Optimización Robusta Distribucional con métrica de Wasserstein y algunas aplicaciones

Diego Fernando Fonseca Valero

Universidad de los Andes
Facultad de Ciencias, Departamento de Matemáticas
Bogotá, Colombia
2018

Optimización Robusta Distribucional con métrica de Wasserstein y algunas aplicaciones

Diego Fernando Fonseca Valero

Trabajo de grado presentado como requisito parcial para optar al título de:
Mágister en Matemáticas

Director(a): Mauricio Junca Peláez
Ph.D. en Industrial Engineering and Operations Research

Línea de Investigación:
Estadística, control y optimización.

Universidad de los Andes
Facultad de Ciencias, Departamento de Matemáticas
Bogotá, Colombia
2018

A mis padres y hermano

Agradecimientos

Me gustaría agradecer a muchas personas por ayudarme durante los últimos años. Primero y ante todo me gustaría agradecer a mi director de trabajo de grado, el profesor Mauricio Junca, por guiarme y proponerme esta interesante línea de investigación.
También me gustaría dar las gracias a mi madre, mi padre y mi hermano por ser un sustento para mis propósitos y por tenerme paciencia sobretodo estos últimos meses, cuando la investigación a menudo superó mi vida.
Finalmente, pero no menos importante, me gustaría agradecer a la Universidad de los Andes por darme la oportunidad de integrar esta institución y brindar condiciones ideales para el desarrollo del presente trabajo.

Resumen

En el presente trabajo estudiaremos los problemas de Optimización Robusta Distribucional DRO, estos son problemas de optimización estocástica formulados desde una visión robusta, esta visión consiste en asumir que la distribución verdadera de la variable aleatoria involucrada en el problema pertenece a un conjunto de distribuciones llamado conjunto de ambigüedad, para este caso tal conjunto se define usando la métrica de Wasserstein. Asumiendo ciertas condiciones en la función objetivo, menos restrictivas que las impuestas hasta ahora en la literatura, demostraremos que un DRO de este tipo se puede reformular como un problema de optimización semi-infinita y que dependiendo de la función objetivo dicho problema se puede formular como un problema de optimización convexa finito. Por último presentaremos una serie de aplicaciones en campos como la estadística y la economía, estas aplicaciones son todas contribuciones propias de este trabajo.

Palabras clave: Optimización, probabilidad, distribuciones, Wasserstein.

Abstract

In this thesis, we will study the problems of Robust Distributional Optimization DRO, these are problems of stochastic optimization formulated from a robust vision, this vision consists of assuming that the true distribution of the random variable involved in the problem belongs to a set of distributions called ambiguity set, for this case that set is defined using the Wasserstein metric. Assuming certain conditions in the objective function, less restrictive than those imposed so far in the literature, we will show that a DRO of this type can be reformulated as a semi-infinite optimization problem and that depending on the objective function, this problem can be formulated as a finite convex optimization problem. Finally, we will present a series of applications belonging to fields such as statistics and economics, these applications are all contributions of this work.

Keywords: Optimization, probability, distributions, Wasserstein

Contenido

1 Introducción

En el presente trabajo pretendemos abordar los problemas de optimización estocástica desde una visión robusta, estos problemas son útiles para describir efectivamente muchos problemas de toma de decisiones en entornos inciertos. Los problemas de optimización estocástica tienen su origen en los problema de optimización del tipo

$$\min_{x \in \mathbb{X}} f(x, \xi)$$

donde $\mathbb{X}$ es un conjunto de soluciones factibles y $f(x, \xi)$ es una función de costo u objetivo donde ξ es un vector de parámetros. Las dificultades emergen cuando se asume que ξ es una variable o un vector aleatorio, es en este caso en donde aparece el apelativo de estocástico. Tal condición permite modelar diversas situaciones, por ejemplo, la necesidad de un inversionista por maximizar las ganancias de su portafolio basado en datos históricos de los retornos de sus bienes, tales retornos son aleatorios; este problema tendrá especial atención en este trabajo. En ese sentido, si se conociera la distribución $\mathbb{P}$ de ξ entonces el problema anterior se puede generalizar como

$$\min_{x \in \mathbb{X}} \mathbb{E}_P[f(x, \xi)].$$

Pero en la práctica $\mathbb{P}$ no se conoce exactamente, entonces es en este punto donde emergen otras formas de abordar este último problema, todas éstas basadas en datos, es decir, en muestras de la variable aleatoria ξ. Uno de los primeros métodos presentados se basa en aproximaciones vía Monte Carlo presentado en [36] y [34], pero tales aproximaciones son costosas computacionalmente. Otra opción es tomar un enfoque estadístico del problema y asumir que $\mathbb{P}$ tiene determinada forma, este camino es quizá mas incierto que el mismo problema que se desea resolver ya que una mala elección puede conducir a un resultado diferente al resultado correcto. Así pues, emerge el enfoque robusto distribucional, este consiste en determinar un conjunto $\mathcal{D}$ de distribuciones de probabilidad de tal manera que contenga a $\mathbb{P}$, bajo esta consideración un problema de Optimización Robusta Distribucional DRO se formula como

$$\min_{x \in \mathbb{X}} \sup_{\mathbb{Q} \in \mathcal{D}} \mathbb{E}_{\mathbb{Q}}[f(x, \xi)].$$

En este enfoque la función objetivo pasa a ser el peor costo esperado para la elección de una distribución en este conjunto. El origen de este enfoque no es claro, se atribuye sus orígenes a la Teoría de Juegos de Von Neummann, pero el primer trabajo en donde se emplean estas ideas es en [30] en el marco de la teoría de investigación de operaciones.

Dependiendo del conjunto $\mathcal{D}$ el problema DRO se puede convertir en un problema tratable o en uno NP, en ese sentido, la elección del conjunto $\mathcal{D}$ es un factor determinante en la tratabilidad del problema. Existen varias formas de definir $\mathcal{D}$, por ejemplo, en [20] y [33] definen $\mathcal{D}$ como un conjunto de distribuciones que se soportan en un único punto, mientras que en [9], [26], [30] y [35] definen $\mathcal{D}$ como el conjunto de distribuciones que satisfacen ciertas restricciones en sus momentos o de las distribuciones pertenecientes a una determinada familia de distribuciones paramétrizada, por ejemplo, las distribuciones normales. En cada una de estas caracterizaciones no existe garantía que $\mathcal{D}$ contengan la distribución verdadera $\mathbb{P}$ a no ser que se tenga información sobre los momentos o la forma de $\mathbb{P}$, situación que en la practica no suele ocurrir. Por lo tanto, ante esta situación emerge otra forma de definir $\mathcal{D}$, esta consiste en dotar el conjunto de las distribuciones de probabilidad con una noción de distancia para luego establecer $\mathcal{D}$ como una bola respecto a dicha noción de distancia, por lo general la bola se centra en la distribución empírica[1] y el radio es elegido de tal manera que la distribución $\mathbb{P}$ pertenezca a la bola, de nuevo, la tratabilidad del DRO resultante depende de la noción de distancia adoptada.

Dependiendo de la distancia que se adopte se obtiene un DRO particular que por lo general se puede reformular como un problema de optimización que pertenecerá a un contexto mayormente explorado, algunas elecciones de nociones de distancia frecuentemente usadas son la entropía de Burg usada en [43], la divergencia de Kullback-Leibler usada en [16] y la distancia de Variación Total usada en [39]. En este trabajo optaremos por emplear una noción de distancia diferente a las ya mencionadas, nos referimos a *la distancia o métrica de Wasserstein*, es decir, definiremos $\mathcal{D}$ como una bola centrada respecto a la métrica de Wasserstein con centro en una distribución empírica y un radio elegido adecuadamente.

De acuerdo lo anterior, la métrica de Wasserstein juega un papel importante en este trabajo. Uno de los primeros trabajos en los que se define esta noción de distancia es en [40], aunque esta noción de distancia surge en diferentes campos de la ciencia casi simultáneamente. Además, esta métrica también suele llamarse la distancia de Mallows,

[1] Dada una muestra $\widehat{\xi}_1, \ldots, \widehat{\xi}_N$ de una variable aleatoria ξ se define la *distribución empírica* de ξ respecto a esta muestra como la medida de probabilidad definida por $\widehat{\mathbb{P}}_N := \frac{1}{N}\sum_{i=1}^{N} \delta_{\widehat{\xi}_i}$ donde δ_x es la función delta de Dirac soportada en x.

siendo estos dos últimos apelativos los nombres más comunes, sin embargo, dependiendo del contexto en el que se emplee se suele denominar de otra manera, por ejemplo, en ciencias de la computación se le llama Earth mover's distance, en el campo de la física se le denomina distancia de Monge-Kantorovich-Rubinstein y en contexto de la optimización algunos investigadores la denominaron distancia de transporte óptimo. Por otro lado, existen características de la distancia de Wasserstein que la hacen particularmente atractiva en diversas aplicaciones y sobre todo para nuestros propósitos al considerar $\mathcal{D}$ como una bola respecto a esta métrica. La cualidad mas significativa dentro del contexto de los problemas DRO es que la bola respecto a la metrica de Wasserstein es más rica en distribuciones relevantes, lo que no ocurre con las demás nociones de distancia, por ejemplo, cualquier bola respecto a la divergencia de Kullback-Leibler y centrada en una medida μ contiene únicamente medidas que son absolutamente continuas respecto a μ, de modo que solo contiene medidas soportadas en puntos donde μ también esta soportada, ademas, si μ es discreta entonces la bola respecto a la divergencia de Killback-Leibler no contiene distribuciones continuas, esto no ocurre con la métrica de Wasserstein.
Otras cualidades de la métrica de Wasserstein son que ésta se puede interpretar como un problema de transporte óptimo, siendo este último un campo de las matemáticas de continuo crecimiento y ampliamente explorado. Además, la distancia de Wasserstein se comporta muy bien capturando la percepción humana de similitud, concepto que es abordado en [28], siendo esta cualidad importante en el momento de comparar imágenes.

En este trabajo, asumiendo ciertas condiciones en la función f, se presenta una reformulación del DRO con $\mathcal{D}$ como una bola respecto a la métrica de Wasserstein centrada en la distribución empírica, tal reformulación es un problema de óptimización semi-infinita, este problema se puede reformular como un problema de optimización convexa finito para casos específicos de f. En [11] y [22] se expone la reformulación expuesta en este trabajo pero ésta es demostrada para un universo de funciones f más restrictivo que el universo de funciones que determina las condiciones que se imponen a f en este trabajo, por ejemplo, [11] asume que f es el máximo de funciones cóncavas y que el soporte de ξ es cerrado y convexo, mientras que [22] asume que f es una función lipschitziana y que el soporte de ξ es compacto. En el caso del presente trabajo no se imponen condiciones en el soporte de ξ, solo se asumen condiciones sobre f (ver Suposición 3.3.1), condiciones que satisfacen las funciones lipschitzianas y algunas funciones cóncavas. En resumen, en esta parte del trabajo se extienden los resultados de [11] y [22] a u universo de funciones más amplio.

También se exhiben algunas aplicaciones en campos como la estadística y la economía. Concretamente, en el contexto de la estadística se presenta un método para generar ban-

das que contengan con alta probabilidad la función de distribución acumulativa de una variable aleatoria, esto permitirá estimar probabilidades y de acuerdo a la forma de la banda se puede deducir información acerca de la función de distribución acumulada. Siguiendo con otra aplicación y permaneciendo en el campo de la estadística, una forma de estimar funciones de densidad de probabilidad de una variable aleatoria es el método de estimación por núcleos, el estimador que produce este método depende de un parámetro conocido como *ancho de banda*, el parámetro que genera el mejor estimador se obtiene minimizando una expresión conocida como MISE (Mean Integrated Squared Error), en este trabajo se formula una versión robusta distribucional de ese problema de minimización. Por último, en el campo de la economía se exhibe una versión robusta distribucional del problema de optimización de portafolios en el contexto de la teoría de Markowitz, este problema consiste en encontrar el portafolio que minimice la volatilidad sujeto a que el retorno esperado sea mayor a una cantidad establecida por el inversionista. Las aplicaciones que se exhiben en este documento son contribuciones e ideas propias de este trabajo.

Para lograr los propósitos expuestos anteriormente, se empieza en el Capítulo 2 definiendo todas las nociones preliminares que ayudan a justificar los resultados posteriores, entre esas nociones se encuentra la definición de la métrica de Wasserstein y su relación con la teoría de transporte óptimo, la convergencia en esta métrica y el concepto de dualidad en problemas convexos y en problemas cónicos lineales. Luego, en el Capítulo 3 se expone una reformulación del problema de Optimización Robusta Distribucional DRO con conjunto de ambigüedad siendo una bola respecto a la métrica de Wasserstein como un problema de optimización semi-infinito, este capítulo también expone la relación entre los problemas de optimización Robusta y los de Optimización Robusta Distribucional. Por último, el Capítulo 4 expone las aplicaciones en los campos de la estadística y la economía antes mencionados.

2 Preliminares

En este capítulo se presentan los conceptos fundamentales para el desarrollo de este trabajo, nos referimos, en primer lugar, a la métrica de Wasserstein, este concepto se enmarca en el ámbito de la teoría de la medida, específicamente en las medidas de probabilidad. Otro aspecto son los problemas de optimización convexa, su formulación y el concepto de dualidad para estos problemas. Por último, exploraremos el concepto de dualidad en problemas cónicos lineales.

2.1. La métrica de Wasserstein y sus caracterizaciones

Existen varias nociones de distancia en el espacio de las distribuciones de probabilidad, algunas de ellas son la divergencia Kullback-Leibler [19], Variación Total [29], Entropía de Burg [8] y la métrica de Prokhorov [27], siendo estas las más populares, no ahondaremos en detalles respecto a estas nociones de distancia, nuestra atención se centra en una en especial, nos referimos a la métrica de Wasserstein.

Definición 2.1.1 (Métrica de Wassertein)**.** *La distancia de Wasserstein* $W_p(\mu,\nu)$ *entre* $\mu,\nu \in \mathcal{P}_p(\Xi)$[1] *es definida por*

$$W_p^p(\mu,\nu) := \inf_{\Pi\in\mathcal{P}(\Xi\times\Xi)} \left\{ \int_{\Xi\times\Xi} d^p(\xi,\zeta)\Pi(d\xi,d\zeta) \ : \ \Pi(\cdot\times\Xi)=\mu(\cdot),\ \Pi(\Xi\times\cdot)=\nu(\cdot) \right\}$$

donde

$$\mathcal{P}_p(\Xi) := \left\{ \mu\in\mathcal{P}(\Xi) \ : \ \int_{\Xi} d^p(\xi,\zeta_0)\mu(d\xi) < \infty \text{ para algun } \zeta_0\in\Xi \right\}$$

donde d *es una métrica en* Ξ*.*

La distancia Wasserstein W_p define una métrica en $\mathcal{P}_p(\Xi)$ para $p\in[1,\infty)$, esto es formulado y demostrado en el Teorema 7.3 de [42]. Adicionalmente, la distancia de Wasserstein tiene una representación dual debida a Kantorovich:

[1] La métrica p-Wasserstein también esta definida para distribuciones fuera de $\mathcal{P}_p(\Xi)$, lo que probablemente podría ocurrir es que ese conjunto la métrica de Wasserstein sea infinito.

Proposición 2.1.1 (Representación dual de la distancia de Wasserstein, Teorema 5.10 en [41])**.**

$$W_p^p(\mu,\nu) = \sup_{u\in L^1(\mu),v\in L^1(\nu)} \left\{ \int_\Xi u(\xi)\mu(d\xi) + \int_\Xi v(\zeta)\nu(d\zeta) \ : \ u(\xi)+v(\zeta) \leqslant d^p(\xi,\zeta),\ \forall \xi,\zeta\in\Xi \right\}$$

donde $L^1(\nu)$ representa el L^1 espacio de funciones ν-medibles (análogo para $L^1(\mu)$).

Dado que la distancia de Wasserstein entre dos medidas es un problema de optimización, la dualidad que se alude en la proposición anterior hace referencia a un concepto análogo a la dualidad que se explorará en las secciones 2.2 y 2.3. Por otro lado, para hacer énfasis en el valor de p en la definición 2.1.1 a dicha noción de distancia se le denomina métrica p-Wasserstein, en algunos textos, por ejemplo en [41], en el caso $p = 1$ se le suele llamar métrica de Kantorovich.

Un objeto matemático importante en el contexto del presente trabajo son las bolas respecto a alguna métrica p-Wasserstein, en el contexto de $\mathcal{P}_p(\Xi)$ la bola de radio $\varepsilon > 0$ con centro en $\mu \in \mathcal{P}_p(\Xi)$ es

$$\mathcal{B}_\varepsilon^p(\mu) = \left\{ \nu \in \mathcal{P}_p(\Xi) \ \middle|\ W_p^p(\mu,\nu) \leqslant \varepsilon^p \right\}. \tag{2-1}$$

Pero de igual manera también se puede definir la bola en $\mathcal{P}(\Xi)$, el conjunto de todas las medidas de probabilidad soportadas en Ξ, como

$$\mathcal{B}_\varepsilon(\mu) = \left\{ \nu \in \mathcal{P}(\Xi) \ \middle|\ W_p^p(\mu,\nu) \leqslant \varepsilon^p \right\}. \tag{2-2}$$

Para finalizar esta sección demostraremos que $\mathcal{B}_\varepsilon(\mu)$ es convexa.

Proposición 2.1.2. *El conjunto $\mathcal{B}_\varepsilon(\mu)$ es un conjunto convexo de medidas de probabilidad.*

Demostración. Por la definición de $\mathcal{B}_\varepsilon(\mu)$ esta es un conjunto de medidas de probabilidad, resta demostrar que es un conjunto convexo. En efecto, sea $\lambda \in [0,1]$ y $\nu_1,\nu_2 \in \mathcal{B}_\varepsilon(\mu)$, entonces $W_p^p(\mu,\nu_1) \leqslant \varepsilon^p$ y $W_p^p(\mu,\nu_2) \leqslant \varepsilon^p$. Definimos

$$K = \left\{ \lambda \int_{\Xi\times\Xi} d^p(\xi,\zeta)\Pi_1(d\xi,d\zeta) + (1-\lambda)\int_{\Xi\times\Xi} d^p(\xi,\zeta)\Pi_2(d\xi,d\zeta) \ \middle|\ \begin{array}{l} \Pi_1,\Pi_2 \in \mathcal{P}(\Xi\times\Xi) \\ \Pi_1(\cdot\times\Xi) = \Pi_2(\cdot\times\Xi) = \mu(\cdot) \\ \Pi_1(\Xi\times\cdot) = \nu_1(\cdot) \\ \Pi_2(\Xi\times\cdot) = \nu_2(\cdot) \end{array} \right\}.$$

Es claro que

$$\lambda W_p^p(\mu,\nu_1) + (1-\lambda)W_p^p(\mu,\nu_2) = \text{ínf}\, K.$$

Pero para cualesquiera $\Pi_1, \Pi_2 \in \mathcal{P}(\Xi \times \Xi)$ tales que $\Pi_1(\cdot \times \Xi) = \Pi_2(\cdot \times \Xi) = \mu(\cdot)$, $\Pi_1(\Xi \times \cdot) = \nu_1(\cdot)$, $\Pi_1(\Xi \times \cdot) = \nu_2(\cdot)$, definiendo $\Pi_{1,2} = \lambda\Pi_1 + (1-\lambda)\Pi_2$, se sigue que $\Pi_{1,2}(\cdot \times \Xi) = \mu(\cdot)$, $\Pi_{1,2}(\Xi \times \cdot) = \lambda\nu_1(\cdot) + (1-\lambda)\nu_2(\cdot)$ y

$$\int_{\Xi\times\Xi} d^p(\xi,\zeta)\Pi_{1,2}(d\xi,d\zeta) = \lambda \int_{\Xi\times\Xi} d^p(\xi,\zeta)\Pi_1(d\xi,d\zeta) + (1-\lambda)\int_{\Xi\times\Xi} d^p(\xi,\zeta)\Pi_2(d\xi,d\zeta).$$

De donde se infiere que

$$\begin{aligned} W_p^p(\mu, \lambda\nu_1 + (1-\lambda)\nu_2) &\leqslant \text{ínf}\, K = \lambda W_p^p(\mu,\nu_1) + (1-\lambda)W_p^p(\mu,\nu_2) \\ &\leqslant \lambda\varepsilon^p + (1-\lambda)\varepsilon^p = \varepsilon^p. \end{aligned}$$

Por lo tanto, $W_p^p(\mu, \lambda\nu_1 + (1-\lambda)\nu_2) \leqslant \varepsilon^p$, esto es, $\lambda\nu_1 + (1-\lambda)\nu_2 \in \mathcal{B}_\varepsilon(\mu)$. □

2.1.1. Relación con la teoría de transporte óptimo

El problema de transporte óptimo consiste en encontrar la forma más eficiente de transformar una distribución de masa a otra, todo esto relativo a una función de costo dada. La mejor forma de entender esta situacion es con un ejemplo, el siguiente ejemplo fue propuesto en [41]. Considere un gran número de panaderías en una ciudad, obviamente produciendo panes, esta deberá transportarlos cada mañana a los cafés de la ciudad donde serán consumidos. La cantidad de pan que puede ser producido en cada panadería y la cantidad que puede ser consumido es conocido previamente, y puede ser modelado como medidas de probabilidad (existe una 'densidad de producción' y una 'densidad de consumo') en un cierto espacio, el cual en nuestro caso sera la ciudad (equipada con la métrica natural en la cual la distancia entre dos puntos es la longitud del camino más corto que los une). El problema es determinar a donde deberá ir cada unidad de pan de tal manera que minimice el costo de transporte total, la Figura **2-1** ilustra este ejemplo.

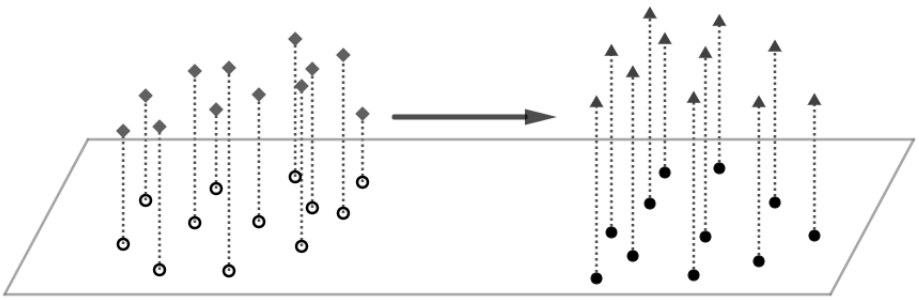


Figura 2-1: Ilustración del problema de transporte óptimo: ○ representa los centros de producción, • los centros de demanda, ♦ la producción y ▲ la demanda.

De manera formal y general un problema de transporte óptimo se puede formular de la siguiente forma, conocida como *formulación de Kantorovich*. Dadas dos medidas de probabilidad μ y ν definidas en espacios de medida X y Y respectivamente, esta formulación considera el problema de transporte óptimo como un problema de optimización sobre todos los planes de transporte, donde un plan de transporte es una medida de probabilidad $\Pi \in \mathcal{P}(X \times Y)$ con marginales μ y ν, recordemos que $\mathcal{P}(X \times Y)$ es el conjunto de todas las medidas de probabilidad definidas en $X \times Y$. Se denota $\mathcal{S}(\mu, \nu)$ al conjunto de transportes entre μ y ν, es decir

$$\mathcal{S}(\mu, \nu) := \{\Pi \in \mathcal{P}(X \times Y) \mid \Pi(\cdot \times Y) = \mu(\cdot),\ \Pi(X \times \cdot) = \nu(\cdot)\}. \tag{2-3}$$

Sea $c : X \times Y \to \mathbb{R}^+$ la función de costo, entonces el problema de transporte óptimo según Kantorovich consiste en encontrar un plan de transporte $\Pi^* \in \mathcal{S}(\mu, \nu)$ que minimice la siguiente función objetivo $\kappa : \mathcal{S}(\mu, \nu) \to \mathbb{R}^+$ dada por

$$\kappa(\Pi) := \int_{X \times Y} c(\xi, \zeta)\Pi(d\xi, d\zeta).$$

En ese sentido, la formulación de Kantorovich puede ser escrita como

$$K(\mu, \nu) = \inf_{\Pi \in \mathcal{S}(\mu,\nu)} \int_{X \times Y} c(\xi, \zeta)\Pi(d\xi, d\zeta).$$

En particular, $W_p^p(\mu, \nu)$ la métrica p-Wasserstein a la p entre dos medidas μ y ν es un problema de trasporte óptimo en el sentido de Kantorovich con $X = Y = \Xi$ y función de costo $c(x, y) = d^p(x, y)$.

2.1.2. £Por qué usar las métricas de Wasserstein?

Existen razones teóricas, muchas expuestas en [41], y otras prácticas que surgen en situaciones puntuales en contextos como la estadística y el aprendizaje automático. Algunas de estas razones son:

- La definición de las distancias de Wasserstein hace conveniente su uso en problemas de transporte óptimo en donde está naturalmente involucrada.
- Sea μ una medida de probabilidad y $\varepsilon > 0$, consideremos dos bolas $\mathcal{B}_w$ y $\mathcal{B}_\phi$ en el espacio de medidas de probabilidad centradas en μ y de radio ε, donde $\mathcal{B}_w$ es tomada respecto a alguna métrica de Wasserstein y $\mathcal{B}_\phi$ respecto a cualquier otra noción de distancia ϕ que no es Wasserstein nombradas al inicio de la Sección 2.1. Un defecto conocido de la nociones de distancia ϕ es que la bola $\mathcal{B}_\phi$ no es rica en

distribuciones relevantes [13], por ejemplo, si ϕ es la divergencia Kullback-Leibler y consideramos un conjunto medible A tal que $\mu(A) = 0$, entonces para toda $\nu \in \mathcal{B}_\phi$ se tiene $\nu(A) = 0$, es decir, $\mathcal{B}_\phi$ contiene únicamente medidas que son absolutamente continuas respecto a μ, de modo que solo contiene medidas soportadas en puntos donde μ también esta soportada. Ademas, si μ es discreta entonce para varias nociones ϕ la bola $\mathcal{B}_\phi$ no contiene distribuciones continuas. Todas estas situaciones descritas asociadas a ϕ son desventajas en el contexto de este trabajo y que no se evidencian en $\mathcal{B}_w$.

- Las distancias de Wasserstein cuentan con un formulación dual, este último termino será explorado en las siguientes secciones, no obstante tal hecho es evidenciado en la Proposición 2.1.1, tener una formulación equivalente como esta es siempre una ventaja ya que abre la posibilidad de que tal formulación, que es un problema de optimización, sea técnicamente más conveniente.
- Las distancias de Wasserstein están definidas por un ínfimo, esto es una ventaja ya que permiten calcular cotas superiores relativamente fácil, por ejemplo, el Teorema 6.15 en [41] demuestra que las métricas de Wasserstein pueden ser controladas superiormente por la distancia de Variación Total, esto será relevante en caso que la distancia de Wasserstein no se pueda calcular explícitamente pero sí la que la controla superiormente.
- Las métricas p-Wasserstein dotan al espacio $(\mathcal{P}_p(\Xi), W_p)$ de una estructura geométrica adicional [18]. En particular, para cualquier $p > 1$ y cualesquiera medidas $\mu, \nu \in \mathcal{P}_p(\Xi)$ existe un camino continuo entre μ y ν cuya longitud es la distancia p-Wassserstein entre μ y ν, esto permite introducir el concepto de geodésica. Adicionalmente, en [1] y [25] se demostró que $(\mathcal{P}_p(\Xi), W_p)$ con $p = 2$ es una variedad formal, infinito dimensional, Riemanniana.
- La percepción humana de similaridad entre imágenes es capturada de mejor manera por las métricas p-Wasserstein por medio de la cercanía entre las imágenes comparadas, recordemos que toda imagen tiene asociado un histograma, las nociones de distancia son aplicadas a esos histogramas, en [28] se presenta un ejemplo de este fenómeno.

2.1.3. Convergencia en el sentido de Wasserstein

Existe una relación entre la métrica de Wasserstein y la convergencia débil, relación que exploraremos en esta parte del trabajo. Recordemos que una sucesión $\{\nu_k\}_{k=1}^{\infty}$ de medidas de probabilidad converge débilmente a ν otra medida de probabilidad si $\int_\Xi \varphi(\xi)\nu_k(d\xi)$

converge a $\int_\Xi \varphi_\xi \nu(\xi)$ para cualquier función φ medible acotada en Ξ donde este último conjunto es el soporte de todas las medidas involucradas, a tal convergencia se denotará como $\nu_k \to \nu$. Este concepto de convergencia débil en el contexto de $\mathcal{P}_p(\Xi)$ requiere de un ajuste, en ese sentido la siguiente definición propuesta en [41] es la definición adecuada para los propósitos del presente trabajo.

Definición 2.1.2 (Convergencia débil en $\mathcal{P}_p(\Xi)$)**.** *Sea (Ξ, d) un espacio Polaco*[2] *y $p \in [1, \infty)$. Sea $\{\nu_k\}_{k=1}^{\infty}$ una sucesión de medidas de probabilidad en $\mathcal{P}_p(\Xi)$ y sea ν otro elemento de $\mathcal{P}_p(\Xi)$. Entonces $\{\nu_k\}_{k=1}^{\infty}$ se dice que converge débilmente en $\mathcal{P}_p(\Xi)$ si cualquiera de las siguientes propiedades equivalentes*[3] *es satisfecha para algún (y entonces cualquier) $\xi_0 \in \Xi$:*

(i) $\nu_k \to \nu$ *y* $\int_\Xi d(\xi_0, \xi)^p \nu_k(d\xi) \to \int_\Xi d(\xi_0, \xi)^p \nu(d\xi)$.

(ii) $\nu_k \to \nu$ *y* $\limsup_{k\to\infty} \int_\Xi d(\xi_0, \xi)^p \nu_k(d\xi) \leqslant \int_\Xi d(\xi_0, \xi)^p \nu(d\xi)$.

(iii) $\nu_k \to \nu$ *y* $\lim_{R\to\infty} \limsup_{k\to\infty} \int_{d(\xi_0,\xi)\geqslant R} d(\xi_0, \xi)^p \nu_k(d\xi) = 0$.

(iv) Para toda función continua ψ con $|\psi(\xi)| \leqslant C(1 + d(\xi_0, \xi)^p)$, $C \in \mathbb{R}$, se tiene

$$\int_\Xi \psi(\Xi)\nu_k(d\xi) \to \int_\Xi \psi(\xi)\nu(d\xi).$$

El siguiente teorema establece la relación entre la convergencia débil en $\mathcal{P}_p(\Xi)$ y la métrica p-Wasserstein, este teorema es propuesto en [41] y en síntesis lo que sentencia es que W_p metriza $\mathcal{P}_p(\Xi)$.

Teorema 2.1.1 (W_p metriza $\mathcal{P}_p(\Xi)$)**.** *Sea (Ξ, d) un espacio Polaco, y $p \in [1, \infty)$; entonces la métrica p-Wasserstein W_p metriza la convergencia débil en $\mathcal{P}_p(\Xi)$. Es decir, si $\{\nu_k\}_{k=1}^{\infty}$ una sucesión de medidas en $\mathcal{P}_p(\Xi)$ y ν es otra medida en $\mathcal{P}_p(\Xi)$, entonces ν_k converge débilmente a ν en $\mathcal{P}_p(\Xi)$ si y solo si $W_p(\nu_k, \nu) \to 0$.*

Los anteriores resultados consideran (Ξ, d) un espacio Polaco, en adelante lo seguiremos asumiendo a no ser que se aclare lo contrario, tal imposición no es restrictiva para los objetivos del presente trabajo ya que todas las situaciones que abordaremos, sobretodo en el capitulo de aplicaciones, consideran Ξ como $\mathbb{R}^m$ para algún $m \in \mathbb{N}$ con métrica euclidiana o subconjuntos compactos de $\mathbb{R}^m$ con la métrica euclidiana inducida, ambos casos son espacios Polacos.

