-Weed. Plug-in estimation of schr\.ºdinger bridges. *arXiv preprint arXiv:2408.11686*, 2024.

[33] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[34] Austin Stromme. Sampling from a schrödinger bridge. In *International Conference on Artificial Intelligence and Statistics*, pages 4058–4067. PMLR, 2023.

[35] Austin J Stromme. Minimum intrinsic dimension scaling for entropic optimal transport. In *International Conference on Soft Methods in Probability and Statistics*, pages 491–499. Springer, 2024.

[36] Cédric Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

Índice de notación

$\hat{b}_t$ *Drift* estimado

$b_t^\star$ *Drift* óptimo

C Conjunto de pares (p_t, v_t) que cumplen ecuación de continuidad

$a \asymp b$ Existen $c, C > 0$ tal que $cb \leq a \leq Cb$

$a \lesssim b$ Existe $C > 0$ tal que $a \leq Cb$

$a \vee b$ $\max\{a, b\}$

$c_2(x, y)$ Función de coste cuadrática $\frac{1}{2} \mid\mid x - y \mid\mid^2$

$C_2[(T_t x)]$ Coste de desplazamiento

$I[\pi]$ Coste de transporte

$\dot{z}$ Derivada temporal

η Paso de discretización

$\mathcal{D}(\mathbb{R}^d)$ Espacio de funciones test en $\mathbb{R}^d$

ε Parámetro de regularización

F Conjunto de pares (p_t, v_t) que cumplen ecuación de Fokker-Planck

$\mathcal{H}_s Q(z)$ Semigrupo del calor $\Lambda_s \int \exp(-\frac{1}{2s} \mid\mid x - z \mid\mid^2) Q(\,\mathrm{d}x)$, Q una medida

$H(\mu)$ $\int \log(\,\mathrm{d}\mu)\,\mathrm{d}\mu$

$\mathrm{KL}(\mu|\nu)$ Divergencia de Kullback-Leibler entre μ y ν

L_t Constante de Lipschitz del *drift*

P_t Ley marginal en t

$M_+(\mathcal{C})$ Espacio de medidas positivas

ℓ Medida de Lebesgue en $\mathbb{R}^d$

$M_+(\mathbb{R}^d)$ Espacio de las medidas positivas en $\mathbb{R}^d$

W Proceso de Wiener

$\mathcal{C}$ Espacio de funciones continuas reales en $[0, 1]$; $\mathcal{C}([0, 1], \mathbb{R}^d)$

$\mathcal{P}(Y)$ Espacio de las medidas de probabilidad en Y

$\mathcal{P}_2(\mathbb{R}^d)$ Espacio de probabilidades con momento de orden 2 finito

$\Pi(\mu, \nu)$ Conjunto de planes de transporte de μ a ν

$P(\cdot|x_t)$ Probabilidad condicionada regular

R Ley del Movimiento Browniano reversible en $\mathbb{R}^d$

∂ f Subdiferencial de una función f

$\mathcal{T}(\mu, \nu)$ Plan de transporte óptimo de μ a ν

$\mathcal{T}_\varepsilon(\mu, \nu)$ Plan de transporte óptimo entrópico de μ a ν

τ Parámetro que controla el tamaño total del trayecto del puente

TV Variación total

f^* Transformada de Legendre de una función f

$T\#\mu$ Medida engendrada por T

$T_\varepsilon(X)$ Proyección baricéntrica

$\mathcal{W}_2$ Distancia 2 de Wasserstein

Índice alfabético