[2] Un espacio Polaco es un espacio topológico separable y completamente metrizable.
[3] La equivalencia de estas propiedades se tienen por el Teorema 7.12 en [42].

Convergencia de la distribución empírica en la métrica de Wassertsein

Dada una variable aleatoria $\xi \in \mathbb{R}^m$ con distribución $\mathbb{P}$ y $\widehat{\Xi}_N := \left\{\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_N\right\}$ una muestra de tamaño N, note que $\widehat{\Xi}_N$ se puede considerar como un vector aleatorio con distribución $\mathbb{P}^N$, esta última es la medidas producto entre $\mathbb{P}$, N veces. A partir de esta muestra $\widehat{\Xi}_N$ se define la *distribución empírica* asociada a ξ como

$$\widehat{\mathbb{P}}_N := \frac{1}{N}\sum_{i=1}^{N} \delta_{\widehat{\xi}_i}$$

donde $\delta_{\widehat{\xi}_i}$ es la distribución delta de Dirac[4]. Del Teorema 2.1.1 se sabe que W_p metriza la convergencia débil en el espacio $\mathcal{P}_p(\Xi)$ y es sabido que la distribución empírica $\widehat{\mathbb{P}}_N$ converge débilmente en $\mathcal{P}_p(\Xi)$ a la distribución $\mathbb{P}$ cuando $N \to \infty$, entonces esta convergencia también se tiene en términos de métrica p-Wasserstein, con esta certeza es natural indagar por el orden de esta última convergencia. Dado $\varepsilon > 0$ la estrategia en [12] es encontrar una sucesión $\{\eta_N(\varepsilon)\}_{N=1}^{\infty}$ tal que

$$\mathbb{P}^N\left(W_p^p(\widehat{\mathbb{P}}_N, \mathbb{P}) \geqslant \varepsilon\right) \leqslant \eta_N(\varepsilon) \qquad \forall N \geqslant 1.$$

Los resultados expuestos en [12] dependen de establecer una serie de condiciones en los momentos de $\mathbb{P}$ que relacionan m y p, todo esto es resumido en el siguiente teorema:

Teorema 2.1.2 (Teorema 2 en [12]). *Sea $\mathbb{P} \in \mathcal{P}(\mathbb{R}^m)$ y sea $p > 0$. Suponga que $\mathbb{P}$ satisface una de las siguientes tres condiciones:*

$$\exists\, a > p,\ \exists\, \gamma > 0,\ \textit{tales que} \int_{\mathbb{R}^m} e^{\gamma\|\xi\|^a}\mathbb{P}(d\xi) < \infty, \tag{2-4}$$

$$\textit{ó}\ \exists\, a \in (0,p),\ \exists\, \gamma > 0,\ \textit{tales que} \int_{\mathbb{R}^m} e^{\gamma\|\xi\|^a}\mathbb{P}(d\xi) < \infty, \tag{2-5}$$

$$\textit{ó}\ \exists q > 2p\ \textit{tal que} \int_{\mathbb{R}^m} \|\xi\|^q\,\mathbb{P}(d\xi) < \infty. \tag{2-6}$$

Entonces para todo $N \geqslant 1$ y todo $\varepsilon \in (0,\infty)$ se tiene[5]

$$\mathbb{P}^N\left(W_p^p(\widehat{\mathbb{P}}_N, \mathbb{P}) \geqslant \varepsilon\right) \leqslant \alpha_N(\varepsilon)\mathbb{1}_{\{\varepsilon\leqslant 1\}} + \beta_N(\varepsilon),$$

donde

$$\alpha_N(\varepsilon) = C_1 \begin{cases} \exp(-C_2 N\varepsilon^2) & \textit{si } p > \frac{m}{2} \\ \exp\left(-C_2 N\left(\frac{\varepsilon}{\log\left(2+\frac{1}{\varepsilon}\right)}\right)^2\right) & \textit{si } p = \frac{m}{2} \\ \exp\left(-C_2 N\varepsilon^{\frac{m}{p}}\right) & \textit{si } p \in \left(0, \frac{m}{2}\right) \end{cases}$$

[4] La función delta de Dirac en un punto p se define como $\delta_p(x) = 1$ si $x = p$ y cero en otro caso.
[5] La función indicadora de un conjunto A se define como $\mathbb{1}_A(x) = 1$ si $x \in A$ y cero en otro caso.

y

$$\beta_N(\varepsilon) = C_1 \begin{cases} \exp\left(-C_2 N \varepsilon^{\frac{a}{p}}\right) \mathbb{1}_{\{\varepsilon>1\}} & \text{bajo (2-4),} \\ \exp\left(-C_2 (N\varepsilon)^{\frac{a-\delta}{p}}\right) \mathbb{1}_{\{\varepsilon\leqslant 1\}} + \exp\left(-C_2 (N\varepsilon)^{\frac{a}{p}} \mathbb{1}_{\{\varepsilon>1\}}\right) & \forall \delta \in (0,a) \text{ bajo (2-5),} \\ N(N\varepsilon)^{\frac{\delta-q}{p}} & \forall \delta \in (0,q) \text{ bajo (2-6).} \end{cases}$$

Las constantes C_1 y C_2 son positivas y dependen únicamente de p, m siempre y adicionalmente dependen de a, γ, $\mathbb{E}_{\mathbb{P}}\left[\exp\left(\gamma \|\xi\|^a\right)\right]$ (bajo (2-4)) ó de a, γ, $\mathbb{E}_{\mathbb{P}}\left[\exp\left(\gamma \|\xi\|^a\right),\right]$, δ (bajo (2-5)) ó de q, $\mathbb{E}_{\mathbb{P}}\left[\|\xi\|^q\right]$, δ (bajo (2-6)).

2.2. Optimización convexa, dualidad y el teorema de Weierstrass

Dada una función f de valor real con dominio D_f, un problema de optimización consiste en encontrar un elemento x^* en un conjunto determinado C, cuya intersección con el dominio de f sea no vacío, de tal manera que $f(x^*)$ sea el valor mínimo de f en C, a tal intersección se le llama región factible y a f función objetivo o función de costo. No obstante, si f no es acotada inferiormente en C el problema no podrá ser solucionado, pero si f es acotada en C no siempre es posible encontrar tal x^* que minimiza f en C pero esto no implica que no se pueda calcular dicho valor mínimo, en este caso tal valor no será alcanzado por ningún elemento de C pero existirán elementos en C cuyos imágenes bajo f tendrán valores muy cercanos al valor mínimo. Esta descripción de un problema de optimización no excluye las situaciones en donde el objetivo es maximizar una función debido a que esta es equivalente a minimizar el negativo de dicha función.
Por otro lado, un problema de optimización ya de por sí abarca una gran variedad de situaciones, de modo que no existe un método general para solucionar tales problemas. En muchos casos no es posible encontrar el valor óptimo explícitamente, es en ese punto en que los algoritmos de aproximación toman un papel relevante, la mayoría son diseñados para solucionar problemas de optimización en donde la función objetivo y la región factible son convexas, precisamente estos problemas son conocidos como *problemas de optimización convexa* y son importantes dentro de la familia de los problemas de optimización ya que en muchos problemas no convexos sus valores óptimos pueden ser aproximados por valores óptimos de problemas convexos adecuados.

Dada la importancia ya mencionada de la convexidad en los problemas de optimización un *problema de optimización convexa* en su versión general se presenta de la siguiente

forma:

$$(P)\begin{cases} \text{minimizar} & f(x) \\ \text{sujeto a} & g_i(x) \leqslant 0 \quad \forall i = 1,\dots,n \\ & h_i(x) = 0 \quad \forall i = 1,\dots,p \\ & x \in C \end{cases} \tag{2-7}$$

donde $C \subset \mathbb{X}$ es un conjunto convexo en $\mathbb{X}$ donde $\mathbb{X}$ es un espacio vectorial con producto interno $\langle \cdot, \cdot \rangle$, f, g_i : $\mathbb{X} \to \mathbb{R}$ son funciones convexas para cada $i = 1,\dots,n$ y h_i son funciones afines, es decir, existen $a_i \in \mathbb{X}$, $b_i \in \mathbb{R}$ tal que $h_i(x) = \langle a_i, x \rangle + b_i$. El dominio $\mathcal{D}$ del problema (2-7) es el conjunto de todos los puntos en los cuales las funciones f, g_i y h_i están definidas, es decir

$$\mathcal{D} := \mathrm{dom} f \cap \bigcap_{i=1}^{n} \mathrm{dom} g_i \cap \bigcap_{i=1}^{p} \mathrm{dom} f_i.$$

El problema (2-7) es llamado el problema primal y se denota con la letra (P).

El problema dual

Todo problema primal (P), así no sea convexo, tiene asociado un problema de optimización que se conoce como *dual* que se denota con la letra (D). Antes de describir la relación entre los dos problemas es más conveniente definir en primera medida el problema dual, en tal sentido se define la *función Lagrangiano* $\mathcal{L} : \mathbb{X} \times \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ asociada al problema (2-7) como

$$\mathcal{L}(x, \lambda, \mu) := f(x) + \sum_{i=1}^{n} \lambda_i g_i(x) + \sum_{i=1}^{p} \mu_i h_i(x).$$

A partir de esta última función se define una nueva llamada *función dual de Lagrange* $g : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ dada por

$$g(\lambda, \mu) := \inf_{x \in \mathcal{D}} \mathcal{L}(x, \lambda, \mu).$$

Esta función es cóncava respecto a (λ, μ) aún cuando el problema (2-7) no fuera convexo.

Se define el *problema dual* de (2-7) como el siguiente problema de optimización:

$$(D)\begin{cases} \sup_{\lambda,\mu} & g(\lambda, \mu) \\ \text{sujeto a} & \lambda \geqslant 0. \end{cases} \tag{2-8}$$

Sean p^* y d^* los valores óptimos de (2-7) y (2-8) respectivamente, entonces se dice que (2-7) satisface *dualidad débil* si se tiene

$$d^* \leqslant p^*.$$

La dualidad débil siempre se satisface independientemente de la convexidad del problema. Por otro lado, se dice que (2-7) satisface *dualidad fuerte* si se tiene

$$d^* = p^*.$$

No todos los problemas de optimización satisfacen dualidad fuerte pero en el caso de los problemas convexos si estos cumplen ciertas condiciones entonces la satisfacen. Precisamente, una de las condiciones más frecuentes para que un problema convexo satisfaga dualidad fuerte es conocido como *condición de Slater*, el problema (2-7) satisface esta condición si existe $x \in \mathbf{relint}(\mathcal{D})$ tal que

$$g_i(x) < 0 \quad \forall i = 1, 2, \ldots, n \quad \text{y} \quad h_i(x) = 0 \quad \forall i = 1, 2, \ldots, n$$

donde $\mathbf{relint}(\mathcal{D})$ es el *interior relativo* de $\mathcal{D}$ el cual se define como

$$\mathbf{relint}(\mathcal{D}) := \{y \in \mathcal{D} \mid \exists \varepsilon > 0 \text{ such that } \mathcal{B}_\varepsilon(y) \cap \mathrm{Aff}(\mathcal{D}) \subseteq \mathcal{D}\}$$

donde

$$\mathrm{Aff}(\mathcal{D}) := \left\{ \sum_{i=1}^{k} \alpha_i x_i \;\middle|\; k > 0,\ x_i \in \mathcal{D},\ \alpha_i \in \mathbb{R}, \sum_{i=1}^{k} \alpha_i = 1 \right\}$$

y $\mathcal{B}_\varepsilon(y)$ la bola respecto a la norma euclidiana de radio ε y centro y.

Si (2-7) satisface la condición de Slater entonces satisface dualidad fuerte.

El teorema de Weierstrass en optimización

Antes de presentar el resultado principal debemos definir una serie de conceptos fundamentales para formular el teorema. Sea $f : X \to \overline{\mathbb{R}}$ una función[6] donde $X \subseteq \mathbb{R}^m$, decimos que f es *cerrada* si el epigrafo de f es cerrado, es decir, si el conjunto

$$\mathrm{epi}(f) := \{(x, w) \mid x \in X,\ w \in \mathbb{R},\ f(x) \leqslant w\}$$

es cerrado en $\mathbb{R}^{m+1}$. Se dice que f es *propia* si $f(x) < \infty$ para al menos un $x \in X$ y $f(x) > -\infty$ para todo $x \in X$. Por último, decimos que f es *coerciva sobre* X si para cada sucesión $\{x_k\}_{k=1}^{\infty} \subset X$ tal que $\lim_{k\to\infty} \|x_k\| = \infty$ se tiene $\lim_{k\to\infty} f(x_k) = \infty$. En el caso en que $X = \mathbb{R}^m$ simplemente decimos que f es *coerciva*. Ademas decimos que f es inferiormente semicontinua en $x_0 \in X$ si $f(x_0) \leqslant \liminf_{x\to x_0} f(x)$.

[6] El conjunto $\overline{R}$ está dado por $\overline{R} := \mathbb{R} \cup \{-\infty, \infty\}$.

Teorema 2.2.1. *Sea $f : \mathbb{R}^m \to \mathbb{R}$ una función cerrada, propia y tal que satisface una de las siguientes tres condiciones:*

(1) El dominio de f es acotado.

(2) Existe un escalar γ tal que el conjunto de nivel $\{x \mid f(x) \leqslant \gamma\}$ es no vacío y acotado.

(3) f es coerciva.

Entonces el conjunto de elementos en $\mathbb{R}^m$ que alcanzan el valor mínimo de f en $\mathbb{R}^m$ es no vacío.

Este teorema es la Proposición 2.1.1 en [3]. Para propósitos de este trabajo el resultado que emplearemos posteriormente es el siguiente:

Corolario 2.2.1.1 (Teorema de Weierstrass)**.** *Sea $X \subseteq \mathbb{R}^m$ cerrado y $f : \mathbb{R}^m \to \mathbb{R}$ una función inferiormente semicontinua en X tales que satisfacen una de las siguientes tres condiciones:*

(1) X es acotado.

(2) Existe un escalar γ tal que el conjunto de nivel $\{x \in X \mid f(x) \leqslant \gamma\}$ es no vacío y acotado.

(3) f es coerciva en X.

Entonces el conjunto de elementos en X que alcanzan el valor mínimo de f en X es no vacío y compacto.

Otro resultado importante para los propósitos del presente trabajo es el siguiente:

Corolario 2.2.1.2 (Proposición 5.5.4 en [2])**.** *Sea $f : X \times Z \to \mathbb{R}$ una función tal que $f(\cdot, z) : X \to \mathbb{R}$ es convexa y cerrada para cada $z \in Z$, $-f(x, \cdot) : Z \to \mathbb{R}$ es convexa y cerrada para cada $x \in X$, donde X y Z son subconjuntos convexos de $\mathbb{R}^n$ y $\mathbb{R}^m$ respectivamente. Consideramos la función ϕ dada por*

$$\phi(x) := \begin{cases} \sup_{z \in Z} f(x, z) & si\ x \in X \\ \infty & si\ x \notin X. \end{cases}$$

Si ϕ es una función propia y los conjuntos de nivel $\{x \mid \phi(x) \leqslant \gamma\}$, $\gamma \in \mathbb{R}$, son compactos, entonces

$$\sup_{z \in Z} \operatorname{ínf}_{x \in X} f(x, z) = \operatorname{ínf}_{x \in X} \sup_{z \in Z} f(x, z).$$

2.3. Problemas cónicos lineales y su dualidad

Antes de introducir la noción de problemas cónicos lineales es fundamental definir los conceptos que determinan el contexto donde estos problemas emergen:

Definición 2.3.1. *Sea X un espacio vectorial (sobre $\mathbb{R}$) y $G \subset X$, se define lo siguiente:*

1. *G es un cono si para todo $\alpha \geqslant 0$ y $x \in C$ se tiene $\alpha x \in G$.*
2. *G es un cono convexo si es un cono y ademas es un conjunto convexo, es decir, para cualquier $x, y \in G$ y $\lambda \in [0,1]$ se tiene $\lambda x + (1-\lambda)y \in G$.*
3. *Dado un conjunto $C \subset X$ se define el cono convexo generado por C como* $\operatorname{cone}(A) = \bigcap_{\alpha \in \mathcal{I}} G_\alpha$ *donde $\{G_\alpha\}_{\alpha \in \mathcal{I}}$ es la familia de todos los conos convexos que contiene a C.*
4. *Sea otro espacio vectorial X', si existe una forma bilineal $\langle \cdot, \cdot \rangle_X : X' \times X \to \mathbb{R}$ entonces se define el cono polar de G como*

$$G^* = \{x^* \in X' \mid \langle x^*, x \rangle_X \geqslant 0,\ \forall x \in G\}.$$

La siguiente proposición, que reúne varios de los resultados acerca de conos expuestos en [7], establece formas equivalentes de expresar los conceptos de la definición anterior.

Proposición 2.3.1. *Sea X un espacio vectorial (sobre $\mathbb{R}$), $C \subset X$ y $G \subset X$, entonce se tiene lo siguiente:*

1. *G es un cono convexo si y solo si $G + G \subseteq G$ y $\alpha G \subseteq G$ para todo $\alpha \geqslant 0$.*
2. $\operatorname{cone}(C)$ *se puede expresar como*

$$\left\{ \sum_{i \in I} \lambda_i x_i \;\middle|\; \{x_i\}_{i \in I} \subset C,\ \lambda_i > 0\ \forall i \in I,\ I \text{ es finito} \right\}.$$

3. *Si C es convexo entonces* $\operatorname{cone}(C) = \bigcup_{\lambda > 0} \lambda C$.

Se dice que un problema de optimización es un *problema cónico lineal* si es de la forma

$$(P) \begin{cases} \min_{x \in G} & \langle c, x \rangle_X \\ \text{sujeto a} & Ax + \mathbf{b} \in K \end{cases} \tag{2-9}$$

donde X y Y espacios vectoriales (sobre $\mathbb{R}$), $G \subseteq X$ y $K \subseteq Y$ son conos convexos, $\mathbf{b} \in Y$ y $A : X \to Y$ es una aplicación lineal. Ademas se asume que existen X' y Y' espacios lineales para los cuales existe las formas bilineales $\langle \cdot, \cdot \rangle_X : X' \times X \to \mathbb{R}$ y $\langle \cdot, \cdot \rangle_Y : Y' \times Y \to \mathbb{R}$

y $c \in X'$.

La formulación dual está dada por la expresión

$$(D)\begin{cases} \text{máx}_{y \in -K^*} & \langle y, \mathbf{b} \rangle_Y \\ \text{sujeto a} & A^* y + c \in G^* \end{cases} \tag{2-10}$$

donde $A^* : Y^* \to X^*$ es la aplicación adjunta de A y G^* y K^* son los conos polares de G y K respectivamente.

Llamando $\text{val}(P)$ y $\text{val}(D)$ los valores óptimos de (2-9) y (2-10) respectivamente, una relación que siempre se tiene es

$$\text{val}(D) \leqslant \text{val}(P). \tag{2-11}$$

Adicionalmente, considerando $\text{sol}(D)$ el conjunto de soluciones óptimas de (2-10), la dualidad fuerte en un problema cónico lineal está dada por el siguiente teorema:

Teorema 2.3.1 (Teorema 2.8 (ii) y (iii) de [32])**.** *Suponga que* $\text{val}(P)$ *es finito, que el espacio* Y *es finito dimensional y que* Y' *es el espacio dual de* Y*, entonces se tiene:*

(i) Si $-b \in \text{int}\,[A(G) - K]$*, entonces* $\text{val}(P) = \text{val}(D)$ *y* $\text{sol}(D)$ *es no vacío, y, además, si* Y *es un espacio de Banach entonces* $\text{sol}(D)$ *acotado.*

(ii) Si Y *es un espacio de Banach y* $\text{sol}(D)$ *es no vacío y acotado, entonces se tiene* $\text{val}(P) = \text{val}(D)$ *y* $-b \in \text{int}\,[A(G) - K]$.

Si tratamos de identificar coincidencias con los problemas de optimización convexa de la sección anterior se observa que la condición $-b \in \text{int}\,[A(G) - K]$ es una especie de análogo de la condición de Slater en este contexto.

3 Optimización Distribucional Robusta (DRO) con métrica Wasserstein

En este capítulo abordaremos el concepto de Optimización Distrisbucional Robusta que denotaremos por sus siglas en ingles DRO. En primera medida se evidenciará que un DRO constituye un intento de generalización de un tipo de problemas de optimización con algunas limitantes que el DRO supera, después se aclara lo que significa un DRO con métrica Wasserstein que denotaremos DROW y se demostrará una formulación equivalente de dicho DROW, en realidad su formulación dual provisto de la justificación de que se satisface dualidad fuerte, tal formulación en muchos casos resulta ser conveniente. Se justificará su conveniencia en este capítulo en unos resultados en los cuales asumiendo condiciones sobre la función objetivo del problema de optimización DROW tendrá una formulación equivalente que resulta ser un problema de optimización convexa finito dimensional.

3.1. £Qué es un DRO?

El concepto de la incertidumbre es recurrente en la naturaleza, la mayoría de procesos de la naturaleza se rigen bajo modelos aleatorios, en muchos casos tal comportamiento no se puede describir exactamente lo que hace del proceso algo incierto. Ante tal situación una aspiración realista es estimar el comportamiento en términos probabilísticos, pero de nuevo surge la inquietud acerca de qué modelo probabilístico se le puede asociar al proceso, de tal dilema será difícil escapar así el proceso natural sea determinista, es decir, así se pueda modelar con una función que depende de unos parámetros constantes con total ausencia de aleatoriedad, en este caso el problema recae en calcular tales parámetros, en muchos casos estos se calculan con las herramientas tecnológicas actuales, pero estas establecen limitantes que se traducen en la cantidad de dígitos que se pueden calcular, si el parámetro es un numero irracional entonces es imposible calcularlo exactamente, el problema emerge si la función que modela el proceso que depende de dicho parámetro es

sensible respecto a la exactitud del calculo del parámetro, en tal punto la falta de dígitos es un problema importante y dicha ausencia de dígitos genera incertidumbre. Ante tal barrera podemos contemplar dos opciones, la primera considerar el parámetro dentro de un conjunto que con certeza sabemos que lo contiene, la otra opción es considerar el parámetro como una variable aleatoria, siendo ambas opciones enfoques más realistas de la situación, no obstante veremos que la opción que involucra aleatoriedad predomina (ver Teorema 3.1.1). Note que este último enfoque permite llevar un modelo determinista a uno aleatorio.

Supongamos que el proceso que se quiere estudiar es modelado por una función $f(\xi) : \mathbb{R}^n \to \mathbb{R}$ que depende de un parámetro $\xi \in \mathbb{R}^m$ desconocido (o que no se conoce exactamente) y se desea solucionar el siguiente problema de optimización

$$\min_{x \in \mathbb{X}} f(x, \xi) \tag{3-1}$$

donde $\mathbb{X} \subseteq \mathbb{R}^n$. Ante el desconocimiento de ξ se tiene las siguientes dos estrategias:

1. Asumir que el parámetro ξ pertenece a un conjunto $\Xi \subseteq \mathbb{R}^m$ lo que permite reformular el problema (3-1) como

$$\min_{x \in \mathbb{X}} \max_{\xi \in \Xi} f(x, \xi). \tag{3-2}$$

   A los problemas de optimización de la forma expuesta en (3-2) es a lo que se lo conoce como *optimización Robusta* (RO por sus siglas en inglés). El valor óptimo de (3-2) constituye una cota superior del valor óptimo del problema (3-1) asumiendo que se conociera exactamente ξ, lo ideal es que la brecha entre los dos valores óptimos sea mínima pero dado el desconocimiento de ξ hace que esa brecha sea desconocida, además, ésta también depende de la forma y tamaño del conjunto Ξ, de modo que a no ser que se tenga un conocimiento a priori de ξ que permita inferir un conjunto Ξ de tal manera que (3-2) sea un problema de optimización tratable y la brecha se pueda controlar, entonces la formulación (3-2) en diferentes contextos no es la mejor opción.

2. Una alternativa a la formulación (3-2) es considerar ξ como una variable aleatoria con distribución $\mathbb{P}$ soportada en $\Xi \subseteq \mathbb{R}^m$, si esta distribución fuera conocida se tendria la siguiente formulación:

$$\min_{x \in \mathbb{X}} \mathbb{E}_{\mathbb{P}} \left[f(x, \xi) \right]. \tag{3-3}$$

   Este último problema es conocido como un problema de *optimización estocástica*. No obstante, en la mayoría de casos $\mathbb{P}$ es desconocida, en tales casos la estrategia

consiste en considerar un conjunto de distribuciones $\mathcal{D}$ que contiene a $\mathbb{P}$, con esta consideración se obtiene la siguiente formulación[1]:

$$\min_{x\in\mathbb{X}} \max_{\mathbb{Q}\in\mathcal{D}} \mathbb{E}_{\mathbb{Q}}\left[f(x,\xi)\right]. \tag{3-4}$$

Los problemas de optimización de la forma (3-4) son conocidos como problemas de *Optimización Robusta Distribucional* (DRO), en la literatura tambie se le suele denominar *Optimización estocástica robusta distribucional* (DRSO), en ambos casos el término "distribucional"proviene de la inclusión en el problema de un conjunto de distribuciones, en este caso $\mathcal{D}$. Precisamente a este último se le conoce como *conjunto de ambigüedad*, además una primera conclusión que se puede inferir es que el valor optimo de (3-4) es una cota superior del valor optimo de (3-3).

La reformulación (3-4) pretende capturar los efectos del desconocimiento de ξ, que tal pretensión se transforme en un hecho depende de $\mathcal{D}$, en ese sentido $\mathcal{D}$ se puede asumir de tres formas:

- En caso de que se tenga alguna información acerca de $\mathbb{P}$ como sus momentos o certeza que pertenece a alguna familia de distribuciones parametrizada, por ejemplo la familia de distribuciones normales, entonces se establece $\mathcal{D}$ como el conjunto de distribuciones que tienen los momentos iguales a los momentos conocidos de $\mathbb{P}$ o que pertenezcan a la familia de distribuciones donde se sabe que pertenece $\mathbb{P}$.
- Usar las nociones de distancia en el espacio de las distribuciones para definir $\mathcal{D}$ como una bola respecto a dicha noción de distancia de tal manera que este centrada en una distribución de referencia que por lo general es la distribución empírica y radio elegido de tal manera que $\mathbb{P}$ pertenezca a la bola con alta probabilidad; algunas de las distancias usadas son las presentadas en la Sección 2.1 en especial para efectos de este trabajo centraremos la atención en las métricas p-Wasserstein.
- Como combinación de los dos casos anteriores.

Anteriormente mencionamos que un problema de Optimización Robusta Distribucional DRO es una especie de generalización de un problema de Optimización Robusta RO aunque esto no es evidente, en principio un DRO parece una alternativa a un RO que intenta abordar la misma situación que un RO pero desde una perspectiva más conservadora y consciente del desconocimiento de información, no obstante, el siguiente teorema formulado y demostrado en [44] establece que un RO es un caso particular de un DRO.

[1] La expresión $\mathbb{E}_{\mathbb{Q}}\left[f(x,\xi)\right]$ se interpreta como $\int_{\Xi} f(x,\xi)\mathbb{Q}(d\xi)$.

Teorema 3.1.1 (Corolario 2.1 en [44]). *Dada una función* $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ *y* $\mathbb{X} \in \mathbb{R}^n$ *tal que* $f(x,\cdot) : \mathbb{R}^n \to \mathbb{R}$ *medible para todo* $x \in \mathbb{X}$*, y* $Z \subset \mathbb{R}^m$ *un conjunto Borel medible, entonces se tien:*

$$\min_{x\in\mathbb{X}} \sup_{\xi\in Z} f(x,\xi) = \min_{x\in\mathbb{X}} \sup_{\mathbb{Q}\in\mathcal{D}} \mathbb{E}_{\mathbb{Q}}\left[f(x,\xi)\right]$$

donde $\mathcal{D} = \{\mathbb{Q} \in \mathcal{P}(\mathbb{R}^m) \mid \mathbb{Q}(Z) = 1\}$[2].

Teniendo claro la relación entre los problemas de tipo DRO y RO es momento de presentar el rol de las métricas p-Wasserstein en el contexto de los problemas de tipo DRO, en ese sentido la primera situación a abordar es el problema con formulación dada por:

$$J^* := \inf_{x\in\mathbb{X}} \mathbb{E}_{\mathbb{P}}[f(x,\xi)] \tag{3-5}$$

donde $\mathbb{X}$, f, ξ y Ξ fueron definidos en la descripción de (3-3) y $\mathbb{P}$ es la distribución de ξ la cual es desconocida. La intención es aproximar lo mejor posible J^*, en ese sentido, la idea es determinar un conjunto $\mathcal{D}$ en el espacio de las distribuciones tal que el hecho $\mathbb{P} \in \mathcal{D}$ sea altamente probable y con dicha probabilidad se pueda garantizar que

$$J^* \leqslant \mathbb{E}_{\mathbb{P}}[f(x_{\mathcal{D}},\xi)] \leqslant J_{\mathcal{D}} \tag{3-6}$$

donde

$$J_{\mathcal{D}} := \inf_{x\in\mathbb{X}} \sup_{\mathbb{Q}\in\mathcal{D}} \mathbb{E}_{\mathbb{Q}}[f(x,\xi)] \tag{3-7}$$

y $x_{\mathcal{D}}$ es un punto óptimo del problema (3-7), es decir, $x_{\mathcal{D}}$ satisface $J_{\mathcal{D}} = \sup_{\mathbb{Q}\in\mathcal{D}} \mathbb{E}_{\mathbb{Q}}[f(x_{\mathcal{D}},\xi)]$.

La estrategia inicia considerando N realizaciones de la variable aleatoria ξ, que se denotan por $\widehat{\Xi}_N := \left\{\widehat{\xi}_i\right\}_{i=1}^N$, por ejemplo, si ξ representara un evento, entonces $\widehat{\Xi}_N$ sería un conjunto de N datos históricos de ese evento. Es sabido que $\widehat{\Xi}_N$ se puede ver como un vector aleatorio con distribución $\mathbb{P}^N$ soportada en Ξ^N. A partir de estas realizaciones de la variable aleatoria, $\widehat{\Xi}_N$ determina la distribución empírica

$$\widehat{\mathbb{P}}_N := \frac{1}{N}\sum_{i=1}^N \delta_{\widehat{\xi}_i}$$

donde $\delta_{\widehat{\xi}_i}$ es la función delta de Dirac. En este punto se está en capacidad de proponer una aproximación de J^* que llamamos *aproximación por muestras promediadas* (SAA) tal aproximación es el valor óptimo del siguiente problema de optimización

$$\widehat{J}_{SAA} := \inf_{x\in\mathbb{X}} \left\{ \mathbb{E}_{\widehat{\mathbb{P}}_N}\left[f(x,\xi)\right] = \frac{1}{N}\sum_{i=1}^N f(x,\widehat{\xi}_i) \right\}. \tag{3-8}$$

[2] Recordemos que $\mathcal{P}(\mathbb{R}^m)$ es el conjunto de distribuciones de probabilidad soportadas en $\mathbb{R}^m$.

En este problema de optimización la función objetivo no tiene parámetros desconocidos, pero esto no garantiza que dicho problema sea fácil de solucionar, eso dependerá de la forma de la función objetivo y de $\mathbb{X}$, además, esta logrará ser una buen aproximación si el tamaño de la muestra N es suficientemente grande esto debido a a la ley de los grandes números, sin embargo, en diversas situaciones acceder a muestras de tamaño grande no es posible, es difícil o es costoso sumado a que $\widehat{J}_{SAA}$ es altamente sensible a muestras contaminadas con influencia de otra variable aleatoria, en tales casos la aproximación SAA no es una buena opción. Lo anterior nos lleva a pensar que la aproximación por medio del valor óptimo del DRO en (3-7) es una opción que tiene en cuenta las problemáticas donde SAA no logra un desempeño óptimo, en ese sentido, este trabajo dirige la atención al problema (3-7) con conjunto de ambigüedad

$$\mathcal{D} := \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N) := \left\{ \mathbb{Q} \,\middle|\, W_p^p(\widehat{\mathbb{P}}_N, \mathbb{Q}) \leqslant \varepsilon \right\}.$$

Es decir, nuestro objeto de estudio es el problema de Optimización Robusta Distribucional

$$\widehat{J}_N := \inf_{x \in \mathbb{X}} \sup_{\mathbb{Q} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}}[f(x, \xi)]. \tag{3-9}$$

Para efectos de notación y hacer remarcar que el conjunto de ambigüedad depende de alguna métrica p-Wasserstein denominaremos al tipo de problemas de la forma (3-9) como DROW.

3.2. Garantía de contenencia de la distribución empírica y consistencia asintótica

Retomando el contexto de la sección anterior, para asegurar que $\widehat{J}_N$ sea una cota superior de J^* es ideal garantizar que $\mathbb{P} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$, lo que se traduce en determinar ε de tal manera que $\mathbb{P} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$, este objetivo resulta ser ambicioso, un propósito más realista es determinar ε tal que que $\mathbb{P}$ pertenezca a $\mathbb{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ con una probabilidad mayor o igual a $1-\beta$ donde $\beta \in (0,1)$, lo ideal es elegir β de tal manera que $1-\beta$ sea grande; si dicho ε es encontrado entonces dependerá de N y de β, para hacer notoria esa dependencia lo notaremos como $\varepsilon_N(\beta)$, para determinar esta expresión es necesario imponer condiciones sobre la variable aleatoria ξ y asumir que su soporte es $\mathbb{R}^m$. Los siguientes resultados fueron expuestos inicialmente en [11], el siguiente teorema establece la realación entre ε y la probabilidad de que la bola $\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ no contenga a $\mathbb{P}$.

Teorema 3.2.1 (Teorema 3.4 en [11]). *Asumiendo que la variable aleatoria $\xi \in \mathbb{R}^m$ es tal que existe $a > p$ tal que*

$$\int_{\mathbb{R}^m} e^{\|\xi\|^a} \mathbb{P}(d\xi) < \infty. \tag{3-10}$$

Entonces

$$\mathbb{P}^N \left(W_p^p(\widehat{\mathbb{P}}_N, \mathbb{P}) \geqslant \varepsilon \right) \leqslant \begin{cases} C_1 \exp\left(-C_2 N \varepsilon^{\max\left\{2, \frac{m}{p}\right\}}\right) & si\ \varepsilon \leqslant 1 \\ C_1 \exp\left(-C_2 N \varepsilon^{\frac{a}{p}}\right) & si\ \varepsilon > 1 \end{cases} \tag{3-11}$$

para todo $N \geqslant 1$, $m \neq p$, y $\varepsilon > 0$, donde C_1 y C_2 son constantes positivas que únicamente dependen de a, (3-10) y m.

Este teorema puede emplearse para estimar el radio ε mas pequeño de tal manera que la bola respecto a la métrica p-Wasserstein con dicho radio centrada en $\widehat{\mathbb{P}}_N$ contenga a $\mathbb{P}$ con un nivel de confianza de $1 - \beta$ para un $\beta \in (0, 1)$ preestablecido, tal ε es

$$\varepsilon_N(\beta) := \begin{cases} \left(\frac{\log(C_1 \beta^{-1})}{C_2 N}\right)^{1/\max\left\{2, \frac{m}{p}\right\}} & \text{si } \frac{\log(C_1\beta^{-1})}{C_2} \leqslant N \\ \left(\frac{\log(C_1 \beta^{-1})}{C_2 N}\right)^{\frac{p}{a}} & \text{si } \frac{\log(C_1\beta^{-1})}{C_2} > N \end{cases}$$

Considerando $\widehat{J}_N$ y $\widehat{x}_N$ el valor y una solución optima del problema (3-9) respectivamente, $\varepsilon_N(\beta)$ garantiza, bajo las hipótesis del Teorema 3.2.1, que

$$\mathbb{P}^N \left\{ \mathbb{E}_{\mathbb{P}} \left[f(\widehat{x}_N, \xi) \right] \leqslant \widehat{J}_N \right\} \geqslant 1 - \beta.$$

Por otro lado, llamando β_N las cotas expuestas en (3-11) es claro que $\beta_N \to 0$ cuando $N \to \infty$, pero si esta convergencia se da a una velocidad adecuada entonces la solución de (3-9) convergerá a la solución de (3-5) cuando $N \to \infty$, todo esto es formalizado en el siguiente teorema:

Teorema 3.2.2 (Teorema 3.6 en [11]). *Asumiendo que la variable aleatoria $\xi \in \mathbb{R}^m$ es tal que existe $a > p$ tal que*

$$\int_{\mathbb{R}^m} e^{\|\xi\|^a} \mathbb{P}(d\xi) < \infty \tag{3-12}$$

y que $\beta_N \in (0, 1)$, $N \in \mathbb{N}$, satisfacen $\sum_{N=1}^{\infty} \beta_N < \infty$ y $\lim_{N\to\infty} \varepsilon_N(\beta_N) = 0$. Sea $\widehat{J}_N$ y $\widehat{x}_N$ el valor y una solución optima del problema (3-9) con $\varepsilon = \varepsilon_N(\beta_N)$, entonces se tiene

(i) Si $f(x, \xi)$ semicontinua superiormente en ξ y existe $L \geqslant 0$ con $|f(x, \xi)| \leqslant L(1 + \|\xi\|)$ para todo $x \in \mathbb{X}$ y $\xi \in \Xi$, entonces $\mathbb{P}^\infty$-casi seguramente tenemos $\widehat{J}_N \to J^*$ cuando $N \to \infty$.

(ii) Si las hipótesis del ítem (i) se tienen y adicionalmente $\mathbb{X}$ es cerrado y $f(x,\xi)$ es semicontinua inferiormente en x para cada $\xi \in \Xi$, entonces cualquier punto de acumulación de $\{\widehat{x}_N\}_{N\in\mathbb{N}}$ es $\mathbb{P}^\infty$-casi seguramente una solución óptima para (3-5).

3.3. Formulación equivalente de un DRO con métrica p-Wasserstein

Dada la imposibilidad de abordar el problema (3-5) directamente debido al desconocimiento de la distribución verdadera $\mathbb{P}$ se dijo en el capítulo anterior que una pretensión menos ambiciosa pero más realista y consciente de las vicisitudes es aproximar superiormente a J^* por medio de $\widehat{J}_N$, el valor optimo del problema en (3-9). Volver realidad esta pretensión inevitablemente conduce a la búsqueda de una forma de formular (3-9) de otra manera, y que tal reformulación sea un problema tratable, con esto último nos referimos a un problema de optimización que se ubique dentro de un contexto familiar o por lo menos más explorado en el ámbito de la optimización, específicamente en este trabajo dicho contexto será el de los problemas de optimización semi-infinita y en algunos casos dependiendo de la forma de f será el contexto de los problemas convexos finitos. En ese sentido, este capítulo pretende establecer y justificar dicha reformulación, para tal fin se recurrirá a los resultados expuestos en la Sección 2.3 evidenciando que (3-9) es un problema cónico lineal. Dependiendo de la función objetivo y la función de costo que se use para definir la métrica p-Wassesrtein la reformulación será un problema de optimización convexa. Todas estas pretensiones son alcanzadas en el Teorema principal de esta sección y de este capítulo, nos referimos al Teorema 3.3.1.

Volviendo a contextualizar, en esta parte del trabajo nos situaremos en el contexto de la Sección 3.1, como se dijo al final de esa sección el objeto de estudio es el problema de Optimización Robusta Distribucional (DRO)

$$\widehat{J}_N := \inf_{x\in\mathbb{X}} \sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}}[f(x,\xi)]. \tag{3-13}$$

donde $\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ es la bola respecto a una métrica p-Wasserstein de radio ε con centro en la distribución empírica $\widehat{\mathbb{P}}_N$ que se construye a partir de una muestra $\widehat{\xi}_1,\ldots,\xi_N$ de la variable aleatoria ξ.
Recordemos que para efectos de simplificar notación y remarcar que el conjunto de ambigüedad depende de alguna métrica p-Wasserstein denominaremos al tipo de problemas de la forma (3-13) como problemas del tipo DROW. El objetivo es encontrar una formulación equivalente de (3-13) y demostrar que en algunos casos tal formulación será un problema

de optimización tratable y carente de incertidumbre.
Momentáneamente centraremos nuestra atención en el problema de maximización interno de (3-13) y ya que x en los siguienetes resultados solo actúa como una constantes omitiremos la presencia de x en f, es decir, estudiaremos el problema de maximización

$$\sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}}[f(\xi)]. \tag{3-14}$$

Antes de avanzar es importante analizar la siguiente cuestión. La mayoría de resultados acerca de la métrica p-Wasserstein en la sección 2.1 asumen que W_p actúa sobre medidas en $\mathcal{P}_p$, esto lleva a indagar acerca de la relación entre $\mathcal{B}_\varepsilon^p\left(\widehat{\mathbb{P}}_N\right)$ y $\mathcal{B}_\varepsilon\left(\widehat{\mathbb{P}}_N\right)$, relación que es esclarecida en el siguiente lema.

Lema 3.3.1. *Dado $\varepsilon > 0$ se tiene $\mathcal{B}_\varepsilon^p\left(\widehat{\mathbb{P}}_N\right) = \mathcal{B}_\varepsilon\left(\widehat{\mathbb{P}}_N\right)$.*

Demostración. Por un lado, ya que $\mathcal{P}_p(\Xi) \subset \mathcal{P}(\Xi)$, entonces

$$\mathcal{B}_\varepsilon^p\left(\widehat{\mathbb{P}}_N\right) \subseteq \mathcal{B}_\varepsilon\left(\widehat{\mathbb{P}}_N\right).$$

Por otro lado, sea $\mathbb{Q} \in \mathcal{B}_\varepsilon\left(\widehat{\mathbb{P}}_N\right)$, entonces $W_p(\mathbb{Q}, \widehat{\mathbb{P}}_N) \leqslant \varepsilon$. Por lo tanto, por la definición de ínfimo, para $\delta > 0$ fijo existe $\Pi \in \mathcal{P}(\Xi \times \Xi)$ que satisface $\Pi(\cdot \times \Xi) = \mathbb{Q}(\cdot)$, $\Pi_k(\Xi \times \cdot) = \widehat{\mathbb{P}}_N(\cdot)$ y

$$\int_{\Xi\times\Xi} d^p(\xi, \zeta)\Pi(d\xi, d\zeta) \leqslant \varepsilon^p + \delta. \tag{3-15}$$

Adicionalmente, por ser $\widehat{\mathbb{P}}_N$ una medida discreta se sigue

$$\int_{\Xi} d^p(\zeta, \xi_0)\widehat{\mathbb{P}}_N(d\zeta) < \infty. \tag{3-16}$$

Así pues, de lo anterior, la desigualdad triangular (DT), la convexidad de la función x^p

(CO) y del hecho que x^p es una función creciente para $x \geqslant 0$ (CR) se obtiene lo siguiente

$$\begin{aligned}
\int_{\Xi} d^p(\xi,\xi_0)\mathbb{Q}(d\zeta) &= \int_{\Xi\times\Xi} d^p(\xi,\xi_0)\Pi(d\xi,d\zeta) \\
&= 2^p \int_{\Xi\times\Xi} \left(\frac{d(\zeta,\xi_0)}{2}\right)^p \Pi(d\xi,d\zeta) \\
&\leqslant 2^p \int_{\Xi\times\Xi} \left(\frac{d(\zeta,\xi_0)}{2} + \frac{d(\xi,\zeta)}{2}\right)^p \Pi(d\xi,d\zeta) \quad \leftarrow \text{ por (DT) y (CR)} \\
&\leqslant 2^{p-1} \int_{\Xi\times\Xi} d^p(\zeta,\xi_0)\Pi(d\xi,d\zeta) + 2^{p-1} \int_{\Xi\times\Xi} d^p(\xi,\zeta)\Pi(d\xi,d\zeta) \quad \leftarrow \text{por (CO)} \\
&= 2^{p-1} \int_{\Xi} d^p(\zeta,\xi_0)\widehat{\mathbb{P}}_N(d\zeta) + 2^{p-1} \int_{\Xi\times\Xi} d^p(\xi,\zeta)\Pi(d\xi,d\zeta) \\
&\leqslant 2^{p-1} \int_{\Xi} d^p(\zeta,\xi_0)\widehat{\mathbb{P}}_N(d\zeta) + 2^{p-1}(\varepsilon^p + \delta) \quad \leftarrow \text{ por (3-15)} \\
&< \infty. \qquad \leftarrow \text{ por (3-16).}
\end{aligned}$$

Por lo tanto, $\mathbb{Q} \in \mathcal{P}_p(\Xi)$, lo que permite concluir que $\mathbb{Q} \in \mathcal{B}_\varepsilon^p\left(\widehat{\mathbb{P}}_N\right)$. □

Una consecuencia inmediata del lema anterior es el siguiente corolario.

Corolario 3.3.0.1. *El problema (3-14) es equivalente al problema de optimización*

$$\sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon^p(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}}[f(\xi)] \tag{3-17}$$

donde $\mathcal{B}_\varepsilon^p(\widehat{\mathbb{P}}_N)$ es la bola en $\mathcal{P}_p(\Xi)$, ésta fue definida en (2-1).

La proposición anterior lo que sentencia es que solucionar el problema (3-14) es lo mismo que solucionar el problema (3-17). Esto justifica que (3-14) no sea incompatible con los resultados expuestos en la Sección 2.1.

Volviendo a nuestro propósito inicial, en primera instancia intentaremos reformular (3-14), para tal fin impondremos la siguiente condición en la función f.

Suposición 3.3.1. *Asumimos sobre f lo siguiente:*

1. *Si f es continua se asume que existe $C > 0$ y $\xi_0 \in \Xi$ tal que $|f(\xi)| \leqslant C(1 + d^p(\xi,\xi_o))$ para todo $\xi \in \Xi$.*

2. *Si f no es continua asumimos que es acotada.*

El resultado central de esta sección es el Teorema 3.3.1, la reformulación allí expuesta es la misma que se propone en [11] y [22], la diferencia radica en que dicha reformulación en [11] y [22] es demostrada bajo supuestos acerca de f que solo abarcan un universo de funciones más restrictivo que el universo de funciones que abarca la Suposición 3.3.1, concretamente [11] asume que f es el máximo de funciones cóncavas y [22] asume que f es lipschitziana[3] en todo Ξ y que Ξ es compacto. Tales suposiciones para los propósitos del presente trabajo son muy restrictivas, los casos propuestos en el Capitulo 4 de aplicaciones quedan excluidos, por ejemplo, la función $f(\xi) := (\xi - a)^2 - b\xi + c$ definida para $\Xi = \mathbb{R}$ donde $a, b, c \in \mathbb{R}$ son constantes no es lipschitziana ni concava, de modo que no satisface las suposiciones de [11] y [22] pero si la Suposición 3.3.1 para $p = 2$.

Para formular y demostrar al resultado central de esta sección y sus consecuencias se requiere establecer algunos conceptos y demostrar los siguientes resultados, en ese sentido denotamos por Val(3-14) el valor óptimo del problema (3-14).

Proposición 3.3.1. *Asumiendo que f satisface la Suposición 3.3.1 el valor optimo de (3-14) es finito, es decir,* Val(*3-14*) $< \infty$.

Demostración. De acuerdo a la Suposición 3.3.1 si f no es continua entonces se asume que es acotada en Ξ, entonces es inmediato Val(3-14) $< \infty$.
Si f es continua entonces de la Suposición 3.3.1 existe $C > 0$ tal que $|f(\xi)| \leqslant C(1 + d^p(\xi_0, \xi))$ para todo $\xi \in \Xi$, por lo tanto, para $\delta > 0$ fijo y para cualquier $\mathbb{Q} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ existe $\Pi \in \mathcal{P}(\Xi \times \Xi)$ tal que $\Pi(\cdot \times \Xi) = \mathbb{Q}(\cdot)$ y $\Pi(\Xi \times \cdot) = \widehat{\mathbb{P}}_N$

$$\int_{\Xi \times \Xi} d^p(\xi, \zeta)\Pi(d\xi, d\zeta) \leqslant W_p^p(\widehat{\mathbb{P}}_N, \mathbb{Q}) + \delta \leqslant \varepsilon^p + \delta.$$

[3] Se dice que f es p-lipschitziana si existe una constante K tal que $|f(\xi) - f(\zeta)| \leqslant K d^p(\xi - \zeta)$ para todo $\xi, \zeta \in \Xi$ donde f sea continua.

De modo que se tienen

$$\begin{aligned}
\mathbb{E}_{\mathbb{Q}}[f(\xi)] &= \int_{\Xi} f(\xi)\mathbb{Q}(d\xi) \\
&= \int_{\Xi\times\Xi} f(\xi)\Pi(d\xi, d\zeta) \\
&\leqslant C\int_{\Xi\times\Xi}(1 + d^p(\xi_0, \xi))\Pi(d\xi, d\zeta) \\
&= C\int_{\Xi\times\Xi}\Pi(d\xi, d\zeta) + C\int_{\Xi\times\Xi} d^p(\xi_0, \xi)\Pi(d\xi, d\zeta) \\
&= C + 2^pC\int_{\Xi\times\Xi}\left(\frac{d(\xi_0, \xi)}{2}\right)^p \Pi(d\xi, d\zeta) \\
&\leqslant C + 2^{p-1}C\int_{\Xi\times\Xi} d^p(\xi_0, \zeta)\Pi(d\xi, d\zeta) + 2^{p-1}C\int_{\Xi\times\Xi} d^p(\xi, \zeta)\Pi(d\xi, d\zeta) \\
&\leqslant C + 2^{p-1}C\int_{\Xi} d^p(\xi_0, \zeta)\widehat{\mathbb{P}}_N(\zeta) + 2^{p-1}C(\varepsilon^p + \delta) \\
&= C + \frac{2^{p-1}C}{N}\sum_{i=1}^{N} d^p(\xi_0, \widehat{\xi}_i) + 2^{p-1}C(\varepsilon^p + \delta) < \infty
\end{aligned}$$

Notando $M := C\left(1 + \frac{2^{p-1}}{N}\sum_{i=1}^{N} d^p(\xi_0, \widehat{\xi}_i) + 2^{p-1}(\varepsilon^p + \delta)\right)$ se sigue que $\mathbb{E}_{\mathbb{Q}}[f(\xi)] \leqslant M$ para todo $\mathbb{Q} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$, por lo tanto

$$\text{Val}(3\text{-}14) = \sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}}[f(\xi)] \leqslant M < \infty.$$

□

La siguiente proposición no depende de la Suposición 3.3.1, es valida en general, las ideas de esta proposición son expuestas en [11] pero los detalles son presentados en el presente trabajo, antes de seguir es importante recordar que los resultados expuestos se desarrollan en el espacio medible $(\Xi, \mathscr{E})$ y denotamos por $\mathscr{E}\otimes\mathscr{E}$ la σ-álgebra producto.

Proposición 3.3.2. *El problema (3-14) es equivalente al problema*

$$\begin{cases} \sup\limits_{\mathbb{Q}_i\in\mathcal{P}(\Xi)} & \frac{1}{N}\sum\limits_{i=1}^{N}\int_{\Xi} f(\xi)\mathbb{Q}_i(d\xi) \\ \textit{sujeto a} & \frac{1}{N}\sum\limits_{i=1}^{N}\int_{\Xi} d^p\left(\xi, \widehat{\xi}_i\right)\mathbb{Q}_i(d\xi) \leqslant \varepsilon^p. \end{cases} \tag{3-18}$$

Demostración. Por la definición de la métrica p-Wasserstein se sigue que (3-14) es equivalente al problema

$$\begin{cases} \sup\limits_{\Pi\in\mathcal{P}(\Xi\times\Xi),\, \mathbb{Q}\in\mathcal{P}(\Xi)} & \int_{\Xi} f(\xi)\mathbb{Q}(d\xi) \\ \text{sujeto a} & \int_{\Xi} d^p\left(\xi,\widehat{\xi}_i\right)\Pi(d\xi,d\zeta) \leqslant \varepsilon^p. \\ & \Pi \in \mathcal{S}(\mathbb{Q},\widehat{\mathbb{P}}_N). \end{cases} \tag{3-19}$$

Donde $\mathcal{S}(\mathbb{Q},\widehat{\mathbb{P}}_N)$ fue definido en la Sección 2.1.1, concretamente en (2-3). Así pues, definimos los siguientes conjuntos:

$$\Gamma := \left\{ (\mathbb{Q},\Pi) \;\middle|\; \mathbb{Q}\in\mathcal{P}(\Xi),\ \Pi\in\mathcal{S}(\mathbb{Q},\widehat{\mathbb{P}}_N),\ \int_{\Xi} d^p\left(\xi,\widehat{\xi}_i\right)\Pi(d\xi,d\zeta)\leqslant\varepsilon^p \right\}$$

y

$$\mathcal{K} := \left\{ \left(\frac{1}{N}\sum_{i=1}^N \mathbb{Q}_i,\ \frac{1}{N}\sum_{i=1}^N \mathbb{Q}_i\otimes\delta_{\widehat{\xi}_i} \right) \;\middle|\; \mathbb{Q}_i\in\mathcal{P}(\Xi),\ \frac{1}{N}\sum_{i=1}^N\int_{\Xi} d^p\left(\xi,\widehat{\xi}_i\right)\mathbb{Q}_i(d\xi)\leqslant\varepsilon^p \right\}.$$

Notemos que (3-19) es equivalente a $\sup\limits_{(\mathbb{Q},\Pi)\in\Gamma} \mathbb{E}_{\mathbb{Q}}[f(\xi)]$. El primer objetivo es demostrar que $\Gamma=\mathcal{K}$. En efecto, por un lado es inmediato que $\mathcal{K}\subseteq\Gamma$, probaremos la otra continencia, esta es consecuencia de la Ley de probabilidad total (PT). Sea $(\mathcal{Q},\Pi)\in\Gamma$, definiendo $\Xi^* := \Xi\backslash\bigcup_{i=1}^N\left\{\widehat{\xi}_i\right\}$ se sigue que $\Pi(\Xi\times\Xi^*)=0$, de modo que para cualesquiera $A,B\in\mathscr{E}$ se tiene

$$\begin{aligned} \Pi(A\times B) &= \Pi((A\times\Xi)\cap(\Xi\times B)) \\ &= \sum_{i=1}^N \Pi\left((A\times\Xi)\cap(\Xi\times B)\,\middle|\,\Xi\times\left\{\widehat{\xi}_i\right\}\right)\Pi\left(\Xi\times\left\{\widehat{\xi}_i\right\}\right) \\ &\quad +\Pi\left((A\times\Xi)\cap(\Xi\times B)|\Xi\times\Xi^*\right)\Pi(\Xi\times\Xi^*) \quad \leftarrow \text{por (PT)} \\ &= \sum_{i=1}^N \Pi\left((A\times\Xi)\cap(\Xi\times B)\,\middle|\,\Xi\times\left\{\widehat{\xi}_i\right\}\right)\Pi\left(\Xi\times\left\{\widehat{\xi}_i\right\}\right). \\ &= \frac{1}{N}\sum_{i=1}^N \Pi\left((A\times\Xi)\cap(\Xi\times B)\,\middle|\,\Xi\times\left\{\widehat{\xi}_i\right\}\right) \end{aligned}$$

Pero notemos que

$$\begin{aligned}
\Pi\left((A\times\Xi)\cap(\Xi\times B)\,\middle|\,\Xi\times\left\{\widehat{\xi}_i\right\}\right) &= \frac{\Pi\left((A\times\Xi)\cap(\Xi\times B)\cap\left(\Xi\times\left\{\widehat{\xi}_i\right\}\right)\right)}{\Pi\left(\Xi\times\left\{\widehat{\xi}_i\right\}\right)}\\
&= \Pi\left(A\times\Xi\,\middle|\,(\Xi\times B)\cap\Xi\times\left\{\widehat{\xi}_i\right\}\right)\frac{\Pi\left((\Xi\times B)\cap\left(\Xi\times\left\{\widehat{\xi}_i\right\}\right)\right)}{\Pi\left(\Xi\times\left\{\widehat{\xi}_i\right\}\right)}\\
&= \Pi\left(A\times\Xi\,\middle|\,\Xi\times\left(B\cap\left\{\widehat{\xi}_i\right\}\right)\right)\frac{\Pi\left(\Xi\times\left(B\cap\left\{\widehat{\xi}_i\right\}\right)\right)}{\Pi\left(\Xi\times\left\{\widehat{\xi}_i\right\}\right)}\\
&= \Pi\left(A\times\Xi\,\middle|\,\Xi\times\left(B\cap\left\{\widehat{\xi}_i\right\}\right)\right)\frac{\widehat{\mathbb{P}}_N\left(B\cap\left\{\widehat{\xi}_i\right\}\right)}{\widehat{\mathbb{P}}_N\left(\left\{\widehat{\xi}_i\right\}\right)}\\
&= \Pi\left(A\times\Xi\,\middle|\,\Xi\times\left(B\cap\left\{\widehat{\xi}_i\right\}\right)\right)\delta_{\widehat{\xi}_i}(B)\\
&= \Pi\left(A\times\Xi\,\middle|\,\Xi\times\left\{\widehat{\xi}_i\right\}\right)\delta_{\widehat{\xi}_i}(B).
\end{aligned}$$

Por lo tanto, considerando $\overline{\mathbb{Q}}_i\in\mathcal{P}(\Xi)$ definida por $\overline{\mathbb{Q}}_i(\cdot):=\Pi\left(\cdot\times\Xi\,\middle|\,\Xi\times\left\{\widehat{\xi}_i\right\}\right)$ se infiere que

$$\Pi(A\times B)=\frac{1}{N}\sum_{i=1}^{N}\overline{\mathbb{Q}}_i(A)\delta_{\widehat{\xi}_i}(B).$$

Así pues, $\Pi=\dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\overline{\mathbb{Q}}_i\otimes\delta_{\widehat{\xi}_i}$, y ya que $\Pi(\cdot\times\Xi)=\mathbb{Q}$ entonces se sigue $\mathbb{Q}=\frac{1}{N}\sum_{i=1}^{N}\overline{\mathbb{Q}}_i$, por lo tanto, $(\mathcal{Q},\Pi)\in\mathcal{K}$ y así queda demostrado $\Gamma\subseteq\mathcal{K}$.

Del hecho $\Gamma=\mathcal{K}$ se sigue que (3-19) es equivalente al problema

$$\sup_{\left(\frac{1}{N}\sum_{i=1}^{N}\mathbb{Q}_i,\frac{1}{N}\sum_{i=1}^{N}\mathbb{Q}_i\otimes\delta_{\widehat{\xi}_i}\right)\in\mathcal{K}}\mathbb{E}_{\frac{1}{N}\sum_{i=1}^{N}\mathbb{Q}_i}[f(\xi)]$$

$$=\begin{cases}\displaystyle\sup_{\mathbb{Q}_i\in\mathcal{P}(\Xi)} & \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\int_{\Xi}f(\xi)\mathbb{Q}_i\\ \text{sujeto a} & \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\int_{\Xi}d^p(\xi,\widehat{\xi}_i)\mathbb{Q}_i(d\xi)\leqslant\varepsilon^p.\end{cases}\tag{3-20}$$

Como (3-19) es equivalente a (3-14) entonces este último es equivalente a (3-20) que es lo que se quería demostrar. □

El siguiente teorema es el resultado central de esta sección ya que proporciona la formulación equivalente deseada de (3-14), esta formulación resulta ser un problema de opti-

mización semi-infinita. Es muy común que la demostración de un resultado sea valorada incluso más que el mismo resultado, esto debido a que las demostraciones involucran ideas y estrategias que podrían servir como hoja de ruta para abordar otros problemas o demostrar resultados similares, en ese sentido, esta demostración recurre a las ideas de la Sección 2.3, camino que fue sugerido en [11] para un universo de funciones diferente al que comprende la Suposición 3.3.1 y ademas no se incurrió en detalles ni demostración alguna.

Teorema 3.3.1 (Teorema principal)**.** *El problema (3-14) es equivalente al problema de optimización*

$$\begin{cases} \inf\limits_{\lambda,s} & \lambda\varepsilon^p + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \textit{sujeto a} & \sup\limits_{\xi\in\Xi}\left(f(\xi) - \lambda d^p(\xi,\widehat{\xi}_i)\right) \leqslant s_i \quad \forall i = 1,\dots,N, \\ & \lambda \geqslant 0. \end{cases} \tag{3-21}$$

Demostración. En primera medida por la Proposición 3.3.2 se tiene que (3-14) es equivalente a (3-18), de modo que es suficiente con reformular (3-18) para obtener una reformulación de (3-14).
Teniendo en cuento esto último definimos las funciones φ_i como

$$\varphi_i(\xi) := \begin{cases} \frac{1}{N} f(\xi) & \text{para } i = 0 \\ \frac{1}{N} d^p(\xi,\widehat{\xi}_i) & \text{para } i = 1,\dots,N \\ 1 & \text{para } i = N+1. \end{cases}$$

Considerando $\mathcal{P}(\Xi)$ el conjunto de medidas de probabilidad no-negativas en $(\Xi,\mathscr{E})$, intentaremos recrear el contexto de la Sección 2.3, en ese sentido, considerando $\widetilde{X}$ el espacio lineal (sobre $\mathbb{R}$) de todas las *medidas signadas finitas* en Ξ generadas[4] por $\mathcal{P}(\Xi)$, asumimos

$$X := \left\{(\mathbb{Q}_1,\dots,\mathbb{Q}_N) \mid \mathbb{Q}_i \in \widetilde{X} \text{ para } i = 1,\dots,N\right\}.$$

Ya que $\widetilde{X}$ es un espacio lineal entonces X también lo es. Asumimos G como el cono convexo en X dado por

$$G := \{(\mathbb{Q}_1,\dots,\mathbb{Q}_N) \in X \mid \mathbb{Q}_1 \geqslant 0,\ i = 1,\dots,N\}$$

donde $\geqslant 0$ significa que la medida es positiva. Consideramos X' el espacio lineal dado por

$$X' := \{(h_1,\dots,h_N) \mid h_i \in \operatorname{span}\{\varphi_0,\varphi_1,\dots,\varphi_{N+1}\} \text{ para } i = 1,\dots,N\}.$$

[4]El termino 'generadas' hace referencia a que las medidas son combinaciones lineales de elementos de $\mathcal{P}(\Xi)$.

La forma bilineal entre X y X' es dada por

$$\begin{aligned} \langle \cdot, \cdot \rangle_X : X' \times X &\longrightarrow \mathbb{R} \\ ((h_1, \ldots, h_N), (\mathbb{Q}_1, \ldots, \mathbb{Q}_N)) &\longmapsto \sum_{i=1}^{N} \int_{\Xi} h_i(\xi) \mathbb{Q}_i(d\xi). \end{aligned}$$

Ademas consideramos $Y = \mathbb{R}^{N+1}$, Y' el espacio dual de Y que en este caso es $\mathbb{R}^{N+1}$ y $\langle \cdot, \cdot \rangle_Y$ es la tradicional forma bilineal entre un espacio y su dual que en este caso resulta ser el producto punto vectorial en $\mathbb{R}^{N+1}$. Siguiendo con las caracterizaciones asumimos $K = (-\infty, 0] \times \{0\}^N$, adicionalmente $\mathbf{b}$ es un vector en $\mathbb{R}^{N+1}$ tal que $\mathbf{b}_i = 1$ para cada $i = 2, \ldots, N+1$ y $\mathbf{b}_1 = \varepsilon^p$, es decir, $\mathbf{b} = (\varepsilon^p, 1, \ldots, 1)$. Por ultimo, definimos las funciones $\psi_i \in X'$ dadas por

$$\psi_i := \begin{cases} \frac{1}{N}(f, \ldots, f) & \text{para } i = 0 \\ (\varphi_1, \ldots, \varphi_N) & \text{para } i = 1 \\ (0, \cdots, 0, \underset{\substack{\uparrow \\ i-1 \text{ ésima posición}}}{1}, 0, \cdots, 0) & \text{para } i = 2, \ldots, N+1. \end{cases}$$

A partir de estas últimas funciones consideramos la aplicación lineal $A : X \to Y$ dada por

$$A(\mathbb{Q}_1, \ldots, \mathbb{Q}_N) := \left(\langle \psi_1, (\mathbb{Q}_1, \ldots, \mathbb{Q}_N) \rangle_X, \langle \psi_2, (\mathbb{Q}_1, \ldots, \mathbb{Q}_N) \rangle_X, \ldots, \langle \psi_{N+1}, (\mathbb{Q}_1, \ldots, \mathbb{Q}_N) \rangle_X \right).$$

Usando toda la caracterización introducida hasta aquí el problema (3-18) se puede reescribir como

$$\begin{cases} \sup\limits_{\Pi} & \langle \psi_0, \Pi \rangle_X \\ \text{sujeto a} & A(\Pi) - \mathbf{b} \in K \\ & \Pi \in G. \end{cases} \tag{3-22}$$

De lo consignado en la Sección 2.3 sobre el dual de problemas cónicos lineales, específicamente el Teorema 2.3.1 que establece condiciones para la dualidad fuerte para problemas cónicos lineales del tipo (2-10), condiciones que al final de esta demostración probaremos que (3-22) satisface, se sigue que el problema dual de (3-22) y a su vez de (3-14) es

$$\begin{cases} \inf\limits_{w} & -\mathbf{b}^T \cdot w \\ \text{sujeto a} & A^*(w) + \psi_0 \in -G^* \\ & w \in K^*. \end{cases} \tag{3-23}$$

Donde K^* y G^* son el cono dual de K y G respectivamente, entonces siguiendo la Definición 2.3.1 que define el cono dual se tiene que $K^* = (-\infty, 0] \times \mathbb{R}^{N+1}$ y

$$-G^* = \left\{ (h_1, \ldots, h_N) \in X' \mid \langle (h_1, \ldots, h_N), (\mathbb{Q}_1, \ldots, \mathbb{Q}_N) \rangle_X \leqslant 0, \ \forall\, \mathbb{Q}_i \in \widetilde{X} \text{ con } \mathbb{Q}_i \geqslant 0 \right\}$$

Ademas A^* es el operador adjunto $A^* : \mathbb{R}^{N+2} \to X'$ que en este caso es dado por

$$A^*(w) = (w_1\varphi_1 + w_2, w_1\varphi_2 + w_3, \ldots, w_1\varphi_N + w_{N+1}). \tag{3-24}$$

Esto último puede no ser evidente, recordemos que A^* el operador adjunto de A es aquel que satisface

$$\langle w, A(\mathbb{Q}_1, \ldots, \mathbb{Q}_N)\rangle = \langle A^*(w), (\mathbb{Q}_1, \ldots, \mathbb{Q}_N)\rangle_X \tag{3-25}$$

El A^* propuesto en (3-24) satisface (3-25), en efecto, por un lado se tiene

$$\begin{aligned}
\langle w, A(\mathbb{Q}_1, \ldots, \mathbb{Q}_N)\rangle &= \sum_{i=1}^{N+1} \langle \psi_i, (\mathbb{Q}_1, \ldots, \mathbb{Q}_N)\rangle_X \, w_i \\
&= w_1 \sum_{j=1}^{N} \int_\Xi \varphi_j(\xi)\mathbb{Q}_j + \sum_{i=2}^{N+1} \langle \psi_i, (\mathbb{Q}_1, \ldots, \mathbb{Q}_N)\rangle_X \, w_i \\
&= \frac{w_1}{N} \sum_{j=1}^{N} \int_\Xi d^p(\xi, \widehat{\xi}_j)\mathbb{Q}_j + \sum_{i=2}^{N+1} w_i \mathbb{Q}_{i-1}(\Xi).
\end{aligned}$$

Pero por otro lado

$$\begin{aligned}
\langle A^*(w), (\mathbb{Q}_1, \ldots, \mathbb{Q}_N)\rangle_X &= \langle (w_1\varphi_1 + w_2, w_1\varphi_2 + w_3, \ldots, w_1\varphi_N + w_{N+1}), (\mathbb{Q}_1, \ldots, \mathbb{Q}_N)\rangle_X \\
&= \sum_{i=1}^{N} \int_\Xi (w_1\varphi_i(\xi) + w_{i+1})\mathbb{Q}_i(d\xi) \\
&= w_1 \sum_{i=1}^{N} \int_\Xi \varphi_i(\xi)\mathbb{Q}_i(d\xi) + \sum_{i=1}^{N} \int_\Xi w_{i+1}\mathbb{Q}_i(d\xi) \\
&= \frac{w_1}{N} \sum_{j=1}^{N} \int_\Xi d^p(\xi, \widehat{\xi}_j)\mathbb{Q}_j + \sum_{i=2}^{N+1} w_i \mathbb{Q}_{i-1}(\Xi).
\end{aligned}$$

De modo que se satisface (3-25).Por lo tanto, con todas la certezas y caracterizaciones anteriores el problema dual de (3-22), es decir (3-23), puede ser expresado como sigue:

$$\begin{cases}
\inf\limits_{w} & -w_1\varepsilon^p - \sum\limits_{i=2}^{N+1} w_i \\
\text{sujeto a} & \left(w_1\varphi_1 + w_2 + \dfrac{1}{N}f, w_1\varphi_2 + w_3 + \dfrac{1}{N}f, \ldots, w_1\varphi_N + w_{N+1} + \dfrac{1}{N}f\right) \in -G^*, \\
& w_1 \leqslant 0.
\end{cases}$$

$$
= \begin{cases} \inf\limits_{w} & -w_1\varepsilon^p - \sum\limits_{i=2}^{N+1} w_i \\ \text{sujeto a} & \sum\limits_{i=1}^{N} \int_{\Xi} \left(w_1\varphi_i + w_{i+1} + \dfrac{1}{N} f(\xi) \right) \mathbb{Q}_i(d\xi) \leqslant 0 \quad \forall\, (\mathbb{Q}_1, \ldots, \mathbb{Q}_N) \succcurlyeq 0, \\ & w_1 \leqslant 0. \end{cases}
$$

$$
= \begin{cases} \inf\limits_{w} & -w_1\varepsilon^p - \sum\limits_{i=2}^{N+1} w_i \\ \text{sujeto a} & \sum\limits_{i=1}^{N} \int_{\Xi} \left(\dfrac{w_1}{N} d^p(\xi, \widehat{\xi}_i) + w_{i+1} + \dfrac{1}{N} f(\xi) \right) \mathbb{Q}_i(d\xi) \leqslant 0 \quad \forall\, (\mathbb{Q}_1, \ldots, \mathbb{Q}_N) \succcurlyeq 0, \\ & w_1 \leqslant 0. \end{cases} \tag{3-26}
$$

La primera restricción de (3-26) puede ser reemplazada por las siguientes N restricciones

$$
\sup_{\xi \in \Xi} \left\{ w_1 d^p(\xi, \widehat{\xi}_i) + f(\xi) \right\} \leqslant -N w_i \quad \forall i = 1, \ldots, N. \tag{3-27}
$$

En efecto, por un lado, ya que en la primera restricción de (3-26) se toma $(\mathbb{Q}_1, \ldots, \mathbb{Q}_N) \succcurlyeq 0$, esto obliga a que si w satisface dicha restricción entonces debe producir que las funciones que allí se están integrando sean negativas para todo $\xi \in \Xi$, así pues w debe satisfacer que

$$
\sup_{\xi \in \Xi} \left\{ \frac{w_1}{N} d^p(\xi, \widehat{\xi}_i) + w_{i+1} + \frac{1}{N} f(\xi) \right\} \leqslant 0, \quad \forall i = 1, \ldots, N. \tag{3-28}
$$

Esta última es equivalente a (3-27), así hemos demostrado que la primera restricción de (3-26) implica (3-27), la otra implicación es consecuencia de escribir (3-27) como (3-28). Por lo tanto, (3-26) es equivalente al problema

$$
\begin{cases} \inf\limits_{w} & -w_1\varepsilon^p - \sum\limits_{i=2}^{N+1} w_i \\ \text{sujeto a} & \sup\limits_{\xi \in \Xi} \left\{ w_1 d^p(\xi, \widehat{\xi}_i) + f(\xi) \right\} \leqslant -N w_i \quad \forall\, i = 1, \ldots, N, \\ & w_1 \leqslant 0. \end{cases} \tag{3-29}
$$

Así pues, reescribiendo las variables w_1 como $-\lambda$ y w_i como $-\frac{s_i}{N}$ para $i = 2, \ldots, N+1$ se obtiene que (3-29) se puede reescribir como

$$
\begin{cases} \inf\limits_{\lambda, s} & \lambda\varepsilon^p + \dfrac{1}{N} \sum\limits_{i=1}^{N} s_i \\ \text{sujeto a} & \sup\limits_{\xi \in \Xi} \left(f(\xi) - \lambda d^p(\xi, \widehat{\xi}_i) \right) \leqslant s_i \quad \forall\, \xi \in \Xi,\ i = 1, \ldots, N, \\ & \lambda \geqslant 0. \end{cases} \tag{3-30}
$$

Y de acuerdo a todas las equivalencias que se han evidenciado se concluye que este último problema es equivalente a (3-14) que es lo que se deseaba demostrar.
Ahora debemos verificar que la dualidad fuerte en realidad se satisface, para tal fin por el Teorema 2.3.1 (i) es suficiente con demostrar que $b = (\varepsilon^p, 1, \dots, 1)$ es punto interior de $\mathbf{int}\,(A(G) - K)$, pero esto se satisface ya que $A(G) - K = \mathbb{R}^{N+1}_+$, en efecto, ya que $-K = [0, \infty) \times \{0\}^N$ entonces

$$\begin{aligned} A(G) - K &= \{(\langle \psi_1, (\mathbb{Q}_1, \dots, \mathbb{Q}_N)\rangle_X + a, \langle \psi_2, (\mathbb{Q}_1, \dots, \mathbb{Q}_N)\rangle_X, \dots, \langle \psi_{N+1}, (\mathbb{Q}_1, \dots, \mathbb{Q}_N)\rangle_X) \mid a \geqslant 0,\ \mathbb{Q}_i \geqslant 0\} \\ &= \left\{ \left(\sum_{i=1}^N \int_\Xi \psi_1(\xi)\mathbb{Q}_i(d\xi) + a, \mathbb{Q}_1(\Xi), \dots, \mathbb{Q}_N(\Xi) \right) \middle| \ a \geqslant 0,\ \mathbb{Q}_i \geqslant 0 \right\} \\ &= \left\{ \left(\sum_{i=1}^N \int_\Xi d^p(\xi, \widehat{\xi}_i)\mathbb{Q}_i(d\xi) + a, \mathbb{Q}_1(\Xi), \dots, \mathbb{Q}_N(\Xi) \right) \middle| \ a \geqslant 0,\ \mathbb{Q}_i \geqslant 0 \right\}. \end{aligned}$$

Entonces, es claro que $A(G) - K \subseteq \mathbb{R}^{N+1}_+$, para demostrar la otra contenencia sea $(c_0, c_1, \dots, c_N) \in \mathbb{R}^{N+1}_+$, entonces consideramos las medidas positivas dadas por

$$\mathbb{Q}_i := c_i \delta_{\widehat{\xi}_i} \text{ para cada } i = 1, \dots, N,$$

y asumimos $a = c_0$, bajo estas caracterizaciones se tiene que $\sum_{i=1}^N \int_\Xi d^p(\xi, \widehat{\xi}_i)\mathbb{Q}_i(d\xi) + a = c_0$ y $\mathbb{Q}_i(\Xi) = c_i$, por lo tanto, se sigue

$$\left(\sum_{i=1}^N \int_\Xi d^p(\xi, \widehat{\xi}_i)\mathbb{Q}_i(d\xi) + a, \mathbb{Q}_1(\Xi), \dots, \mathbb{Q}_N(\Xi) \right) = (c_0, c_1, \dots, c_N).$$

Lo que permite concluir que $(c_0, c_1, \dots, c_N) \in A(G) - K$, de modo que $\mathbb{R}^{N+1}_+ = A(G) - K$. □

Un caso particular del Teorema 3.3.1 es el siguiente corolario cuya idea principal es el gran aporte de [11], dependiendo de la forma de la función f este resultado conduce a una reformulación que será un problema de optimización convexo finito.

Corolario 3.3.1.1. *Asumiendo $\Xi \subseteq \mathbb{R}^m$ convexo, d igual a una norma $\|\cdot\|$ en $\mathbb{R}^m$ y $\|\cdot\|_*$ su norma dual, $p = 1$, y para cada $\lambda > 0$ se define $\mathbb{C}_\lambda := \{z \in \mathbb{R} \ \|z\|_* \leqslant \lambda\}$ y las funciones $F^i_\lambda : \mathbb{C}_\lambda \times \Xi \to \mathbb{R}$ dadas por $F^i_\lambda(z, \xi) := f(\xi) - \langle z, \xi - \widehat{\xi}_i\rangle$. Se asume que f satisface la Suposición 3.3.1 y también satisfaciendo que $-F^i_\lambda(z, \cdot) : \Xi \to \mathbb{R}$ y $F^i_\lambda(\cdot, \xi) : \mathbb{C}_\lambda \to \mathbb{R}$ son convexas y cerradas para cada $z \in \mathbb{C}_\lambda$, $\xi \in \Xi$ y cada $i = 1, \dots, N$, y adicionalmente se asume que el valor optimo de (3-14) es finito y que los conjuntos de nivel $\left\{ z \in \mathbb{C}_\lambda \mid \sum_{\xi \in \Xi} \left(f(\xi) - \langle z, \xi - \widehat{\xi}_i\rangle \right) \leqslant \gamma \right\}$ son compactos. Entonces el problema (3-14) es*

equivalente al problema de optimización

$$\begin{cases} \inf\limits_{\lambda,s,z} & \lambda\varepsilon + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \textit{sujeto a} & \sup\limits_{\xi\in\Xi} \left(f(\xi) - \langle z, \xi\rangle\right) + \left\langle z, \widehat{\xi}_i \right\rangle \leqslant s_i \quad \forall i = 1,\ldots,N, \\ & \|z\|_* \leqslant \lambda. \end{cases} \tag{3-31}$$

donde $\langle\cdot,\cdot\rangle$ *es el producto interno en* $\mathbb{R}^m$.

Demostración. Por el Teorema 3.3.1 se obtiene la reformulación

$$\begin{cases} \inf\limits_{\lambda,s} & \lambda\varepsilon^p + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{sujeto a} & \sup\limits_{\xi\in\Xi} \left(f(\xi) - \lambda \left\|\xi - \widehat{\xi}_i\right\|\right) \leqslant s_i \quad \forall i = 1,\ldots,N, \\ & \lambda \geqslant 0. \end{cases} \tag{3-32}$$

Dado que la norma dual de $\|\cdot\|$ se define como $\|z\|_* := \sup_{\|y\|\leqslant 1} \langle z, y\rangle$ entonces es sabido que para este caso $\lambda\|\cdot\|$ se puede expresar en términos de $\lambda\|\cdot\|_*$ como $\|y\|_* := \sup_{\|z\|_*\leqslant 1} \langle z, y\rangle$, de este hecho se sigue que (3-32) es equivalente a

$$\begin{aligned} &\begin{cases} \inf\limits_{\lambda,s} & \lambda\varepsilon^p + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{sujeto a} & \sup\limits_{\xi\in\Xi} \left(f(\xi) - \sup\limits_{\|z\|_*\leqslant\lambda} \left\langle z, \xi - \widehat{\xi}_i \right\rangle\right) \leqslant s_i \quad \forall i = 1,\ldots,N, \\ & \lambda \geqslant 0. \end{cases} \\ =&\begin{cases} \inf\limits_{\lambda,s} & \lambda\varepsilon^p + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{sujeto a} & \sup\limits_{\xi\in\Xi} \inf\limits_{\|z\|_*\leqslant\lambda} \left(f(\xi) - \left\langle z, \xi - \widehat{\xi}_i \right\rangle\right) \leqslant s_i \quad \forall i = 1,\ldots,N, \\ & \lambda \geqslant 0. \end{cases} \\ =&\begin{cases} \inf\limits_{\lambda,s} & \lambda\varepsilon^p + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{sujeto a} & \inf\limits_{\|z\|_*\leqslant\lambda} \sup\limits_{\xi\in\Xi} \left(f(\xi) - \left\langle z, \xi - \widehat{\xi}_i \right\rangle\right) \leqslant s_i \quad \forall i = 1,\ldots,N, \\ & \lambda \geqslant 0. \end{cases} \end{aligned} \tag{3-33}$$

$$
= \begin{cases} \inf\limits_{\lambda,s,z} & \lambda\varepsilon^p + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{sujeto a} & \sup\limits_{\xi\in\Xi}\left(f(\xi) - \left\langle z, \xi - \widehat{\xi}_i \right\rangle\right) \leqslant s_i \quad \forall i = 1,\ldots,N, \\ & \|z\|_* \leqslant \lambda. \end{cases}
$$

$$
= \begin{cases} \inf\limits_{\lambda,s,z} & \lambda\varepsilon + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{sujeto a} & \sup\limits_{\xi\in\Xi}\left(f(\xi) - \langle z, \xi\rangle\right) + \left\langle z, \widehat{\xi}_i \right\rangle \leqslant s_i \quad \forall i = 1,\ldots,N, \\ & \|z\|_* \leqslant \lambda. \end{cases}
$$

La igualdad (3-33) se obtiene aplicando N veces el Corolario 2.2.1.2 con $X = \mathbb{C}_\lambda$, $Z = \Xi$. □

Por ultimo, no podemos finalizar este capítulo sin presentar la siguiente variación del problema (3-13)

$$
\begin{cases} \sup\limits_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} & \mathbb{E}_{\mathbb{Q}}\left[f(\xi)\right] \\ \text{sujeto a} & \mathbb{E}_{\mathbb{Q}}\left[g_i(\xi)\right] = b_i. \quad \forall\, i = 1,\ldots,k. \end{cases} \tag{3-34}
$$

donde $b_i \in \mathbb{R}$ y $g_1,\ldots,g_k$ son funciones integrables respecto a cada medida en $\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$. En el capitulo de aplicaciones quedara en evidencia la importancia de este problema ya que permite abordar diferentes situaciones, precisamente en tales contextos emerge la necesidad de saber si este problema satisface dualidad fuerte, como vimos en las secciones 2.2 y 2.3 satisfacerla tiene ciertas ventajas, el siguiente teorema garantiza tal condición.

Teorema 3.3.2. *Asumiendo que el valor optimo del problema (3-34) es finito, si ademas se satisface alguna de las siguientes condiciones*

i) El punto $(b_1,\ldots,b_k,1)$ *es punto interior del conjunto*

$$
\left\{\lambda\left(\langle g_1,\mathbb{Q}\rangle_X,\ldots,\langle g_k,\mathbb{Q}\rangle_X,1\right) \mid \lambda > 0,\ \mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)\right\}.
$$

ii) El conjunto de distribuciones optimas[5] *del problema (3-34) es no vacío y acotado.*

Entonces (3-34) satisface dualidad fuerte, es decir, (3-34) es equivalente al problema

$$
\inf_{a_1,\ldots,a_k}\left\{\sum_{i=1}^{k} a_i b_i + \sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)}\int_\Xi\left(f(\xi) - \sum_{i=1}^{k} a_i g_i(\xi)\right)\mathbb{Q}(d\xi)\right\}.
$$

[5] Distribuciones que alcanzan el valor óptimo.

Demostración . Considerando $\overline{\mathcal{P}}(\Xi)$ el conjunto de medidas de probabilidad no-negativas en $(\Xi, \mathscr{E})$ tales que $f, g_1, g_2, \ldots, g_k$ son integrables respecto a cada medida en $\overline{\mathcal{P}}(\Xi)$, intentaremos recrear el contexto de la Sección 2.3, en ese sentido asumimos X el espacio lineal (sobre $\mathbb{R}$) de las medidas signadas generadas por $\overline{\mathcal{P}}(\Xi)$. X' el espacio lineal de funciones $h : \Xi \to \mathbb{R}$ generadas por $f, g_1, \ldots, g_k$, es decir, funciones formadas por combinaciones lineales de $f, g_1, \ldots, g_k$. La forma bilineal entre X y X' es dada por

$$\begin{array}{rcl} \langle \cdot, \cdot \rangle_X : X' \times X & \longrightarrow & \mathbb{R} \\ (h, \mathbb{Q}) & \longmapsto & \int_\Xi h(\xi)\mathbb{Q}(d\xi). \end{array}$$

Ademas consideramos $Y = \mathbb{R}^{k+1}$, Y' el espacio dual de Y que en este caso es $\mathbb{R}^{k+1}$ y $\langle \cdot, \cdot \rangle_Y$ es la tradicional forma bilineal entre un espacio y su dual que en este caso resulta ser el producto punto vectorial en $\mathbb{R}^{k+1}$. Siguiendo con las caracterizaciones asumimos $K = \{0\}$ donde 0 es el vector cero de $\mathbb{R}^{k+1}$, G es el cono convexo generado por $\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$, de la Proposición 2.1.2 se sabe que $\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ es convexo, entonces por la Proposición 2.3.1 3 se tiene $G = \cup_{\lambda>0}\lambda\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$, adicionalmente $\mathbf{b}$ es un vector en $\mathbb{R}^{k+1}$ tal que $\mathbf{b}_i = b_i$ para cada $i = 1, \ldots, k$ y $\mathbf{b}_{k+1} = 1$, y consideramos $\varphi = f$, $\psi_i = g_i$ para $i = 1, \ldots, k$ y $\psi_{k+1}(x) = 1$ para todo $x \in \Xi$. Por ultimo, consideramos la aplicación lineal $A : X \to Y$ dada por $A(\mathbb{Q}) = (\langle \psi_1, \mathbb{Q} \rangle_X, \ldots, \langle \psi_k, \mathbb{Q} \rangle_X, \langle \psi_{k+1}, \mathbb{Q} \rangle_X)$.
Bajo estas consideraciones el problema (2-9) para este contexto es

$$\begin{cases} \sup\limits_{\mathbb{Q} \in G} & \langle f, \mathbb{Q} \rangle_X \\ \text{sujeto a} & \langle g_i, \mathbb{Q} \rangle_X = b_i \quad \forall i \leqslant k \\ & \langle 1, \mathbb{Q} \rangle_X = 1. \end{cases} \tag{3-35}$$

Pero dado que $G = \cup_{\lambda>0}\lambda\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ y que $\langle 1, \mathbb{Q} \rangle_X = 1$ para toda $\mathbb{Q} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$, entonces el problema (3-35) es igual al siguiente

$$\begin{cases} \sup\limits_{\mathbb{Q} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} & \lambda \langle f, \mathbb{Q} \rangle_X \\ \text{sujeto a} & \lambda \langle g_i, \mathbb{Q} \rangle_X = b_i \quad \forall i \leqslant k \\ & \lambda \langle 1, \mathbb{Q} \rangle_X = 1 \\ & \lambda > 0. \end{cases} = \begin{cases} \sup\limits_{\mathbb{Q} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} & \langle f, \mathbb{Q} \rangle_X \\ \text{sujeto a} & \langle g_i, \mathbb{Q} \rangle_X = b_i \quad \forall i \leqslant k. \end{cases}$$

El último problema de la igualdad anterior es precisamente (3-34), es decir, (3-34) es igual a (3-35). Por lo tanto, dado que $\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ es convexo entonces

$$G^* = \left\{ f \in X' \mid \lambda \langle f, \mathbb{Q} \rangle_X \geqslant 0,\ \mathbb{Q} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N),\ \lambda > 0 \right\} = \left\{ f \in X' \mid \langle f, \mathbb{Q} \rangle_X \geqslant 0,\ \mathbb{Q} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N) \right\}.$$

De esto ultimo, del hecho que $-K^* = -\mathbb{R}^{k+1} = \mathbb{R}^{k+1}$ y de lo consignado en la Sección 2.3 sobre el dual de problemas cónicos lineales, específicamente el Teorema 2.3.1 que establece

condiciones para la dualidad fuerte del problema (2-10), las condiciones de dicho teorema son satisfechas por (3-35) debido a que este satisface las condiciones *i)* ó *ii)*, de modo que se satisface dualidad fuerte, en ese sentido se sigue que el problema dual de (3-35) y a su vez de (3-34) es

$$
\begin{aligned}
&\begin{cases} \inf\limits_{a\in\mathbb{R}^{k+1}} & \langle a, \mathbf{b}\rangle_Y \\ \text{sujeto a} & \sum\limits_{i=1}^{k} a_i \langle g_i, \mathbb{Q}\rangle_X + a_{k+1} \geqslant \langle f, \mathbb{Q}\rangle_X \quad \forall \mathbb{Q} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N). \end{cases} \\
= &\begin{cases} \inf\limits_{a\in\mathbb{R}^{k+1}} & \sum\limits_{i=1}^{k} a_i b_i + a_{k+1} \\ \text{sujeto a} & \sum\limits_{i=1}^{k} a_i \int_\Xi g_i(\xi)\mathbb{Q}(d\xi) + a_{k+1} \geqslant \int_\Xi f(\xi)\mathbb{Q}(d\xi) \quad \forall \mathbb{Q} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N). \end{cases} \\
= &\begin{cases} \inf\limits_{a\in\mathbb{R}^{k+1}} & \sum\limits_{i=1}^{k} a_i b_i + a_{k+1} \\ \text{sujeto a} & \int_\Xi \left(f(\xi)\mathbb{Q}(d\xi) - \sum\limits_{i=1}^{k} a_i g_i(\xi) \right) \mathbb{Q}(d\xi) \leqslant a_{k+1} \quad \forall \mathbb{Q} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N). \end{cases} \\
= &\inf_{a_1,\ldots,a_k} \left\{ \sum_{i=1}^{k} a_i b_i + \sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \int_\Xi \left(f(\xi) - \sum_{i=1}^{k} a_i g_i(\xi) \right) \mathbb{Q}(d\xi) \right\}.
\end{aligned}
$$

□

Corolario 3.3.2.1. *Asumiendo que las funciones f y g_i del problema (3-34) son tales que la función $F_a(\xi) := f(\xi) - \sum_{i=1}^{k} a_i(g_i(\xi) - b_i)$ satisface la Suposición 3.3.1 para todo $a \in \mathbb{R}^k$, y asumiendo que se satisfacen alguna de las condiciones i) o ii) del Teorema 3.3.2. Entonces el problema (3-34) se puede reformular como*

$$
\begin{cases} \inf\limits_{a_1,\ldots,a_k,\lambda} & \lambda\varepsilon + \dfrac{1}{N}\sum\limits_{i=1}^{N} s_i \\ \textit{sujeto a} & \sup\limits_{\xi\in\Xi} \left(f(\xi) - \sum\limits_{i=1}^{k} a_i(g_i(\xi) - b_i) - \lambda d^p(\xi, \widehat{\xi}_i) \right) \leqslant s_i \quad \forall\, i = 1, \ldots, N, \\ & \lambda \geqslant 0. \end{cases} \tag{3-36}
$$

Este corolario es consecuencia inmediata del Teorema 3.3.1 y el Teorema 3.3.2.
Para finalizar, al inicio de esta sección convenimos omitir la parecencia de x en la función f, tal omisión obedece a la pretensión de hacer menos densa la notación, no obstante, al introducir x las reformulaciones anteriores siguen siendo validas con la salvedad que en los ínfimos de cada reformulación se debe introducir la variable x.

4 Aplicaciones

En este capitulo se presentan una serie de situaciones pertenecientes a diversos campos de las matemáticas que incluyen la estadística y la optimización de portafolios que pueden ser formuladas como un problema de optimización del tipo (3-3) el cual, de acuerdo al capitulo anterior, se puede aproximar por un problema del tipo (3-9) que a su vez es equivalente a un problema de optimización semi-infinita. Cada una de las siguientes aplicaciones son contribuciones e ideas propias de este trabajo.

4.1. Bandas de confianza para funciones de distribución acumulada

Sea ξ una variable aleatoria con distribución desconocida $\mathbb{P}$, en esta sección presentaremos una banda en $\mathbb{R}^2$ tal que la función de distribución acumulada $F(x) := \mathbb{P}(\xi \leqslant x)$ pertenece a dicha banda con probabilidad $1-\beta$, a esto último es a lo que se le conoce como *nivel de confianza*, para tal fin estimaremos puntualmente las fronteras de la banda, siendo específicos estas fronteras son gráficos de funciones.
Nuestra pretensión se traduce en estimar superior e inferiormente la función de distribución, tal aproximación se realizará puntualmente, en ese orden de ideas, dado un $x \in \mathbb{R}$ el objetivo es estimar superior e inferiormente la cantidad $\mathbb{P}(\xi \leqslant x)$, pero dado que esto último se puede expresar como $\mathbb{E}_{\mathbb{P}}\left[\mathbb{1}_{\{\xi \leqslant x\}}(\xi)\right]$, entonces la estrategia es que tales estimaciones serán el valor óptimo de un DROW, para tal fin asumimos que con la única información con la que contamos acerca de ξ es una muestra $\widehat{\xi}_1, \ldots \widehat{\xi}_N$, esta permite definir la bola $\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ respecto a la métrica 1-Wasserstein de centro en la medida empírica $\mathbb{P}_N$ y radio $\varepsilon > 0$.
Así pues, continuación procederemos estimando la frontera superior de la banda propuesta y luego la frontera inferior, en cada caso emerge un problema DROW diferente.

Estimación superior

Supongamos que hemos encontrado un $\varepsilon > 0$ tal que $\mathbb{P} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ con probabilidad[1] $1-\beta$, dicho la justificaremos más adelante $\varepsilon > 0$. Bajo estas consideraciones tenemos con probabilidad $1-\beta$ la siguiente desigualdad

$$\mathbb{P}(\xi \leqslant x) = \mathbb{E}_{\mathbb{P}}\left[\mathbb{1}_{\{\xi \leqslant x\}}(\xi)\right] \leqslant \sup_{\mathbb{Q} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}}\left[\mathbb{1}_{\{\xi \leqslant x\}}(\xi)\right].$$

La expresión de la parte derecha de la desigualdad es un DROW cuya función objetivo es una función acotada, de modo que esta dentro del contexto de la Suposición 3.3.1, aquí asumimos $\Xi = \mathbb{R}$ ya que en principio no conocemos el soporte de la distribución. Llamaremos $\widehat{U}_{N,\varepsilon}$ la función que describe la frontera superior de la banda, es decir

$$\widehat{U}_{N,\varepsilon}(x) := \sup_{\mathbb{Q} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}}\left[\mathbb{1}_{\{\xi \leqslant x\}}(\xi)\right].$$

Entonces, por el Teorema 3.3.1 considerando $p = 1$ y d como la métrica en $\mathbb{R}$ inducida por el valor absoluto se obtiene lo siguiente

$$\widehat{U}_{N,\varepsilon}(x) := \begin{cases} \inf\limits_{\lambda,s} & \lambda\varepsilon + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{sujeto a} & \sup\limits_{\xi \in \mathbb{R}} \left(\mathbb{1}_{\{\xi \leqslant x\}}(\xi) - \lambda\left|\xi - \widehat{\xi}_i\right|\right) \leqslant s_i \quad \forall i = 1, \ldots, N, \\ & \lambda \geqslant 0. \end{cases}$$

Pero la función $\mathbb{1}_{\{\xi \leqslant x\}}$ se puede expresar como $\mathbb{1}_{\{\xi \leqslant x\}}(\xi) = \max\{f_1(\xi), f_2(\xi)\}$ donde f_1 y f_2 son definidas como $f_1(\xi) := 1 - \mathcal{X}_{\{\xi \leqslant x\}}$ y $f_2(\xi) := 0$, donde, en general, dado un conjunto A la función[2] $\mathcal{X}_A$ se define como

$$\mathcal{X}_A(\xi) := \begin{cases} 0 & \text{si } \xi \in A, \\ \infty & \text{si } \xi \notin A. \end{cases}$$

Bajo estas consideraciones se sigue

$$\widehat{U}_{N,\varepsilon}(x) := \begin{cases} \inf\limits_{\lambda,s} & \lambda\varepsilon + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{sujeto a} & \sup\limits_{\xi \in \mathbb{R}} \left(\max\{f_1(\xi), f_2(\xi)\} - \lambda\left|\xi - \widehat{\xi}_i\right|\right) \leqslant s_i \quad \forall i = 1, \ldots, N, \\ & \lambda \geqslant 0. \end{cases}$$

[1] La aleatoriedad que permite hablar de probabilidad es en el sentido expuesto en la Sección 3.2.
[2] La función $\mathcal{X}_A$ es habitual en el área de optimización.

$$
= \begin{cases} \underset{\lambda,s}{\text{ínf}} & \lambda\varepsilon + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{sujeto a} & \sup\limits_{\xi\in\mathbb{R}} \left(f_1(\xi) - \lambda \left| \xi - \widehat{\xi}_i \right| \right) \leqslant s_i \quad \forall i = 1, \ldots, N, \\ & \sup\limits_{\xi\in\Xi} \left(f_2(\xi) - \lambda \left| \xi - \widehat{\xi}_i \right| \right) \leqslant s_i \quad \forall i = 1, \ldots, N, \\ & \lambda \geqslant 0. \end{cases}
$$

$$
= \begin{cases} \underset{\lambda,s}{\text{ínf}} & \lambda\varepsilon + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{sujeto a} & \sup\limits_{\xi\in\mathbb{R}} \left(f_1(\xi) - \lambda \left| \xi - \widehat{\xi}_i \right| \right) \leqslant s_i \quad \forall i = 1, \ldots, N, \\ & 0 \leqslant s_i \qquad\qquad\qquad\qquad\quad \forall i = 1, \ldots, N, \\ & \lambda \geqslant 0. \end{cases}
$$

Ya que $|\cdot|$ se puede expresar como $\lambda|\xi| = \max_{|z|_*\leqslant\lambda}\{z\xi\}$ y como la norma dual $|\cdot|_*$ de $|\cdot|$ es $|\cdot|$ entonces $\widehat{F}_s$ se puede expresar como

$$
\widehat{U}_{N,\varepsilon}(x) := \begin{cases} \underset{\lambda,s}{\text{ínf}} & \lambda\varepsilon + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{sujeto a} & \sup\limits_{\xi\in\mathbb{R}} \left(f_1(\xi) - \max\limits_{|z_i|\leqslant\lambda} \left\{ z_i \left(\xi - \widehat{\xi}_i \right) \right\} \right) \leqslant s_i \quad \forall i = 1, \ldots, N, \\ & 0 \leqslant s_i \qquad\qquad\qquad\qquad\quad \forall i = 1, \ldots, N, \\ & \lambda \geqslant 0. \end{cases}
$$

$$
= \begin{cases} \underset{\lambda,s}{\text{ínf}} & \lambda\varepsilon + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{sujeto a} & \sup\limits_{\xi\in\mathbb{R}} \min\limits_{|z_i|\leqslant\lambda} \left(f_1(\xi) - z_i \left(\xi - \widehat{\xi}_i \right) \right) \leqslant s_i \quad \forall i = 1, \ldots, N, \\ & 0 \leqslant s_i \qquad\qquad\qquad\qquad\quad \forall i = 1, \ldots, N, \\ & \lambda \geqslant 0. \end{cases}
$$

El Corolario 2.2.1.2 permite intercambiar el supremo y el mínimo de la primera restricción lo que conduce a lo siguiente:

$$
\widehat{U}_{N,\varepsilon}(x) := \begin{cases} \underset{\lambda,s}{\text{ínf}} & \lambda\varepsilon + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{sujeto a} & \min\limits_{|z_i|\leqslant\lambda} \sup\limits_{\xi\in\mathbb{R}} \left(f_1(\xi) - z_i \left(\xi - \widehat{\xi}_i \right) \right) \leqslant s_i \quad \forall i = 1, \ldots, N, \\ & 0 \leqslant s_i \qquad\qquad\qquad\qquad\quad \forall i = 1, \ldots, N, \\ & \lambda \geqslant 0. \end{cases}
$$

$$= \begin{cases} \inf\limits_{\lambda,s,z} & \lambda\varepsilon + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i & \\ \text{sujeto a} & \sup\limits_{\xi\in\mathbb{R}} \left(f_1(\xi) - z_i \left(\xi - \widehat{\xi}_i \right) \right) \leqslant s_i & \forall i = 1, \ldots, N, \\ & |z_i| \leqslant \lambda & \forall i = 1, \ldots, N, \\ & 0 \leqslant s_i & \forall i = 1, \ldots, N, \\ & \lambda \geqslant 0. & \end{cases}$$

Pero el supremo de la primera restricción se puede expresar como

$$\sup_{\xi\in\mathbb{R}} \left(f_1(\xi) - z_i \left(\xi - \widehat{\xi}_i \right) \right) = \begin{cases} 1 - z_i \left(\xi - \widehat{\xi}_i \right) & \text{si } z_i \in [-\lambda, 0] \\ \infty & \text{si } z_i \in [0, \lambda]. \end{cases}$$

Por lo tanto, realizando el cambio de variable $w_i = -z_i$ tenemos

$$\widehat{U}_{N,\varepsilon}(x) := \begin{cases} \inf\limits_{\lambda,s,w} & \lambda\varepsilon + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i & \\ \text{sujeto a} & 1 + w_i \left(x - \widehat{\xi}_i \right) \leqslant s_i & \forall i = 1, \ldots, N, \\ & 0 \leqslant w_i \leqslant \lambda & \forall i = 1, \ldots, N, \\ & s_i \geqslant 0 & \forall i = 1, \ldots, N, \\ & \lambda \geqslant 0. & \end{cases}$$

Así pues, $\widehat{U}_{N,\varepsilon}(x)$ es el valor óptimo de un problema de optimización convexo finito, siendo esta la formulación deseada.

Estimación inferior

Asumimos $\varepsilon > 0$ como en el caso anterior. Bajo estas consideraciones tenemos con probabilidad $1 - \beta$ la siguiente desigualdad

$$\mathbb{P}(\xi \leqslant x) = \mathbb{E}_{\mathbb{P}} \left[\mathbb{1}_{\{\xi \leqslant x\}}(\xi) \right] \geqslant \inf_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}} \left[\mathbb{1}_{\{\xi \leqslant x\}}(\xi) \right] = - \sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}} \left[-\mathbb{1}_{\{\xi \leqslant x\}}(\xi) \right].$$

La expresión de la parte derecha de la desigualdad es -1 por el valor óptimo de un DROW cuya función objetivo es una función acotada, de modo que esta dentro del contexto de la Suposición 3.3.1, como en el caso anterior aquí asumimos $\Xi = \mathbb{R}$. Llamaremos $\widehat{L}_{N,\varepsilon}$ la función que describe la frontera inferior de la banda, es decir

$$\widehat{L}_{N,\varepsilon}(x) := - \sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}} \left[-\mathbb{1}_{\{\xi \leqslant x\}}(\xi) \right].$$

Entonces, por el Teorema 3.3.1 considerando $p = 1$ y d como la métrica en $\mathbb{R}$ inducida por el valor absoluto se obtiene lo siguiente

$$\widehat{L}_{N,\varepsilon}(x) := -\begin{cases} \inf\limits_{\lambda,s} & \lambda\varepsilon + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{sujeto a} & \sup\limits_{\xi\in\mathbb{R}} \left(-\mathbb{1}_{\{\xi\leqslant x\}}(\xi) - \lambda\left|\xi - \widehat{\xi}_i\right|\right) \leqslant s_i \quad \forall i = 1,\ldots,N, \\ & \lambda \geqslant 0. \end{cases}$$

Pero la función $-\mathbb{1}_{\{\xi\leqslant x\}}$ se puede expresar como $-\mathbb{1}_{\{\xi\leqslant x\}}(\xi) = \max\{f_1(\xi), f_2(\xi)\}$ donde f_1 y f_2 son definidas como $f_1(\xi) := -\mathcal{X}_{\{\xi>x\}}$ y $f_2(\xi) := -1$, donde $\mathcal{X}_A$ la definimos en el caso anterior. Entonces, bajo estas consideraciones se sigue

$$\begin{aligned} \widehat{L}_{N,\varepsilon}(x) &:= -\begin{cases} \inf\limits_{\lambda,s} & \lambda\varepsilon + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{sujeto a} & \sup\limits_{\xi\in\mathbb{R}} \left(\max\{f_1(\xi), f_2(\xi)\} - \lambda\left|\xi - \widehat{\xi}_i\right|\right) \leqslant s_i \quad \forall i = 1,\ldots,N, \\ & \lambda \geqslant 0. \end{cases} \\ &= -\begin{cases} \inf\limits_{\lambda,s} & \lambda\varepsilon + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{sujeto a} & \sup\limits_{\xi\in\mathbb{R}} \left(f_1(\xi) - \lambda\left|\xi - \widehat{\xi}_i\right|\right) \leqslant s_i \quad \forall i = 1,\ldots,N, \\ & \sup\limits_{\xi\in\Xi} \left(f_2(\xi) - \lambda\left|\xi - \widehat{\xi}_i\right|\right) \leqslant s_i \quad \forall i = 1,\ldots,N, \\ & \lambda \geqslant 0. \end{cases} \\ &= -\begin{cases} \inf\limits_{\lambda,s} & \lambda\varepsilon + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{sujeto a} & \sup\limits_{\xi\in\mathbb{R}} \left(f_1(\xi) - \lambda\left|\xi - \widehat{\xi}_i\right|\right) \leqslant s_i \quad \forall i = 1,\ldots,N, \\ & -1 \leqslant s_i \qquad\qquad\qquad\qquad \forall i = 1,\ldots,N, \\ & \lambda \geqslant 0. \end{cases} \end{aligned}$$

De nuevo, ya que $|\cdot|$ se puede expresar como $\lambda|\xi| = \max_{|z|_*\leqslant\lambda}\{z\xi\}$ y como la norma dual $|\cdot|_*$ de $|\cdot|$ es $|\cdot|$ entonces $\widehat{F}_I$ se puede expresar como

$$\widehat{L}_{N,\varepsilon}(x) := -\begin{cases} \inf\limits_{\lambda,s} & \lambda\varepsilon + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{sujeto a} & \sup\limits_{\xi\in\mathbb{R}} \left(f_1(\xi) - \max\limits_{|z_i|\leqslant\lambda}\left\{z_i\left(\xi - \widehat{\xi}_i\right)\right\}\right) \leqslant s_i \quad \forall i = 1,\ldots,N, \\ & -1 \leqslant s_i \qquad\qquad\qquad\qquad\qquad\qquad \forall i = 1,\ldots,N, \\ & \lambda \geqslant 0. \end{cases}$$

$$= - \left\{ \begin{array}{lll} \inf\limits_{\lambda,s} & \lambda\varepsilon + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i & \\ \text{sujeto a} & \sup\limits_{\xi\in\mathbb{R}} \min\limits_{|z_i|\leqslant\lambda} \left(f_1(\xi) - z_i \left(\xi - \widehat{\xi}_i \right) \right) \leqslant s_i & \forall i = 1, \ldots, N, \\ & -1 \leqslant s_i & \forall i = 1, \ldots, N, \\ & \lambda \geqslant 0. & \end{array} \right.$$

El Corolario 2.2.1.2 permite intercambiar el supremo y el mínimo de la primera restricción lo que conduce a lo siguiente:

$$\widehat{L}_{N,\varepsilon}(x) := - \left\{ \begin{array}{lll} \inf\limits_{\lambda,s} & \lambda\varepsilon + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i & \\ \text{sujeto a} & \min\limits_{|z_i|\leqslant\lambda} \sup\limits_{\xi\in\mathbb{R}} \left(f_1(\xi) - z_i \left(\xi - \widehat{\xi}_i \right) \right) \leqslant s_i & \forall i = 1, \ldots, N, \\ & -1 \leqslant s_i & \forall i = 1, \ldots, N, \\ & \lambda \geqslant 0. & \end{array} \right.$$

$$= \left\{ \begin{array}{lll} \inf\limits_{\lambda,s,z} & \lambda\varepsilon + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i & \\ \text{sujeto a} & \sup\limits_{\xi\in\mathbb{R}} \left(f_1(\xi) - z_i \left(\xi - \widehat{\xi}_i \right) \right) \leqslant s_i & \forall i = 1, \ldots, N, \\ & |z_i| \leqslant \lambda & \forall i = 1, \ldots, N, \\ & -1 \leqslant s_i & \forall i = 1, \ldots, N, \\ & \lambda \geqslant 0. & \end{array} \right.$$

Pero el supremo de la primera restricción se puede expresar como

$$\sup_{\xi\in\mathbb{R}} \left(f_1(\xi) - z_i \left(\xi - \widehat{\xi}_i \right) \right) = \left\{ \begin{array}{ll} -z_i \left(\xi - \widehat{\xi}_i \right) & \text{si } z_i \in [0, \lambda] \\ \infty & \text{si } z_i \in [-\lambda, 0]. \end{array} \right.$$

Por lo tanto, realizando el cambio de variable $w_i = -z_i$ tenemos

$$\widehat{L}_{N,\varepsilon}(x) := - \left\{ \begin{array}{lll} \inf\limits_{\lambda,s,w} & \lambda\varepsilon + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i & \\ \text{sujeto a} & w_i \left(x - \widehat{\xi}_i \right) \leqslant s_i & \forall i = 1, \ldots, N, \\ & -\lambda \leqslant w_i \leqslant 0 & \forall i = 1, \ldots, N, \\ & s_i \geqslant -1 & \forall i = 1, \ldots, N, \\ & \lambda \geqslant 0. & \end{array} \right.$$

$$= \begin{cases} \sup\limits_{\lambda,s,w} & \lambda\varepsilon + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{sujeto a} & w_i\left(x - \widehat{\xi}_i\right) \leqslant s_i \quad \forall i = 1,\ldots,N, \\ & -\lambda \leqslant w_i \leqslant 0 \qquad \forall i = 1,\ldots,N, \\ & s_i \geqslant -1 \qquad\quad \forall i = 1,\ldots,N, \\ & \lambda \geqslant 0. \end{cases}$$

Así pues, $\widehat{L}_{N,\varepsilon}(x)$ es el valor óptimo de un problema de optimización convexa finito, siendo esta la formulación deseada.

Elección de ε

Antes de proceder con la elección adecuada de $\varepsilon > 0$ debemos determinar la confianza que tiene un ε dado, este se determina mediante los siguientes pasos:

1. Se estima x_m la media muestral de ξ, es decir, $x_m = \frac{1}{N}\sum_{i=1}^{N}\widehat{\xi}_i$.

2. Se divide la muestra $\widehat{\xi}_1,\ldots,\widehat{\xi}_N$ en dos partes, una de tamaño $N_T < N$ que llamaremos conjunto de entrenamiento, los elementos de este conjunto son tomados aleatoriamente, la otra parte será el conjunto conformado por los elementos restantes, este conjunto sera de tamaño $N_V = N - N_T$ y lo llamaremos conjunto de validación. Procuramos tomar $N_T > N/2$.

3. Usando el conjunto de entrenamiento calculamos $\widehat{U}_{N_T,\varepsilon}(x_m)$ y $\widehat{L}_{N_T,\varepsilon}(x_m)$.

4. Usando el conjunto de validación calculamos $\widehat{\mathbb{E}}_{N_v}$ la versión muestral de $\mathbb{E}_{\mathbb{P}}[\mathbb{1}_{\{\xi\leqslant x\}}(\xi)]$, es decir, si $\left\{\widehat{\xi}_{i_i},\ldots,\widehat{\xi}_{i_{N_V}}\right\}$ es el conjunto de validación, entonces $\widehat{\mathbb{E}}_{N_v} = \dfrac{1}{N_V}\displaystyle\sum_{j=1}^{N_V}\mathbb{1}_{\{\xi\leqslant x\}}(\widehat{\xi}_{i_j})$.

5. Fijamos un número natural $K > 0$ suficientemente grande y fijamos N_T, con esta consideración repetimos K veces los pasos 2, 3 y 4, contamos el numero de las veces en las que se obtuvo $\widehat{L}_{N_T,\varepsilon}(x_m) \leqslant \widehat{\mathbb{E}}_{N_v} \leqslant \widehat{U}_{N_T,\varepsilon}(x_m)$. El porcentaje respecto a K que representa el numero de veces que so obtuvo lo anterior es a lo que denominamos la confianza de ε y la denotamos $1 - \beta_\varepsilon$.

Dado un nivel de confianza $1-\beta$ decimos que ε es una *elección ideal* si es el menor $\varepsilon > 0$ tal que su confianza es $1 - \beta_\varepsilon$ es $1 - \beta$, para elegir dicho ε generamos un conjunto $E := \{\varepsilon_1, \varepsilon_2 \ldots, \varepsilon_M\}$ de posibles candidatos a ser ε, por ejemplo, $E = \left\{0, \frac{1}{10}, \frac{2}{10}, \ldots, 2\right\}$, lo ideal es considerar E como una partición de un intervalo. Por cada $\varepsilon \in E$ calculamos su

confianza $1-\beta_\varepsilon$ siguiendo los pasos descritos anteriormente. El mejor ε en E será aquel que tenga la confianza $1-\beta_\varepsilon$ mas cercana a la confianza deseada $1-\beta$.
Si E es una partición de un intervalo entonces entre mas fina sea la partición más cercano sera el mejor ε en E al de la elección adecuada, ademas, entre mas grande sea ε su confianza $1-\beta_\varepsilon$ tendera a 1, pero la banda sera mas ancha lo cual no es deseable.
La Figura 4-1 ilustra la confianza obtenida para cada ε en E siendo este último una partición de 50 puntos del intervalo $[0, 0.08]$; el procedimiento anterior es recreado para cada ε en E donde la muestra es generada por una distribución normal con media igual a 9 y varianza igual a 2,25, además se consideró $N = 500$, $N_T = 300$ y $K = 50$. El valor de ε ideal que produce el procedimiento anterior para este caso es $\varepsilon = 0{,}026$, la gráfica ilustra la tendencia antes mencionada, a medida que crece ε la confianza es prácticamente 1.

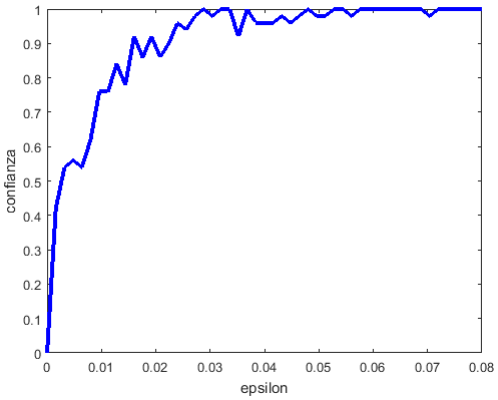
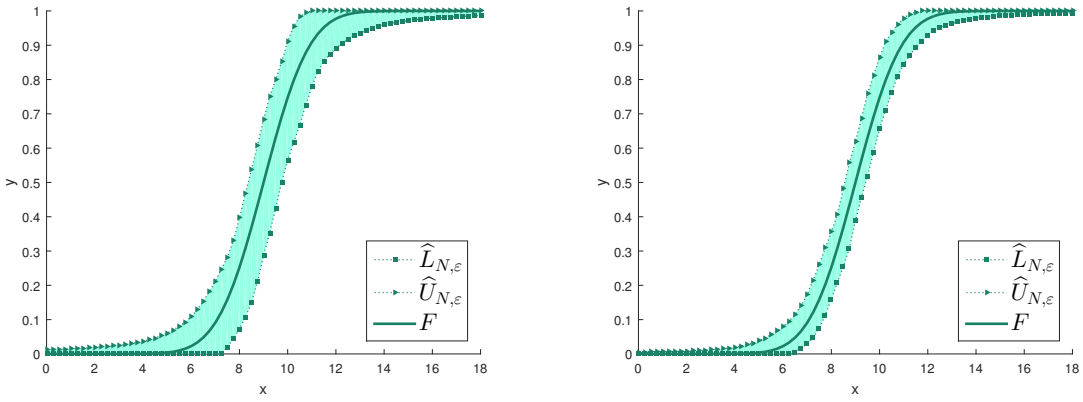


Figura 4-1: Gráfica de ε frente a la confianza para una muestra de tamaño $N = 500$ de una distribución normal con media 9 y varianza 2,25.

Resultados

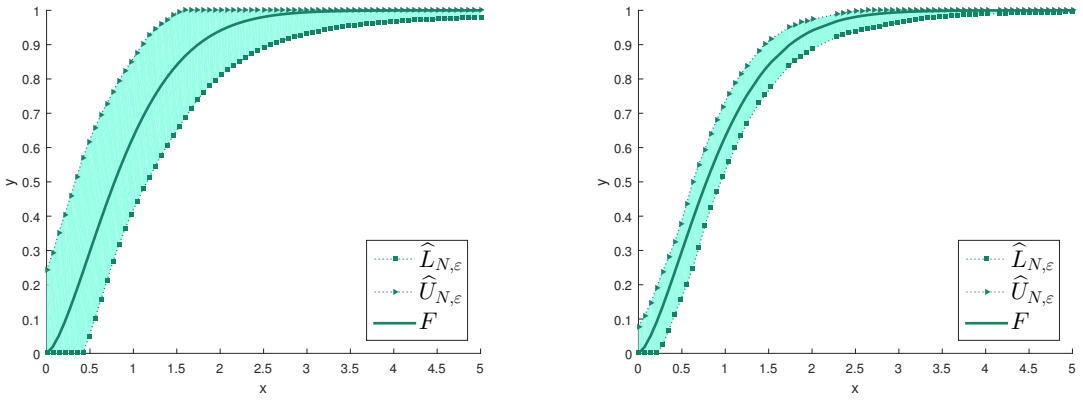
Las siguientes gráficas corresponden a simulaciones hechas para diferentes muestras de distribuciones conocidas, cada una de las bandas contiene el grafo de la función de distribución verdadera con un nivel de confianza de 0,95, las funciones $\widehat{L}_{N,\varepsilon}(x)$ y $\widehat{U}_{N,\varepsilon}(x)$ fueron calculadas usando el solver para problemas de optimización convexa CVX[3]. El valor de ε para cada N fue calculado siguiendo los pasos expuestos anteriormente, en las siguientes gráficas denotamos por F la función de distribución acumulativa verdadera.

[3]CVX es un solver para problemas de optimización convexa creado por Michael C. Grant y Stephen P. Boyd, este solver se soporta en MatLab.



(a) (b)

Figura 4-2: Banda de confianza para una distribución normal con media 9 y varianza 1,5. El tamaño de la muestra y el valor de ε de (a) es $N = 100$, $\varepsilon = 0{,}07$, de (b) es $N = 500$ y $\varepsilon = 0{,}026$.



(a) (b)

Figura 4-3: Banda de confianza para una distribución Weibull con parámetro de escala 1 y de forma 1,5. El tamaño de la muestra y el valor de ε de (a) es $N = 100$, $\varepsilon = 0{,}053$, de (b) es $N = 500$ y $\varepsilon = 0{,}009$.

En la Figura **4-2** se presenta el caso en el que la muestra es generada por una distribución normal con media 9 y varianza 2,25, mientras que en la Figura **4-3** se presenta el caso en el que la muestra es generada por una distribución Weibull de parámetro de escala 1 y de forma 1,5. De ambos casos se observa que la banda se encoje a medida que la muestra se hace mas grande, otro aspecto que también depende de N es ε, entre mas grande N

menor sera ε, esto también repercute en el ancho de la banda.
Desde una perspectiva estadística y contando con acceso a muestras de tamaño considerable, la forma de la banda puede dar indicios acerca de la forma de la función de distribución acumulativa verdadera y así poder catalogar la distribución de la variable aleatoria dentro una familia de distribuciones cuya función de distribución acumulada tengan formas similares a la de la banda, esto podría mejorar pruebas estadísticas que se fundamentan en supuesto acerca de la distribución de la variable aleatoria.

4.2. Estimación por núcleos de funciones de densidad

En esta sección exploraremos el método de estimación de densidades por núcleos, este método consiste en estimar la función de densidad de una variable aleatoria, función que es desconocida, a partir de una función de densidad conocida y una muestra de la variable aleatoria. Sea ξ una variable aleatoria y f su función de densidad, dada una muestra $\widehat{\xi}_1, \ldots, \widehat{\xi}_N$ este método de estimación propone el estimador de f dado por

$$\widehat{f}_h(x) := \frac{1}{Nh} \sum_{i=1}^{N} \mathcal{K}\left(\frac{x - \widehat{\xi}_i}{h}\right).$$

Este estimador depende del parámetro $h > 0$ comúnmente conocido como *ancho de banda* o *parámetro de suavidad*, también depende de $\mathcal{K}$ la cual es una función de densidad simétrica al rededor de cero y tal que $\int x^2 \mathcal{K}(x)dx = 1$, estas dos condiciones garantizan que si una variable aleatoria tiene como función de densidad a $\mathcal{K}$ entonces la varianza de dicha variable es uno y su valor esperado es cero. A la función $\mathcal{K}$ se le llama *núcleo* (kernel). Frecuentemente se considera $\mathcal{K}$ como la función de densidad normal estándar, no obstante $\mathcal{K}$ puede ser cualquier función de densidad que satisfaga las condiciones de simetría y varianza.

El parámetro h es fundamental en este método, este se puede elegir de dos formas, la primera consiste en elegir $h > 0$ de tal manera que minimice el *error cuadrático integrado* (integrate square error)

$$\mathrm{ISE}(h) := \int \left(\widehat{f}_h(x) - f(x)\right)^2 dx. \tag{4-1}$$

El h producto de esta elección dependerá de la muestra.
La otra forma consiste en encontrar un $h > 0$ tal que minimice el *error cuadrático integrado medio* (mean integrated squared error)

$$\mathrm{MISE}(h) := \mathbb{E}_{\mathbb{P}^N}\left[\int \left(\widehat{f}_h(x) - f(x)\right)^2 dx\right]. \tag{4-2}$$

La aleatoriedad en esta expresión esta en el vector aleatorio $\left(\widehat{\xi}_1, \ldots, \widehat{\xi}_N\right)$ cuya distribución es $\mathbb{P}^N = \mathbb{P} \times \cdots \times \mathbb{P}$, donde $\mathbb{P}$ es la distribución inducida por la función de densidad f, de modo que $\mathbb{P}^N$ también es desconocida. El h producto de esta elección no depende de de la muestra.
Ambas formas de elegir h comparten el mismo inconveniente, el desconocimiento de f, en la practica, de acuerdo a [17], apuntar al ancho de banda óptimo de ISE es más apropiado que apuntar al ancho de banda óptimo de MISE. Existen varios métodos para estimar h, están los que intentan minimizar MISE, estos en realidad minimizan una versión asintótica de esta expresión conocida como AMISE que proviene de una expansión en series de Taylor de f, para esto requieren del conocimiento de algunas derivadas de f, de modo que estos métodos se enfocan en estimar esas derivadas, conociendo esas derivadas en [31] se da una expresión para el h que minimiza AMISE, y bajo el supuesto de que f y $\mathcal{K}$ son funciones de densidad normales en [6] se da una expresión de h que minimiza MISE. Otros métodos se concentran en minimizar ISE, siendo el mas conocido y a la vez más usado el propuesto en [5] el cual se basa en el principio de validación cruzada.

En esta parte del trabajo exhibiremos que los problemas de optimización que constituyen el hecho de minimizar ISE y MISE respectivamente se pueden formular como un problema de optimización estocástica cada uno, de modo que dichos problemas se pueden aproximar por su versión robusta distribucional DROW.

Minimización de ISE como un DROW

Por la linealidad de la integral la expresión (4-1) se puede reescribir como

$$\begin{aligned}\mathrm{ISE}(h) &:= \int \left(\widehat{f}_h(x)\right)^2 dx - 2\int \widehat{f}_h(x) f(x) dx + \int \left(f(x)\right)^2 dx \\ &= \int \left(\widehat{f}_h(x)\right)^2 dx - 2\mathbb{E}_{\mathbb{P}}\left[\widehat{f}_h(\xi)\right] + \int \left(f(x)\right)^2 dx\end{aligned}$$

Considerando $\overline{J}$ la parte que depende de h en la expresión anterior, es decir,

$$\overline{J}(h) := \int \left(\widehat{f}_h(x)\right)^2 dx - 2\mathbb{E}_{\mathbb{P}}\left[\widehat{f}_h(\xi)\right], \tag{4-3}$$

entonces se infiere que h minimiza la expresión ISE si y solo si minimiza $\overline{J}$, de modo que enfocamos nuestra atención en minimizar $\overline{J}$, minimizar $\overline{J}$ ya es un problema de optimización estocástica de la forma (3-5), por lo tanto, se pueden emplear las técnicas de optimización robusta distribucional con métrica de Wasserstein vistas en el capitulo

anterior. Por un lado, el problema de interés es

$$\overline{J}^* := \min_{h\geqslant 0} \overline{J}(h) = \min_{h\geqslant 0} \left(\int \left(\widehat{f}_h(x) \right)^2 dx - 2\mathbb{E}_{\mathbb{P}} \left[\widehat{f}_h(\xi) \right] \right) \tag{4-4}$$

donde $\mathbb{P}$ es desconocida, como se mencionó en el capítulo anterior, tal desconocimiento motiva abordar este problema desde la perspectiva de la Optimización Robusta Distribucional, específicamente como un DROW, con el fin de lograr una estimación superior del valor óptimo del problema, que en este caso notamos J^*, en ese sentido, dada una muestra $\widehat{\xi}_1, \ldots, \widehat{\xi}_N$ de ξ, tal aproximación superior es el valor óptimo del problema

$$\widehat{\overline{J}}_N := \min_{h\geqslant 0} \left(\int \left(\widehat{f}_h(x) \right)^2 dx + 2 \sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}} \left[-\widehat{f}_h(\xi) \right] \right). \tag{4-5}$$

Donde $\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ es la bola respecto a una métrica p-Wasserstein de radio $\varepsilon > 0$ y centro en la distribución empírica $\widehat{\mathbb{P}}_N$, esta ultima es construida respecto la muestra. Vamos a considerar $\mathcal{K}$ como una función de densidad acotada, esto implica que $-\widehat{f}_h(\xi)$ es acotada, entonces satisface la Suposición 3.3.1 de modo que se satisfacen las hipótesis del Teorema 3.3.1, luego, considerando $\Xi = \mathbb{R}$, $p = 2$ y d como la distancia euclidiana, dicho Teorema permite reescribir el problema interno $\sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}} \left[-\widehat{f}_h(\xi) \right]$ como

$$\begin{cases} \inf\limits_{\lambda,s} & \lambda\varepsilon^p + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{sujeto a} & \sup\limits_{\xi\in\mathbb{R}} \left(-\widehat{f}_h(\xi) - \lambda \left(\xi - \widehat{\xi}_i \right)^2 \right) \leqslant s_i \quad \forall i = 1, \ldots, N, \\ & \lambda \geqslant 0. \end{cases} \tag{4-6}$$

Por lo tanto, (4-5) es equivalente al problema de optimización semi-infinito

$$\begin{cases} \inf\limits_{h,\lambda,s} & \displaystyle\int \left(\widehat{f}_h(x) \right)^2 dx + 2\lambda\varepsilon^p + \frac{2}{N}\sum_{i=1}^{N} s_i \\ \text{sujeto a} & \sup\limits_{\xi\in\mathbb{R}} \left(-\widehat{f}_h(\xi) - \lambda \left(\xi - \widehat{\xi}_i \right)^2 \right) \leqslant s_i \quad \forall i = 1, \ldots, N, \\ & \lambda \geqslant 0. \\ & h > 0. \end{cases} \tag{4-7}$$

La integral $\int \left(\widehat{f}_h(x) \right)^2 dx$ se puede calcular explícitamente dependiendo del núcleo $\mathcal{K}$ que se use. Por ejemplo, consideremos $\mathcal{K}$ como la función de densidad dada por

$$\mathcal{K}(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}. \tag{4-8}$$

Esta última es la función de densidad de una variable aleatoria que se distribuye normal estándar, esta es la caracterización de $\mathcal{K}$ mas empleada. Entonces se tiene

$$\int \left(\widehat{f}_h(x)\right)^2 dx = \frac{1}{N^2 h 2\sqrt{\pi}} \sum_{i=1}^{N}\sum_{j=1}^{N} e^{-\frac{(\hat{\xi}_i - \hat{\xi}_j)^2}{4h^2}}.$$

Minimización de MISE como un DROW

Por la linealidad del valor esperado la expresión (4-2) se pude reformular de la siguiente manera:

$$\text{MISE}(h) = \mathbb{E}_{\mathbb{P}^N}\left[\int \left(\widehat{f}_h(x)\right)^2 dx\right] - 2\mathbb{E}_{\mathbb{P}^N}\left[\int \widehat{f}_h(x) f(x) dx\right] + \int (f(x))^2 dx.$$

Considerando J la parte que depende de h en la expresión anterior, es decir

$$J(h) := \mathbb{E}_{\mathbb{P}^N}\left[\int \left(\widehat{f}_h(x)\right)^2 dx\right] - 2\mathbb{E}_{\mathbb{P}^N}\left[\int \widehat{f}_h(x) f(x) dx\right], \tag{4-9}$$

entonces, al igual que con ISE, se infiere que h minimiza la expresión MISE si y solo si minimiza J, de modo que enfocamos nuestra atención en minimizar J, para tal fin intentaremos reescribir J de tal manera que minimizar J sea un problema de optimización estocástica de la forma (3-5) y así poder emplear las técnicas de optimización robusta distribucional con métrica de Wasserstein vistas en el capitulo anterior. En ese sentido, por un lado se observa que

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}^N}\left[\int \left(\widehat{f}_h(x)\right)^2 dx\right] &= \int \cdots \int\int \left(\widehat{f}_h(x)\right)^2 dx \mathbb{P}(d\xi_1)\cdots\mathbb{P}(d\xi_N) \\
&= \int \cdots \int\int \tfrac{1}{N^2h^2}\sum_{i=1}^{N}\sum_{j=1}^{N} \mathcal{K}\left(\tfrac{x-\xi_i}{h}\right)\mathcal{K}\left(\tfrac{x-\xi_j}{h}\right) dx \mathbb{P}(d\xi_1)\cdots\mathbb{P}(d\xi_N) \\
&= \tfrac{1}{N^2h^2}\sum_{i=1}^{N}\sum_{j=1}^{N} \int \cdots \int\int \mathcal{K}\left(\tfrac{x-\xi_i}{h}\right)\mathcal{K}\left(\tfrac{x-\xi_j}{h}\right) dx \mathbb{P}(d\xi_1)\cdots\mathbb{P}(d\xi_N) \\
&= \tfrac{1}{Nh^2}\int\int \left(\mathcal{K}\left(\tfrac{x-\xi}{h}\right)\right)^2 dx \mathbb{P}(d\xi) \\
&\quad + \tfrac{N-1}{Nh^2}\int\int\int \mathcal{K}\left(\tfrac{x-\xi}{h}\right)\mathcal{K}\left(\tfrac{x-\zeta}{h}\right) dx \mathbb{P}(d\xi)\mathbb{P}(d\zeta) \\
&= \tfrac{1}{Nh^2}\mathbb{E}_{\mathbb{P}\sim\xi}\left[\int \left(\mathcal{K}\left(\tfrac{x-\xi}{h}\right)\right)^2 dx\right] + \tfrac{N-1}{Nh^2}\mathbb{E}_{\mathbb{P}\times\mathbb{P}\sim(\xi,\zeta)}\left[\int \mathcal{K}\left(\tfrac{x-\xi}{h}\right)\mathcal{K}\left(\tfrac{x-\zeta}{h}\right) dx\right] \\
&= \mathbb{E}_{\mathbb{P}\times\mathbb{P}\sim(\xi,\zeta)}\left[\tfrac{1}{Nh^2}\int \left(\mathcal{K}\left(\tfrac{x-\xi}{h}\right)\right)^2 dx + \tfrac{N-1}{Nh^2}\int \mathcal{K}\left(\tfrac{x-\xi}{h}\right)\mathcal{K}\left(\tfrac{x-\zeta}{h}\right) dx\right].
\end{aligned}$$

Por otro lado, se tiene

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}^N}\left[\mathbb{E}_{\mathbb{P}}\left[\widehat{f}_h(x)\right]\right] &= \int\cdots\int\int \widehat{f}_h(x)f(x)dx\mathbb{P}(d\xi_1)\cdots\mathbb{P}(d\xi_N)\\
&= \int\cdots\int\int \frac{1}{Nh}\sum_{i=1}^{N}\mathcal{K}\left(\frac{x-\xi_i}{h}\right)f(x)dx\mathbb{P}(d\xi_1)\cdots\mathbb{P}(d\xi_N)\\
&= \frac{1}{Nh}\sum_{i=1}^{N}\int\cdots\int\int \mathcal{K}\left(\frac{x-\xi_i}{h}\right)f(x)dx\mathbb{P}(d\xi_1)\cdots\mathbb{P}(d\xi_N)\\
&= \frac{1}{Nh}\sum_{i=1}^{N}\int\int \mathcal{K}\left(\frac{x-\xi_i}{h}\right)f(x)dx\mathbb{P}(d\xi_i)\\
&= \frac{1}{h}\int\int \mathcal{K}\left(\frac{x-\xi}{h}\right)f(x)dx\mathbb{P}(d\xi)\\
&= \mathbb{E}_{\mathbb{P}\times\mathbb{P}\sim(\xi,\zeta)}\left[\frac{1}{h}\mathcal{K}\left(\frac{\zeta-\xi}{h}\right)\right].
\end{aligned}$$

Por lo tanto, la expresión que define J dada en (4-9) es igual a

$$J(h) = \mathbb{E}_{\mathbb{P}\times\mathbb{P}\sim(\xi,\zeta)}\left[\frac{1}{Nh^2}\int\left(\mathcal{K}\left(\frac{x-\xi}{h}\right)\right)^2 dx + \frac{N-1}{Nh^2}\int\mathcal{K}\left(\frac{x-\xi}{h}\right)\mathcal{K}\left(\frac{x-\zeta}{h}\right)dx - \frac{2}{h}\mathcal{K}\left(\frac{\zeta-\xi}{h}\right)\right]. \quad (4\text{-}10)$$

Para efectos de notación definimos

$$F(h,\xi,\zeta) := \frac{1}{Nh^2}\int\left(\mathcal{K}\left(\frac{x-\xi}{h}\right)\right)^2 dx + \frac{N-1}{Nh^2}\int\mathcal{K}\left(\frac{x-\xi}{h}\right)\mathcal{K}\left(\frac{x-\zeta}{h}\right)dx - \frac{2}{h}\mathcal{K}\left(\frac{\zeta-\xi}{h}\right).$$

De modo que minimizar $J(h)$ es en realidad un problema de optimización estocástica expresado como

$$J^* := \min_{h\geqslant 0} J(h) = \min_{h\geqslant 0}\mathbb{E}_{\mathbb{P}\times\mathbb{P}\sim(\xi,\zeta)}\left[F(h,\xi,\zeta)\right] \quad (4\text{-}11)$$

donde $\mathbb{P}$ es desconocida, de nuevo, tal desconocimiento motiva abordar este problema desde la perspectiva de la Optimización Robusta Distribucional, específicamente como un DROW, con el fin de lograr una estimación superior del valor óptimo del problema, que en este caso notamos J^*, en ese sentido tal aproximación superior es el valor óptimo del problema

$$\widehat{J}_N := \min_{h\geqslant 0}\sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)}\mathbb{E}_{\mathbb{Q}}[F(h,\xi,\zeta)]. \quad (4\text{-}12)$$

Donde $\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ es la bola respecto a la métrica 2-Wasserstein de radio $\varepsilon > 0$ y centro en la distribución empírica $\widehat{\mathbb{P}}_N$, esta ultima es construida respecto a una muestra del vector aleatorio (ξ,ζ) el cual tiene distribución $\mathbb{P}\times\mathbb{P}$, de esto se sigue que ζ se distribuye $\mathbb{P}$, de modo

que si $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_N, \widehat{\xi}_1, \ldots, \widehat{\xi}_{2N}$ es una muestra de ξ, entonces $(\widehat{\xi}_1, \widehat{\xi}_{N+1}), \ldots, (\widehat{\xi}_N, \widehat{\xi}_{2N})$ es una muestra de tamaño N de (ξ, ζ) lo que permite definir

$$\widehat{\mathbb{P}}_N = \sum_{i=1}^{N} \delta_{(\widehat{\xi}_i, \widehat{\xi}_{N+i})}.$$

En resumen, la tarea de encontrar el valor de h que minimiza la expresión MISE se traduce en encontrar el valor de h que es punto óptimo del problema (4-12), aunque no necesariamente es el mismo h si es una buena aproximación y cuenta con la ventaja que es encontrada sin asumir condiciones sobre la distribución de ξ, es el resultado de un análisis completamente no paramétrico.

El objetivo es reformular (4-12), para tal fin asumimos que $\mathcal{K}$ es acotada, en la practica siempre se usan kernel acotados, ademas, debido al desconocimiento de la distribución de ξ asumiremos que el soporte de $\mathbb{P} \times \mathbb{P}$ es $\mathbb{R}^2$, de estas consideraciones se obtiene la siguiente proposición:

Proposición 4.2.1. *Para cada $h > 0$ se considera la función $\Psi_h : \mathbb{R}^2 \to \mathbb{R}$ definida por $\Psi_h(\xi, \zeta) := F(h, \xi, \zeta)$ para cada $(\xi, \zeta) \in \mathbb{R}^2$. Entonces Ψ_h es acotada.*

Demostración. Ya que $\mathcal{K}$ es acotada, entonces el término $\mathcal{K}\left(\frac{\zeta - \xi}{h}\right)$ es acotado como función de (ξ, ζ), por lo tanto, es suficiente con demostrar que la expresión

$$\frac{1}{Nh^2} \int \left(\mathcal{K}\left(\frac{x-\xi}{h}\right)\right)^2 dx + \frac{N-1}{Nh^2} \int \mathcal{K}\left(\frac{x-\xi}{h}\right) \mathcal{K}\left(\frac{x-\zeta}{h}\right) dx \tag{4-13}$$

es acotada como función de (ξ, ζ). En efecto, para cada ξ se tiene que $\frac{1}{h}\mathcal{K}\left(\frac{x-\xi}{h}\right)$ es una función de densidad, entonces esta define una medida de probabilidad que notaremos $\mathbb{P}_\xi$ y que es definida para cualquier conjunto medible A como

$$\mathbb{P}_\xi(A) = \int_A \frac{1}{h} \mathcal{K}\left(\frac{x-\xi}{h}\right) dx.$$

Entonces, considerando δ una variable aleatoria con distribución $\mathbb{P}_\xi$ se infiere que (4-13) es igual a

$$\frac{1}{Nh} \mathbb{E}_{\mathbb{P}_\xi}\left[\mathcal{K}\left(\frac{\delta - \xi}{h}\right)\right] + \frac{N-1}{Nh} \mathbb{E}_{\mathbb{P}_\xi}\left[\mathcal{K}\left(\frac{\delta - \zeta}{h}\right)\right]. \tag{4-14}$$

Esta ultima expresión es acotada inferiormente por cero. Respecto a la cota superior, sea M la cota superior de $\mathcal{K}$, entonces por ser $\mathbb{P}_\xi$ una medida de probabilidad se concluye que (4-14) es acotada superiormente por

$$\frac{1}{Nh} M + \frac{N-1}{Nh} M.$$

□

De esta última proposición se sigue que $F(h,\cdot)$ es acotada, de modo que se satisface la Suposición 3.3.1 y por ende se satisfacen las hipótesis del Teorema 3.3.1, luego, considerando $p=2$ y d como la distancia euclidiana en $\mathbb{R}^2$ se sigue que el problema (4-12) es equivalente a

$$\begin{cases} \inf\limits_{h,\lambda,s} & \lambda\varepsilon^p + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{sujeto a} & \sup\limits_{(\xi,\zeta)\in\mathbb{R}^2}\left(F(h,\xi,\zeta) - \lambda\left\|(\xi,\zeta) - (\widehat{\xi}_i,\widehat{\xi}_{N+i})\right\|^2\right) \leqslant s_i \quad \forall i=1,\dots,N, \\ & \lambda \geqslant 0, \\ & h \geqslant 0, \end{cases} \tag{4-15}$$

donde $\|\cdot\|$ es la norma euclidiana en $\mathbb{R}^2$.
Ya que el núcleo más empleado es el proveniente de la distribución normal estándar, centraremos nuestra atención en núcleos $\mathcal{K}$ de la forma (4-8). Para esta caracterización de $\mathcal{K}$ se sigue

$$\int \left(\mathcal{K}\left(\frac{x-\xi}{h}\right)\right)^2 dx = \frac{1}{2\sqrt{\pi}h}$$

y

$$\int \mathcal{K}\left(\frac{x-\xi}{h}\right)\mathcal{K}\left(\frac{x-\zeta}{h}\right) dx = \frac{1}{2\sqrt{\pi}h} e^{-\frac{(\xi-\zeta)^2}{4h^2}}.$$

Por lo tanto, para este caso se tiene

$$F(h,\xi,\zeta) := \frac{1}{2\sqrt{\pi}Nh^3} + \frac{N-1}{2\sqrt{\pi}Nh^3} e^{-\frac{(\xi-\zeta)^2}{4h^2}} - \frac{2}{\sqrt{2\pi}h} e^{-\frac{(\xi-\zeta)^2}{2h^2}}.$$

En este caso es más evidente que $F(h,\cdot,\cdot):\mathbb{R}^2\to\mathbb{R}$ es una función acotada, aunque esto ya se tiene por la Proposición 4.2.1. La siguiente proposición establece una reformulación de (4-15) para este contexto.

Proposición 4.2.2. *Para el kernel normal estándar el problema (4-15) es equivalente a*

$$\begin{cases} \inf\limits_{h,\lambda,s} & \lambda\varepsilon^p + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \textit{sujeto a} & \sup\limits_{\xi\in\mathbb{R}}\left(F(h,\xi,-\xi+\widehat{\xi}_i+\widehat{\xi}_{N+i}) - 2\lambda(\xi-\widehat{\xi}_i)^2\right) \leqslant s_i \quad \forall i=1,\dots,N, \\ & \lambda \geqslant 0, \\ & h \geqslant 0. \end{cases} \tag{4-16}$$

Demostración. El objetivo es demostrar que

$$\sup_{(\xi,\zeta)\in\mathbb{R}^2}\left(F(h,\xi,\zeta) - \lambda\left\|(\xi,\zeta) - (\widehat{\xi}_i,\widehat{\xi}_{N+i})\right\|^2\right) = \sup_{\xi\in\mathbb{R}}\left(F(h,\xi,-\xi+\widehat{\xi}_i+\widehat{\xi}_{N+i}) - 2\lambda(\xi-\widehat{\xi}_i)^2\right).$$

En efecto, por un lado se tiene que

$$\sup_{(\xi,\zeta)\in\mathbb{R}^2}\left(F(h,\xi,\zeta)-\lambda\left\|(\xi,\zeta)-(\widehat{\xi}_i,\widehat{\xi}_{N+i})\right\|^2\right)=\sup_{c\in\mathbb{R}}\sup_{\xi\in\mathbb{R}}\left(F(h,\xi,\xi+c)-\lambda\left((\xi-\widehat{\xi}_i)^2+(\xi+c-\widehat{\xi}_{N+i})^2\right)\right)$$

Pero $F(h,\xi,\xi+c)$ ya no depende de ξ pues

$$g(c):=F(h,\xi,\xi+c)=\frac{1}{2\sqrt{\pi}Nh^3}+\frac{N-1}{2\sqrt{\pi}Nh^3}e^{-\frac{c^2}{4h^2}}-\frac{2}{\sqrt{2\pi}h}e^{-\frac{c^2}{2h^2}}.$$

Por lo tanto, se tiene

$$\sup_{(\xi,\zeta)\in\mathbb{R}^2}\left(F(h,\xi,\zeta)-\lambda\left\|(\xi,\zeta)-(\widehat{\xi}_i,\widehat{\xi}_{N+i})\right\|^2\right)=\sup_{c\in\mathbb{R}}\left(g(c)-\frac{\lambda}{2}\left(c+\widehat{\xi}_i-\widehat{\xi}_{N+i}\right)^2\right)$$

Realizando el cambio de variable $c=-2\xi+\widehat{\xi}_i+\widehat{\xi}_{N+i}$ se obtiene el resultado deseado. □

Progresos y limitantes numéricas

Asumiendo $\mathcal{K}$ como el núcleo normal estándar, todo lo expuesto anteriormente conduce a la necesidad de solucionar los problemas de optimización semi-infinita (4-7) y (4-16). Ante este escenario emergen diversas limitantes que impiden su calculo exacto, una forma de solucionarlo es por medio de algoritmos, no obstante, los algoritmos para solucionar problemas de optimización semi-infinita no son tan eficientes como se desea, en ese sentido, los resultados que se exponen en esta parte del trabajo son obtenidos mediante un procedimiento intuitivo e ilustrativo pero poco eficiente sobretodo para valores de ε del orden de 10^{-3}.

Centraremos nuestra atención en (4-7), para este problema estimamos su valor óptimo y solucion óptima, el procedimiento para estimar dichos valores se enfoca en tomar grillas de valores de λ y h, es decir, notemos que (4-7) se puede reescribir como

$$\inf_{h,\lambda}\left(\frac{1}{N^2h2\sqrt{\pi}}\sum_{i=1}^{N}\sum_{j=1}^{N}e^{-\frac{(\widehat{\xi}_i-\widehat{\xi}_j)^2}{4h^2}}+2\lambda\varepsilon^p+\frac{2}{N}\sum_{i=1}^{N}\sup_{\xi\in\mathbb{R}}\left(-\widehat{f}_h(\xi)-\lambda\left(\xi-\widehat{\xi}_i\right)^2\right)\right)\qquad(4\text{-}17)$$

Considerando

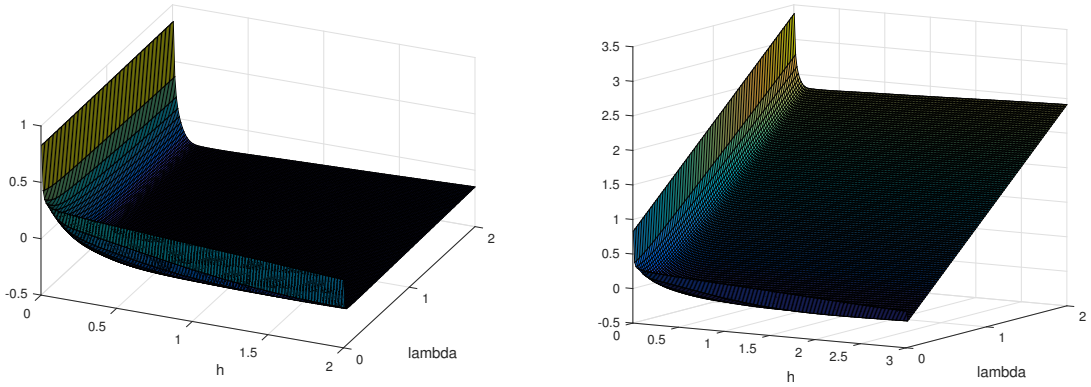
$$g_\varepsilon(h,\lambda):=\frac{1}{N^2h2\sqrt{\pi}}\sum_{i=1}^{N}\sum_{j=1}^{N}e^{-\frac{(\widehat{\xi}_i-\widehat{\xi}_j)^2}{4h^2}}+2\lambda\varepsilon^p+\frac{2}{N}\sum_{i=1}^{N}\sup_{\xi\in\mathbb{R}}\left(-\widehat{f}_h(\xi)-\lambda\left(\xi-\widehat{\xi}_i\right)^2\right)$$

la idea es considerar un grilla $\{(h_i,\lambda_j)\}_{\substack{k=1,\dots,M\\ j=1,\dots,L}}$ en $(0,h_s]\times[0,\lambda_s]\subset\mathbb{R}^2_+$ donde h_s y λ_s son valores preestablecidos, seleccionando una grilla de 10.000 puntos en la Figura 4-4

se ilustra la gráfica de g_ε en (a) para $\varepsilon = 0{,}1$ y $h_s = \lambda_s = 2$, y en (b) para $\varepsilon = 0{,}8$ y $h_s = 3$ y $\lambda_s = 2$. De estas gráficas se evidencia que la función objetivo g_ε es convexa, dicha percepción es reafirmada realizando el mismo procedimiento para varios valores de ε, no obstante, dicha percepción no puede ser justificada teóricamente, la dificultad esta en la expresión

$$\sup_{\xi\in\mathbb{R}} \left(-\widehat{f}_h(\xi) - \lambda \left(\xi - \widehat{\xi}_i \right)^2 \right)$$

que aparece en g_ε, para estimar este supremo se considera un intervalo $[a, b] \subset \mathbb{R}$ tal que la muestra $\widehat{\xi}_1, \ldots, \widehat{\xi}_N$ este contenida en $[a, b]$, se seleccionan b y a alejados del máximo y mínimo valor de la muestra respectivamente, luego se elige una partición finita $\{c_i\}_{i=1}^k$ de $[a, b]$, entonces una estimación del supremo anterior es $\max\limits_{j=1,\ldots,K} \left(-\widehat{f}_h(c_j) - \lambda \left(c_j - \widehat{\xi}_i \right)^2 \right)$.



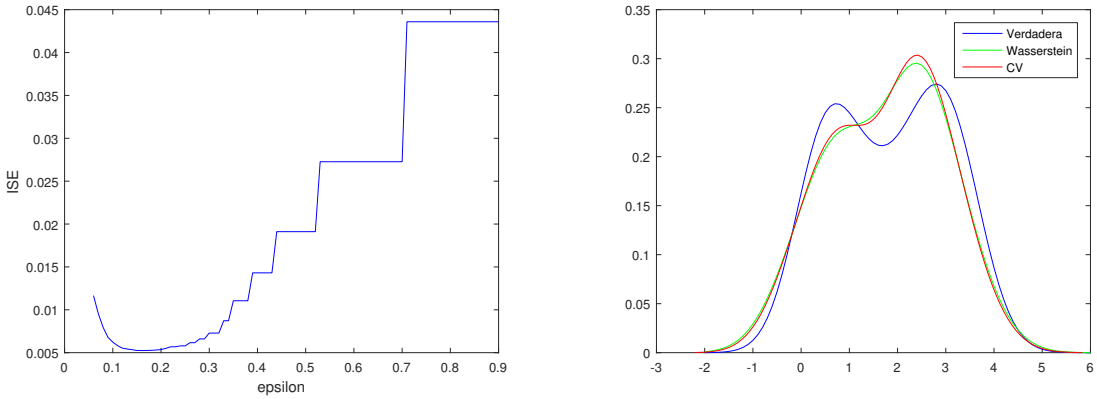
(a) (b)

Figura 4-4: Gráfica de la función g_ε para (a) $\varepsilon = 0{,}1$ y (b) es $\varepsilon = 0{,}8$.

Análogamente, el valor óptimo de (4-17) es estimado por $\max\limits_{\substack{k=1,\ldots,M \\ j=1,\ldots,L}} g_\varepsilon(h_k, \lambda_s)$. Esta forma de encontrar valores y soluciones óptimas es poco eficiente cuando ε es menor a 10^{-3} ya que λ toma valores significativamente grandes de tal manera que no hay una forma analítica para determinar en que parte de la recta real se ubica λ, conocer esta información permite saber en que parte se puede establecer la grilla.

El experimento anterior se realizo con una muestra de tamaño $N = 100$ generada por una distribución de dos picos, esta distribución se define a partir de dos muestras de dos variables aleatorias Weibull diferentes, es decir, estas muestras están ligeramente alejadas entre si, a partir de estas muestra se genera una función de densidad vía estimación por

kernels, esta función de densidad tiene dos picos.
La Figura **4-5** (a) expone la relación entre ISE y ε, esta gráfica se genera para la muestra descrita anteriormente, en esta figura se evidencia que a medida que ε toma valores cercanos a 1 ó a 0 entonces ISE crece, para este caso $\varepsilon^* = 0{,}16$ minimiza ISE.



(a) (b)

Figura 4-5: En (a) la gráfica de ε vs ISE. En (b) la representación de las funciones de densidad estimada vía Wasserstein, la verdadera y una estimada mediante validación cruzada que notamos como CV.

Para ε^* se tiene que $h^* = 0{,}604$ minimiza (4-17), la Figura **4-5** (b) ilustra la gráfica del estimador por kernels $\widehat{f}_{h^*}$ para h^*, también se exponen las las gratificas de la estimaciones por kernel para h elegido mediante validación cruzada y la gráfica de la función de densidad verdadera, se observa que el resultado es muy similar al obtenido por validación cruzada aunque tiene una leve diferencia, recordemos que la estimación por validación cruzada es la mas empleada en la practica.

4.3. El modelo de Markowitz robusto distribucional respecto a W_2 para optimización de portafolios

Un inversionista tiene un monto de dinero que desea invertir en m bienes, denotamos por ξ_i el retorno del bien i-ésimo y x_i la proporción del monto inicial invertida en el bien i-ésimo, en este sentido, dado que los retornos de cada bien son en la practica aleatorios entonces $\xi = (\xi_1, \ldots, \xi_m) \in \mathbb{R}^m$ es un *vector aleatorio*, adicionalmente, $x = (x_1, \ldots, x_m) \in \mathbb{R}^m$ es un vector conocido como el vector de pesos, este satisface la relación $\sum_{i=1}^m x_i = 1$. El conjunto de los m bienes se le conoce como el *portafolio* y la expresión $R := \sum_{i=1}^m x_i \xi_i$

es el *retorno total del portafolio*, considerando $\langle \cdot, \cdot \rangle$ el producto interno usual de $\mathbb{R}^m$ se infiere que $R = \langle x, \xi \rangle$. Dado que ξ es un vector aleatorio con distribución $\mathbb{P}$ entonces los siguientes conceptos son los que tomaran relevancia en esta sección:

$$\text{Retorno esperado} := \mathbb{E}_{\mathbb{P}}\left[\langle x, \xi \rangle\right].$$
$$\text{Volatilidad} := \text{Var}_{\mathbb{P}}\left[\langle x, \xi \rangle\right] = \mathbb{E}_{\mathbb{P}}\left[\left(\langle x, \xi \rangle - \mathbb{E}_{\mathbb{P}}\left[\langle x, \xi \rangle\right]\right)^2\right].$$

Lo ideal para un inversionista es encontrar un vector de pesos x que le garantice un retorno esperado alto pero con una volatilidad baja, en tal virtud el inversionista establece un nivel de retorno esperado mínimo μ, esto significa que los únicos vectores de pesos x que considerará son aquellos que garantizan un retorno esperado mayor o igual a μ. Esta visión es representada en el siguiente modelo:

$$J := \begin{cases} \min\limits_{x \in \mathbb{R}^m} & \text{Var}_{\mathbb{P}}\left[\langle x, \xi \rangle\right] \\ \text{sujeto a} & \mathbb{E}_{\mathbb{P}}\left[\langle x, \xi \rangle\right] \geqslant \mu, \\ & \sum_{i=1}^m x_i = 1. \end{cases} \tag{4-18}$$

Si se conociera la matriz de covarianza E y el vector de valores esperados $\mathbf{m}$ del vector aleatorio ξ entonces (4-18) es equivalente al problema de optimización

$$\begin{cases} \min\limits_{x \in \mathbb{R}^m} & x^T E x \\ \text{sujeto a} & \mathbf{m}^T x \geqslant \mu, \\ & \sum_{i=1}^m x_i = 1. \end{cases} \tag{4-19}$$

Pero en la practica E y $\mathbf{m}$ no son conocidos, ante esta situación es común considerar E y $\mathbf{m}$ como las versiones muéstrales. Para que (4-19) se pueda solucionar y tenga una solución óptima se requiere que E no sea singular, tal condición no se satisface en general ni para E ni para sus versiones muéstrales, ademas la versión muestral es muy inestable respecto a la muestra. Por lo tanto, un enfoque mas conservador es aproximar J superiormente por medio del valor optimo de un problema de Optimización Robusta Distribucional DROW modificado.

En el intento por formular una contraparte robusta para (4-18) siempre emerge el problema de estimar la varianza de una variable aleatoria sujeto a que conocemos su media, tal estimación se puede realizar de manera robusta, esto motiva a dedicar una subsección a este problema antes de avanzar.

Estimación DROW de la varianza de una variable aleatoria con media conocida

Sea ζ una variable aleatoria con distribución $\mathbb{P}$ desconocida soportada en $\Xi = \mathbb{R}$, asumimos que se conoce el valor esperado de ζ, es decir, se sabe que $\mathbb{E}_{\mathbb{P}}[\zeta] = \eta$ y ademas

se conoce una muestra $\widehat{\zeta}_1, \ldots, \widehat{\zeta}_N$ de ζ. Deseamos estimar superiormente la varianza de ζ de manera robusta, es decir, considerando $\widehat{\mathbb{P}}_N$ la distribución empírica inducida por la muestra anterior y $\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ la bola con centro en $\widehat{\mathbb{P}}_N$ y radio ε tomada respecto a la métrica 2-Wasserstein con función de costo d como la distancia euclidiana en $\mathbb{R}$, y eligiendo ε de manera adecuada de tal manera que $\mathbb{P} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ con alta probabilidad, entonces con alta probabilidad se tiene

$$\mathrm{Var}_{\mathbb{P}}[\xi] \leqslant \begin{cases} \sup\limits_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} & \mathbb{E}_{\mathbb{Q}}\left[(\zeta-\eta)^2\right] \\ \text{sujeto a} & \mathbb{E}_{\mathbb{Q}}\left[\zeta\right] = \eta. \end{cases} \tag{4-20}$$

Para ciertos valores de ε el problema de la derecha puede ser no factible, la siguiente proposición establece dichos valores.

Proposición 4.3.1. *Si* $\varepsilon < \left|\eta - \frac{1}{N}\sum_{i=1}^N \widehat{\zeta}_i\right|$ *entonces*

$$\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N) \cap \{\mathbb{Q} \in \mathcal{P}(\mathbb{R}) \mid \mathbb{E}_{\mathbb{Q}}[\zeta] = \eta\} = \phi.$$

Demostración. Sea $\mathbb{Q} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$, debemos demostrar que $\mathbb{E}_{\mathbb{Q}}[\zeta] \neq \eta$. En efecto, por la Observación 6.6[4] en [41] se sabe que

$$p \leqslant q \quad \Longrightarrow \quad W_p \leqslant W_q.$$

En particular se tiene $W_1 \leqslant W_2$, lo que implica que

$$W_1\left(\mathbb{Q}, \widehat{\mathbb{P}}_N\right) \leqslant W_2\left(\mathbb{Q}, \widehat{\mathbb{P}}_N\right) \leqslant \varepsilon < \left|\eta - \frac{1}{N}\sum_{i=1}^N \widehat{\zeta}_i\right|. \tag{4-21}$$

Luego, existe $\Pi \in \mathcal{S}(\mathbb{Q}, \widehat{\mathbb{P}}_N)$, siendo este último el conjunto de couplings entre $\mathbb{Q}$ y $\widehat{\mathbb{P}}_N$, tal que

$$\int_{\Xi\times\Xi} |\zeta - \delta| \Pi(d\xi, d\zeta) < \left|\eta - \frac{1}{N}\sum_{i=1}^N \widehat{\zeta}_i\right|.$$

Pero también se tiene

$$\int_{\Xi\times\Xi} \zeta \Pi(d\zeta, d\delta) = \int_{\Xi} \zeta \mathbb{Q}(d\zeta) = \mathbb{E}_{\mathbb{Q}}[\zeta] \quad \text{y} \quad \int_{\Xi\times\Xi} \delta \Pi(d\zeta, d\delta) = \int_{\Xi} \delta \widehat{\mathbb{P}}_N(d\delta) = \mathbb{E}_{\widehat{\mathbb{P}}_N}[\delta].$$

Entonces

$$\left|\mathbb{E}_{\mathbb{Q}}[\zeta] - \frac{1}{N}\sum_{i=1}^N \widehat{\zeta}_i\right| = \left|\int_{\Xi\times\Xi} (\zeta - \delta)\Pi(d\xi, d\zeta)\right| \leqslant \int_{\Xi\times\Xi} |\zeta - \delta| \Pi(d\xi, d\zeta).$$

[4] Esta es una consecuencia de la desigualdad de Hölder.

Por lo tanto, por lo último y (4-21) se tiene

$$\left|\mathbb{E}_{\mathbb{Q}}[\zeta] - \frac{1}{N}\sum_{i=1}^{N}\widehat{\zeta}_i\right| < \left|\eta - \frac{1}{N}\sum_{i=1}^{N}\widehat{\zeta}_i\right|.$$

De lo inmediatamente anterior y la desigualdad triangular inversa se sigue

$$|\eta - \mathbb{E}_Q[\zeta]| = \left|\eta - \tfrac{1}{N}\textstyle\sum_{i=1}^{N}\widehat{\zeta}_i - \left(\mathbb{E}_Q[\zeta] - \tfrac{1}{N}\textstyle\sum_{i=1}^{N}\widehat{\zeta}_i\right)\right| \geqslant \left|\left|\eta - \tfrac{1}{N}\textstyle\sum_{i=1}^{N}\widehat{\zeta}_i\right| - \left|\mathbb{E}_{\mathbb{Q}}[\zeta] - \tfrac{1}{N}\textstyle\sum_{i=1}^{N}\widehat{\zeta}_i\right|\right| > 0.$$

Lo que permite concluir $\mathbb{E}_Q[\zeta] \neq \eta$. □

En el siguiente Teorema se establece una expresión explícita para el valor óptimo del problema de optimización a la derecha de (4-20).

Teorema 4.3.1. *Sea* $\varepsilon > 0$ *con*[5] $\varepsilon^2 \geqslant \left(\frac{1}{N}\sum_{i=1}^{N}(\eta - \widehat{\zeta}_i)\right)^2$ *y tal que* $\mathbb{P} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ *con alta probabilidad. Entonces con alta probabilidad se tiene (4-20) y el valor óptimo del problema de optimización*

$$\begin{cases} \sup\limits_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} & \mathbb{E}_{\mathbb{Q}}\left[(\zeta-\eta)^2\right] \\ \textit{sujeto a} & \mathbb{E}_{\mathbb{Q}}\left[\zeta\right] = \eta. \end{cases} \tag{4-22}$$

es igual a

$$\left(\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i)^2 - \left(\frac{1}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i)\right)^2} + \sqrt{\varepsilon^2 - \left(\frac{1}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i)\right)^2}\right)^2.$$

Para demostrar el Teorema anterior requerimos de la siguiente proposición.

Proposición 4.3.2. *Sea* $\{a_i\}_{i=1}^N \subset \mathbb{R}$, *entonces*

$$\left(\sum_{i=1}^{N} a_i\right)^2 \leqslant N\sum_{i=1}^{N} a_i^2.$$

Demostración. Por la desigualdad de CauchySchwarz se tiene

$$\left|\sum_{i=1}^{N} a_i\right| = |\langle(1,\ldots,1),(a_1,\ldots,a_N)\rangle|$$

[5]Imponer esta condición sobre ε es natural ya que por la Proposición 4.3.1 garantiza que (4-22) sea factible.

$$\begin{aligned}&\leqslant \|(1,\ldots,1)\|_2 \, \|(a_1,\ldots,a_N)\|_2\\ &= \sqrt{N}\sqrt{\sum_{i=1}^{N} a_i^2}.\end{aligned}$$

Elevando al cuadrado ambos lados de la desigualdad anterior se obtiene el resultado esperado. □

Demostración del Teorema 4.3.1. Por el Teorema 3.3.2 se sigue que el lado derecho de (4-22) satisface dualidad fuerte y es equivalente al problema

$$\inf_{\beta} \sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}}\left[(\zeta-\eta)^2 - \beta\zeta + \beta\eta\right].$$

Ya que la función $g(\zeta) := (\zeta-\eta)^2 - \beta\zeta + \beta\eta$ satisface la Suposición 3.3.1 en su parte 1, entonces por el Teorema 3.3.1 esta última formulación es equivalente al problema de optimización semi-infinita

$$\begin{cases} \inf\limits_{\beta,\,\lambda,\,s_i} & \lambda\varepsilon^2 + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{sujeto a} & \sup\limits_{\zeta\in\mathbb{R}} \left((\zeta-\eta)^2 - \beta\zeta + \beta\eta - \lambda\left|\zeta-\widehat{\zeta}_i\right|^2\right) \leqslant s_i \quad \forall i = 1,\ldots,N \\ & \lambda \geqslant 0. \end{cases} \tag{4-23}$$

Si $\lambda \leqslant 1$ entonces λ no es un valor óptimo ya que el conjunto

$$\left\{(\zeta-\eta)^2 - \beta\zeta + \beta\eta - \lambda\left|\zeta-\widehat{\zeta}_i\right|^2 \mid \zeta\in\mathbb{R}\right\}$$

no será acotado, por otro lado, si $\lambda > 1$ pasa todo lo contrario y

$$\sup_{\zeta\in\mathbb{R}} \left((\zeta-\eta)^2 - \beta\zeta + \beta\eta - \lambda\left|\zeta-\widehat{\zeta}_i\right|^2\right)$$

puede ser determinado explícitamente ya que es el supremo de un polinomio cuadrático cóncavo, en ese sentido (4-23) es equivalente a

$$\begin{cases} \inf\limits_{\beta,\,\lambda,\,s_i} & \lambda\varepsilon^2 + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{sujeto a} & \dfrac{\beta^2}{4(\lambda-1)} + \dfrac{\lambda}{\lambda-1}\left(\beta(\eta-\widehat{\zeta}_i) + (\eta-\widehat{\zeta}_i)^2\right) \leqslant s_i \quad \forall i = 1,\ldots,N \\ & \lambda \geqslant 1. \end{cases} \tag{4-24}$$

En esta ultima formulación podemos suprimir las variables s_i lo que conduce al siguiente problema de optimización

$$\begin{cases} \inf\limits_{\lambda,\,\beta} & \lambda\varepsilon^2 + \dfrac{\beta^2}{4(\lambda-1)} + \dfrac{\lambda}{\lambda-1}\left(\dfrac{\beta}{N}\displaystyle\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i) + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i)^2\right) \\ \text{sujeto a} & \lambda \geqslant 1. \end{cases} \tag{4-25}$$

Este problema anterior se puede simplificar aun más por medio del análisis de la función objetivo, en ese sentido, de la Proposición 4.3.2 se sigue que para un $\beta \in \mathbb{R}$ fijo la función que tiene a λ como variable dada por

$$\lambda\varepsilon^2 + \frac{\beta^2}{4(\lambda-1)} + \frac{\lambda}{\lambda-1}\left(\frac{\beta}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i) + \frac{1}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i)^2\right) \tag{4-26}$$

es convexa para $\lambda \geqslant 1$. En efecto, es suficiente con demostrar que la segunda derivada respecto a λ de (4-26) es positiva para $\lambda \geqslant 1$, en ese sentido tenemos que la segunda derivada es dada por

$$\frac{\beta^2 + 4\frac{\beta}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i) + \frac{4}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i)^2}{2(\lambda-1)^3}.$$

Dado que $\lambda \geqslant 1$ entonces el signo de la última expresión es determinado por el signo de $\beta^2 + 4\frac{\beta}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i) + \frac{4}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i)^2$, pero esta expresión es positiva para cualquier β ya que en términos de β esta es un polinomio cuyo discriminarte es

$$\left(\frac{1}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i)\right)^2 - \frac{1}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i)^2,$$

el cual, por la Proposición 4.3.2, es negativo, de modo que dicho polinomio es siempre positivo. Con estas certezas y dado que la expresión (4-26) tiende a infinito cuando $\lambda \to 1^+$ o $\lambda \to \infty$, entonces esto garantiza que la función respecto a λ en (4-26) tiene un valor mínimo en la región $\lambda \geqslant 1$, este valor se calcula de la manera habitual, derivando para determinar los puntos críticos para luego identificar el punto critico que esta en la región de interés y evaluarlo en la función objetivo, en ese orden de ideas, después de desarrollar dichos pasos el valor mínimo es dado por

$$\varepsilon^2 + \frac{\beta}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i) + \frac{1}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i)^2 + \sqrt{\varepsilon^2\left(\beta^2 + 4\frac{\beta}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i) + \frac{4}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i)^2\right)}.$$

Por lo tanto, (4-25) se puede reescribir como

$$\inf_{\beta\in\mathbb{R}}\left(\varepsilon^2+\tfrac{\beta}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i)+\tfrac{1}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i)^2+\sqrt{\varepsilon^2\left(\beta^2+4\tfrac{\beta}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i)+\tfrac{4}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i)^2\right)}\right). \quad (4\text{-}27)$$

Pero el ínfimo anterior se puede calcular explícitamente analizando la función objetivo que allí aparece, la cual es diferenciable para todo $\beta\in\mathbb{R}$, tal expresión de este ínfimo es su vez una expresión concreta del valor óptimo de (4-25) que a su vez es valor óptimo de (4-22), por lo tanto, para efectos de notación llamamos $A:=\frac{1}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i)$ y $B:=\frac{1}{N}\sum_{i=1}^{N}(\eta-\widehat{\zeta}_i)^2$, entonces de dicho calculo del ínfimo en (4-27) se sigue que el valor óptimo de de (4-22) es

$$\begin{cases}\left(\sqrt{B-A^2}+\sqrt{\varepsilon^2-A^2}\right)^2 & \text{si } A^2\leqslant\varepsilon^2\\ -\infty & \text{si } A^2>\varepsilon^2.\end{cases} \quad (4\text{-}28)$$

Pero por hipótesis el caso $A^2>\varepsilon$ no se tiene, de modo que se tiene el resultado deseado. □

Contraparte DROW del modelo variaza-media de Markowitz

Antes de proponer una contraparte DROW de (4-18) se establecen las siguientes convenciones. Fijando $x\in\mathbb{R}^m$ definimos $\zeta^x:=\langle x,\xi\rangle$ la cual es una variable aleatoria, llamamos $\mathbb{P}^x$ su distribución la cual depende de $\mathbb{P}$, de modo que también es desconocida, luego, dada $\widehat{\xi}_1,\ldots,\widehat{\xi}_N$ una muestra de $\mathbb{P}$, entonces $\widehat{\zeta}_1^x,\ldots,\widehat{\zeta}_N^x$ definida por $\widehat{\zeta}_i^x:=\langle x,\widehat{\xi}_i\rangle$ es una muestra de ζ^x, esto permite definir la distribución empírica $\widehat{\mathbb{P}}_N^x$ asociada a ζ^x la cual es dada por

$$\widehat{P}_N^x:=\sum_{i=1}^{N}\delta_{\widehat{\zeta}_i^x}.$$

Una primera inquietud es conocer la relación entre $W_2^2(\widehat{\mathbb{P}}_N,\mathbb{P})$ y $W_2^2(\widehat{\mathbb{P}}_N^x,\mathbb{P}^x)$, en la siguiente proposición se establece esa relación:

Proposición 4.3.3. $W_2^2(\widehat{\mathbb{P}}_N^x,\mathbb{P}^x)\leqslant\|x\|^2W_2^2\left(\widehat{\mathbb{P}}_N,\mathbb{P}\right)$.

Demostración. Sea $\widehat{\widehat{\xi}}_1,\ldots,\widehat{\widehat{\xi}}_M$ otra muestra de ξ, entonces sea $\widehat{\widehat{\mathbb{P}}}_M$ la distribución empírica determinada por esta muestra. Esta última muestra de ξ determina la muestra $\widehat{\widehat{\zeta}}_1^x,\ldots,\widehat{\widehat{\zeta}}_M^x$ de ζ^x dada por $\widehat{\widehat{\zeta}}_i^x:=\left\langle x,\widehat{\widehat{\xi}}_i\right\rangle$, considerammos $\widehat{\widehat{\mathbb{P}}}_M^x$ como la distribución empírica generada por la muestra $\left\{\widehat{\zeta}_i^x\right\}_{i=1}^{M}$.

Ya que $\widehat{\widehat{\mathbb{P}}}_M \to \mathbb{P}$ y $\widehat{\widehat{\mathbb{P}}}^x_M \to \mathbb{P}^x$ débilmente, entonces por el Corolario 6.11 en [41] se tiene

$$W_2^2\left(\widehat{\mathbb{P}}_N, \widehat{\widehat{\mathbb{P}}}_M\right) \overset{M\to\infty}{\longrightarrow} W_2^2\left(\widehat{\mathbb{P}}_N, \mathbb{P}\right) \quad \text{y} \quad W_2^2\left(\widehat{\mathbb{P}}^x_N, \widehat{\widehat{\mathbb{P}}}^x_M\right) \overset{M\to\infty}{\longrightarrow} W_2^2\left(\widehat{\mathbb{P}}^x_N, \mathbb{P}^x\right). \tag{4-29}$$

Pero por otro lado se tiene

$$\begin{aligned}
W_2^2\left(\widehat{\mathbb{P}}^x_N, \widehat{\widehat{\mathbb{P}}}^x_M\right) &= \text{ínf}\left\{ \sum_{i=1}^N \sum_{j=1}^M \lambda_{i,j} \left|\widehat{\zeta}^x_i - \widehat{\widehat{\zeta}}^x_j\right|^2 \;\middle|\; \begin{array}{l} \sum_{i=1}^N \lambda_{i,j} = \frac{1}{M}, \\ \sum_{j=1}^M \lambda_{i,j} = \frac{1}{N}, \\ \lambda_{i,j} \geqslant 0, \\ i = 1, \ldots, N, \\ j = 1, \ldots, M \end{array} \right\} \\
&= \text{ínf}\left\{ \sum_{i=1}^N \sum_{j=1}^M \lambda_{i,j} \left|\left\langle x, \widehat{\xi}_i \right\rangle - \left\langle x, \widehat{\widehat{\xi}}_j \right\rangle\right|^2 \;\middle|\; \begin{array}{l} \sum_{i=1}^N \lambda_{i,j} = \frac{1}{M}, \\ \sum_{j=1}^M \lambda_{i,j} = \frac{1}{N}, \\ \lambda_{i,j} \geqslant 0, \\ i = 1, \ldots, N, \\ j = 1, \ldots, M \end{array} \right\} \\
&\leqslant \text{ínf}\left\{ \sum_{i=1}^N \sum_{j=1}^M \lambda_{i,j} \|x\|^2 \left\|\widehat{\xi}_i - \widehat{\widehat{\xi}}_j\right\|^2 \;\middle|\; \begin{array}{l} \sum_{i=1}^N \lambda_{i,j} = \frac{1}{M}, \\ \sum_{j=1}^M \lambda_{i,j} = \frac{1}{N}, \\ \lambda_{i,j} \geqslant 0, \\ i = 1, \ldots, N, \\ j = 1, \ldots, M \end{array} \right\} \quad \begin{array}{l} \text{por la desigualdad} \\ \text{Hölder} \end{array} \\
&= \|x\|^2 W_2^2\left(\widehat{\mathbb{P}}_N, \widehat{\widehat{\mathbb{P}}}_M\right).
\end{aligned}$$

Por lo tanto, por (4-29) se concluye

$$W_2^2\left(\widehat{\mathbb{P}}^x_N, \mathbb{P}^x\right) \leqslant \|x\|^2 W_2^2\left(\widehat{\mathbb{P}}_N, \mathbb{P}\right).$$

□

Así pues, por esta última proposición, si $\varepsilon > 0$ es tal que $\mathbb{P} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ donde la bola es tomada respecto a la métrica 2-Wasserstein para medidas soportadas en $\mathbb{R}^m$, entonces $\mathbb{P}^x \in \mathcal{B}_{\|x\|\varepsilon}(\widehat{\mathbb{P}}^x_N)$ donde la bola es tomada respecto a la métrica 2-Wasserstein para medidas soportadas en $\mathbb{R}$. Por lo tanto, definimos el conjunto

$$\begin{aligned}
\mathbb{X} &:= \left\{ x \in \mathbb{R}^m \;\middle|\; \sum_{i=1}^m x_i = 1,\ x_i \geqslant 0,\ \mathbb{E}_{\mathbb{Q}}[\zeta^x] \geqslant \mu \ \ \forall\, \mathbb{Q} \in \mathcal{B}_{\|x\|\varepsilon}(\widehat{\mathbb{P}}^x_N) \right\} \\
&= \left\{ x \in \mathbb{R}^m \;\middle|\; \sum_{i=1}^m x_i = 1,\ x_i \geqslant 0,\ \inf_{\mathbb{Q}\in\mathcal{B}_{\|x\|\varepsilon}(\widehat{\mathbb{P}}^x_N)} \mathbb{E}_{\mathbb{Q}}[\zeta^x] \geqslant \mu \right\}
\end{aligned} \tag{4-30}$$

El siguiente Teorema permite reescribir $\mathbb{X}$ de manera mas simple.

Teorema 4.3.2. *Sea ζ una variable aleatoria con distribución $\mathbb{P}$ desconocida y sea $\widehat{\zeta}_1, \ldots, \widehat{\zeta}_N$ una muestra de ζ. Dado $\varepsilon > 0$ se considera $\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ la bola respecto a la métrica 2-Wasserstein de radio ε centrada en $\widehat{\mathbb{P}}_N$ la distribución empírica respecto a la muestra anterior, entonces*

$$\inf_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}}[\zeta] = \frac{1}{N}\sum_{i=1}^{N} \widehat{\xi}_i - \varepsilon.$$

Demostración. Es claro que

$$\inf_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}}[\zeta] = - \sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}}[-\zeta].$$

Entonces, la función objetivo de este problema es $g(\zeta) := -\zeta$ la cual satisface la Suposición 1 en su parte 1, entonces por el Teorema 3.3.1 de la tesis esta última formulación es equivalente al problema de optimización semi-infinita

$$-\begin{cases} \inf\limits_{\lambda\geqslant 0} & \lambda\varepsilon^2 + \frac{1}{N}\sum_{i=1}^{N} s_i \\ \text{sujeto a} & \sup_{\zeta\in\mathbb{R}}\left(-\zeta - \lambda\left(\zeta - \widehat{\zeta}_i\right)^2\right) \leqslant s_i \quad \forall\, i = 1, \ldots, N, \end{cases}$$

$$= -\begin{cases} \inf\limits_{\lambda\geqslant 0} & \lambda\varepsilon^2 + \frac{1}{N}\sum_{i=1}^{N} s_i \\ \text{sujeto a} & \frac{1}{4\lambda} - \widehat{\zeta}_i \leqslant s_i \qquad \forall\, i = 1, \ldots, N, \end{cases}$$

$$= -\inf_{\lambda\geqslant 0}\left(\lambda\varepsilon^2 + \frac{1}{4\lambda} - \frac{1}{N}\sum_{i=1}^{N}\widehat{\zeta}_i\right)$$

$$= -\left(\varepsilon - \frac{1}{N}\sum_{i=1}^{N}\widehat{\zeta}_i\right)$$

$$= \frac{1}{N}\sum_{i=1}^{N}\widehat{\zeta}_i - \varepsilon.$$

□

Entonces, de acuerdo a este Teorema anterior $\mathbb{X}$ es de la forma

$$\begin{aligned} \mathbb{X} &= \left\{ x\in\mathbb{R}^m \;\middle|\; \sum_{i=1}^{m} x_i = 1,\ x_i \geqslant 0,\ \frac{1}{N}\sum_{i=1}^{N}\widehat{\zeta}_i^x - \varepsilon\,\|x\| \geqslant \mu \right\} \\ &= \left\{ x\in\mathbb{R}^m \;\middle|\; \sum_{i=1}^{m} x_i = 1,\ x_i \geqslant 0,\ \frac{1}{N}\sum_{i=1}^{N}\langle x, \widehat{\xi}_i\rangle - \varepsilon\,\|x\| \geqslant \mu \right\} \end{aligned} \tag{4-31}$$

Así pues, tenemos el siguiente DROW modificado:

$$\begin{aligned}\widehat{J}_N &:= \underset{x\in\mathbb{X}}{\text{minimizar}} \sup_{\mathbb{Q}\in\mathcal{B}_{\|x\|\varepsilon}(\widehat{\mathbb{P}}_N^x)} \operatorname{Var}_{\mathbb{Q}}\left[\zeta^x\right] \\ &= \underset{x\in\mathbb{X}}{\text{minimizar}} \sup_{\eta\geqslant\mu} \begin{cases} \sup_{\mathbb{Q}\in\mathcal{B}_{\|x\|\varepsilon}(\widehat{\mathbb{P}}_N^x)} & \operatorname{Var}_{\mathbb{Q}}\left[\zeta^x\right] \\ \text{sujeto a} & \mathbb{E}_{\mathbb{Q}}\left[\zeta^x\right] = \eta. \end{cases}\end{aligned} \tag{4-32}$$

Pero de la Proposición 4.3.1 se concluye

$$\widehat{J}_N = \underset{x\in\mathbb{X}}{\text{minimizar}} \sup_{\substack{\eta\geqslant\mu, \\ \left(\eta-\frac{1}{N}\sum_{i=1}^N \widehat{\zeta}_i^x\right)^2\leqslant\varepsilon^2\|x\|^2}} \begin{cases} \sup_{\mathbb{Q}\in\mathcal{B}_{\|x\|\varepsilon}(\widehat{\mathbb{P}}_N^x)} & \operatorname{Var}_{\mathbb{Q}}\left[\zeta^x\right] \\ \text{sujeto a} & \mathbb{E}_{\mathbb{Q}}\left[\zeta^x\right] = \eta. \end{cases} \tag{4-33}$$

La estrategia es reescribir el primer problema de maximización interno en (4-32) usando los resultados expuestos en el capítulo anterior, en ese sentido, por el Teorema 4.3.1 se infiere que (4-33) es equivalente al problema de optimización:

$$\widehat{J}_N = \underset{x\in\mathbb{X}}{\text{minimizar}} \begin{cases} \sup\limits_{\eta\geqslant\mu} & \left(\sqrt{\frac{1}{N}\sum\limits_{i=1}^N\langle x,\widehat{\xi}_i\rangle^2 - \frac{1}{N^2}\left(\sum\limits_{i=1}^N\langle x,\widehat{\xi}_i\rangle\right)^2} + \sqrt{\varepsilon^2\|x\|^2 - \left(\mu - \frac{1}{N}\sum\limits_{i=1}^N\langle x,\widehat{\xi}_i\rangle\right)^2}\right)^2 \\ \text{sujeto a} & \left(\eta - \frac{1}{N}\sum\limits_{i=1}^N\widehat{\zeta}_i^x\right)^2 \leqslant \varepsilon^2\|x\|^2 \end{cases} \tag{4-34}$$

Pero el problema de maximización interno de (4-34) puede solucionarse explícitamente, en realidad dicho problema alcanza su valor óptimo en $\eta^* = \frac{1}{N}\sum_{i=1}^N \widehat{\zeta}_i^x$, por lo tanto, (4-34) se puede reescribir como

$$\begin{aligned}\widehat{J}_N &= \underset{x\in\mathbb{X}}{\text{minimizar}} \left(\sqrt{\frac{1}{N}\sum_{i=1}^N\langle x,\widehat{\xi}_i\rangle^2 - \frac{1}{N^2}\left(\sum_{i=1}^N\langle x,\widehat{\xi}_i\rangle\right)^2} + \varepsilon\|x\|\right)^2 \\ &= \begin{cases} \underset{x\in\mathbb{R}^m}{\text{minimizar}} & \left(\sqrt{\frac{1}{N}\sum\limits_{i=1}^N\langle x,\widehat{\xi}_i\rangle^2 - \frac{1}{N^2}\left(\sum\limits_{i=1}^N\langle x,\widehat{\xi}_i\rangle\right)^2} + \varepsilon\|x\|\right)^2 \\ \text{sujeto a} & \sum\limits_{i=1}^N\langle x,\widehat{\xi}_i\rangle - \varepsilon\|x\| \geqslant \mu, \\ & \sum\limits_{i=1}^m x_i = 1. \end{cases}\end{aligned} \tag{4-35}$$

Note que $\widehat{J}_N$ depende de ε así que cuando sea importante recalcar dicha dependencia escribiremos $\widehat{J}_N(\varepsilon)$ en lugar de $\widehat{J}_N$, de igual manera las soluciones óptimas de (4-35) las denotaremos $\widehat{x}_N(\varepsilon)$.
El problema (4-35) se puede simplificar aun más, dicha simplificación es con la intención de llevarlo a un contexto en el cual existan mas herramientas para abordar dicho problema, en ese sentido el siguiente resultado cumple esa tarea.

Proposición 4.3.4. *Sea M la matriz de tamaño $m\times N$ cuyas columnas son los vectores de la muestra $\widehat{\xi}_1,\dots,\widehat{\xi}_N$ y sean $\mathbf{0},\mathbf{e}\in\mathbb{R}^N$ los vectores columna de ceros y unos respectivamente A partir de estas convenciones se definen las matrices*

$$E := \frac{1}{N}MM^T - \frac{1}{N^2}(M\mathbf{e})(M\mathbf{e})^T \quad y \quad L := \frac{1}{N}(Me)^T.$$

Por lo tanto (4-35) es equivalente al problema de optimización

$$\begin{cases} \inf\limits_{x\in\mathbb{R}^m} & \left(\left\|K^Tx\right\| + \varepsilon\left\|x\right\|\right)^2 \\ \text{sujeto a} & Lx - \varepsilon\left\|x\right\| \geqslant \mu, \\ & e^Tx = 1. \end{cases} \tag{4-36}$$

Demostración. Después de unos cálculos extensos pero sin dificultad se puede concluir

$$\frac{1}{N}\sum_{i=1}^N\langle x,\widehat{\xi}_i\rangle^2 - \frac{1}{N^2}\left(\sum_{i=1}^N\langle x,\widehat{\xi}_i\rangle\right)^2 = x^TEx \quad \text{y} \quad \frac{1}{N}\sum_{i=1}^N\langle x,\widehat{\xi}_i\rangle = Lx.$$

Pero por la Proposición 4.3.2 se infiere que E es semi-definida positiva, entonces E tiene una factorización LDL, es decir, existe L matriz triangular inferior, D una matriz diagonal y P una matriz de permutación tales que $E = (P^{-1})^TLDL^TP^{-1}$, entonces consideramos[6] $K := (P^{-1})^TLD^{1/2}$. Por lo tanto, (4-35) es equivalente a (4-36). □

Se debe tener en cuenta que (4-36) puede no ser factible para algunos valores de ε, concretamente, dada L y μ se tiene que (4-36) es factible si

$$\varepsilon < \widehat{\epsilon}_N(\mu) := \begin{cases} \sup\limits_{x\in\mathbb{R}^m} & \dfrac{Lx-\mu}{\left\|x\right\|} \\ \text{sujeto a} & \sum\limits_{i=1}^m x_i = 1. \end{cases}$$

La dependencia de N en $\widehat{\epsilon}_N(\mu)$ se debe a que L depende de la muestra. En adelante llamaremos *radio extremo factible* a la expresión $\widehat{\epsilon}_N(\mu)$.

[6] Si E fuera definida positiva entonces tendríamos la factorización de Cholesky.

Elección de ε

En esta parte del trabajo presentamos dos formas de elegir ε, cada una de las formas que se presentan dependen de las preferencias del inversionista, con esto nos referimos a dos posibilidades, la primera es que se priorice que el portafolio que se elija posea en promedio una volatilidad baja lo mas cercana superiormente a la mínima volatilidad posible, esta volatilidad mínima es el valor óptimo de (4-18) de modo que es desconocida, al priorizar lo anterior el inversionista esta dispuesto a que el retorno de dicho portafolio sea inferior a μ, en ese sentido, denotamos por ε_{var} el menor valor posible de ε que satisface

$$V(\varepsilon) := \mathbb{E}_{\mathbb{P}^N}\left[\widehat{J}_N(\varepsilon) - \text{Var}_{\mathbb{P}}\left[\langle \widehat{x}_N(\varepsilon), \xi\rangle\right]\right] \geqslant 0. \tag{4-37}$$

El portafolio que se elige para este caso es $\widehat{x}_N(\varepsilon_{\text{var}})$ que es la solución óptima de $\widehat{J}_N(\varepsilon_{\text{var}})$. La segunda posibilidad es que se priorice que el portafolio que se elija posea en promedio un retorno esperado lo mas cercano superiormente a μ, al priorizar esto el inversionista esta dispuesto a obtener en promedio volatilidades ligeramente altas, aunque veremos en los experimentos que no son tan elevadas, en ese sentido, denotamos por ε_{ret} el menor valor de ε que satisface

$$R(\varepsilon) := \mathbb{E}_{\mathbb{P}^N}\left[\mathbb{E}_{\mathbb{P}}\left[\langle \widehat{x}_N(\varepsilon), \xi\rangle\right]\right] \geqslant \mu. \tag{4-38}$$

El portafolio que se elige para esta segunda preferencia es $\widehat{x}_N(\varepsilon_{\text{ret}})$ que es la solución óptima de $\widehat{J}_N(\varepsilon_{\text{ret}})$.
En ambos casos emerge la dificultad del desconocimiento de $\mathbb{P}$ lo que obliga a abordar la búsqueda de ε_{var} y ε_{ret} desde un enfoque estadístico ya que con lo único con que se cuneta es con muestras del vector de retornos. Para $N \in \mathbb{N}$ fijo y dada una muestra $\widehat{\xi}_1, \ldots, \widehat{\xi}_N$ el procedimiento para elegir ε_{var} y ε_{ret} es el siguiente: Se considera un conjunto $\mathcal{E}$ de posibles valores de ε, por ejemplo,

$$\mathcal{E} = \bigcup_{i=0}^{5} \left\{k10^{-i} \mid k = 1, \ldots, 10\right\}.$$

Para cada ε en $\mathcal{E}$ se realizan los siguientes pasos:

1. Usando la muestra $\widehat{\xi}_1, \ldots, \widehat{\xi}_N$ sea $\widehat{E}$ y $\widehat{\mathbf{m}}$ la versión muestral de la matriz de covarianza y el vector de esperanzas del vector aleatorio ξ.

2. Usando la muestra $\widehat{\xi}_1, \ldots, \widehat{\xi}_N$ se estima vía estimación por núcleos la función de densidad conjunta de ξ usando núcleos multinormales estándar (es decir, el vector

de ceros como el vector de esperanzas y la matriz identidad como la matriz de covarianza), llamemos $\widehat{f}_{N,h}$ a dicha estimación[7].

3. Se generan K muestras de tamaño N con distribución determinada por la función de densidad conjunta $\widehat{f}_{N,h}$, se sugieren valores de K como 200 o 1000 dependiendo de la capacidad de computo.

4. Para cada una de las muestras producidas en el ítem anterior se soluciona $\widehat{J}_N(\varepsilon)$ y su solución optima $\widehat{x}_N(\varepsilon)$, ademas se calcula $(\widehat{x}_N(\varepsilon))^T\widehat{E}\widehat{x}_N(\varepsilon)$ y $\widehat{\mathbf{m}}\widehat{x}_N(\varepsilon)$.

5. Ya que el proceso del ítem 4 es una rutina de K instancias, por cada $i = 1, \ldots, K$ se calcula $\tau_i^{\text{var}}(\varepsilon) := \widehat{J}_N(\varepsilon) - (\widehat{x}_N(\varepsilon))^T\widehat{E}\widehat{x}_N(\varepsilon)$ usando la i-esima muestra, luego, de define el numero $\nu_\varepsilon = \sum_{i=1}^{K}\tau_i(\varepsilon)$.
De manera análoga, por cada $i = 1, \ldots, K$ se calcula $\tau_i^{\text{ret}}(\varepsilon) := \widehat{\mathbf{m}}\widehat{x}_N(\varepsilon) - \mu$, llamamos dicho numero γ_ε.

Por cada $\varepsilon \in \mathcal{E}$ tendremos un ν_ε y un γ_ε. Si deseamos estimar ε_{var} entonces elegimos el menor ε de $\mathcal{E}$ tal que $\frac{\nu_\varepsilon}{K} \geqslant 0$.
Si deseamos estimar ε_{ret} entonces elegimos el menor ε de $\mathcal{E}$ tal que $\frac{\gamma_\varepsilon}{K} \geqslant 0$.
Cabe resaltar que el método elegido para el remuestreo es sensible al valor de h y para vectores aleatorios con funciones de densidad con muchos picos la estimación no es la mejor, pero aun con dichas características su desempeño es mejor que re-muestreos vía boostrap.

Resultados numéricos

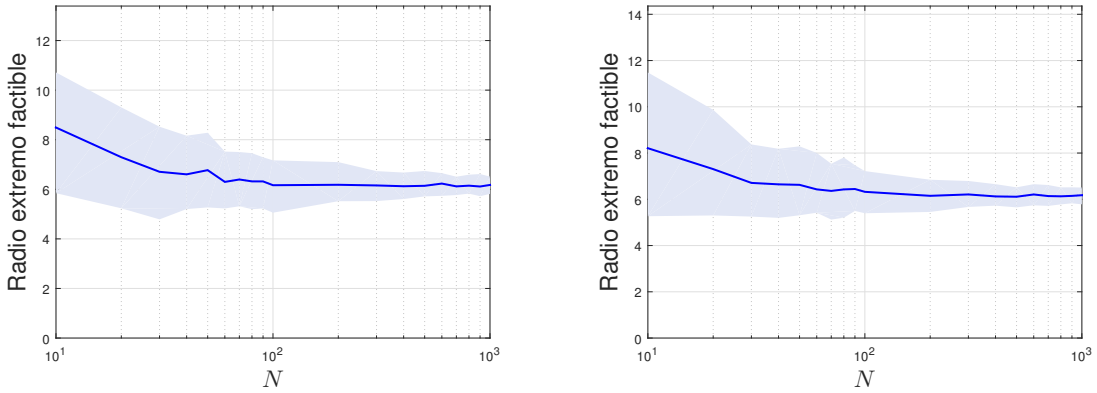
Los resultados numéricos que se traducen en las siguiente gráficas son producto de simulaciones realizadas para un portafolio compuesto de cuatro bienes, es decir, $m = 4$, la distribución de ξ es multinormal con matriz de covarianza C y vector de medias $\mathbf{m}$ dados por

$$C = \begin{bmatrix} 185 & 86{,}5 & 80 & 20 \\ 86{,}5 & 196 & 76 & 13{,}5 \\ 80 & 76 & 411 & -19 \\ 20 & 13{,}5 & -19 & 25 \end{bmatrix} \quad \text{y} \quad \mathbf{m} = (14, 12, 15, 17).$$

En la Figura **4-6** se representa el radio extremo factible como función de N para (a) $\mu = 20$ y (b) $\mu = 40$, en esta gráfica, por cada N, se consideran 200 muestras independientes de ξ

[7] El valor de las componentes del vector h es $h_i := \sigma_i \left(\frac{4}{(m+2)N}\right)^{\frac{1}{m+4}}$ donde σ_i es la desviación estándar de las componentes i-ésima de la muestra, esta forma de elegir h es sugerida en [37].

de tamaño N, por cada muestra se calcula $\widehat{\epsilon}_N(\mu)$, el área sombreada representa la banda entre el 20 % y 80 % quantiles de $\widehat{\epsilon}_N(\mu)$ y la linea solida es el 50 % quantil (la mediana) de $\widehat{\epsilon}_N(\mu)$. De esta gráfica, para ambos casos (a) y (b), se observa que una elección de ε que asegure que (4-36) sea factible debe ser considerada satisfaciendo $\varepsilon \leqslant 4$. Esta gráfica también evidencia que $\widehat{\epsilon}_N(\mu)$ no es tan dependiente de μ. El hecho de que el máximo valor de ε que se puede considerar sea 4 no es un aspecto restrictivo ya que en la practica los valores de ε que garantizan que $\mathbb{P} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ con alta probabilidad no serán tan grandes, por lo general son valores inferiores a 4.



(a) (b)

Figura 4-6: Radio extremo factible $\widehat{\epsilon}_N(\mu)$ como función de N. En (a) para $\mu = 20$ y (b) $\mu = 40$.

Con lo anterior en mente y eligiendo ε con el método expuesto anteriormente, deseamos comparar el desempeño del enfoque robusto distribucional DROW que se le dio al problema (4-18) comparado con el enfoque muestral expuesto en (4-19), en ese sentido, para ambos enfoques procedemos a analizar el *desempeño fuera de muestra*, es decir $\mathrm{Var}_{\mathbb{P}}[\langle \widehat{x}_N, \xi \rangle]$ (caso DROW) donde $\widehat{x}_N$ es la solución óptima de $\widehat{J}_N(\varepsilon)$, y $\mathrm{Var}_{\mathbb{P}}[\langle \widehat{x}_N^{muest}, \xi \rangle]$ (caso muestral) donde $\widehat{x}_N^{muest}$ es la solución (4-19) de manera muestral. También se analiza el *certificado*, es decir, $\widehat{J}_N(\varepsilon)$ (caso DROW) y $(\widehat{x}_N^{muest} =^T \widehat{E}\widehat{x}_N^{muest}$ (caso muestral) donde $\widehat{E}$ es la matriz de covarianza muestral. Y el último aspecto es el *retorno esperado*, es decir, $\mathbb{E}_{\mathbb{P}}[\langle \widehat{x}_N, \xi \rangle]$ (caso DROW) y $\mathbb{E}_{\mathbb{P}}[\langle \widehat{x}_N^{muest}, \xi \rangle]$ (caso muestral).
La figura **4-7** muestra el caso en el que se prioriza la volatilidad, es decir, ε es ε_{var}, en la figura las áreas sombreadas representan las bandas delimitadas por los quantiles 20 % y 80 %, y las lineas solidas son las medias, esto a partir de 200 simulaciones, en verde se representa el caso DROW y en azul el caso muestral, la linea negra en la figura (a) y (b) es el valor óptimo verdadero de (4-18) y ya que en este ejemplo consideramos $\mu = 20$ la linea

roja en (c) representa dicho valor. De esta gráfica se evidencia que se obtiene en promedio una volatilidad superior a la mínima posible, lo cual era de esperarse, sin embargo los portafolios óptimos del método DROW no satisfacen tener un retorno esperado menor a 20, aunque esto tampoco lo satisfacen los portafolios óptimos obtenidos muestralmente (ver (c)), no obstante, el retorno esperado de los portafolios DROW tienen un mayor retorno y su volatilidad no es tan elevada (ver (a)) siendo esto una ventaja respecto al enfoque muestral.

En la Figura **4-8** se muestra el caso en el que se prioriza el retorno esperado, es decir, ε es ε_{ret}, las convenciones son las mismas que en la Figura **4-7**. De esta gráfica se evidencia que se obtiene en promedio un retorno esperado superior a lo que estipulamos de inicio que es $\mu = 20$, el enfoque muestral nunca logra obtener dichos retornos (ver (c)), además en promedio la volatilidad de dichos portafolios no es tan elevada (ver (a)) teniendo en cuenta que a mayor retorno mayor es el riesgo, consideramos que este es un aspecto positivo aunque gran parte del análisis depende de las preferencias del inversionista, un inversionista arriesgado priorizara el retorno y uno moderado priorizara minimizar el riesgo, en esta aplicación se ofrecen dos opciones que tiene en cuenta esas preferencias.

Otro aspecto que se evidencia de las figuras es que el enfoque DROW tiene un mejor desempeño para valores de N mayores a 50 ya que para esos valores es menos inestable.

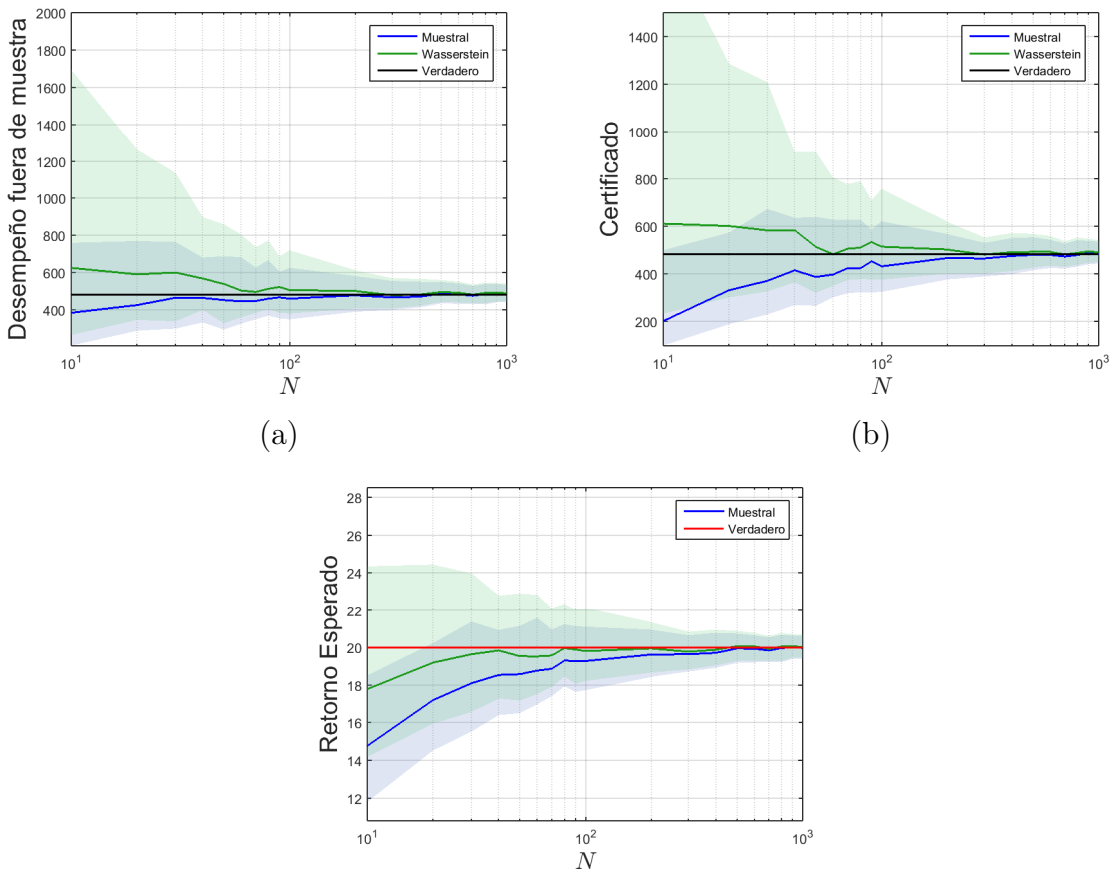


Figura 4-7: (a) Desempeño fuera de muestra $\mathrm{Var}_{\mathbb{P}}[\langle \widehat{x}_N, \xi \rangle]$ (linea verde y área sombreada verde) y $\mathrm{Var}_{\mathbb{P}}[\langle \widehat{x}_N^{muest}, \xi \rangle]$ (linea azul y área sombreada azul) donde $\widehat{x}_N^{muest}$ es la solución muestral. (b) certificado $\widehat{J}_N$ (linea verde y área sombreada verde) y $\mathrm{Var}_{\widehat{\mathbb{P}}_N}[\langle \widehat{x}_N^{muest}, \xi \rangle]$ (linea azul y área sombreada azul). (c) Retorno esperado $\mathbb{E}_{\mathbb{P}}[\langle \widehat{x}_N, \xi \rangle]$ (linea verde y área sombreada verde) y $\mathbb{E}_{\mathbb{P}}[\langle \widehat{x}_N^{muest}, \xi \rangle]$ (linea azul y área sombreada azul). Se tomó $\varepsilon = \varepsilon_{\mathrm{var}}$.

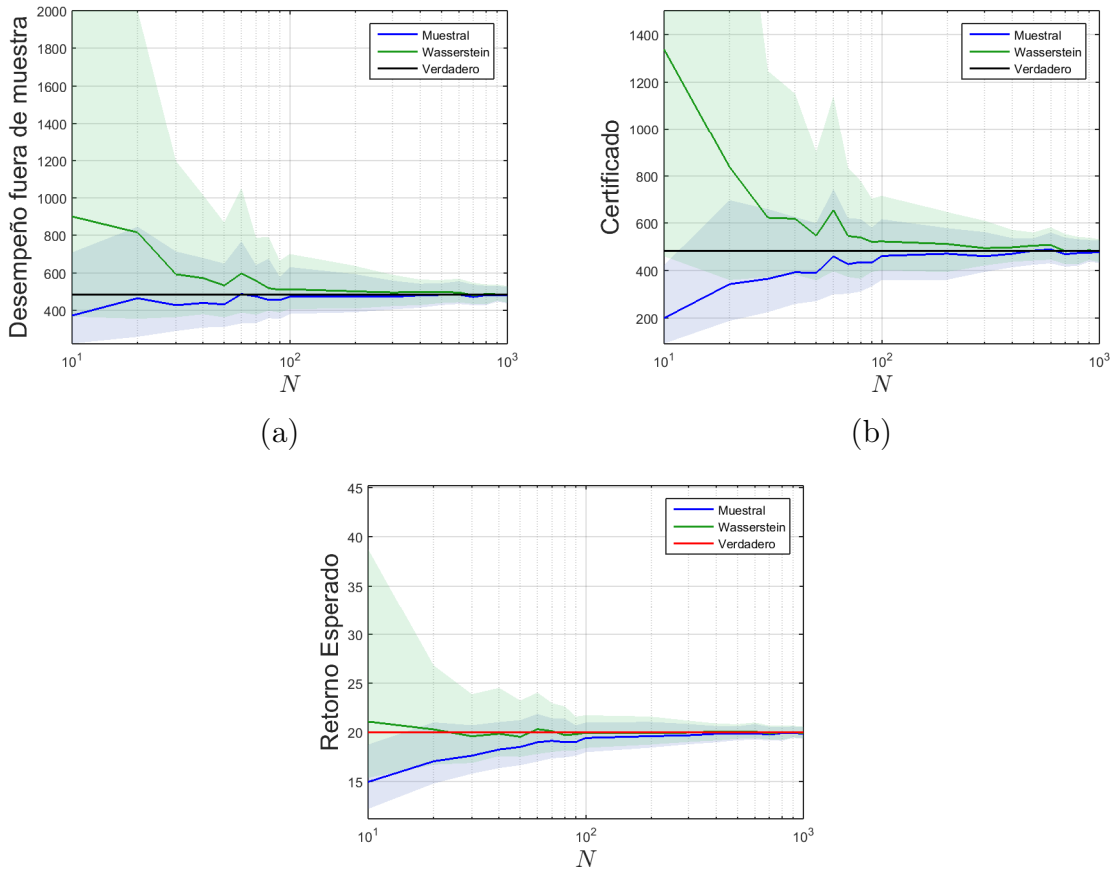


Figura 4-8: (a) Desempeño fuera de muestra $\mathrm{Var}_{\mathbb{P}}[\langle \widehat{x}_N, \xi \rangle]$ (linea verde y área sombreada verde) y $\mathrm{Var}_{\mathbb{P}}[\langle \widehat{x}_N^{muest}, \xi \rangle]$ (linea azul y área sombreada azul) donde $\widehat{x}_N^{muest}$ es la solución muestral. (b) certificado $\widehat{J}_N$ (linea verde y área sombreada verde) y $\mathrm{Var}_{\widehat{\mathbb{P}}_N}[\langle \widehat{x}_N^{muest}, \xi \rangle]$ (linea azul y área sombreada azul). (c) Retorno esperado $\mathbb{E}_{\mathbb{P}}[\langle \widehat{x}_N, \xi \rangle]$ (linea verde y área sombreada verde) y $\mathbb{E}_{\mathbb{P}}[\langle \widehat{x}_N^{muest}, \xi \rangle]$ (linea azul y área sombreada azul). Se tomó $\varepsilon = \varepsilon_{\mathrm{ret}}$.

5 Conclusiones y trabajo futuro

En este trabajo, comenzamos abordando los problemas de optimización estocástica desde una perspectiva Robusta Distribucional, esta perspectiva da origen a otro problema de optimización que llamamos DROW ya que el conjunto de ambigüedad $\mathcal{D}$ es considerado como una bola respecto a la métrica de Wasserstein centrada en la distribución empírica, asumiendo ciertas condiciones en la función f, se presentó una reformulación del DROW (ver Teorema 3.3.1), tal reformulación es un problema de óptimización semi-infinita, este problema se puede reformular como un problema de optimización convexa finito para casos específicos de f. Las condiciones impuestas a f son presentadas en la Suposición 3.3.1, demostrar que dicha reformulación semi-infinita es valida para el universo de funciones comprendido en la Suposición 3.3.1 implica ampliar ese conjunto de funciones en donde esa reformulación ya era valida, recordemos que en [11] y [22] se expone la reformulación expuesta en este trabajo pero en esos trabajos esta es demostrada para un universo de funciones f mas restrictivo que el universo de funciones f comprendido en este trabajo, en [11] se asume que f es el máximo de funciones cóncavas y que el soporte de ξ es cerrado y convexo, mientras que [22] asume que f es una función lipschitziana y que el soporte de ξ es compacto, no obstante, de acuerdo a lo demostrado en este trabajo podemos concluir que para justificar dicha reformulación no es necesario imponer condiciones en el soporte de la variable aleatoria involucrada, solo se asumen condiciones sobre f (ver Suposición 3.3.1), condiciones que satisfacen las funciones lipschitzianas y algunas funciones cóncavas.

Por otro lado, en materia teórica, en el contexto de la sección 3.2, un problema que emerge en este trabajo y requiere de mayor atención es la determinación de ε de tal manera que la distribución verdadera $\mathbb{P}$ pertenezca a la bola $\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ con alta probabilidad; las cotas dadas en la sección 3.2 para estimar ε son existenciales, es decir, ε dependerá de unas constantes que no se pueden determinar explícitamente, ante tal dificultad se recurre a estimaciones que se basan en métodos estadísticos basados en principios como validación cruzada, la idea de aplicar esta forma de estimar ε surge cuando se aplica esta teoría a problemas concretos como se hace en [11], [21] y en la Sección 4.1 de este trabajo, no obstante, no se ha justificado la relación del ε determinado por estos métodos estadísticos y el ε desconocido que garantiza que $\mathbb{P}$ pertenezca a la bola $\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ con alta probabili-

dad, pareciera que dicha forma de elegir ε en realidad lo que trata de contrarrestar es el sobreajuste (overfitting) del modelo; esta situacion queda como una temática abierta a estudio.

De la sección 4.3 se obtiene un enfoque Robusto Distribucional Robusto Distribucional con métrica de Wasserstein DROW del modelo variaza-media de Markowitz, se concluye que dependiendo de las preferencias del inversionista, arriesgado o moderado, el enfoque DROW satisface mejor sus necesidades que como lo hace el enfoque muestral sobretodo cuando se cuenta con muestras de tamaño superior a 50, no obstante, queda una asignatura pendiente que creemos que su mejora requiere de mayor investigación y tiempo, este aspecto pendiente consiste en hacer el enfoque DROW más robusto, esto se evidencia en las áreas sombreadas verdes de las Figuras **4-7** y **4-8**, la intención es modificar el enfoque DROW propuesto de tal manera que las bandas determinadas por esas áreas sean mas pequeñas, tal modificación haría del método DROW mas estable ya que en dicho aspecto con el actual enfoque DROW no se obtiene ninguna ventaja visible respecto al enfoque muestral.

Por ultimo, de la Sección 4.2 queda de manifiesto que la reformulación semi-infinita del DROW en varios contextos no son problemas fáciles de solucionar, creemos que el avance y divulgación del enfoque Robusto Distribucional con métrica de Wasserstein DROW para problemas de optimización estocástica esta estrechamente ligado con el avance de la investigación en materia de optimización semi-infinita, por ejemplo, en [21] se propone una versión Robusta Distribucional del método de clasificación de datos conocido como Maquinas de Soporte Vectorial (SVM), el cual llamamos DR-SVM, los resultados allí expuesto son positivos, muestra una mejor estabilidad e indices de clasificación excelentes superando al método tradicional de SVM, no obstante, a pesar de los resultados positivos DR-SVM no ha tenido una repercusión importante en la comunidad que trabaja problemas de clasificación, la razón de esta situación radica en la forma de abordar DR-SVM que se traduce en solucionar un problema de óptimización semi-infinita, el algoritmo existente para abordar este problema tiene un numero de subrutinas que dependen del numero de datos de entrenamiento, dado que en la actualidad estos problemas de clasificación se presentan en un contexto en el cual la cantidad de datos es considerablemente grande, esta situación hace el algoritmo ineficiente. En ese sentido, es fundamental enfocar la atención en los problemas de optimización semi-infinta provenientes de un DROW, en particular, una asignatura pendiente es idear una forma de abordar de manera eficiente el problema de la sección 4.2. En resumen, dada las propiedades de la métrica de Wasserstein y su inclusión a la visión Robusta Distribucional es natural que todo problema de óptimización

estocástica se pretenda abordar desde esta perspectivas, este enfoque tiene ventajas pero también tiene limitantes, esta tesis es un intento por evidenciar ambos aspectos y tratar de brindar un panorama en el cual se pueda sugerir hacia donde se podrían destinar esfuerzos para superar las dificultades.

Bibliografía

[1] Ambrosio, L. ; Gigli, N. ; Savare, G.: Gradient flows in metric spaces and in the space of probability measures. En: *Lectures in Mathematics ETH Zurich* (2008)

[2] Bertsekas, D.: *Convex Optimization Theory.* Athena Scientific, 2009

[3] Bertsekas, D. ; Nedic, A. ; Ozdaglar, AE.: *Convex Analysis and Optimization.* Athena Scientific, 2003

[4] Birge, JR. ; Louveaux, F.: *Introduction to Stochastic Programming.* Springer, 2006

[5] Bowman, A. W.: An alternative method of cross-validation for the smoothing of kernel density estimates. En: *Biometrika.* 271 (1984), p. 353360

[6] Bowman, A. W.: A comparative study of some kernel-based nonparametric density estimators. En: *Journal of Statistical Computation and Simulation.* 21 (1985), p. 3–4

[7] Boyd, S. ; Vandenberghe, L.: *Convex optimization.* Cambridge University Press, 2009

[8] Burg, J.: The Relationship between Maximum Entropy Spectra and Maximum Likelihood Spectra. En: *Geophysics.* 97 (1972), p. 375–376

[9] Delage, E. ; Ye, Y.: Distributionally robust optimizatión under moment uncertainty with application to data-driven problems. En: *Oparations Research* 58 (2010), Nr. 3, p. 595–612

[10] El Ghaoui, L. ; Oks, M. ; Oustry, F: Worst-case value-at-risk and robust portafolio optimization: A conic programming approach. En: *Operations Research* 51 (2006), Nr. 4

[11] Esfahani, PM. ; Kuhn, D.: Data-driven Distributionally Robust Optimization Using the Wasserstein Metric: Performance Guarantees and Tractable Reformulations. En: *arXiv preprint arXiv:1505.05116v2* (2016)

[12] Fournier, N. ; Guillin, A.: On the rate of convergence in Wsserstein distence of the empirical measure. En: *Probability Theory and Related Fields* (2014), p. 1–32

[13] Gao, R. ; Kleywegt, AJ.: Distributionally Robust Stochastic Optimization with Wasserstein Distance. En: *arXiv preprint arXiv:1604.02199v2* (2016)

[14] Goh, J. ; Sim, M.: Distributionally robust optimization and its tratable approximations. En: *Operations Research* (2013)

[15] Hanasusanto, GA. ; Kuhn, D.: Robust data-driven dynamic programing. En: *Advances in Natural Information Processing Systems* (2013)

[16] Jiang, R. ; Guan, Y.: Data-driven chance costrained stochastic program. En: *Mathematical Programming* (2015), p. 1–37

[17] Jones, M. c.: The roles of ISE and MISE in density estimation. En: *Statistics & Probability Letters.* 12 (1991), p. 5156

[18] Kolouri, S. ; Park, S. ; Thorpe, M. ; Slepev, D. ; Rohde, GK.: Transport-based analysis, modeling, and learning from signal and data distributions. En: *arXiv preprint arXiv:arXiv:1609.04767v1* (2016)

[19] Kullback, S.: Letter to the Editor: The KullbackLeibler distance. En: *The American Statistician.* 41 (1987), Nr. 4, p. 340–341

[20] Lagoa, C. M. ; Barmish, R. B.: Distributionally robust Monte Carlo simulation. En: *In Proceedings of the International Federation of Automatic Control World Congress* (2002), p. 1–12

[21] Lee, C. ; Mehrotra, S.: A distributionally-Robust Optimization approach for finding support vector machines. En: *Optimization Online.* (2015)

[22] Luo, F. ; Mehrotra, S: Decomposition Algorithm for Distributionally Robust Optimization using Wasserstein Metric. En: *preprint arXiv:1704.03920v1* (2017)

[23] Nesterov, Y.: *Introductory Lectures on Convex Optimization.* Springer, 2004

[24] Nocedal, J. ; Wright, S.: *Numerical Optimization.* Springer, 2006

[25] Otto, F.: The geometry of dissipative evolution equations: the porous medium equation. En: *Communications in Partial Differential Equations* 26 (2001), Nr. 1-2, p. 101174

[26] Popescu, I.: Robust mean-covariance solutions for stochastic optimization. En: *Oparations Research* 55 (2007), Nr. 1, p. 98–112

[27] Prokhorov, Y. V.: Convergence of Random Processes and Limit Theorems in Probability Theory. En: *Theory of Probability and its Applications.* 1 (1956), Nr. 2, p. 157–214

[28] Rubner, Y. ; Tomasi, L. J.: The Earth Mover's Distance as a Metric for Image Retrieval. En: *Probl. Peredachi Inf.* 40 (2000), Nr. 2, p. 99121

[29] Saks, S.: *Theory Of The Integral.* Hafner publishing company, 1937

[30] Scarf, H. ; Arrow, K. ; Karlin, S.: A min-max solution of an inventory problem. En: *Studies in the Mathematical Theory of Inventory and Production* 10 (1958), p. 201–209

[31] Scott, D. W.: *Multivariate Density Estimation. Theory, Practice and Visualization.* Wiley, 1992

[32] Shapiro, A.: On duality theory of conic linear problems. En: *In: Goberna M.Á., López M.A. (eds) Semi-Infinite Programming. Nonconvex Optimization and Its Applications* (2001), p. 135–365

[33] Shapiro, A.: Worst-case distribution analysis of stochastic programs. En: *Mathematical Programming* 107 (2006), Nr. 1, p. 91–96

[34] Shapiro, A. ; Dentcheva, D.: Lectures on Stochastic programming: modeling and theory. En: *SIAM* (2016)

[35] Shapiro, A. ; Kleywegt, A.: Minimax analysis of stochastic problems. En: *Optimizations Methods and Software* 17 (2002), Nr. 3, p. 523–542

[36] Shapiro, A. ; Homem-de Mello, T.: On the rate of convergence of optimal solutions of Monte Carlo approximations of stochastic programs . En: *SIAM Journal on Optimization* 11 (2000), Nr. 1, p. 70–86

[37] Silverman, B. W.: Density Estimation for Statistics and Data Analysis. En: *Chapman & Hall.* (1986)

[38] Spall, JC.: *Introduction to Stochastic Search and Optimization.* Wiley, 2003

[39] Sun, H. ; Xu, H.: Convergence analysis for distributionally robust optimization and equilibrium problems. En: *Mathematicas of Operations Research* (2015)

[40] Vasershtein, L. N.: Markov processes over denumerable products of spaces describing large system of automata. En: *Probl. Peredachi Inf.* 5 (1969), Nr. 3, p. 64–72

[41] Villani, C.: *Optimal transport: old and new*. Vol. 338. Springer Science & Business Media, 2003

[42] Villani, C.: *Topics in optimal transportation*. American Mathematical Soc, 2003

[43] Wang, Z. ; Glynn, PW. ; Ye, Y.: Likelihood robust optimization for data-driven problems. En: *Computational Magmagement Science* (2015), p. 1–21

[44] Xu, M. ; Caramanis, C. ; Mannor, S.: A Distributional Interpretation of Robust Optimization. En: *Mathematics of Operations Research* (2012), Nr. 37, p. 95–110